Stefan E. Huber (Hrsg.)

Anwendung statistischer Verfahren am Computer

mit IBM SPSS Statistics

–

Übungsaufgaben & Lösungen

(Version 1.42)

I am happy to share these materials openly. The content of this Open Educational Resource is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. The CC Attribution 4.0 International license means that you can copy, redistribute, remix, transform and build upon the content for any purpose, even commercially, as long as you give appropriate credit to the source, provide a link to the license, and indicate if changes were made.

Recommended citation: Huber, S. E. (2026). *Anwendung statistischer Verfahren am Computer mit IBM SPSS Statistics – Übungsaufgaben & Lösungen (Version 1.42)*. Retrieved November 16, 2026, from https://osf.io/9tcx3/

Inhalt

Kapitel 1 Worum handelt es sich bei diesem Dokument und wie kann es verwendet werden? 5

Kapitel 2 SPSS. Was ist das und wie kann ich es verwenden? 37

Kapitel 3 Datenmanagement und deskriptive Statistiken 63

Kapitel 4 Parameterschätzung und Testen statistischer Hypothesen über Populationsmittelwerte 107

Kapitel 5 Schätzung und Testung von Mittelwertsunterschieden zwischen zwei Gruppen 133

Kapitel 6 Einfaktorielle Varianzanalyse ohne Messwiederholung 165

Kapitel 7 Zweifaktorielle Varianzanalyse ohne Messwiederholung 195

Kapitel 8 Varianzanalysen mit Messwiederholung 221

Kapitel 9 Einführung in die Regressionsanalyse: Einfache und multiple lineare Regression 249

Kapitel 10 Regressionsdiagnostik, Effektstärken, Stichprobenplanung, Kollinearität 279

Kapitel 11 Regressionsmodelle mit diskreten Prädiktoren und Interaktionen 321

Lösungen zu den Übungsaufgaben 347

Nachwort zur ersten überarbeiteten Fassung dieses Manuskript 441

Nachwort zur ersten Fassung dieses Manuskripts 443

Literatur 449

Kapitel 1

Worum handelt es sich bei diesem Dokument und wie kann es verwendet werden?

Stefan E. Huber

Der Untertitel dieses Dokuments ist durchaus treffend: in erster Linie handelt es sich dabei um eine Sammlung von Aufgaben zum Üben; zum Selbst-Üben, zum Üben in der Lern- oder Kleingruppe (2-5 Personen), zum Üben in mittelgroßen Gruppen (6-50 Personen), etwa in prüfungsimmanenten Lehrveranstaltungen, im Idealfall auch zum Üben in Großgruppen (> 50 Personen), etwa zur Begleitung oder Vertiefung einer Vorlesung. Warum? Weil Üben sowohl für die Konsolidierung als auch die Verfeinerung von Wissen von grundlegender Bedeutung ist (Roelle & Richter, 2025).

Das gemeinsame Ziel der in diesem Dokument gesammelten Übungen ist deshalb sowohl die Übung von der als auch die Einübung in die Anwendung grundlegender statistischer Verfahren mittels geeigneter Software am Computer. Insbesondere sollte die Beschäftigung mit den Übungen Übende dazu befähigen,

- geeignete, grundlegende statistische Verfahren für gegebene psychologische Fragestellungen *auszuwählen*,
- die behandelten statistischen Verfahren auf gegebene Datensätze mittels geeigneter Computerprogramme *anzuwenden*,
- die Ergebnisse der Anwendung statistischer Verfahren auf gegebene Datensätze einer gegebenen Fragestellung angemessen zu *berichten* und (u.a. mittels passender Grafiken) *darzustellen*,
- sowie verschiedene statistische Verfahren hinsichtlich ihrer Limitationen und ihrer Eignung für bestimmte Fragestellungen zu *vergleichen*.

Ferner handelt es sich bei diesem Dokument aber auch um Lernmaterial, das als Grundlage für Lehrveranstaltungen wie die „Anwendung statistischer Verfahren am Computer“ (semestral im Entstehungszeitraum dieses Dokuments an der Universität Graz angeboten und u.a. dort vom Verfasser abgehalten) dienen können soll. Besagte Lehrveranstaltung war im Entstehungszeitraum dieses

Dokuments als begleitende bzw. anwendungsvertiefende Übung zu den (ebenfalls an der Universität Graz angebotenen) Vorlesungen „Psychologische Statistik 1 & 2“ konzipiert und sollte Datenmanagement und statistische Datenanalyse anhand der Statistiksoftware SPSS illustrieren und demonstrieren. In Abstimmung mit den Vorlesungsinhalten sollten dabei grundlegende statistische Verfahren wie Punkt- und Intervallschätzungen von Populationsmittelwerten, Vergleich zweier und mehrerer abhängiger und unabhängiger Stichproben, oder Regressionsverfahren behandelt werden. Studierende sollten zudem in die Berichterstellung statistischer Ergebnisse nach den Richtlinien der *American Psychological Association* (APA) eingeführt werden.

Um der Möglichkeit der Verwendung dieses Dokuments als Lernmaterial für eine solche Lehrveranstaltung gerecht werden zu können, wird daher in diesem Dokument einerseits ein großes Gewicht auf die inhaltliche Nähe zu den genannten Vorlesungen und andererseits auf den Übungscharakter der Lehrveranstaltung „Anwendung statistischer Verfahren am Computer“ gelegt. Konkret bedeutet dies, dass Vorlesungsinhalte anhand einer Vielzahl von Übungsbeispielen illustriert werden. Dies soll insbesondere die Konsolidierung von Wissen aus der Vorlesung fördern, gleichzeitig aber auch Raum für Transferleistungen vom Abstrakten zum Konkreten und vice versa, sowie für anschließende Reflexionsprozesse (Elvira et al., 2017) als Grundlage zur Entwicklung statistischer Expertise schaffen.

Die Übungsbeispiele können größtenteils mit der Statistiksoftware SPSS ausgearbeitet werden (aber eine Beschränkung auf diese Software ist gleichzeitig keinesfalls notwendig). Für manche Übungsbeispiele werden auch andere geeignete Computerprogramme verwendet (z.B. die Software G*Power zum Zweck der Stichprobenumfangsplanung). Da auch das Erlernen eines grundlegenden Umgangs mit diesen Programmen zum Inhalt der Lehrveranstaltung „Anwendung statistischer Verfahren am Computer“ gehört, werden wiederum anhand praktischer Beispiele Schritt-für-Schritt-Anleitungen angeboten, die ein Erlernen dieses Umgangs einerseits erst ermöglichen und in weiterer Folge durch wiederholtes Üben fördern sollen.

Aufbau des Lernmaterials

Das gesamte Lernmaterial, das in diesem Dokument gesammelt vorliegt, ist in einzelne Kapitel unterteilt, die jeweils einen bestimmten Themenkomplex (oder mehrere kleinere Themenbereiche) grundlegender statistischer Datenanalyse behandeln. Kapitel 2 gibt beispielsweise eine knappe Einführung in die Verwendung und den Aufbau der Software SPSS, sowie die Dateneingabe bzw. das Einlesen externer Dateiformate. Dieser Einführung in die Software wird etwas mehr Platz bzw. Detail eingeräumt, da gerade die Bedienung einer neuen Software anfangs eine Hürde darstellen kann, die sich erfahrungsgemäß äußerst negativ auf die Lernmotivation auswirken kann. Ein Ziel dieses Kapitels ist bzw. war es u.a. auch Studierenden der Universität Graz zu ermöglichen, die Software SPSS von zu Hause aus verwenden und erproben zu können, ohne diese gleich käuflich erwerben zu müssen. Kapitel 3 widmet sich anschließend dem Datenmanagement bereits vorhandener Datensätze (z.B. dem Umkodieren von Variablen oder der Erzeugung neuer Variablen aus bereits vorhandenen) sowie der Ermittlung und Darstellung üblicher deskriptiver Statistiken mittels Tabellen, Maßzahlen (Modalwert, Median, Mittelwert, Varianz, Standardabweichung), und grafischen Darstellungsformen.

Kapitel 4 befasst sich erstmals mit inferenzstatistischen Inhalten am Beispiel von Punkt- und Intervallschätzungen des Populationsmittelwerts sowie ungerichteten und gerichteten Hypothesen über den Populationsmittelwert auf der Grundlage einer einfachen Zufallsstichprobe einer normalverteilten Zufallsvariable. Diese grundlegenden Verfahren werden dann in Kapitel 5 zu den entsprechenden Analysen für Mittelwertsunterschiede zwischen zwei abhängigen und unabhängigen Stichproben weiterentwickelt. Bis hierhin entsprechen die Inhalte der oben erwähnten Vorlesung „Psychologische Statistik 1“. Das heißt ab Kapitel 6 werden Inhalte des zweiten Teils der erwähnten Vorlesung in Übungsbeispielen illustriert. In den Kapiteln 6-8 werden Varianzanalysen behandelt. Kapitel 9-12 befassen sich schließlich mit Regressionsanalysen.

In den meisten Kapiteln führen nach einer Wiederholung wesentlicher theoretischer Grundkonzepte einige vollständig ausgearbeitete Übungsbeispiele durch die jeweiligen Lerninhalte. Dabei wird darauf Wert gelegt, dass die Anwendung des jeweiligen Verfahrens Schritt-für-Schritt erläutert und beschrieben wird. Der Detailgrad der Ausarbeitungen soll den Transfer auf andere Fragestellungen, die prinzipiell mit denselben Verfahren beantwortet werden können, erleichtern. Um

diesen Transfer erproben bzw. einüben zu können, wird am Ende jedes Kapitels eine Reihe weiterer Übungsaufgaben angeboten, die jedenfalls mit den Inhalten der ausgearbeiteten Beispiele des jeweiligen oder der vorhergehenden Kapitel selbständig lösbar sein sollten.

Mithilfe dieses Dokuments sollte es also prinzipiell möglich sein, sich die Anwendung grundlegender statistischer Verfahren am Computer (mittels SPSS) selbst anzueignen. Wie bereits oben erwähnt, soll sich aber das Dokument auch als Grundlage für die Lehrveranstaltung „Anwendung statistischer Verfahren“ der Universität Graz eignen können (jedenfalls in der Form, in welcher diese Lehrveranstaltung in den Jahren 2023-2026 üblicherweise abgehalten wurde). Zur Verwendung für diese oder hinreichend ähnliche Lehrveranstaltungen seien mir hier noch einige einführende Anmerkungen gestattet.

Verwendung des Dokuments für die Lehrveranstaltung „Anwendung statistischer Verfahren am Computer“

Für die Verwendung dieses Dokuments als Grundlage für besagte Lehrveranstaltung empfiehlt sich aufgrund des Aufbaus der einzelnen Kapitel die didaktische Methode des sogenannten umgekehrten Klassenzimmers. Die einzelnen Kapitel sind so gestaltet, dass die vollständig ausgearbeiteten Inhalte in etwa 1-2 Stunden konzentrierten Studiums selbständig erarbeitet werden können. Als Lehrender würde ich daher Studierenden die Ausarbeitung jeweils eines Kapitels pro Woche bis zur nächsten Präsenzeinheit empfehlen und die erste Hälfte der folgenden Präsenzeinheit für die gemeinsame Nachbesprechung bzw. ein „Ins-Gedächtnis-Rufen“ der Inhalte und Klärung von Fragen und Unklarheiten nutzen. Die zweite Hälfte würde ich zur Festigung des Lernmaterials mithilfe der am Ende jeden Kapitels zur Verfügung stehenden Übungsbeispiele nutzen. Für diese Übungsbeispiele können zur Vorbereitung der Lehrveranstaltung von den Lehrenden Musterlösungen ausgearbeitet werden (in einem Alternativmodus zur Durchführung der Lehrveranstaltung, der am Ende dieses Kapitels kurz erläutert wird, werden solche Musterlösungen auch von Studierenden zur Vorbereitung auf die jeweilige Präsenzeinheit erarbeitet).

Der Vergleich der von den Studierenden in den Präsenzeinheiten erarbeiteten Lösungen mit diesen Musterlösungen erlaubt dann eine direkte Rückmeldung sowohl für Studierende als auch

Lehrende dahingehend, wo noch Unklarheiten bzw. Nachholbedarf an zusätzlichen Erläuterungen oder Illustrationen des Lernmaterials besteht. Für Studierende ist diese Rückmeldung durch direkten Versuch der Anwendung der jeweiligen Verfahren auf konkrete Inhalte in der Präsenzeinheit nützlich, um beurteilen zu können, wie gut sie sich das Lernmaterial des jeweiligen Kapitels bereits im Selbststudium aneignen konnten und wo noch Schwierigkeiten hinsichtlich des Transfers auf andere Datenanalysekontexte (wie sie durch die Übungsbeispiele repräsentiert werden) bestehen. Die Studierenden können also im Idealfall durch die Übungsbeispiele in jeder Präsenzeinheit beurteilen, wo sie mit ihrem Verständnis des Lernmaterials aktuell stehen und im Bedarfsfall die wöchentlich aufgebrachte Lernzeit oder Intensität des Studiums für die Lehrveranstaltung adjustieren.

Für Lehrende sind die erhaltenen Rückmeldungen nützlich um beurteilen zu können, wo noch Unterstützungsbedarf für die Bildung eines grundlegenden Verständnisses des Lernmaterials besteht. Lehrenden bietet das gemeinsame Bearbeiten der Übungsbeispiele also im Idealfall ein Beurteilungsinstrument, um Verständnisschwierigkeiten zeitnah entgegenwirken und die Gestaltung der Lehrveranstaltung dynamisch anpassen zu können.

Verwendung der Übungsbeispiele

Bei einer angenommenen Dauer von 1,5 Stunden für die wöchentlichen Präsenzeinheiten wird es nur in Ausnahmefällen möglich sein, alle Übungsbeispiele eines bestimmten Kapitels in der Präsenzeinheit mit den Studierenden durchzuarbeiten. Hier bietet es sich an, vorab eine didaktisch überlegte Auswahl vorzunehmen, die aber auf Basis der Rückmeldungen in den Präsenzeinheiten dynamisch auf akute Bedürfnisse abgestimmt werden kann. Übungsbeispiele, die nicht gemeinsam bearbeitet und besprochen werden, können von den Studierenden zur weiteren Vertiefung, aber insbesondere auch zur Klausurvorbereitung (siehe unten) genutzt werden.

Für die Bearbeitung der Übungsbeispiele in den Präsenzeinheiten wird die Ausarbeitung in Kleingruppen von etwa 3-4 Personen empfohlen. Dies erlaubt anfangs leistungsschwächeren Studierenden sich an leistungsstärkeren zu orientieren bzw. von diesen Unterstützung zu erhalten, leistungsstärkere Studierende haben zudem den Bonus durch die Unterstützung anfangs leistungsschwächerer Studierender das Lernmaterial zusätzlich zu festigen (es gibt kaum eine bessere

Methode Lerninhalte selbst zu verinnerlichen als den wiederholten Versuch die Lerninhalte anderen begreiflich zu machen; siehe auch sog. Lernen durch Lehren bei Duran, 2017; interessanterweise funktioniert der Ansatz auch hervorragend ohne Gegenüber, siehe z.B. Lachner et al., 2022). Über die Zeit können so anfängliche Verständnisunterschiede in den Kleingruppen ausgeglichen werden.

Einem etwaigen „Trittbrettfahren" in der Gruppe kann dadurch entgegengewirkt werden, dass die Gruppen dazu aufgefordert werden, dass für die gemeinsame Bearbeitung jedes neuen Übungsbeispiels jeweils ein:e andere:r Studierende:r hauptverantwortlich ist (d.h. die nötigen Schritte am PC durchführt, den Ergebnisbericht schreibt etc.) und die anderen Studierenden unterstützend bzw. beratend zur Seite stehen. Auch der Lerneffekt durch kritisches Beobachten und aufmerksames „Ausschauhalten" nach Fehlern und Irrtümern wie sie grundsätzlich im Laufe jeder Datenanalyse immer wieder passieren, kann kaum überschätzt werden. Ein besonders wichtiges Lernziel im Rahmen statistischer Datenanalyse sollte gerade sein, für möglichst viele Arten möglicher Denkfehler und Irrtümer eine Sensibilität zu entwickeln und zu schärfen, die es schließlich erlaubt, praktische Strategien zu entwickeln, wie solche Fehler dann in eigenen Datenanalysesituationen verlässlich und verhältnismäßig rasch erkannt und korrigiert werden können.

Gerade für einen konstruktiven, lehrreichen Umgang mit Fehlern (Metcalfe, 2017), Missverständnissen und Irrtümern, d.h. für eine in der Praxis nützliche Fehlerkultur, ist es meines Erachtens für die Lehrveranstaltung unerlässlich, dass die Bearbeitung der Übungsbeispiele in den Präsenzeinheiten lediglich der Festigung, dem Üben und dem Ausprobieren der Anwendung statistischer Verfahren gilt und als solche Tätigkeit selbst nicht – etwa zum Zwecke der Notenvergabe – bewertet wird. Die Präsenzeinheiten sollen Lernräume sein, in denen Fehler gemacht, wenn nicht sogar herausgefordert werden sollen (Metcalfe, 2017), um ihnen dann konstruktiv begegnen zu können; d.h. Fehler nutzen zu können, um von ihnen und durch sie zu lernen, um letztlich dafür Sorge zu tragen, die Wahrscheinlichkeit für Fehler in Situationen, in denen es „wirklich darauf ankommt", zu minimieren.

Zu erlernen wie statistische Verfahren anzuwenden sind, heißt auch herausfinden, was vermieden werden sollte, was lieber zu oft als zu selten überprüft werden sollte, und welche Limitationen grundsätzlich mit den (eigenen) Auswertungen einhergehen. Das heißt auch, die

Sicherheit, mit der gewisse Aussagen getroffen werden können, vernünftig einschätzen zu können (eine Fähigkeit, die in der statistischen Datenanalyse in der Tat eine sehr konkrete Bedeutung bekommt).

Auch Negativbeispiele, z.B. von Ergebnisberichten, können dann gleich in der Präsenzeinheit konstruktiv genutzt werden, um im Plenum auf typische Fehler hinzuweisen bzw. Studierende für diese zu sensibilisieren, sowie Strategien zu erarbeiten wie diese Fehler erkannt, vermieden oder korrigiert werden können. Wenn also erste Versuche selbst Ergebnisberichte zu verfassen, noch keine Texte hervorbringen, die man gerne – mit dem eigenen Namen versehen – mit der breiten Öffentlichkeit würde teilen wollen, ist das kein Problem, sondern ganz im Gegenteil eine Gelegenheit, die das Lernen und Üben überhaupt erst ermöglicht.

Das gemeinsame Probieren und Üben in den Präsenzeinheiten soll eben gerade jenen großen Vorteil bieten, dass typische Missverständnisse und Irrtümer von allen Teilnehmenden erlebt und ein konstruktiver Umgang mit ihnen erlernt und geübt werden kann. Die Bewertungsfreiheit der Präsenzeinheit muss dahingehend aber gewährleisten, dass Fehler passieren dürfen und zum Zwecke des Lernens sogar passieren sollen, und keineswegs als stigmatisierend erlebt werden sollten. Darauf hat insbesondere die Lehrperson aktiv zu achten und eine wohlwollende, konstruktive Fehlerkultur sowohl im Umgang mit anderen als auch mit sich selbst zu exemplifizieren. Im Idealfall kann jeder Fehler in den Präsenzeinheiten für alle Teilnehmenden zu einem Fehler werden, der im Ernstfall – etwa in einer Klausur – nicht mehr passieren muss.

Der Vergleich der von den Kleingruppen erarbeiteten Lösungen für die Übungsbeispiele mit von den Lehrenden bereitgestellten Musterlösungen erlaubt den Studierenden zudem Lernen durch selbstständiges Beurteilen der eigens erarbeiteten Lösungen oder der Lösungen, die von anderen Kleingruppen erarbeitet wurden. Insbesondere der Vergleich mit den Lösungen anderer Kleingruppen kann für alternative Lösungswege oder weniger passende Formulierungen von Ergebnisberichten etc. sensibilisieren und damit die Reflexion (und dadurch die Aneignung) des Lernmaterials fördern. Die Vergabe von Punkten (nicht durch die Lehrperson, sondern durch die Studierenden selbst) kann zudem einen Vergleich der Kleingruppen untereinander ermöglichen, der etwa mittels einer wöchentlich aktualisierten Rangliste (Engl.: Leaderboard) rückgemeldet werden kann (wobei eine solche Form der

„Gamifizierung“ aber auch keinesfalls nötig ist). Diese Punkte können dann zum einen als weiterer Rückmeldeprozess dienen, der anzeigt, wo Studierende mit ihrem Verständnis der Inhalte relativ zu anderen Studierenden stehen. Zum anderen kann sich dieses relative und wöchentlich aktualisierte Punktesystem, jedenfalls sofern dieses klar vom Beurteilungsschema für die Lehrveranstaltung, aus dem sich individuelle Noten ableiten, getrennt bleibt, durchaus motivational positiv auswirken. Im Idealfall wird jede Präsenzeinheit ein spielerisches Erlebnis, bei dem die Kleingruppen versuchen, durch Anwendung ihrer statistischen Kenntnisse möglichst viele Punkte zu ergattern. Wenn die Übungsbeispiele das Lernmaterial hinreichend abdecken, bedeutet eine hohe Punktzahl auch ein entsprechend hohes Verständnis, zumindest innerhalb der Kleingruppe, und stellt damit ein geeignetes Rückmeldeinstrument für Studierende und Lehrende dar, und kann und soll daher auch als solches genutzt werden.

Beurteilung: Hausübungen und Abschlussklausur

Um die Präsenzeinheiten als Lern-, Probier- und Spielraum zu gewährleisten, wird ferner empfohlen, zur im Rahmen einer prüfungsimmanenten Lehrveranstaltung notwendigen Beurteilung der individuellen Leistung die wöchentliche Vorbereitung auf die Präsenzeinheiten und das gemeinsame Üben in den Präsenzeinheiten um mehrere Hausübungen und wenigstens eine Abschlussklausur zu ergänzen. Zur Beurteilung der individuellen Leistung werden dann ausschließlich die Leistungen bei diesen Hausübungen und der Abschlussklausur herangezogen. Im Gegensatz zu den gemeinsamen Übungen in den Präsenzeinheiten sind diese dann klarerweise auch individuell zu erbringen und zu bewerten.

In einem Alternativmodus der Abhaltung der Präsenzeinheiten (siehe auch die detailliertere Beschreibung am Ende dieses Kapitels) demonstrieren die Studierenden selbst wöchentlich ihre eigens in der Vorbereitung erarbeiteten Lösungen zu den Übungsaufgaben. D.h. dieser Modus setzt vermehrt auf das Konzept des Lernens durch Lehren (Duran, 2017). Da in diesem Modus allerdings deutlich mehr Leistung außerhalb der Präsenzübungen über das Semester hinweg erbracht werden muss, kann in diesem Alternativmodus auch auf die Hausübungen verzichtet werden. Bei der Vorbereitung auf jede Präsenzeinheit handelt es sich dann ohnehin bereits jeweils um regelmäßige Hausübungen. Für die

Beurteilung der individuellen Leistungen zu Semesterende kann die Abschlussklausur in diesem Fall auch auf zwei Klausuren – eine zu Semestermitte, eine wie gehabt zu Semesterende – aufgeteilt werden.

Für den Modus mit beurteilten Hausübungen werden vier Hausübungen empfohlen, die relativ gleichmäßig über den Zeitraum der Lehrveranstaltung bzw. das Lernmaterial verteilt werden sollten. Das heißt, dass im Rahmen der Hausübungen jeweils ein etwas umfangreicheres Stoffgebiet als für die wöchentliche Vorbereitung auf die Präsenzeinheit zu bearbeiten ist. Dabei sollte es sich in etwa um den Lernstoff von etwa 3-4 Präsenzeinheiten handeln (mit Ausnahme der ersten Hausübung, siehe unten). Die Hausübungen sollten den vollständig ausgearbeiteten sowie gemeinsam erarbeiteten Übungsbeispielen aus den Präsenzeinheiten insofern ähneln, als dass eine selbständige Lösung der Hausübungsbeispiele auf der Grundlage der durch die Übungsbeispiele gelernten Inhalte jedenfalls gut möglich sein sollte. D.h. konkret, wer sich jeweils auf die Präsenzeinheiten regelmäßig vorbereitet und in diesen aktiv mitgewirkt hat, sollte sich bei den Hausübungen keinesfalls vor unlösbare Aufgaben gestellt sehen, sondern im Idealfall Gelegenheiten vorfinden, bei denen das eigene Wissen und Können individuell zur Anwendung gebracht und demonstriert, und dadurch Selbstwirksamkeit erlebt werden kann.

Davon sollte sich prinzipiell auch die Abschlussklausur nicht unterscheiden. Auch diese sollte aus mehreren konkreten Beispielen bestehen, die durch die Aneignung des Lernmaterials individuell und selbständig zu lösen sein sollten. Das heißt wiederum, dass es sich bei der Abschlussklausur im Idealfall um eine Gelegenheit handelt, um zu zeigen, dass man das, was man in den Präsenzeinheiten in der Kleingruppe und bei den Hausübungen individuell, nun auch selbständig und individuell für den gesamten Lernstoff der Lehrveranstaltung umsetzen kann.

In engem Zusammenhang mit der Abschlussklausur offenbaren sich dann auch einige weitere Vorteile des regelmäßigen, gemeinsamen Übens von Übungsaufgaben inklusive des gemeinsamen Bewertens erarbeiteter Lösungen. Erstens sind mit den zahlreichen unterschiedlichen Übungsaufgaben die unterschiedlichen Formate an möglichen Beispielen – sowohl für Hausübungen als auch die Abschlussklausur – bekannt. Bei diesen Formaten handelt es sich zum Teil um geschlossene oder offene Fragen, die hauptsächlich auf das Verständnis oder auch das Einprägen einer statistischen „Grammatik“

abzielen. Wie jeder Fachbereich hat auch die Statistik eine eigene, wenn nicht gar recht eigentümliche Sprache. Ohne jede Einübung in diese bleibt es oft lange schwierig sich in der Statistik zurechtzufinden. Auch gut gemeinte Hilfsangebote (von natürlicher wie auch künstlicher Intelligenz) bleiben dann häufig weit hinter ihrer Intention zurück, wenn ein grundlegendes Sprachverständnis fehlt. Wenn mir jemand den Weg auf Portugiesisch erklärt, wird mir nicht viel geholfen sein, sofern ich Portugiesisch nicht auch verstehen kann. Zudem bieten Aufgaben dieses Typs als Form einer Abrufübung (Heitmann et al., 2018, 2022) den Vorteil der leichten Implementierbarkeit, führen zu deutlich besseren Ergebnissen für die Wissenskonsolidierung als das erneute Durchsehen oder Durcharbeiten des Lernmaterials (Adesope et al., 2017; Roediger & Karpicke, 2006; Roelle & Berthold, 2017), und fördern die Wirksamkeit von Transferaufgaben, bei welchen die durch die Abrufübung konsolidierten Inhalte verwendet werden müssen (Pan & Rickard, 2018).

Ein weiteres typisches Format ist das Vorgeben einer Fragestellung und eines Datensatzes mit der Aufforderung, die (statistische) Fragestellung zu erhellen und einen entsprechenden Ergebnisbericht zu erstellen. Dabei handelt es sich vermutlich um den klassischen Fall eines Übungsbeispiels für die Anwendung statistischer Verfahren. Er entspricht auch einer recht häufigen Situation, wenn man es selbst in der wissenschaftlichen Praxis mit statistischer Datenanalyse zu tun hat. Das regelmäßige Üben dieses Aufgabenformats in Vorbereitung auf und während der Präsenzeinheiten stellt eine Realisierung des sukzessiven Wiederlernens (Bahrick, 1979) dar, das gegenüber herkömmlichen Lehr- und Lernmethoden den Vorteil bietet, dass Inhalte wiederholt gemeistert werden müssen. Gerade in Hochschulkontexten ist es häufig der Fall, dass Inhalte, nachdem sie in der Lehrveranstaltung von Lehrenden behandelt wurden, von Studierenden erst kurz vor Prüfungen oder Klausuren erneut behandelt werden. Ganz im Gegensatz zum sukzessiven Wiederlernen werden Lerninhalte in diesem Fall kein einziges Mal überhaupt gemeistert (Rawson & Dunlosky, 2022). Sukzessives Wiederlernen scheint demgegenüber massive Vorteile in Bezug auf die langfristige Behaltensleistung erbringen zu können (Higham et al., 2022; Rawson et al., 2013; Rawson & Dunlosky, 2011).

Ein weiteres und recht praxisnahes Aufgabenformat besteht darin, dass man eine statistische Analyse mit oder ohne Ergebnisbericht bereits vorliegen hat und diese bzw. den Ergebnisbericht nun überprüfen soll (etwa weil man ganz nach dem – sehr nützlichen – Prinzip „Vier Augen sehen mehr als

zwei“ die Arbeit eines:einer Kollegen:in überprüfen darf oder vielleicht auch bloß Hausübungen in einer entsprechenden Lehrveranstaltung korrigieren soll). Auch dieses Format wird man hier in entsprechenden Übungsbeispielen wieder finden.

Schließlich wird auch das Einüben in das Schreiben von Ergebnisberichten didaktisch stellenweise intensiviert bzw. fokussiert, indem die Ausgabe einer bestimmten Analyse vorgegeben wird und „nur“ noch der Ergebnisbericht zu erstellen ist. Didaktisch kann es durchaus sinnvoll sein, einzelne Teilaspekte zusammengesetzter Tätigkeiten für sich genommen zu üben, und dann, wenn die einzelnen Handlungen gut gefestigt sind, wieder zu einem Ganzen zusammenzusetzen. Dementsprechend wird es auch Übungsbeispiele geben, in welchen man einen Ergebnisbericht zu einer bestimmten Fragestellung bereits teilweise gegeben hat und nur noch die Lücken in diesem auszufüllen sind. Hier geht es also vorrangig darum, zu erkennen, welches Verfahren hier zu verwenden ist, und dieses dann anzuwenden, um die fehlenden Informationen vervollständigen zu können. Dieses Übungsformat steht also wiederum einer Abrufübung näher.

Während also einzelne dieser Formate vielleicht besser oder schlechter zur Förderung eines bestimmten Aspekts der Lernziele dienen, sollte ihre Gesamtheit die Erreichung der Lernziele doch recht gut fördern können. In ihrer Gesamtheit erfüllt die Heterogenität der unterschiedlichen Übungsformate auch die Voraussetzungen des verschachtelten Übens, dessen Wirksamkeit ebenfalls empirisch belegt ist (Brunmair & Richter, 2019). Die Verteilung der verschiedenen Übungsformate über das Semester in unterschiedliche Phasen strenger regulierten Übens während der Präsenzeinheiten sowie autonomeren Übens zwischen den Präsenzeinheiten ist zudem der Lernform des verteilten Übens dienlich, deren Wirksamkeit empirisch gut belegt ist (Ebersbach et al., 2022).

Studierende werden demnach im Verlauf der Lehrveranstaltung schrittweise mit einer Vielzahl unterschiedlicher Übungsformate bekanntgemacht und dadurch sowohl auf entsprechende Hausübungs- und Klausuraufgaben als auch reale bzw. praxisnahe Datenanalysesituationen vorbereitet. Durch das gemeinsame Bewerten erarbeiteter Lösungen von Übungsaufgaben wird ferner der Beurteilungsmodus für Hausübungen und Abschlussklausur transparent. Die Wahrscheinlichkeit für „böse Überraschungen“, sowohl was die Art als auch die Bewertung von Hausübungen und Klausuraufgaben

angeht, sollte dadurch minimiert werden. Unklarheiten und Rückfragen können dadurch im Idealfall im Vorhinein und nicht erst im Nachgang anhand von konkreten Situationen und Beispielen geklärt werden.

Formal sind den Hausübungen und der Klausur selbstverständlich noch Gewichtungen zu vergeben. Zum Beispiel können hier jeder der vier Hausübungen jeweils 10 Punkte und der Abschlussklausur 60 Punkte vergeben werden, was in einer Gesamtpunkteanzahl von 100 Punkten resultiert. Die größere Gewichtung der Abschlussklausur hat zum Hintergrund, dass alle Studierenden auch individuell aufgefordert sind, sich mit der Aneignung der Gesamtheit des Lernmaterials zu befassen, und nicht bloß häppchenweise zu verarbeiten, was erfahrungsgemäß einer längerfristigen Festigung der Inhalte nicht förderlich ist. Letzterer soll auch das Format der Abschlussklausur dienen, auf welches im folgenden Abschnitt eingegangen wird.

Format der Abschlussklausur

Da es bei der Abschlussklausur zum einen um die Überprüfung der individuellen Fähigkeit zur Aneignung und selbständigen Anwendung des Lernmaterials geht, wird die Durchführung der Abschlussklausur als sogenannte Closed-Book Klausur empfohlen. Dies soll u.a. eine tiefere kognitive Verarbeitung der Lerninhalte in der Vorbereitung erfordern und damit auch fördern. Es ist bekannt, dass das Wissen um die Möglichkeit jederzeit gewisse Inhalte nachschlagen zu können, der kognitiven Verarbeitungstiefe zuwiderläuft. Tatsächlich werden Gedächtnisinhalte flüchtiger gespeichert (das Gehirn handelt sozusagen ökonomisch und sagt sich „warum soll ich mich dafür großartig umstrukturieren, wenn das ohnehin jederzeit nachgeschaut werden kann?“), wenn bekannt ist, dass sie leicht zugänglich sind und ohne große Schwierigkeiten nachgeschlagen werden können (siehe z.B. Sparrow et al., 2011).

Eine gewisse kognitive Mindestverarbeitungstiefe von grundlegenden Inhalten (grundsätzlich egal welchen Fachgebiets) einzufordern, scheint allerdings aus mehreren Hinsichten empfehlenswert. Spätere, konkrete Datenanalysesituationen, wie etwa im Rahmen der Bachelor- oder Masterarbeit, werden häufig das Erlernen spezialisierterer Analyseverfahren erfordern, die weit über die grundlegenden Konzepte dieses Dokuments hinausgehen können und werden. Das Erlernen solcher fortgeschrittenen Verfahren setzt meist fundierte und belastbare Grundlagenkenntnisse und ein

Verständnis einer grundlegenden statistischen Denkungsart voraus. Um in der Metapher von oben zu bleiben: es fällt einem entsprechend umstrukturierten Gehirn leichter, sich in diese fortgeschrittenen Verfahren einzuarbeiten bzw. einzudenken. Das heißt, das Umstrukturieren, das sich das Gehirn gerne ersparen würde, ist gerade das Ziel der Übung. Man kann dies auch mit grundlegenden Übungen von Sportler:innen vergleichen. Skispringer:innen etwa führen regelmäßig (im Profibereich auf nahezu täglicher Basis) grundlegende, die Bauchmuskeln und den Rücken stärkende Übungen durch (neben einer Vielzahl anderer Fitnessübungen). Das tun sie aber nicht, weil sie im Wettkampf oder beim Training auf der Schanze diese Übungen „vorzeigen" müssen. Der Grund ist einfach die Bildung eines Muskel- bzw. Bewegungsapparats, der die nötigen Strukturen besitzt und Voraussetzungen erfüllt, die für die Durchführung der eigentlichen Kernaktivitäten ihres Berufs (Skispringen) schlichtweg notwendig ist. Genauso hat die Einübung in die Anwendung grundlegender statistischer Verfahren vorrangig die Bildung eines entsprechenden Denk- und Handlungsapparats zum Ziel, der die Praxis der statistischen Datenanalyse in späteren, „echten" Datenanalysesituationen überhaupt erst ermöglicht.

Kommt es dann zu diesen „echten" Datenanalysesituationen (etwa im Rahmen der Bachelor- oder Masterarbeit, einer Dissertation, oder der ganz normalen Berufspraxis, die selbstverständlich je nach Werdegang durchaus im Aufkommen der Notwendigkeit von statistischer Datenanalyse variieren kann) müssen diese grundlegenden Fähigkeiten dann lediglich reaktiviert werden. Müssen sie stattdessen in diesen Fällen überhaupt erst zum ersten Mal erlernt werden (wird also erst mit dem Bauchmuskel- und Rückentraining begonnen, wenn man bereits auf der Schanze darauf wartet als Nächster hinunterzuspringen), ist eventuell spät ein hoher Preis für ein frühes Versäumnis zu entrichten. Dem soll durch die Einforderung einer ausreichenden Verarbeitungstiefe bei der Aneignung des Lernmaterials zumindest entgegengewirkt werden.

Dabei ist es auch wichtig, sich noch einmal klarzumachen, dass es sich bei der Klausursituation nicht um die Abbildung einer praxisnahen Datenanalysesituation handelt und auch nicht handeln soll (ein Einwand, der immer wieder einmal gerne von Studierenden, aber auch Absolvent:innen gemacht wird). Es ist selbstverständlich in jeder praxisnahen Datenanalysesituation sinnvoll bei Unklarheiten in einschlägigen Quellen nachzuschlagen, und nicht wie bei einer Closed-Book Klausur zu versuchen, alle Herausforderungen bloß mit den Mitteln zu bewältigen, die man noch aktiv im Gedächtnis hat (dies

würde auch nicht der Sorgfaltspflicht, die man als Datenanalyst etwaigen Kund:innen gegenüber, als Wissenschaftler:in der Wahrheitsfindung gegenüber auf sich nimmt, gerecht werden). Allerdings setzt die Fähigkeit sich adäquat in solchen konkreten Situationen zu informieren und überhaupt die eigenen Kenntnisse einschätzen zu können, wie oben bereits bemerkt, bereits ausreichende grundlegende Fähigkeiten und Kenntnisse voraus. Man stelle sich nur einmal vor, jemand bestellt aufgrund eines Heizungsproblems Handwerker zu sich nach Hause und diese beginnen sich dann vor Ort mit Google, Wikipedia und ChatGPT über Schraubenschlüssel, Rohrzangen, typische Maße von Anschlüssen und andere handwerkliche Grundlagen zu informieren. Es ist klar, dass die fiktiven Handwerker mit diesem bestimmten Heizungssystem vielleicht noch nie konfrontiert waren, aber ein grundlegendes Wissen von Heizungssystemen überhaupt und entsprechenden Werkzeugen etc. würden sich wohl die meisten Kund:innen völlig zu Recht erwarten.

Das heißt, das Ziel der Abschlussklausur ist nicht, dass Studierende einzelne Arbeitsschritte auswendig lernen sollen. Das Ziel ist, dass Studierende sich kognitiv intensiv genug mit dem Lernmaterial befassen, dass sie dieses zum Ende der Lehrveranstaltung hin behände anwenden können. Eine angemessen tiefe Verarbeitung hat oft zur Konsequenz, dass grundlegende Schritte leicht oder wie automatisiert von der Hand gehen und es so aussieht, als würde man eine Abfolge von Arbeitsschritten auswendig wissen, auch wenn es sich bei einem grundsätzlichen Verstehen ganz sicher nicht um lexikalisches Wissen handelt. In der Tat sind solche (scheinbaren) Automatismen auch gute Indikatoren für eine ausreichende kognitive Verarbeitungstiefe in der Vorbereitung. Das Ziel bzw. der Zweck einer Closed-Book Klausur bleibt aber immer jene ausreichende kognitive Verarbeitung und nicht die Indikatoren, an denen man diese (u.a. und nicht zweifelsfrei) erkennen kann. Dass Closed-book Formate diese tiefergehende kognitive Verarbeitung auch in der Tat herausfordern und fördern, dürfte auch dem empirischen Befund zugrunde liegen, dass höhere Lernleistungen tatsächlich in Closed-book Implementierungen von Wissensprüfungen erbracht werden als in Open-book Implementierungen (Rummer et al., 2019).

Dass ein grundlegendes Verständnis von Inhalten oft von einem (scheinbaren) hohen Maß an lexikalischem Wissen begleitet wird, kann man sich auch in einem Gedankenexperiment (das zu realisieren ich nur jedem empfehlen kann, der an der Aneignung welchen Lerninhalts auch immer

ernsthaft interessiert ist) leicht selbst ein Bild machen. Stellen Sie sich vor, Sie möchten einem Kollegen oder einer Kollegin helfen, ein gewisses Konzept aus der Statistik zu erklären. Sie treffen sich und beginnen zu erklären. Allerdings kommen Sie bereits beim ersten Satz ins Stocken, unterbrechen Ihre Erläuterung, um Google oder ChatGPT zu konsultieren (z.B. um zu fragen, wie man dieses Konzept jemandem in einfachen Worten erklären kann). Wie hoch würden Sie Ihre Kompetenz in diesem Fall einschätzen? Vermutlich kommen Sie in diesem Fall eher zu einer unbefriedigenden Antwort.

Allerdings lassen Sie sich von dieser Erfahrung nicht entmutigen und beschäftigen sich daraufhin eingehend mit weiterer Literatur (und natürlich auch mit den Antworten, die Sie von ChatGPT und Google erhalten haben; am Einholen derselben ist ja grundsätzlich nichts Verwerfliches, sondern ganz im Gegenteil, häufig sehr viel Nützliches und Sinnvolles). Tatsächlich fragt Sie einige Zeit später wieder ein:e Kolleg:in um Rat. Sie beginnen wieder zu erklären, und bemerken, dass ihre Erklärung schon etwas flüssiger und weniger stockend wirkt als beim letzten Mal. Sie fühlen sich auch deutlich wohler in Ihrer Haut.

So geht es weiter und nach einigen weiteren Versuchen stellen Sie fest, dass Sie das komplizierte Konzept ganz ohne Zuhilfenahme weiterer Hilfsmittel, Personen erklären können, die ganz zu Recht Schwierigkeiten damit haben, weil es nun einmal kein einfaches Konzept ist (wie vieles in der Statistik – daraus braucht man keinen Hehl zu machen – eben einfach nicht einfach oder intuitiv ist). Sie hören sich Sätze sagen wie „Ah ja, ich kann mir vorstellen, was dir dabei schwerfällt. Lass es mich einmal so erklären. Nehmen wir einmal an…“, sehen sich Hilfsskizzen und Diagramme zeichnen und Flächen unter Kurven kennzeichnen, und können ganz den Verständnisschwierigkeiten ihres Gegenübers zugewandt bleiben, weil sie sich nicht abwenden müssen, um sich erstmal selbst über einige Aspekte des Konzepts klar zu werden. Zugegeben: das ist ein sehr hohes Ziel und der Verfasser dieser Zeilen glaubt selbst nicht, diesem Ziel, außer in den aller einfachsten Fällen, auch nur annähernd gerecht werden zu können.

Aber angenommen, Ihnen gelingt es. Wie würden Sie Ihre Kompetenz in diesem Fall einschätzen? Vermutlich höher als in allen zuvor beschriebenen Fällen. Und es wird für Ihr Gegenüber in diesem Fall dann sehr wahrscheinlich auch so aussehen, als würden Sie sehr viele Details oder

einzelne Arbeitsschritte auswendig wissen. Ihnen selbst wird aber klar sein, dass es sich dabei um kein bloßes lexikalisches Wissen handelt, sondern es Ihnen lediglich leicht fällt die nötigen Arbeitsschritte bei der Erläuterung des Konzepts zu reproduzieren. Gerade weil Sie es grundlegend verstanden haben, brauchen Sie sich nicht darum zu bemühen, etwas „auswendig“ zu behalten. Es ist Ihre Expertise, die wie von Geisterhand dafür sorgt, dass grundlegende Zusammenhänge wie automatisch aus den Tiefen Ihres Langzeitgedächtnisses hervorkommen, wenn Sie sie gerade brauchen.

Ein entsprechend fundiertes Verständnis setzt aber u.a. viel Übung und Wiederholung der Inhalte voraus. Dafür soll die Lehrveranstaltung inklusive der Closed-Book Klausur eine kleine Anschubhilfe leisten. Lernen und üben muss aber tatsächlich jede:r selbst (auch wenn Lerngruppen – wie hoffentlich auch die jeweiligen Kleingruppen in der Lehrveranstaltung – dafür sehr gute Motivationshilfen sein können).

Zusammengefasst heißt das, dass die intensive Beschäftigung mit den Übungen, die in diesem Dokument gesammelt vorliegen, die Bildung eines solchen grundlegenden Verständnisses für die Anwendung grundlegender statistischer Verfahren ermöglichen sollen. In den seltensten Fällen werden diese grundlegenden Verfahren ausreichen, um konkrete wissenschaftliche Fragestellungen (statistisch) zu erhellen. Dazu werden häufig Verfahren notwendig sein, die weit über die hier behandelten hinausgehen, in manchen Fällen überhaupt erst entwickelt werden müssen. Ein grundlegendes Verständnis für die hier behandelten Grundlagen soll es Ihnen aber ermöglichen, genau das tun zu können, nämlich weit über das hinauszugehen, was Sie hier beschrieben vorfinden. Das Dokument soll Ihnen sozusagen dabei helfen – frei nach Wittgenstein (2003) – die Leiter wegwerfen zu können, nachdem Sie über sie hinausgestiegen sind. Mit dieser Intention wurde es jedenfalls erstellt. Und falls Ihnen dieses Dokument in der Tat dabei helfen kann, hat es seinen Zweck auch mehr als erfüllt.

Forschungsgeleitete Lehre

Lehrveranstaltungen, die der Einführung und Einübung in die Anwendung statistischer Verfahren dienen, haben häufig mit einer recht hohen Hemmschwelle für die entsprechenden Lerninhalte, gerade unter Studienanfänger:innen zu kämpfen. So entscheiden sich die wenigsten Studierenden in den Sozial- oder Humanwissenschaften wohl für ihr Studium, weil sie sich besonders

gerne mit statistischen Fragestellungen auseinandersetzen. Zwar wird den Studierenden dann erzählt, dass statistische Grundkenntnisse essenziell sind, um später in der Praxis die Ergebnisse wissenschaftlicher Studien verstehen, interpretieren und bewerten zu können. Dazu gehört die Fähigkeit zur Beurteilung, ob sich Ergebnisse wissenschaftlicher Studien beispielsweise erfolgreich in Psychotherapie, Erziehungs- oder Unterrichtskontexten oder in einem Unternehmen oder einer Organisation umsetzen lassen, sowie die Fähigkeit die Wirksamkeit entsprechender Interventionen realistisch einschätzen zu können. Die selbständige, kritische Auseinandersetzung mit Studienergebnissen soll schließlich eine realistische, professionelle Einschätzung der Vielzahl an Ergebnissen ermöglichen, um hochwertige von schlechten Ergebnissen unterscheiden zu können, und nicht alles glauben zu müssen, was von irgendjemandem publiziert wurde, sondern entsprechende Veröffentlichungen selbst auf Stichhaltigkeit prüfen zu können.

In der Tat haben Vorlesende damit völlig recht: zu all dem können fundierte statistische Grundkenntnisse selbstverständlich einen wesentlichen Beitrag leisten. Allerdings besteht für Studienanfänger:innen und der beschworenen Nützlichkeit der Lerninhalte in der späteren Berufspraxis doch meist noch eine gehörige Distanz, die die Ernsthaftigkeit der Auseinandersetzung mit dem Lernmaterial zumindest erschweren kann. Dies kann sich etwa darin äußern, dass der Nutzen der Auseinandersetzung mit statistischen Inhalten bezweifelt wird. Manchmal wird dieser zusätzlich dahingehend in Frage gestellt, dass jemand „doch sowieso klinischer Psychologie“ oder hauptsächlich in einem sehr angewandten Bereich tätig werden wolle, in welchem man sehr selten mit wissenschaftlicher Originalliteratur zu tun habe. Von der Haltbarkeit der impliziten Unterstellungen, die in diesen Beispielen eventuell einzelnen Berufsgruppen gemacht werden, einmal abgesehen, werden die Lerninhalte auf Studierendenseite in diesen Fällen eben dennoch manchmal als abstrakt, praxisfern, oder als bloße akademische und „eigentlich“ unnötige Hürde erlebt. Um die intrinsische Motivation der entsprechenden Studierenden bei der Aneignung dieser Inhalte ist es dann verständlicherweise schlecht bestellt. Auch die eigene Motivation und Begeisterung der Lehrperson für das Lernmaterial kann dann in diesen Fällen nur begrenzt Abhilfe schaffen (ein bisschen kann sie das aber mit Sicherheit).

Zu erwarten, dass alle Studierenden mit der nötigen Weitsicht oder mit der nötigen Kreativität, um selbst motivierende Sinnzusammenhänge zwischen Lerninhalten und persönlichen Zielen des

(beruflichen) Werdegangs herstellen zu können, in die je eigene Lehrveranstaltung kommen, ist naiv und wenig zielführend. Den Studierenden schlichtweg zu sagen, dass etwas „so und so sei (und wer das nicht verstehe sei selbst schuld)“ ist mindestens genauso naiv und didaktisch fragwürdig. Nur weil eine Lehrperson einem Lernenden etwas sagt, ist das noch lange nicht gelernt; davon abgesehen, dass das fraglose Übernehmen von Ausgesprochenem sowieso nicht Ziel akademischer Bildung sein kann (jedenfalls nicht im Sinne eines Leitbilds humanistischer Bildung).

Was hingegen durchaus helfen kann, die inhaltliche Relevanz der Lerninhalte hervorzuheben, ist, genau diese Relevanz anhand konkreter Beispiele in den Unterricht hereinzuholen und auf diese Weise zu etwas Erlebbarem zu machen. Wenn schon behauptet wird, dass statistische Grundkenntnisse so wesentlich für die Bewertung sozialwissenschaftlicher Studienergebnisse sind, dann ist die wohl geradlinigste Art, diese Behauptung zu einer Erfahrungstatsache für Studierende zu machen, das gemeinsame Durchführen und anschließende Beurteilen einer solchen Studie.

Damit der Fokus der Lehrveranstaltung auf den statistischen Inhalten bleiben kann und sich nicht mit allen anderen Aspekten des empirisch-experimentellen Forschens eingehend befassen muss (davon gibt es einige und zur angemessenen Einschätzung einer wissenschaftlichen Studie gehört deutlich mehr als ein bloßes *statistisches* Grundverständnis, siehe z.B. Huber, 2019), ist es dafür wünschenswert, wenn es sich bei einer solchen Studie um eine handelt, die mit relativ einfachen Mitteln durchgeführt werden kann. Beispiele für Studien dieser Art gibt es aber in der Geschichte der Sozialwissenschaften oder der Psychologie mehr als genug. Zwei Möglichkeiten bestehen etwa im Spatial-Cueing Paradigma aus der kognitiven Psychologie oder dem impliziten Assoziationstest aus der Sozialpsychologie.

Beide Experimente lassen sich relativ einfach mit digitalen Hilfsmitteln wie etwa der Software PsychoPy implementieren und mit den Studierenden in einem Online-Versuch umsetzen. Beide Experimente beinhalten Innersubjektfaktoren, die die Behandlung statistischer Analysen für abhängige Stichproben für die Lehrveranstaltung erschließen. Vergleiche zwischen sozialen Geschlechtern erschließen den statistischen Vergleich unabhängiger Stichproben. Die Kombination aus beidem ermöglicht die Besprechung gemischter Designs im Rahmen von Varianzanalysen, die

Berücksichtigung von Kovariaten (etwa dem Alter) erschließt Fragestellungen im Rahmen allgemeiner linearer Modelle und insbesondere der linearen Regression. Insgesamt erlaubt also die Durchführung eines Experiments die Behandlung des gesamten Lernmaterials an Studienergebnissen, die mit den Studierenden selbst hervorgebracht wurden. Auf diese Weise wird nicht nur die Relevanz der statistischen Inhalte für die Beantwortung wissenschaftlicher Fragestellungen erlebbar, sondern auch (allgemein-)psychologische oder sozialwissenschaftliche Inhalte der konkreten Erfahrung zugänglich gemacht. Es wird klar, dass es sich bei wissenschaftlichen Erkenntnissen der Psychologie und der Sozialwissenschaften nicht um eine Sache von rein akademischem Interesse, sondern um Zusammenhänge bzw. Effekte handelt, die dem ganz realen, tagtäglichen Verhalten, Denken, Erleben, und Handeln von Menschen zugrunde liegen.

Natürlich gibt es eine Vielzahl weiterer Experimente, die einen mindestens ebenso großen Mehrwert für die Durchführung einer entsprechenden Lehrveranstaltung bieten können. Beispielsweise können auch moderne experimentelle Ergebnisse, die eine einfache Implementierung und Durchführung im Rahmen der Lehrveranstaltung zulassen, genutzt werden und im Kontext der Lehrveranstaltung auf ihre Stichhaltigkeit geprüft werden. Limitationen (etwa der Generalisierbarkeit oder aufgrund recht überschaubarer Stichprobengrößen) einer solchen Durchführung im Rahmen einer Lehrveranstaltung können dann gleich mit den Studierenden erarbeitet und diskutiert werden. Beispielsweise handelt es sich bei vielen bekannten Effekten in der Psychologie um relativ kleine Effekte, die als Unterschied in den Mittelwerten zweier Experimentalgruppen (oder einer sogenannten Experimental und einer Kontrollgruppe) erst bei hinreichend großen Stichproben mit hoher Wahrscheinlichkeit als statistisch signifikante Unterschiede zutage treten. Das heißt, je nach Lehrveranstaltung, kann man sich etwa in einer Übungsgruppe aus 50 Studierenden gar keine substantielle Bestätigung eines solchen Effekts erwarten. Gerade dies bietet aber eine exzellente Lerngelegenheit um oft schwer greifbare Konzepte wie statistische Teststärke, statistische Signifikanz, Fehlerarten erster und zweiter Art zu diskutieren und durch das konkrete Beispiel der gelebten Erfahrung zugänglich zu machen.

Die Einbeziehung „echter“ Forschung auf diese oder ähnliche Weise kann unter diesen Gesichtspunkten nur empfohlen werden. Die Durchführung eines Experiments im Rahmen einer für die Psychologie (ob aktuell oder zumindest in früheren Zeiten) tatsächlich relevanten Fragestellung und der

anschließenden statistischen Analyse der Ergebnisse durch die Studierenden selbst unter Anleitung der Lehrveranstaltungsleiter:innen lässt die Studierenden im Gegensatz die sonst als „abstrakt“ und praxisfern empfundene Anwendung statistischer Verfahren in einem wissenschaftlich bedeutsamen Sinnzusammenhang erfahren. Zugleich können die Ergebnisse des Experiments selbst einen wesentlichen Erkenntnisfortschritt durch ihre Einbettung in den Lernkontext darstellen. Gerade auf Studien im Bereich der pädagogischen Psychologie trifft dies im Besonderen zu. Ferner zeichnen sich psychologische Studien häufig durch ein hohes Maß an Kontextabhängigkeit aus. Das heißt, ob eine Studie im psychologischen Labor oder im Unterrichtsraum stattfindet, kann selbst ein Faktor sein, der einen bedeutsamen Einfluss auf die Studienergebnisse hat. Zumindest im Rahmen der statistischen Auswertung der Ergebnisse können die Studierenden somit zu aktiven Teilnehmer:innen am Forschungsprozess werden und bekommen Einblicke in Bedeutung und Limitationen statistischer Datenanalyse anhand konkreter Forschungsergebnisse. Damit steht dieser Aspekt der Durchführung der Untersuchung am Übergang zwischen forschungsmotivierter Lehre zu forschendem Lernen (Sonntag et al., 2017) und erlaubt damit Zugriff auf die Vorzüge beider Lehr-/Lernformen im Rahmen forschungsgeleiteter Lehre in der universitären Lehre (Huber, 2014; Rueß et al., 2016).

Dabei ist aber noch einmal zu betonen, dass der Einsatz einer experimentellen Studie innerhalb der Lehrveranstaltung zu diesem Zweck nicht auf Kosten der Einübung in die grundlegenden statistischen Verfahren gehen darf. Im Gegenteil soll dieser Einsatz dieser Einübung dienlich sein und für sie einen Mehrwert darstellen. Das kann dadurch erreicht bzw. gefördert werden, dass die konkrete Durchführung des Experiments auf die erste der oben genannten vier Hausübungen beschränkt bleibt. Das heißt, im Rahmen der ersten von vier Hausübungen führen die Studierenden jeweils selbständig ein Online-Experiment durch. Eventuell akquirieren die Studierenden auch noch zusätzlich jeweils ein oder zwei Freiwillige, die an dem Experiment teilnehmen; dies würde einen Vergleich einer Stichprobe innerhalb des Lernkontexts mit einer Stichprobe außerhalb des Lernkontexts erlauben. Die Dateneingabe bzw. das Datenmanagement kann dann mit den Studierenden anhand dieser eigens generierten Daten in einer der auf die Hausübung folgenden Einheiten besprochen und geübt werden. In den verbleibenden Hausübungen können die in den Präsenzeinheiten besprochenen grundlegenden Verfahren auf diese Daten angewandt werden. Die Besprechungen der Hausübungen werden so

schließlich zusätzlich auch zu einer Besprechung von Forschungsergebnissen. Fehler, Irrtümer, Missverständnisse und deren Bedeutung werden in einem inhaltlich sinnvollen Zusammenhang erlebbar, der nicht nur der Generierung von Punkten zur Beurteilung der Lehrveranstaltung dient. Das heißt, Studierende können erleben, welche Konsequenzen ein Fehler in der statistischen Datenauswertung für die Interpretation „echter“ Forschungsergebnisse hat, jedoch ohne den sicheren Rahmen einer Lehrveranstaltung dafür erst verlassen zu müssen. Das heißt, Fehler können hier gemacht und erlebt werden, um später in der eigenen Forschungspraxis dieselben Fehler nicht mehr machen zu müssen. Der Transfer in die eigene Praxis sollte durch die Einbettung in inhaltliche relevante Forschung innerhalb der Lehrveranstaltung zumindest gefördert werden.

Insgesamt werden die didaktischen Ziele der Lehrveranstaltung dabei durch die Generierung und anschließende Verwendung dieser Forschungsdaten nicht nur nicht beeinträchtigt, sondern gefördert. Alle grundsätzlich in der Lehrveranstaltung behandelten Inhalte, inklusive derer Limitationen und Voraussetzungen, können im Rahmen der Hausübungen an diesen konkreten Forschungsdaten in einem psychologisch bedeutsamen Sinnzusammenhang illustriert und vertieft werden. Die Präsenzeinheiten, die weiterhin der Einführung und Einübung in die Anwendung dieser grundlegenden Verfahren dienen, bleiben davon unberührt. Auch die Beurteilung der individuellen Leistung ist gänzlich von den Ergebnissen des Experiments unabhängig. Die korrekte Anwendung statistischer Verfahren, auf der die individuelle Leistungsbeurteilung beruht, setzt schließlich kein bestimmtes Ergebnis des gemeinsam durchgeführten Experiments voraus.

Ein alternativer Übungsmodus

Eine weniger verspielte, aber vermutlich nicht minder effektive Möglichkeit eine Lehrveranstaltung zur Einübung in die Anwendung statistischer Verfahren zu gestalten, besteht darin, das Konzept des umgekehrten Klassenzimmers noch weiter zu intensivieren. Wie in der oben erläuterten Unterrichtsmethode hätten die Studierenden auch in diesem Fall wöchentlich den Arbeitsauftrag, ein jeweiliges Kapitel dieses Dokuments bis zur nächsten Präsenzübungseinheit vorzubereiten. Diese Vorbereitung würde allerdings ebenfalls die Übungsaufgaben des Kapitels (oder eine Auswahl derselben) umfassen.

Zu Beginn jeder Übungseinheit wäre dann von den Studierenden anzugeben, welche Übungsaufgaben sie so gründlich genug vorbereitet haben, dass sie ihre Lösung anderen präsentieren können (im Idealfall hätten sich natürlich alle auf alle Übungsaufgaben so eingehend vorbereitet, aber es dürfte jeder Person, die jemals gelehrt oder gelernt hat, klar sein, dass dieser Idealfall nur äußerst selten auch der Realität entspricht). Diese Lösungen würden dann in den Präsenzeinheiten von einzelnen Studierenden den anderen Studierenden präsentiert. Dabei hätten die Studierenden die zusätzliche Aufgabe, den übrigen Studierenden so weit Anleitung zu geben, dass diese die Übungsaufgabe begleitend zur Präsentation durchführen können. Dies umfasst insbesondere auch Studierende, die sich selbst nicht auf die jeweilige Aufgabe vorbereiten konnten. Die präsentierenden Studierenden übernehmen dabei also für einzelne Übungsaufgaben die Rolle der Lehrperson, was die Wirkung der Präsentation als Übung zum Lernen durch Lehren intensiviert (Duran, 2017).

Gleichzeitig würde aufgrund der gehobenen Anforderungen während der Präsenzeinheit dieser Übungsmodus auch die Vorbereitungen der Studierenden auf die jeweiligen Präsenzeinheiten intensivieren. Dies würde in Folge das Lernen und Üben über das Semester hinweg über größere Zeiträume verteilen und auf diese Weise sowohl die Einübung in die Inhalte als auch die langfristige Behaltensleistung fördern (Ebersbach et al., 2022).

Auch für diesen Übungsmodus müssten allerdings noch formale Beurteilungskriterien festgelegt werden. Über den Verlauf des Semesters könnte dazu etwa jede:r Studierende verpflichtet werden, mindestens zwei Übungsaufgaben in der beschriebenen Form zu präsentieren. Dabei wurde angenommen, dass ein Kurs 25 Studierende umfasst und pro Präsenzeinheit etwa 5 Übungsaufgaben erschöpfend behandelt werden könnten. Für eine erledigte Präsentation könnte dem:der jeweiligen Studierenden zudem 5 Beurteilungspunkte vergeben werden.

Die übrigen 90 Punkte könnten durch Leistungen bei zwei Klausuren verdient werden. Für die erste Klausur etwa zur Hälfte des Semesters könnten etwa 40 Punkte, für die zweite Klausur am Ende des Semesters 50 Punkte vergeben werden. Bei positivem Abschluss der Lehrveranstaltung ab einer Leistung von 51 Punkten wäre so auch gewährleistet, dass die Lehrveranstaltung einschließlich der zweiten Klausur aktiv besucht werden muss.

Möglicher Ablauf einer prüfungsimmanenten Lehrveranstaltung auf Basis dieses Dokuments

Ein möglicher Ablauf einer Lehrveranstaltung, der die bisherigen Erläuterungen zu diesem Dokument in sich integriert, ist im folgenden Ablaufschema zusammengefasst. Dieses Schema kann zur Gestaltung der eigenen Lehrveranstaltung herangezogen und nötigenfalls entsprechend angepasst werden. Die Themen der einzelnen Kapitel/Wochen können dem Inhaltsverzeichnis dieses Dokuments entnommen werden.

Zeitpunkt/-raum	Kapitel	Aktivität
Woche 1	1	Vorbesprechung und Klärung von Organisatorischem
Wochen 1-2	n.a.	Hausübung 1 (Online-Experiment)
Wochen 1-2	2	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 2	2	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 2-3	3	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 3	3	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 3-4	4	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 4	4	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 4-5	5	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 5	5	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 5-6	2-5	Hausübung 2
Wochen 5-6	6	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 6	6	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 6-7	7	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 7	7	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen (oder erste Klausur im Falle des Alternativmodus)
Wochen 7-8	8	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 8	8	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 8-9	6-8	Hausübung 3
Wochen 8-9	9	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 9	9	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 9-10	10	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 10	10	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 10-11	11	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 11	11	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Wochen 11-12	9-11	Hausübung 4
Wochen 11-12	12	Selbständiges Erarbeiten des jeweiligen Kapitels
Woche 12	12	Präsenzeinheit: Besprechung des jeweiligen Kapitels und Übungen
Woche 13	2-12	Wiederholung, Fragestunde, Übungen aus allen Kapiteln (oder wie bisherige Einheiten im Falle des Alternativmodus)
Wochen 14-15	2-12	Klausur bzw. Ersatzklausur

Eine weitere Möglichkeit der Verwendung dieses Dokuments für eine Lehrveranstaltung

Eine weitere Möglichkeit dieses Dokument zur Gestaltung einer einsemestrigen Lehrveranstaltung zu verwenden, wurde im Studienjahr 2025/2026 an der Universität Graz konzipiert und erprobt. Grundsätzlich baut auch dieses Lehrveranstaltungskonzept auf dem Konzept des umgekehrten Klassenzimmers (Engl.: flipped classroom) auf. Das heißt, die Studierenden (und die Lehrpersonen) erarbeiten jeweils bis zur nächsten Präsenzeinheit ein Kapitel oder Teile eines oder mehrerer Kapitel aus diesem Dokument und die Präsenzeinheiten werden dann zur Wiederholung, Aktivierung, Vertiefung dieser Inhalte verwendet. Zur (Re-)Aktivierung der Lerninhalte werden jeweils zu Beginn der Präsenzeinheiten kurze Quiz durchgeführt, die neben einer Abschlussklausur zur Beurteilung herangezogen werden. Die Quiz sind dabei so konzipiert, dass Personen, die sich zur Vorbereitung auf die Präsenzeinheiten ernsthaft mit den Inhalten auseinandergesetzt haben, belohnt werden, und gleichzeitig die wesentlichen Durchführungsaspekte der Inhalte wiederholt bzw. reaktiviert werden. Die Erfahrungen während der Erprobung dieses Formats zeigten, dass so der Rest der Präsenzeinheiten tatsächlich für inhaltliche Verständnis- und Vertiefungsfragen und Üben der Inhalte an weiteren Beispielen freigemacht werden konnte und die bloße Bedienung der Software sowie rein technische Aspekte hingegen zu größten Teilen bereits in der selbständigen Vorbereitung erlernt werden konnten. Die Quiz waren ferner so konzipiert, dass die Durchführung nur die ersten 10-15 Minuten jeder Einheit in Anspruch nehmen sollte. Danach sollten die Lösungen der Quizfragen in etwa in 15-20 Minuten im Plenum erörtert werden. Das heißt, sowohl die Studierenden als auch die Lehrenden erhielten in jeder Einheit auch Rückmeldung darüber, welche Inhalte bereits gut verstanden werden konnten und welche Inhalte noch weiterer Vertiefung bedurften. Dieser Vertiefung bzw. Ergänzung durch weitere Übungsaufgaben diente dann die verbleibende Zeit jeder Einheit, in welcher Studierende (u.a. miteinander, zu zweit) ausgewählte Übungsaufgaben (aus dem jeweiligen Kapitel) bearbeiteten und sich in der Erstellung von Ergebnisberichten übten. Dies wurden von den Übungsleiter:innen begleitet, die bei individuellen Fragen und Schwierigkeiten unterstützten und Rückmeldung gaben (und bekamen). Ergaben sich Fragen bzw. Einsichten, die für alle wichtig erschienen, wurden diese im Plenum aufgegriffen und erörtert.

Die Erprobung dieses Kursformats ergab auch folgenden Vorschlag für eine Beurteilung bzw. Notenvergabe. Durch jedes Quiz zu Beginn jeder Präsenzeinheit konnten bis zu 5 Punkte verdient werden. Von insgesamt zehn durchgeführten Quiz (siehe auch den im Folgenden im Detail erläuterten Syllabus) wurden für jede:n Studierende:n die acht besten Quiz bis zur Klausur gewertet. Das heißt, bis zur Klausur konnten bis zu maximal 40 Punkte durch Teilnahme an den Quiz erarbeitet werden. Bei der Abschlussklausur konnten bis zu 60 Punkte erreicht werden. Studierenden, die insgesamt mindestens 51 von 100 Punkten erreichten (d.h. aus Quiz und Klausur zusammen), wurden schließlich noch die Punkte aller Quiz, an welchen sie über die acht besten Quiz hinaus teilgenommen hatten, als Bonuspunkte auf die Gesamtpunkte angerechnet. Dadurch sollte ein zusätzlicher Anreiz geboten werden, über die ganze Lehrveranstaltung hinweg engagiert mitzuarbeiten. Die Teilnahme an den Quiz war ausschließlich in Präsenz zu Beginn jeder Einheit möglich.

Bei den Quiz durften sämtliche Lehrveranstaltungsunterlagen verwendet werden (open book Format), bei der Abschlussklausur nicht (closed book Format). Dies hatte zum Hintergrund, dass die Quiz vorrangig Anreize zur Vorbereitung der Inhalte bieten, aber noch nicht deren Beherrschung erfordern sollten, da ja das Erlernen der Inhalte oftmals noch die gemeinsame Erläuterung und Vertiefung erforderte. Bei der Abschlussklausur hingegen ging es tatsächlich um die Prüfung der Fähigkeit sich die Inhalte aneignen und selbstständig ohne Zuhilfenahme weiterer als der erlaubten Hilfsmittel auch wieder zur Anwendung bringen zu können. Die Vorteile eines closed book Formats für diesen Zweck wurden bereits oben erläutert.

Insgesamt gab es im gesamten Kurs demnach 100 Punkte zu verdienen. Ein Bestehen des Kurses erforderte mindestens 51 Punkte (davon maximal 40 durch Teilnahme an den Quiz). Eine Punktezahl von 50 Punkten oder weniger wurde daher mit der Note „Nicht genügend (5)“ bewertet. Bei einer Punktezahl von 51 bis 62 Punkten wurde die Note „Genügend (4)“ vergeben, ab 63 Punkten wurde die Note „Befriedigend (3)“ vergeben, ab 75 Punkten die Note „Gut (2)“, ab 87 Punkten die Note „Sehr gut (1)“. Es wurden lediglich ganz Punkte (d.h., keine Teilpunkte) vergeben.

Im Folgenden ist tabellarisch ein möglicher Ablauf der Lehrveranstaltung im Detail illustriert. Bei der Erstellung wurde dabei von 15 Semesterwochen sowie des Ausfalls einer Woche (etwa durch

gesetzliche Feiertage). Mit einem nötigen Termin für eine Vorbesprechung sowie zwei Terminen für Klausur und Ersatzklausur (ausschließlich im Krankheitsfall oder anderweitig unaufschiebbarer Verhinderung) verbleiben elf Termine. Die ersten zehn dieser Termine sind den Inhalten dieses Dokuments gewidmet. Der verbleibende Termin (Woche vor der Abschlussklausur) ist einer Wiederholung der gesammelten Inhalte z.B. im Format einer Probe- oder Übungsklausur vorbehalten.

Einheit 1 Vorbesprechung / Organisatorisches	Inhalte: • Erklärung Kursablauf, Quiz, Klausur, Beurteilung • Vorstellung SPSS • Bedienung der IT-Infrastruktur (lokale PCs, Netzlaufwerke, Verzeichnisse etc.) • Fernzugriff SPSS Demonstration • Herunterladen der Materialien (dieses Dokument und elektronisches Zusatzmaterial)
Einheit 2 ***Vorzubereiten: Kapitel 2*** ***(insb. S. 37-57)*** Einführung in SPSS	Inhalte: • 10 Minuten Quiz (4 Punkte) • Danach: SPSS Fernzugriff eigenständig öffnen & Überprüfung durch LV-Leitung (1 Punkt) • Wiederholung/Vertiefung (Plenum): Bestandteile von SPSS; SPSS auf Englisch; Datensätze öffnen; Datensätze lesen; Datensätze zusammenfügen • Ausgewählte Übungsaufgaben aus dem Buch • (Optional: Andere Dateitypen (z.B. csv) einlesen)
Einheit 3 ***Vorzubereiten: Kapitel 3*** ***(insb. S. 63-101)*** Datenmanagement & Deskriptive Statistiken	Inhalte: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): Umkodieren von Variablen; Index- oder Skalenbildung; Deskriptive Statistiken: Häufigkeiten, Maßzahlen, Boxplot; Kreuztabelle & Korrelation (Durchführung, ohne Hypothesentest); Formatierung von Ergebnisberichten (APA-Format) • Ausgewählte Übungsaufgaben aus dem Buch • (Optional: Kategorienbildung)
Einheit 4 ***Vorzubereiten: Kapitel 4*** ***(insb. S. 107-126)*** Parameterschätzung und Testen von Hypothesen über Populationsmittelwerte	Inhalt: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): Einstichproben t-Test; Cohens *d*; Stichprobenplanung (Einführung in G*Power) • Ausgewählte Übungsaufgaben aus dem Buch

Einheit 5 ***Vorzubereiten: Kapitel 5*** ***(insb. S. 133-154)*** Schätzung und Testung von Mittelwerts-unterschieden zwischen zwei Gruppen	Inhalte: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): t-Test für abhängige Messungen; t-Test für unabhängige Messungen; Levenes Test; Stichprobenplanung; • Ausgewählte Übungsaufgaben aus dem Buch • (Optional: t-Test für abhängige Messungen als Einstichproben t-Test mit Differenzvariable)
Einheit 6 ***Vorzubereiten: Kapitel 6*** ***(insb. S. 165-183)*** Einfaktorielle ANOVA ohne Messwiederholung	Inhalte: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): Omnibustest; Paarweise post-hoc Vergleiche; • Ausgewählte Übungsaufgaben aus dem Buch • (Optional: Welchs ANOVA; Stichprobenplanung) • [Nicht behandelt (auch nicht vorzubereiten): A-priori Vergleiche (Kontraste)]
Einheit 7 ***Vorzubereiten: Kapitel 7*** ***(insb. S. 195-214)*** Zweifaktorielle ANOVA ohne Messwiederholung	Inhalte: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): Omnibustests; Paarweise post-hoc Vergleiche • Ausgewählte Übungsaufgaben aus dem Buch • (Optional: Stichprobenplanung)
Einheit 8 ***Vorzubereiten: Kapitel 8*** ***(insb. S. 221-237)*** ANOVA mit Messwiederholung	Inhalte: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): Einfaktorielle ANOVA mit Messwiederholung; Mauchlys Test für Sphärizität; Varianzanalyse mit gemischten Design • Ausgewählte Übungsaufgaben aus dem Buch • (Optional: Zweifaktorielle ANOVA mit 2 Messwiederholungsfaktoren; Stichprobenplanung)
Einheit 9 ***Vorzubereiten: Kapitel 9*** ***(insb. S. 249-271)*** Einführung in die Regressionsanalyse: Einfache & Multiple Regression	Inhalte: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): Einfache lineare Regression; Exkurs Zentrierung; Exkurs Korrelation; Multiple lineare Regression • Ausgewählte Übungsaufgaben aus dem Buch • (Optional: Exkurs Standardisierung)

Einheit 10 ***Vorzubereiten: Kapitel 10*** ***(insb. S. 279-297)*** Regressionsdiagnostik & Effektstärken in der MLR	Inhalte: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): Regressionsdiagnostik (Linearität, Normalverteilung, Homoskedastizität); Ausreißeranalyse (Cook'sche Distanz); Effektstärken (R^2, Standardisiertes Regressionsgewicht, Quadrierte Semipartialkorrelation); Stichprobenplanung; • Ausgewählte Übungsaufgaben aus dem Buch • [Nicht behandelt (auch nicht vorzubereiten): Kollinearität; Gerichtete azyklische Graphen (DAGs)]
Einheit 11 ***Vorzubereiten: Kapitel 11*** ***(insb. S. 321-325, 328-329, 334-337)*** Diskrete Prädiktoren und Interaktion in der MLR (Moderation)	Inhalte: • 10 Minuten Quiz (5 Punkte) • Wiederholung/Vertiefung (Plenum): Regressionsanalyse mit diskretem Prädiktor mit 2 Ausprägungen; Dummy-Kodierung; Vergleich zu unabhängigem t-Test; Regressionsanalyse mit Interaktion zwischen 1 stetigen & 1 dichotomen Prädiktor • Ausgewählte Übungsaufgaben aus dem Buch • [Nicht behandelt (auch nicht vorzubereiten): Regressionsanalyse mit diskretem Prädiktor mit mehr als 2 Ausprägungen; Regressionsanalyse mit Interaktion zwischen 2 diskreten Prädiktoren; Regressionsanalyse mit Interaktion zwischen 2 stetigen Prädiktoren]
Einheit 12	Wiederholungseinheit, Fragestunde, Probe-/Übungsklausur
Einheit 13	Schriftliche Klausur (90 Minuten)
Einheit 14	Ersatzklausur

Übungsaufgaben

In diesem Kapitel ging es noch um keine konkreten statistischen Inhalte, sondern lediglich um den konzeptuellen Rahmen und die Verwendungsmöglichkeiten dieses Dokuments. Zur Illustration der künftigen Kapitel werden aber schon hier einige Aufgaben bereitgestellt, die einerseits grundlegende Begriffe in Erinnerung rufen sollen, und andererseits zur Illustration des Ablaufs des Übungsteils der Präsenzeinheiten einer entsprechenden Lehrveranstaltung im Rahmen einer Vorbesprechung in der ersten Präsenzeinheit verwendet werden können. Für letztere Verwendungsart wird empfohlen die erste Einheit auch gleich zur Bildung der Kleingruppen zu verwenden und dafür die nötige Zeit einzuräumen.

Beispiel 1.1

Es kommt immer wieder vor, dass Studierende im Rahmen ihrer Masterarbeit Aussagen wie die folgende treffen: „Es gibt einfach keine Quelle, in der etwas zu Voraussetzungen für Varianzanalysen steht, deshalb habe ich dann einfach diese Internetseite zitiert, weil dort steht, dass…". Dann muss ich

(wenn ich wieder einmal in den Genuss gekommen bin bei der Betreuung einer Masterarbeit unterstützen zu dürfen) wieder lang und breit erklären, weshalb Internetquellen in vielen Fällen keine optimalen Quellen für eine wissenschaftliche Arbeit sind und dass solcherlei grundlegende statistische Inhalte durchaus in den meisten einschlägigen Statistiklehrbüchern zu finden sind. Um dieser Herausforderung zumindest etwas vorzubeugen, nun diese Frage: In welchem der folgenden Bücher könnten Sie bezüglich statistischer Grundkenntnisse fündig werden?

(a) Eid, M., Gollwitzer, M. & Schmitt, M. (2017). Statistik und Forschungsmethoden (5. korrigierte Auflage). Beltz. Permalink für Ebook Version: https://permalink.obvsg.at/UGR/AC15718869.

(b) Bühner, M. & Ziegler, M. (2017) Statistik für Psychologen und Sozialwissenschaftler: Grundlagen und Umsetzung mit SPSS und R (2., aktualisierte und erweiterte Auflage). Pearson.

(c) Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE publications.

(d) Solche Bücher gibt es nicht. Hilfe bei statistischen Fragen bekommt man ausschließlich auf www.statistik-guru.de.

Beispiel 1.2

Ordnen sie die vier Begriffe „Merkmal“, „Merkmalsausprägung“, „Variable“, „Variablenwert“ den passenden Stellen (markiert mit „???“) in der folgenden Abbildung zu.

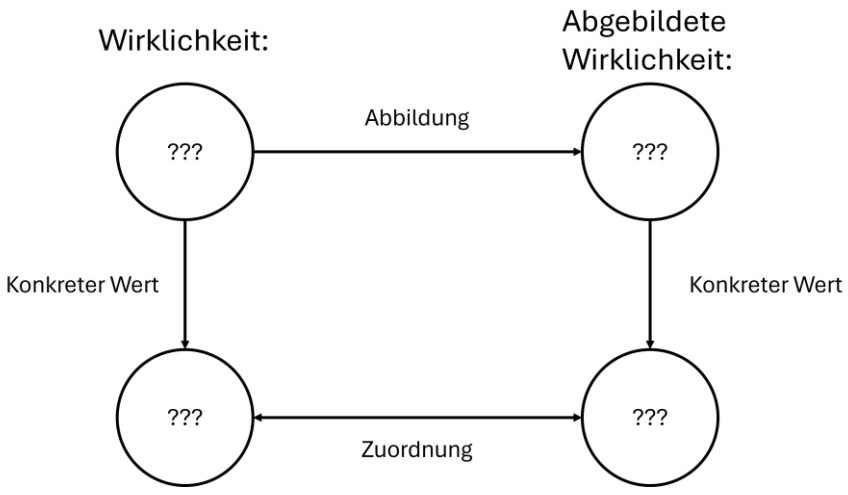


Abbildung 1.1. Welche Begriffe gehören jeweils an die Stelle der „???“?

Beispiel 1.3

Was ist ein Synonym für welchen Begriff?

Begriffe: Merkmalsträger:in, Merkmalsausprägung, Variablenwert.

Synonyme: Untersuchungseinheit, Messwert, Merkmalswert, Untersuchungsobjekt.

Beispiel 1.4

Welche der folgenden Aussagen trifft/treffen zu?

(a) Bei kategorialen Variablen kann es sich um Variablen mit Nominalskalenniveau oder Ordinalskalenniveau handeln.
(b) Bei metrischen Variablen kann es sich um Variablen mit Absolutskalenniveau oder Ordinalskalenniveau handeln.
(c) Bei metrischen Variablen kann es sich um Variablen mit Intervallskalenniveau, Verhältnisskalenniveau oder Absolutskalenniveau handeln.
(d) Die Variablenwerte von Variablen mit Intervallskalenniveau können Zahlen oder Begriffe sein.

Beispiel 1.5

Geben Sie für jedes der folgenden Skalenniveaus mindestens ein Beispiel an: Nominalskalenniveau, Ordinalskalenniveau, Intervallskalenniveau, Verhältnisskalenniveau, Absolutskalenniveau.

Beispiel 1.6

Welche Aussage/n trifft/treffen in Bezug auf die Hypothese „Alkoholkonsum auf Partys steigert die Extraversion“ zu?

(a) Alkoholkonsum ist hier die abhängige Variable (AV).
(b) Alkoholkonsum ist hier die unabhängige Variable (UV).
(c) Extraversion ist hier die abhängige Variable (AV).
(d) Extraversion ist hier die unabhängige Variable (UV).

Beispiel 1.7

Geben Sie jeweils einige Beispiele für diskrete und kontinuierliche Variablen.

Beispiel 1.8

Welche Aussage/n trifft/treffen zu?

(a) Diskrete Variablen müssen kategorial sein.

(b) Metrische Variablen müssen kontinuierlich sein.

(c) Eine Variable kann gleichzeitig kategorial, diskret und abhängig sein.

(d) Eine Variable kann gleichzeitig diskret, metrisch und abhängig sein.

Beispiel 1.9

Formulieren Sie eine Hypothese, die eine kategoriale, diskrete und abhängige Variable beinhält.

Beispiel 1.10

Formulieren Sie eine Hypothese, die eine metrische, diskrete und unabhängige Variable beinhält.

Beispiel 1.11

Beschreiben Sie in eigenen Worten den Begriff „Urliste“.

Beispiel 1.12

Wie kann die relative Häufigkeit $h(x_j)$ der Messwertausprägung x_j aus der absoluten Häufigkeit $H(x_j)$ und der Gesamtanzahl an Messwerten n berechnet werden?

(a) $h(x_j) = H(x_j) \cdot n.$
(b) $h(x_j) = H(x_j) - n.$
(c) $h(x_j) = H(x_j) + n.$
(d) $h(x_j) = H(x_j)/n.$

Beispiel 1.13

Gegeben ist die folgende Häufigkeitstabelle, bei der sinnvollerweise bereits alle aufgetretenen unterschiedlichen Messwertausprägungen in aufsteigender Reihenfolge angeordnet wurden (zur einfacheren Darstellung sind für diese Übung nur die ersten vier Zeilen der Tabelle angeführt).

Anzahl Liegestütz	Absolute Häufigkeit
5	1
9	2
11	4
12	1
...	...

Bemerkung. $n = 50$.

Berechnen Sie (a) die absolute kumulierte Häufigkeit sowie (b) die relative kumulierte Häufigkeit der viertkleinsten Messwertausprägung (d.h. der letzten in der Tabelle noch ersichtlichen Messwertausprägung). Beantworten Sie schließlich noch folgende Frage: (c) Welcher Anteil (in %) der getesteten Schüler:innen führte weniger als 10 Liegestütz durch?

Beispiel 1.14

Berechnen Sie die relative Häufigkeit sowie die absolute und die relative kumulierte Häufigkeit für die Häufigkeitstabelle des vorhergehenden Beispiels auch für die anderen dargestellten Messwertausprägungen und ergänzen Sie die Tabelle um zwei entsprechende Spalten.

Beispiel 1.15

Welche der folgenden Aussagen trifft/treffen zu?

(a) (Mindestens) 50% der Merkmalsträger:innen haben einen Messwert, der kleiner oder gleich dem Median ist.

(b) (Mindestens) 50% der Merkmalsträger:innen haben einen Messwert, der größer oder gleich dem Median ist.

(c) Bei einer unimodalen, rechtsschiefen Verteilung befindet sich der Median üblicherweise links vom Mittelwert.

(d) Bei einer unimodalen, rechtsschiefen Verteilung befindet sich der Median üblicherweise rechts vom Mittelwert.

Kapitel 2

SPSS. Was ist das und wie kann ich es verwenden?

Nadine Schmer, Stefan E. Huber

Zur Verarbeitung großer Datenmengen wird heutzutage auf die Hilfe digitaler Computer zurückgegriffen. Dazu wird häufig auf statistische Analysen spezialisierte Software wie das kommerzielle Programmpaket SPSS, das von IBM vertrieben wird, zurückgegriffen (Blanca et al., 2018). Auch zur Lösung der in diesem Dokument gesammelten Übungsbeispiele wird hauptsächlich (aber nicht ausschließlich) die Software SPSS verwendet. Aus diesem Grund wird in diesem Kapitel eine Einführung in das Programm SPSS gegeben und die grundlegende Bedienung erläutert.

SPSS? Was ist das?

SPSS ist eine kommerzielle Software für statistische Datenanalyse. Die Abkürzung SPSS steht für "Statistical Package for the Social Sciences" und gehört zur IBM-Produktreihe unter dem Namen IBM SPSS Statistics. Sie wird nach wie vor häufig in Forschung, Bildung und kommerziellen Anwendungen eingesetzt (Blanca et al., 2018), insbesondere in den Sozialwissenschaften, aber auch in anderen Disziplinen wie Wirtschaft, Medizin, Marktforschung und der Psychologie.

Die Funktionen von SPSS sind weitreichend, angefangen von Datenmanagement, statistischen Analysen, Visualisierung bis hin zu Prognosen. Außerdem gibt es Erweiterungen für spezielle Anwendungen wie Textanalyse und es ist integrierbar mit Programmiersprachen wie Python und R. Zudem ist SPSS benutzerfreundlich, recht intuitiv, und setzt keine Programmierkenntnisse voraus, was den Einstieg in die Verwendung der Software zur statistischen Datenanalyse erleichtert.

Allerdings handelt es sich bei SPSS um eine kommerzielle Software. Das heißt, wer SPSS verwenden will, muss die Software bzw. eine Lizenz für ihre Verwendung erst käuflich erwerben. Damit Studierende diese zum Erlernen der Verwendung aber nicht gleich kaufen müssen, haben viele Universitäten entsprechende Lizenzen erstanden, die für Studierende einen Fernzugriff auf die Software von zu Hause aus ermöglichen. Für Studierende der Universität Graz (Stand: März, 2026) ist diese Möglichkeit im folgenden Abschnitt beschrieben.

Wie kann man SPSS (von zu Hause aus) verwenden?

Ob für die Vorbereitung auf die Präsenzeinheiten einer entsprechenden Lehrveranstaltung, die Vorbereitung auf Prüfungen oder Klausuren, oder einfach zum Ausprobieren oder Festigen der eigenen Kenntnisse, in jedem Fall brauchen Studierende, die den Umgang mit der Software erlernen sollen, Zugriff auf SPSS. Die bequemste Möglichkeit mag es zwar vielleicht durchaus sein, sich eine relativ günstige Studierendenlizenz zu kaufen (viele Universitäten bieten eine solche Möglichkeit über entsprechende Software-Portale), aber zumindest an der Universität Graz (und auch vielen anderen Universitäten) ist das nicht notwendig. Als Studierende:r der Universität Graz kann SPSS an jedem beliebigen Computer mit Internetverbindung über den Terminalserver der Universität Graz genutzt werden. Dafür muss ein sogenannter Fernzugriff eingerichtet werden. Dieser wird im Folgenden für die Betriebssysteme MS Windows und Mac OS beschrieben.

Fernzugriff auf SPSS für MS Windows

Für den Fernzugriff zu SPSS wird eine VPN-Verbindung benötigt. Eine schrittweise Anleitung der Universität Graz zur Herstellung einer VPN-Verbindung finden Sie unter: https://static.uni-graz.at/fileadmin/uni-it/docs/VPN_Netzzugang_unter_Windows_mit_AnyConnect_secure.pdf.

Wenn Ihr Computer mit dem VPN verbunden ist, besuchen Sie die Website der Universität Graz zu IT-Services für Studierende unter https://it.uni-graz.at/de/. Hier klicken Sie auf das Feld „Ich möchte“. Auf der sich öffnenden Seite finden Sie etwas weiter unten die Schaltfläche „SPSS virtuell verwenden“, siehe Abbildung 2.1. Betätigen Sie diese Schaltfläche.

Durch Betätigung der Schaltfläche öffnet sich eine Seite, über die Sie Software über die Universität Graz beziehen können, und u.a. auch virtuelle Software wie SPSS starten können. Letzteres können Sie tun, indem Sie die Schaltfläche „virtuelle Software starten“ betätigen.

Im nächsten Schritt werden Sie gebeten Ihre E-Mail-Adresse (jene, die Sie als Studierende:r der Universität Graz erhalten haben) und Ihr Benutzerkennwort (ebenfalls jenes für die Universität Graz) einzugeben. Diese Aufforderung ist in Abbildung 2.2 gezeigt.

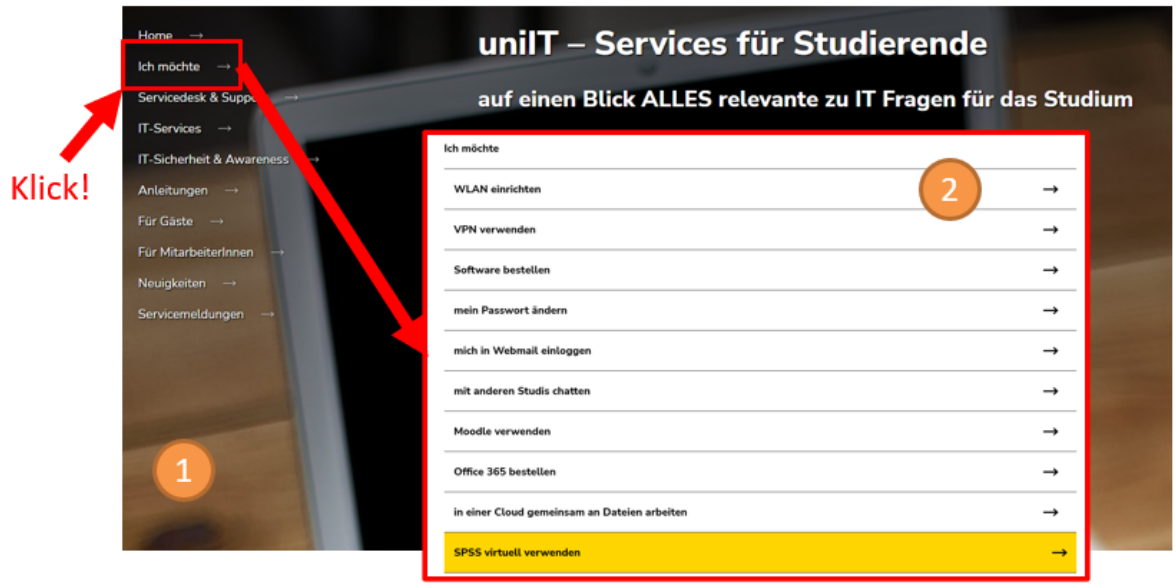


Abbildung 2.1. Website „uniIT – Services für Studierende“ der Universität Graz. Hier finden sich auch Zugang zur Software SPSS über den sogenannten Fernzugriff.

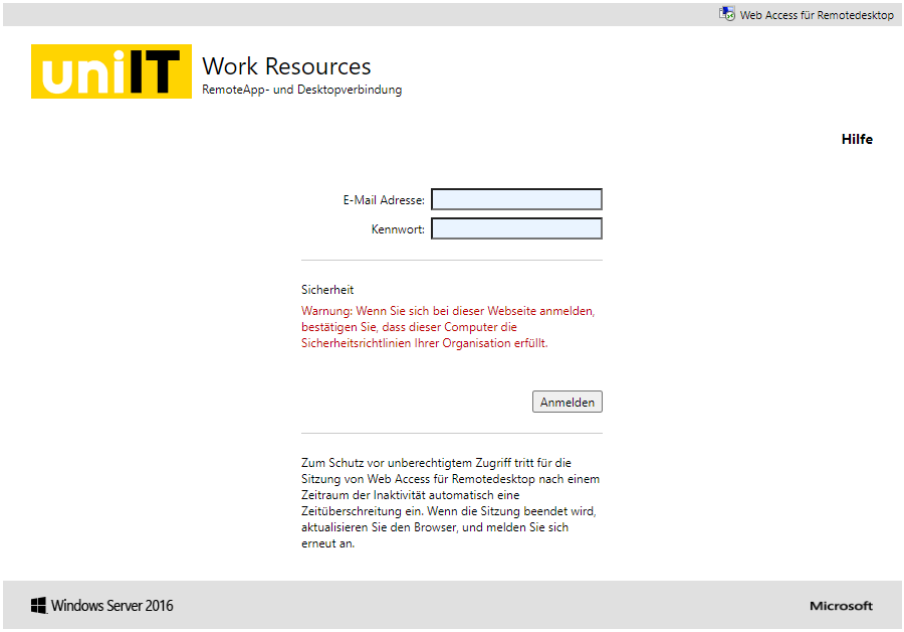


Abbildung 2.2. Eingabeaufforderung (E-Mail und Kennwort) für den Zugriff auf virtuelle Software.

Nachdem Sie sich angemeldet haben, können Sie zwischen zwei Ordnern auswählen. Wählen Sie hier den Ordner „SPSS“ und anschließend im sich öffnenden Unterordner die Anwendung „IBM SPSS Statistics“. Daraufhin wird eine ausführbare Datei (mit Endung „rdp“) heruntergeladen und typischerweise im Ordner „Downloads“ abgespeichert. Führen Sie diese Datei aus und klicken Sie anschließend auf die Schaltfläche „Verbinden“. Daraufhin werden Sie noch einmal gebeten Ihre E-Mail-

Adresse sowie Ihr Kennwort einzugeben. Nach der Eingabe wird der Fernzugriff schließlich gestartet. Die Einrichtung desselben kann allerdings einige Momente dauern. Sobald der Fernzugriff eingerichtet ist, öffnet sich das Programm SPSS.

Fernzugriff auf SPSS für Mac OS

Auf Mac OS benötigen Sie eine VPN-Verbindung, um SPSS zu starten. Für eine Schritt-für-Schritt Anleitung zur Herstellung einer VPN-Verbindung öffnen sie den folgenden Link: https://it.uni-graz.at/de/anleitungen/detail/?tx_news_pi1%5Baction%5D=detail&tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Bnews%5D=99932&cHash=a28185975f10ea776b23360eb7e8fa23.

Wenn ihr Computer mit dem VPN verbunden ist, besuchen Sie die Website der Universität Graz zu IT-Services für Studierende unter https://it.uni-graz.at/de/. Hier klicken Sie auf das Feld für „Anleitungen“. Auf dieser Seite geben Sie „SPSS“ im Suchfeld ein und filtern Sie bei Betriebssystemen nach Mac OS wie in Abbildung 2.3 dargestellt.

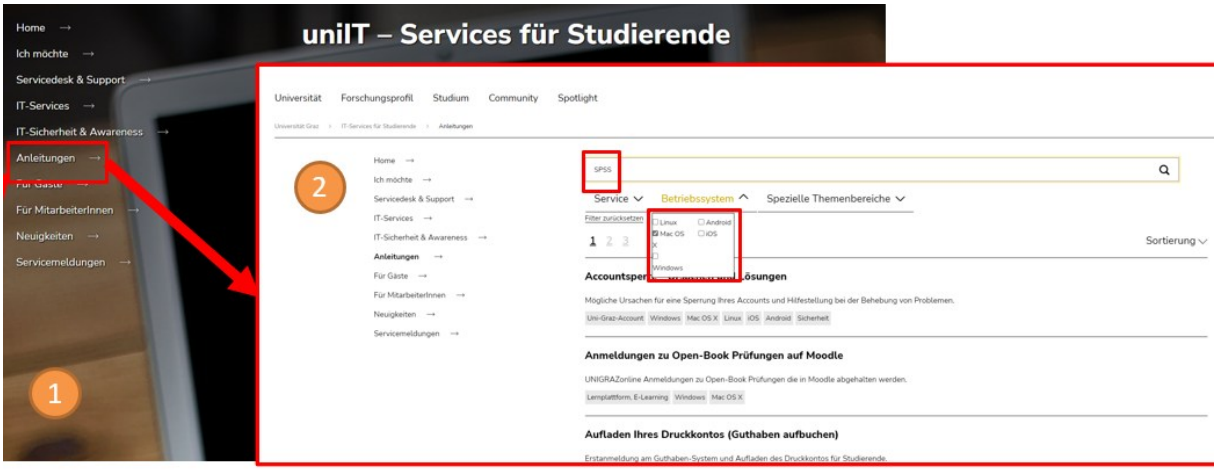


Abbildung 2.3. Suchen nach einer Anleitung für den Fernzugriff auf SPSS für MacOS.

Die Suche sollte nur in einer Anleitung mit dem Titel „RDS (Remote Desktop Services) unter Mac OS“ resultieren. Wählen Sie diese Anleitung aus. Im sich öffnenden Fenster finden Sie einen Link zu einer ausführlichen Anleitung mit detaillierten Screenshots. Wählen Sie diese Anleitung aus und befolgen Sie sie Schritt für Schritt. Dabei kann es sein, dass die in der Anleitung verwendete Bezeichnung „Remote Services“ auf Ihrem System eventuell „Add Workspace“ heißt (siehe auch Abbildung 2.4). Im Laufe der Einrichtung des Fernzugriffs wird es auch für Mac OS (zweimal) nötig sein, Ihre Kenndaten (E-Mail-Adresse und Kennwort) für die Universität Graz einzugeben.

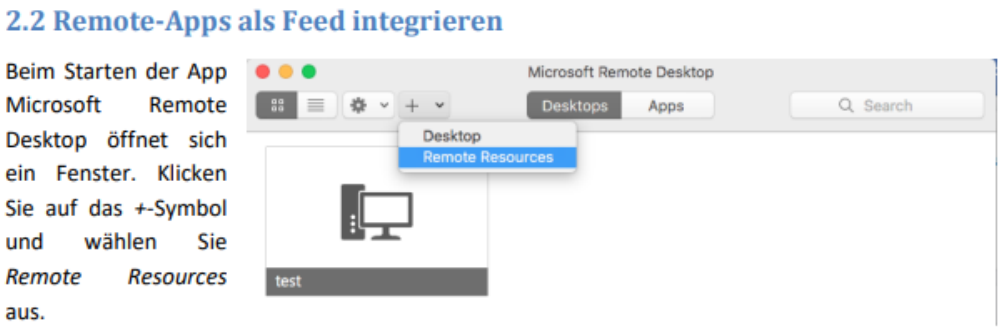


Abbildung 2.4. Es kann sein, dass auf Ihrem System anstelle der Bezeichnung „Remote Resources“ (hier gezeigt) die Bezeichnung „Add Workspace“ verwendet wird. Lassen Sie sich davon nicht verunsichern und wählen Sie „Add Workspace“ aus.

Nützliches im Zusammenhang mit dem Fernzugriff auf SPSS

Beim nächsten Mal, wenn Sie Fernzugriff auf SPSS benötigen, ist es nicht mehr nötig alle oben beschriebenen Schritte durchzuführen. Sobald Sie eine VPN-Verbindung eingerichtet haben, können Sie normalerweise einfach die, wie oben beschrieben, heruntergeladene rdp-Datei ausführen und der Fernzugriff sollte gestartet werden. Falls Sie die Datei gelöscht haben oder der Fernzugriff wider Erwarten nicht gestartet wird, führen Sie einfach die oben beschriebenen Schritte erneut aus.

Beim Fernzugriff auf SPSS kann es zudem eine Herausforderung sein, auf lokale Dateien am Computer zuzugreifen. Prinzipiell sollte dies möglich sein. Wenn Sie auf das Symbol zum Öffnen neuer Dateien in SPSS klicken, befinden Sie sich zwar in Ihrem Homeverzeichnis im Universitätsnetzwerk, Sie sollten aber beispielsweise den lokalen Ordner „Dokumente“ unter „Dieser PC\C auf <Gerätebezeichnung>\Users\<Benutzername>\Documents“ auffinden können. Analog sollte sich der Ordner „Downloads“ finden lassen. Sofern vorhanden, können die benötigten Dateien auch auf einen USB-Stick gespeichert werden, der sich jedenfalls unter den gelisteten, lokalen Laufwerken auffinden lassen sollte.

Zugriff auf SPSS in der Lehrveranstaltung „Anwendung statistischer Verfahren am Computer“

Die Präsenzeinheiten der Lehrveranstaltung „Anwendung statistischer Verfahren am Computer“ finden in Computerräumen statt, die mit Computern ausgestattet sind, auf denen eine lokale SPSS-Installation vorhanden ist. Das heißt, in den Präsenzeinheiten können Sie SPSS verwenden, indem Sie einfach in der Windows-Suchfunktion SPSS eingeben und das daraufhin aufscheinende Programm „IBM SPSS Statistics“ starten.

Erstmalige Verwendung von SPSS

Wenn Sie SPSS zum ersten Mal starten (und auch bei jedem weiteren Start, sofern Sie die Option nicht auswählen, dass Ihnen das Fenster nicht wieder angezeigt wird), werden Sie von einem Dialogfenster willkommen geheißen, in dem Sie u.a. die Option haben, ein neue (leere) Datendatei zu erzeugen oder kürzlich geöffnete Dateien wieder zu öffnen, siehe Abbildung 2.5. Zudem finden Sie hier auch Links zu Hilfe- und Supportseiten oder auch Tutorials im Internet.

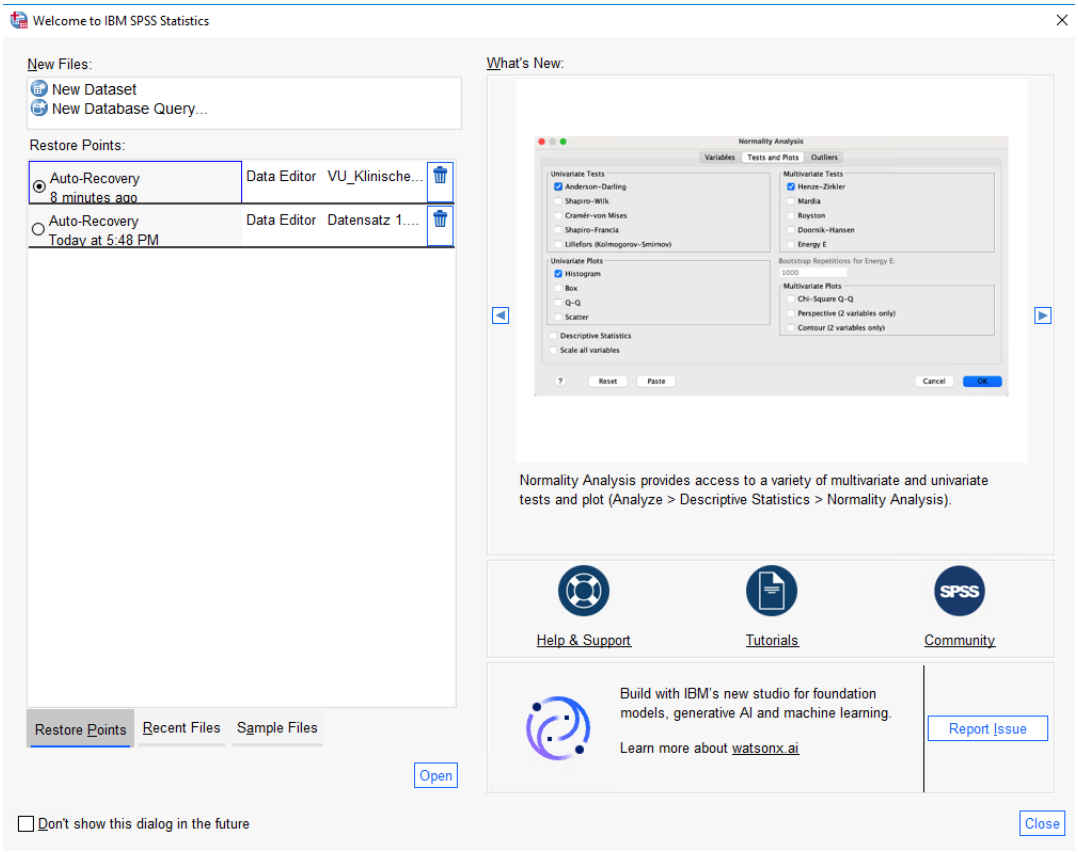


Abbildung 2.5. Dialogfenster beim (erstmaligen) Start von SPSS.

Falls Sie diese Funktionen nicht benötigen, können Sie dieses Fenster einfach schließen. Daraufhin haben Sie eine leere Datendatei im sogenannten Dateneditor vor sich auf dem Bildschirm. Der Dateneditor ist eines von drei wesentlichen Programmfenstern in SPSS. Bei den anderen beiden handelt es sich um die sogenannte Syntax und das Ausgabefenster. Jedes dieser Programmfenster ist auch mit einem eigenen entsprechenden Dateiformat verbunden, die jeweils durch eine eigene Dateiendung ausgezeichnet sind. SPSS-Datendateien, die sie im Dateneditor erstellen, öffnen, bearbeiten und speichern können, haben die Endung „.sav“. Syntaxdateien haben die Endung „.sps“. Ausgabedateien haben die Endung „.spv“. Keine Sorge, falls das alles noch sehr abstrakt klingt, zu Funktionalität und Verwendung der einzelnen Dateiformate bzw. Programmfenster kommen wir bald! Um sich aber einmal einen ersten Überblick über die einzelnen Fenster zu verschaffen, können Sie einfach einmal eine neue Syntaxdatei unter *File >> New >> Syntax* sowie eine neue Ausgabedatei unter *File >> New >> Output* öffnen. (Neue leere Dateien dieser beiden Arten werden Sie im Regelfall nur selten brauchen, wie wir unten noch sehen werden, aber um einmal ein bisschen mit der Software vertraut zu werden, schadet es nicht, sich einmal ein bisschen umzuschauen.)

Womöglich haben Sie bemerkt, dass in der vorhergehenden Anleitung, um die beiden Fenster zu öffnen, die englische Sprache verwendet wurde. Da sehr viele Tutorials und Literatur, auf die Sie im Internet Zugriff haben, auf Englisch zur Verfügung stehen und es sich dabei um wesentlich mehr Ressourcen zur Unterstützung handelt als Sie auf Deutsch finden können, empfiehlt es sich SPSS gleich von Anfang an in englischer Sprache zu nutzen (es kann sein, dass über den Fernzugriff bereits die englische Sprachversion bereitgestellt wird). Daher empfiehlt es sich die Sprache gleich bei der ersten Verwendung umzustellen. Hierzu klicken Sie oben links in der Ecke auf „Bearbeiten“ und dann auf das unterste Feld „Optionen“. Dann wählen Sie im Reiter „Sprache“, sowohl für die Ausgabe als auch für die Benutzerschnittstelle Englisch aus.

Trotz der Umstellung der Sprache auf Englisch bleibt allerdings das Dezimaltrennzeichen in SPSS bei Ein- und Ausgabe ein Komma, in der Syntax wird es hingegen als Punkt dargestellt. In der englischsprachigen Literatur ist es allerdings üblich als Dezimaltrennzeichen (fast) durchwegs einen Punkt zu verwenden, während ein Komma ein sogenanntes Tausendertrennzeichen darstellen kann. Um etwaige Verwirrungen diesbezüglich gleich von Anfang an zu vermeiden, empfiehlt es sich auch diese

Einstellung des Dezimaltrennzeichens gleich zu Beginn zu vereinheitlichen. Dazu kann gleich das bereits geöffnete Syntaxfenster genutzt werden.

Wählen Sie dazu dieses Fenster aus und schreiben Sie in die erste Zeile (ohne den Anführungszeichen): „*Set locale to English (dot as decimal separator).“ Achten Sie dabei darauf, den Stern am Beginn und den Punkt am Ende nicht zu vergessen. Schreiben Sie dann in die nächste Zeile „set locale 'en_us'.“. Beachten Sie hierbei, dass Sie wiederum den Punkt am Ende sowie die eingestrichenen Anführungszeichen innerhalb der Zeichenfolge nicht vergessen. Ihr Syntaxfenster sollte dann so wie in Abbildung 2.6 links aussehen. Markieren Sie nun die beiden Zeilen (prinzipiell genügt es die zweite Zeile zu markieren) und klicken Sie auf die grüne „Abspielen“-Taste wie in Abbildung 2.6 rechts dargestellt. Daraufhin werden die Kommandos in den markierten Zeilen ausgeführt und falls Sie dieser Anleitung bis hierher gefolgt sind: Glückwunsch, Sie haben gerade Ihr erstes eigenes Programm in der SPSS-eigenen Programmiersprache SPSS Syntax geschrieben! Was Sie da genau gemacht haben, wird im nächsten Abschnitt noch etwas weiter erläutert. Sie können aber dieses Erfolgserlebnis jetzt sofort damit feiern, dass Sie Ihr erstes Programm im Syntax-Fenster unter *File >> Save As…* an einem Speicherort und mit einem Dateinamen Ihrer Wahl abspeichern. Wir werden im Rahmen der Übungen nur selten (wenn überhaupt) direkt etwas im Syntax-Editor programmieren, aber wir werden diesen häufig nutzen, um unsere Analysen zu dokumentieren. Darauf wird im nächsten Kapitel noch genauer eingegangen.

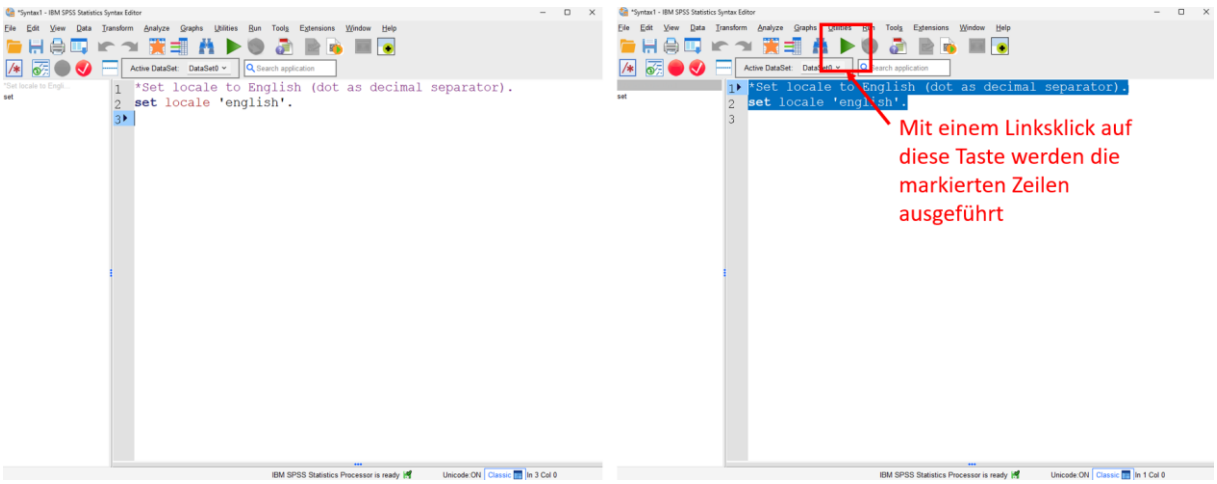


Abbildung 2.6. Ein erstes Programm in der SPSS-eigenen Programmiersprache!

SPSS Syntax und Syntaxdateien

Der Syntax-Editor wird zum Erstellen und Bearbeiten von Programmen in der SPSS-eigenen Programmiersprache SPSS Syntax verwendet. Prinzipiell kann jede Analyse (und überhaupt alles, was man mit SPSS machen kann) durch Eingabe entsprechender Kommandos und deren Ausführung im Syntax-Editor durchgeführt werden.

Der Syntax-Editor ist wie folgt aufgebaut. Links befindet sich der Navigationsbereich, in dem zwischen den einzelnen bereits eingegebenen Syntax-Befehlen hin- und hergesprungen werden kann. Rechts ist der Bereich, in dem die SPSS-Kommandos eigegeben werden können.

Sämtliche SPSS-Kommandos müssen mit einem Punkt („.“) abgeschlossen werden. Zeilen, die mit einem Stern („*“) beginnen, werden von SPSS als Kommentare interpretiert und bei der Ausführung ignoriert. Kommentare dienen lediglich menschlichen Nutzer:innen, um den Programmcode zu erläutern. Auch bei Kommentaren zeigt ein Punkt am Ende SPSS an, dass der Kommentar zu Ende ist. Alternativ kann nach einem Kommentar auch eine Leerzeile gelassen werden. Durch Klicken auf das grüne „Abspielen“-Symbol werden markierte Zeilen ausgeführt. Kommentare werden dabei ignoriert (und können also einfach mitmarkiert werden).

Über *Edit >> Options* gelangen Sie zu den Optionen und können im Reiter Syntax Editor (siehe Abbildung 2.7) u.a. einstellen, welche Farbkodierung Sie für verschiedene Teile der Syntax bevorzugen. Hier wird empfohlen die Farbe zur Darstellung von Kommentaren auf eine sichtbarere als hellgrau zu ändern, welche als Standard eingestellt ist.

Zudem wird zwecks besserer Lesbarkeit auch empfohlen, den im Syntax-Editor verwendeten Font unter *View >> Fonts…* auf Courier New zu ändern.

SPSS Ausgabefenster

Im sog. SPSS-Viewer (Ausgabefenster) werden Ergebnisse in tabellarischer oder grafischer Form ausgegeben. Sämtliche Tabellen und Grafiken sind prinzipiell bearbeitbar. Dieses Fenster werden wir später bei den einzelnen Analysen noch im Detail besprechen. Aktuell sollte es noch sehr leer aussehen, da wir noch keine Datenverarbeitungen durchgeführt haben, bei denen etwas (z.B. Resultate von Signifikanztests) auszugeben war.

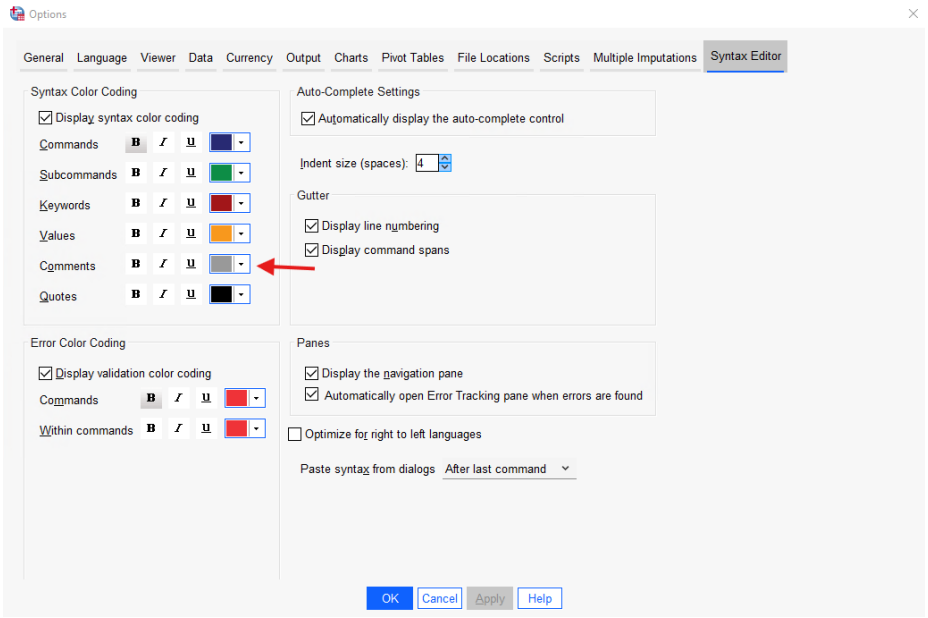


Abbildung 2.7. Einstellungen für den Syntax-Editor. Der rote Pfeil zeigt an, wo die Farbe geändert werden kann, in der Kommentare dargestellt werden.

Der SPSS Daten-Editor

Der Daten-Editor öffnet sich per Voreinstellung mit einer leeren Datendatei, sobald Sie SPSS öffnen (und dabei nicht bereits eine vorhandene Datendatei ausgewählt haben). Im Daten-Editor können Datendateien erstellt, geöffnet, eingesehen, Daten geändert und Variablendefinitionen vorgenommen werden. Sie können mit SPSS mehrere Datendateien gleichzeitig öffnen. Dabei ist zu beachten, dass SPSS immer nur mit dem aktuell aktiven Datensatz arbeitet. Dieser ist an einem (sehr) kleinen roten Plus-Symbol über dem SPSS-Icon (in der oberen linken Ecke) zu erkennen, siehe Abbildung 2.8.

Der Daten-Editor enthält wiederum zwei Modi zur Ansicht von Daten: die sog. Datenansicht und die Variablenansicht. Zwischen diesen beiden Ansichten können mit den beiden Schaltflächen unten links hin- und herwechseln, siehe Abbildung 2.8. Ab Version 30 von SPSS gibt es auch noch einen Übersichtsmodus, auf den hier nicht weiter eingegangen wird.

Variablenansicht. In der Variablenansicht werden die Eigenschaften der Variablen, die im Datensatz enthalten sind, dargestellt bzw. definiert (siehe nächster Abschnitt). In unserem Fall ist die Variablenansicht noch leer (Abbildung 2.8), da noch keine Daten eingegeben wurden bzw. die Variablen noch nicht definiert sind.

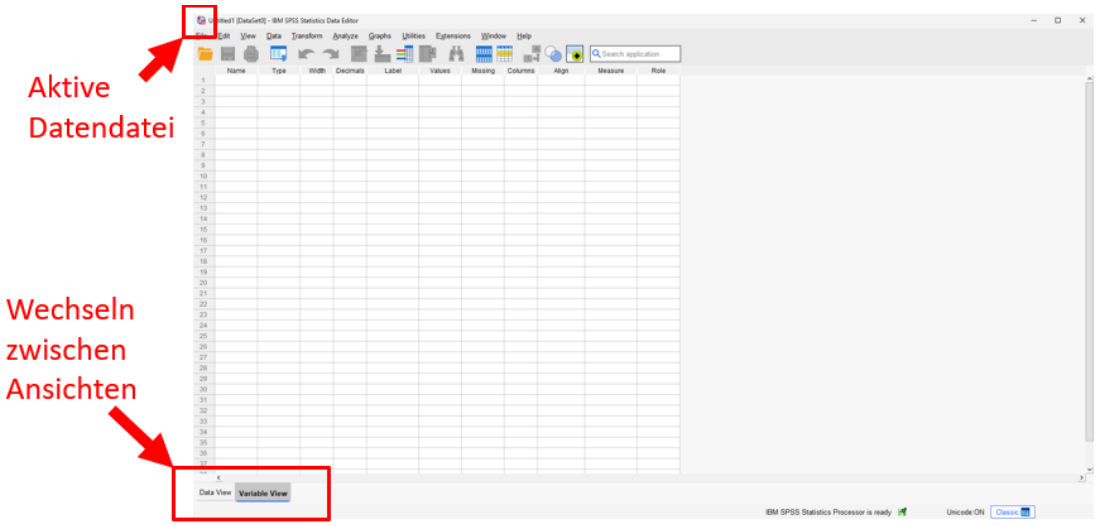


Abbildung 2.8. Eine geöffnete, wenn auch noch sehr übersichtliche Datendatei. Dass dies die aktuell aktive Datendatei ist, erkennt man am roten Plus-Symbol oben links. Aktuell befinden wir uns in der Variablenansicht (dunkel hinterlegter Hintergrund). Zur Datenansicht können wir mit der entsprechenden Schaltfläche unten links umschalten.

Datenansicht. Abbildung 2.9 zeigt die Datenansicht unserer noch leeren Datendatei. Jede Zeile ist jeweils ein Fall (oder eine Beobachtung oder eine Person). Jede Spalte ist jeweils eine Variable (oder ein Messwert). Das bedeutet, alle Informationen zu einem Fall (einer Person) befinden sich in derselben Zeile und die Variablenausprägungen aller Personen in Bezug auf eine Variable befinden sich jeweils in derselben Spalte.

Dateneingabe

Variablen in SPSS können entweder händisch in der Datenansicht eingetippt oder aus anderen Datenquellen (Textdateien, MS Excel-Dateien, Datenbanken, etc.) importiert werden. Aber ganz gleich woher man diese Daten hat, in jedem Fall ist es wichtig diese Variablen zu definieren und deren Eigenschaften festzulegen.

Dazu benötigt man die Variablenansicht. Zur Erklärung der möglichen Eigenschaften werden hier die in Abbildung 2.10 bereits vordefinierten Variablen verwendet. Insgesamt liegen 6 Variablen vor. Jede Variable wird mit ihren Eigenschaften entlang einer Zeile dargestellt. Zur Einsicht oder Änderung der Eigenschaften klickt man mit der Maus (links) in die jeweilige Zelle in der Variablenansicht.

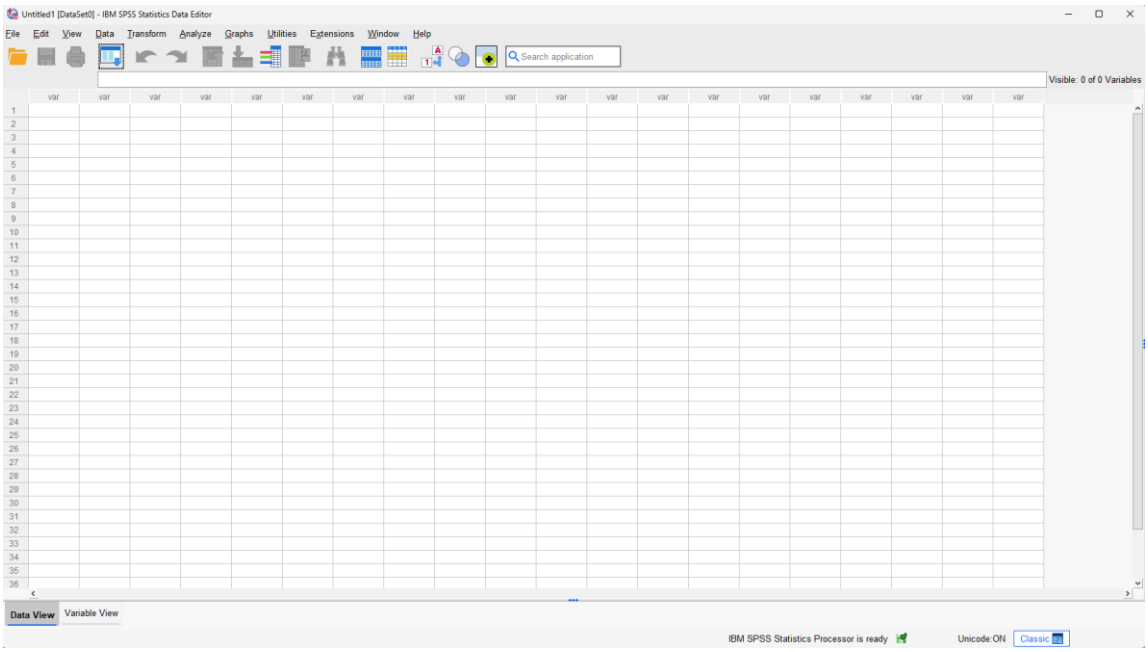


Abbildung 2.9. Eine ebenfalls noch sehr übersichtliche Datenansicht.

Im Folgenden sind die einzelnen Variableneigenschaften kurz erläutert. Vorweg: Sie müssen die Details dieser Erläuterungen nicht auswendig lernen. Wir werden im Rahmen dieses Übungsbuchs noch mit sehr vielen Datendateien zu tun haben, so dass Ihnen die wichtigsten Eigenschaften alleine durch die Übung geläufig werden werden. Es schadet aber nicht, einige Details hier in einer Übersicht versammelt zu haben.

Kap2UE6.sav [DataSet9] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

Search application

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Code	String	10	0	Proband:innen-...	None	None	8	Left	Nominal	Both
2	Sprache	String	1	0	Fließend gespr...	{1, Deutsch}...	None	8	Left	Nominal	Both
3	Geschlecht	Numeric	1	0	Geschlechterzu...	{1, weiblich}...	None	8	Right	Nominal	Both
4	Alter	Numeric	3	0	Alter in Jahren	None	None	8	Right	Scale	Both
5	Größe	Numeric	5	0	Körpergröße in ...	None	None	8	Right	Scale	Both
6	G_score	Numeric	3	0	Intelligenzscore	None	None	8	Right	Scale	Both
7											
8											
9											

Abbildung 2.10. Beispiel für eine Datendatei in der Variablenansicht.

Name. In der ersten Spalte der Variablenansicht steht der Variablenname. Beim Benennen von Variablen sollte man allerdings einige Regeln beachten. Der Name der Variable muss eindeutig sein; doppelt vorkommende Namen innerhalb einer Datendatei sind nicht zulässig. Variablennamen müssen mit einem Buchstaben beginnen. Es gibt allerdings einige SPSS-interne Ausnahmen, die mit einem „@“,

„#“, oder „$“ beginnen können. Dabei handelt es sich um Variablen, die von SPSS (etwa im Rahmen bestimmter Verarbeitungsschritte) erzeugt wurden. Variablennamen dürfen keine Leerzeichen enthalten. Variablennamen dürfen nicht länger als 64 Zeichen sein. Variablennamen sollten nicht mit Punkt oder Unterstrich enden (aber innerhalb eines Namens ist das durchaus in Ordnung). Reservierte Schlüsselwörter (z.B. „ALL“, „AND“, „BY“, „EQ“, „GE“, „GT“, „LE“, „LT“, „NE“, „NOT“, „OR“, „TO“, „WITH“) können nicht als Namen verwendet werden.

Type. In der zweiten Spalte wird die Art einer Variablen definiert. Hierbei kann zwischen 9 verschiedenen Typen gewählt werden. Am häufigsten werden die beiden Arten *Numeric* (Dezimalzahl) und *String* (Zeichenkette bzw. Symbolfolge) verwendet. Die übrigen Variablentypen können bei Bedarf in der Dokumentation unter https://www.ibm.com/docs/en/spss-statistics/31.0.0?topic=view-variable-type (zumindest für die aktuelle Version dieses Manuskripts) nachgelesen werden.

Width. Width bezeichnet die Variablenbreite; d.h. die Anzahl der Zeichen, die die Ausprägungen der Variablen maximal umfassen darf. Eine Variablenbreite von 10 heißt demnach, dass in der Datenansicht eine Variable maximal 10 Zeichen lang sein darf.

Decimals. Diese Spalte gibt an, wie viele Dezimalstellen eine Variable hat. In diesem Beispiel (Abbildung 2.10) haben alle Variablen 0 Dezimalstellen, es handelt sich also ausschließlich um ganze Zahlen ohne Nachkommastellen.

Label. Das Label einer Variablen ist eine zentrale Eigenschaft. Sie dient der Beschreibung der Variablen und die Definition dieses Labels erleichtert die weitere Arbeit mit den Daten ungemein. Variablenlabel können Leerzeichen und reservierte Zeichen enthalten, die in Variablennamen nicht zulässig sind. Bei Fragebogendaten wird beispielsweise empfohlen unter Label den exakten Wortlaut des jeweiligen Items einzugeben. So kann man später bei der Datenanalyse jederzeit schnell nachschauen, was genau mit einem bestimmten Item erfragt wurde (z.B. „Naturschutz ist mir sehr wichtig“ bei einem Fragebogen zur Naturschutzakzeptanz).

Values. Sie können jedem Wert einer Variablen ein beschreibendes Wertelabel (also eine Beschreibung) zuordnen. Das wird hauptsächlich verwendet, wenn die Datendatei numerische Codes zur Darstellung nominaler Kategorien benutzt (zum Beispiel die Zahlen 1 und 2 für „weiblich“ und

„männlich“). Einzelne Werte und deren Labels können in dem Fenster, das sich nach Klicken auf die entsprechende Zelle unter *Values* öffnet, mit dem Plus-Symbol hinzugefügt werden und mit dem Kreuz-Symbol wieder entfernt werden, siehe Abbildung 2.11.

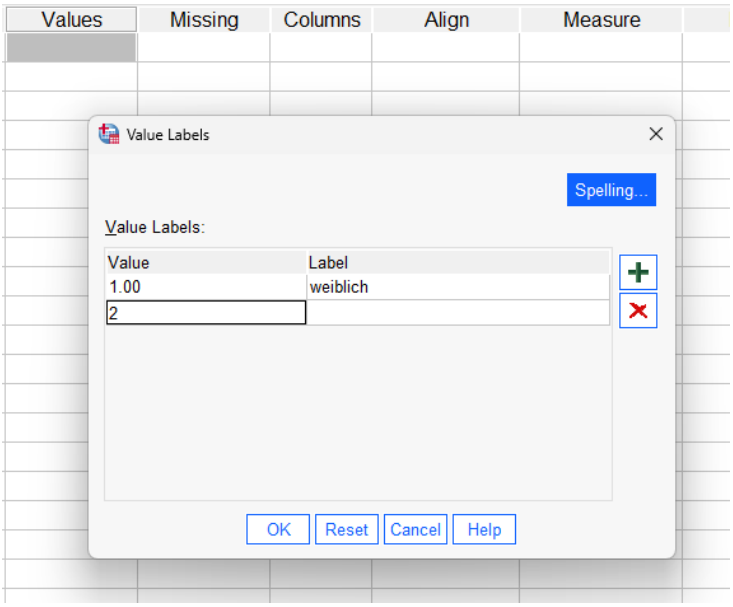


Abbildung 2.11. Hinzufügen von Werten für kategoriale Variablen.

Missing. Hier können bestimmte Datenwerte als benutzerdefiniert fehlende Werte deklariert werden. Datenwerte, die als benutzerdefiniert fehlende Werte angegeben sind, werden zur Sonderbehandlung gekennzeichnet und von den meisten Berechnungen ausgeschlossen. Man kann entweder bis zu drei diskrete (einzelne) fehlende Werte (z.B. 99 oder -1), einen Bereich fehlender Werte oder einen Bereich und einen diskreten Wert eingeben.

Columns. Diese Eigenschaft bezeichnet die Spaltenbreite; dabei handelt sich ausschließlich um eine Sache der visuellen Darstellung der Datenansicht. Unabhängig davon, wie viele Zeichen eine Variablenausprägung maximal umfassen darf (unter *Width*), kann die Breite der Spalte größer oder kleiner sein.

Align. Auch diese Einstellung dient bloß der visuellen Darstellung in der Datenansicht. Hier kann ausgewählt werden, ob Daten links, rechts oder mittig im Feld ausgerichtet sein sollen.

Measure. Die Eigenschaft Measure bzw. Skalenniveau hat zentralen Charakter für die Variablendefinition. Hier ist das Skalenniveau der Variablen einzustellen, also ob eine Variable metrisch

(scale), ordinal oder nominal skaliert ist. Im Gegensatz zur grundsätzlichen Bedeutung des Begriffs in der Statistik fasst SPSS mit dem Begriff *scale* Intervallskalenniveau, Verhältnisskalenniveau, und Absolutskalenniveau in eine Kategorie zusammen. Entscheidend ist also lediglich, ob eine Variable mindestens Intervallskalenniveau aufweist.

Role. Manche Dialogfelder unterstützen vordefinierte Rollen, die zur Vorauswahl von Variablen zur Analyse verwendet werden können. Wenn Sie eines dieser Dialogfelder öffnen, werden in der/den Zielliste/n automatisch Variablen angezeigt, die die Rollenbedingungen erfüllen. Standardmäßig wird allen Variablen die Rolle Input zugewiesen. Manche Autoren (Bühner & Ziegler, 2017) empfehlen, für alle Variablen die Rolle ‚Both' zu vergeben, da jede Variable so gut wie immer sowohl als UV als auch als AV fungieren kann.

Existierende Datendateien öffnen

SPSS-Datendateien. Beim Öffnen von Dateien ist zu beachten, dass man über das Ordner-Symbol (Abbildung 2.12) in den Menüs der Programmfenster nur denselben Dateityp öffnen kann, der dem gerade verwendeten Programmfenster entspricht. Das bedeutet, dass über das Ordnersymbol im Dateneditor nur andere Datendateien geöffnet werden können, über das Ordnersymbol im Syntax-Editor nur Syntax-Dateien und über das Ordner-Symbol im Ausgabefenster nur Ausgabedateien. Über *File >> Open* lassen sich aber aus jedem Programmfenster alle anderen SPSS-Dateitypen öffnen, indem man auswählt, welchen Dateityp man öffnen möchte.

Externe Dateitypen. Mit SPSS lassen sich aber auch Daten einlesen, die in anderen Dateiformaten vorliegen. Um beispielsweise MS Excel Dateien einlesen zu können, können Sie wie folgt vorgehen. Klicken Sie zuerst auf *File >> Open >> Data…* und wählen Sie anschließend in dem sich öffnenden Fenster unter *Files of type* „Excel: (*.xls, *.xlsx, *xlsm)" aus. Danach werden ihnen nur mehr Excel Dateien angezeigt. Wählen Sie die gewünschte Datei und klicken Sie auf Open. Nach dem Importieren ist es meist noch klug, die Eigenschaften der Daten in der Variablenansicht zu überprüfen und falls nötig anzupassen. Zudem empfiehlt sich diese neue und gegebenenfalls ergänzte SPSS-Datendatei dann zusätzlich zur existierenden MS Excel Datei abzuspeichern, um Sie für später (und für eine in jedem Fall empfehlenswerte Dokumentation) verfügbar zu haben.

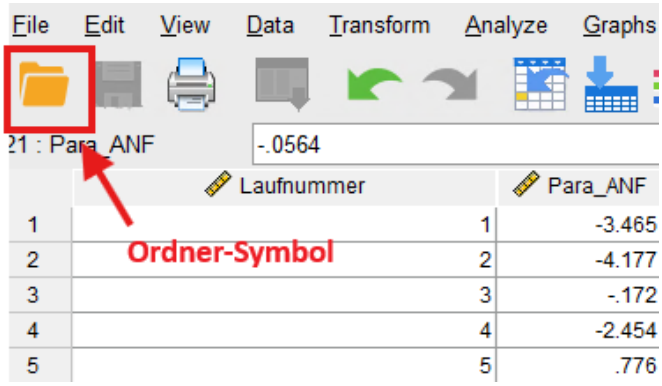


Abbildung 2.12. Öffnen von Dateien mittels des Ordner-Symbols.

Auch CSV-Dateien können mit dem gleichen Prinzip in SPSS importiert werden. Bei CSV-Dateien (CSV steht für „comma separated values“) handelt es sich um ein Dateiformat, dem sie in der Praxis statistischer Datenverarbeitung recht häufig begegnen können, weil es sich dabei um ein sehr einfaches (und damit auch robustes, d.h. eher wenig fehleranfälliges) Dateiformat handelt. Um eine solche Datei in SPSS zu importieren, müssen Sie bei „Files of type“ die Option „CSV (*.csv)“ auswählen. Bei diesem Dateientyp öffnet sich nach Auswahl der Datei ein weiteres Fenster (ein sog. Import-Assistent, d.h. ein kleines Programm, das uns beim Einlesen der Daten aus diesem Dateiformat hilft). Als erstes wird abgefragt, ob die Datei ein vordefiniertes Design aufweist. Falls wir über kein vorgefertigtes Design verfügen, wählen wir hier „No“ und klicken dann auf „Next“. Danach wird abgefragt, ob Variablenwerte mit einem definierten Symbol (z.B. einem Komma) oder durch einen bestimmten Abstand (z.B. vier Leerzeichen) getrennt sind. Zudem ist es wichtig zu wissen, ob die erste Zeile Variablenbezeichnungen enthält oder die Datei sofort mit einer Auflistung von Variablenwerten beginnt. Schließlich muss das verwendete Dezimaltrennzeichen eingegeben werden. Haben wir alle nötigen Informationen angegeben, können wir wieder auf „Next“ klicken. Es folgen einige weitere Fragen. In welcher Zeile beginnt die Auflistung der Variablenwerte? Entspricht jede Zeile einem Fall (z.B. einer Person oder einer Beobachtung aller Messwerte für eine Person)? Wollen wir nur einen bestimmten Teil der Datendatei einlesen? Im nächsten Fenster wählen wir das Trennzeichen zwischen einzelnen Variablenwerten aus und geben an wie Zeichenfolgen gekennzeichnet sind (etwa durch ein- oder zweigestrichene Anführungszeichen). Hier können wir auch angeben, ob Zeichenfolgen, die Leerzeichen ganz zu Beginn oder ganz am Ende beinhalten, um diese Leerzeichen bereinigt werden

sollen. Im nächsten Fenster können wir noch neue Variablennamen wählen, falls wir das möchten. Im anschließenden Fenster können wir dann entscheiden, ob wir diese Prozedur als Design abspeichern wollen, um sie beim nächsten Mal, wenn wir eine Datei dieses Formats einlesen, nicht mehr durchführen zu müssen (das Design kann dann im ersten Schritt ausgewählt werden, siehe oben). Wir können uns auch entscheiden, die einzelnen Schritte in unsere Syntax-Datei einzufügen. Dadurch wird in unsere Syntaxdatei der Programmcode eingefügt (falls wir keine Syntaxdatei geöffnet haben, wird eine neue erzeugt), der der gesamten Prozedur entspricht, die wir soeben durch Point-and-Click in den einzelnen Fenstern ausgewählt haben. Wenn wir diesen Code markieren und ausführen, wird die CSV-Datei eingelesen.

Verschiedene Datendateien zusammenführen

Sollen Daten aus mehreren SPSS-Datendateien in einer Datei zusammengefügt werden, sind dabei grundsätzlich zwei Fälle zu unterscheiden. Entweder möchte man neue Fälle (in der Regel neue Personen) zu bereits existierenden hinzufügen, oder eine Datei um neue Variablen, die von denselben Personen stammen, ergänzen. In beiden Fällen wird aber empfohlen zum Zusammenführen der Daten die entsprechenden Funktionalitäten von SPSS zu verwenden und nicht händisch über Copy-Paste Daten von einer Datei in eine andere kopieren. Letzteres ist gerade bei größeren Datenmengen fehleranfällig und die Gefahr ist groß, dass fehlerhafte Dateneinträge unbemerkt in allen weiteren Auswertungsschritten weiterverwendet werden.

Hinzufügen von weiteren zu bereits existierenden Fällen

Dies ist z.B. der Fall, wenn es zwei Datendateien gibt, die dieselben Variablen enthalten, aber Daten von unterschiedlichen Personen. In Abbildung 2.13 sind zwei geöffnete Datendateien gezeigt, die in den ersten sechs Zeilen der Variablenansicht dieselben Variablen enthalten. In den Zeilen 7 – 9 sind in der oberen der beiden Dateien weitere Variablen enthalten, die in der anderen Datei fehlen.

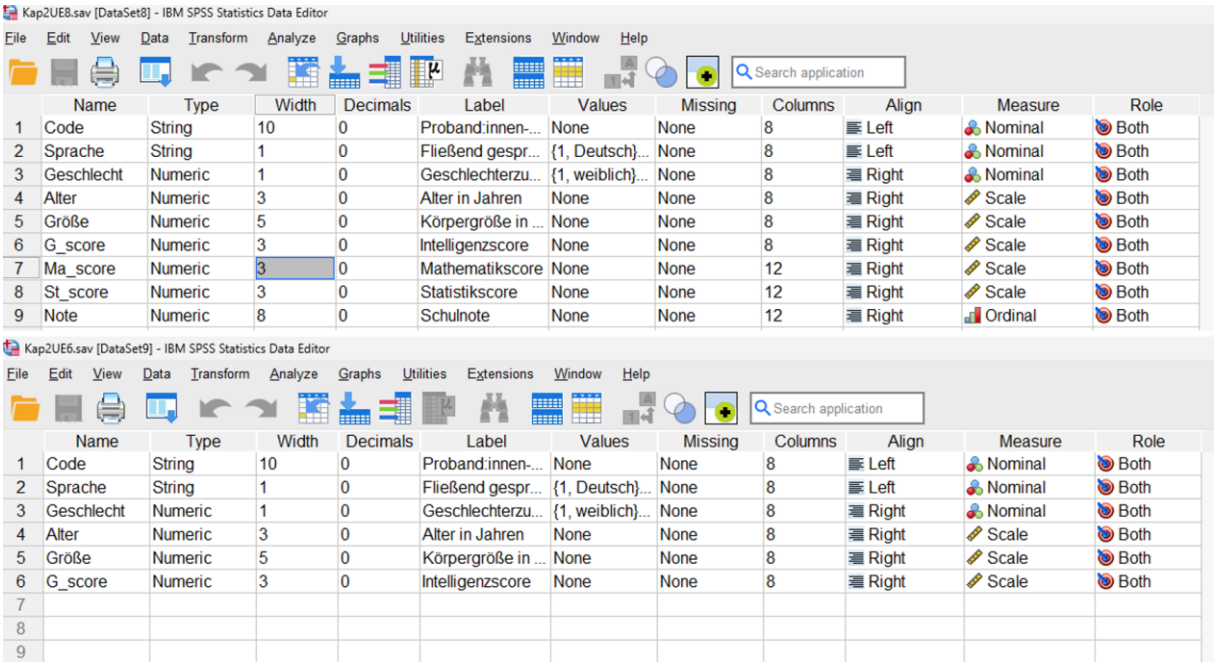


Abbildung 2.13. Ausschnitte aus zwei geöffneten Datendateien.

Zunächst muss geprüft werden, ob die Variablen, die in beiden Dateien vorkommen auch tatsächlich dasselbe erfasst haben und deshalb auch dieselben Eigenschaften in beiden Dateien haben (Name, Typ, Breite, Skalenniveau etc.). Die Variableneigenschaften können falls nötig entweder einzeln manuell oder auch für ganze Spalten auf einmal mittels Copy & Paste angepasst werden. Dazu kann eine Eigenschaft der gewünschten Variablen in der Datei, in der die Eigenschaften bereits korrekt eingegeben sind, markiert werden, auf die Auswahl ein Klick mit der rechten Maustaste getätigt und anschließend *Copy* ausgewählt werden. In der Datei, in der diese Eigenschaften geändert bzw. angepasst werden sollen, können die Eingaben durch den entsprechenden Prozess mittels *Paste* einfügt werden.

Wenn alle Variablen, die in beiden Datendateien vorkommen, mit den korrekten Variableneigenschaften ausgestattet sind, kommt man zum eigentlichen Zusammenfügen. Dazu wird im Menü *Data >> Merge Files >> Add Cases* ausgewählt und im Dialogfenster die zusammenzuführenden Dateien ausgewählt. Dadurch werden Fälle aus einer anderen Datei dem bestehenden Datensatz hinzugefügt.

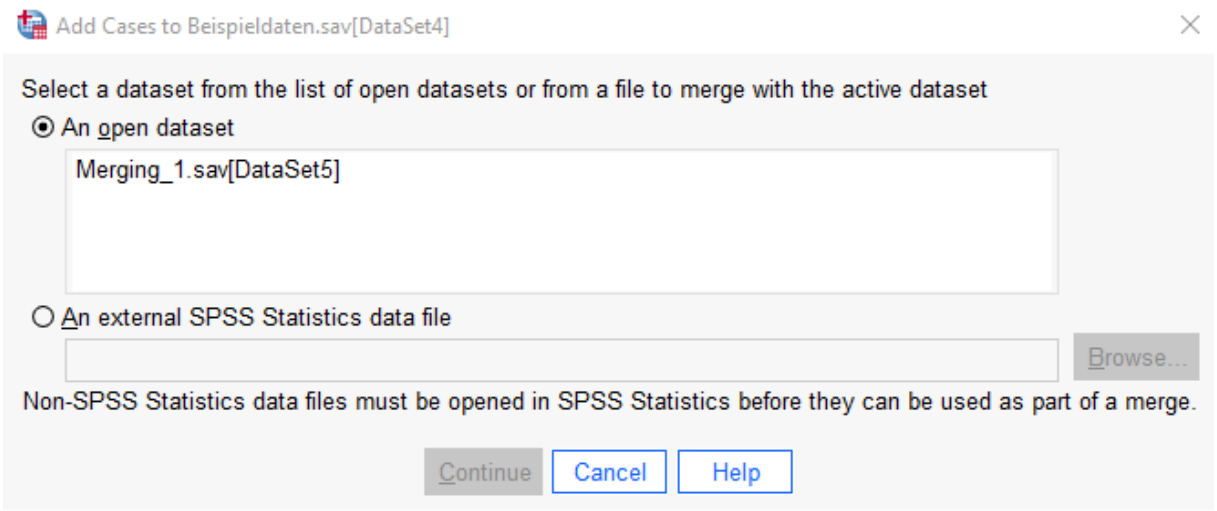


Abbildung 2.14. Fälle zu einer bestehenden Datendatei hinzufügen.

Dieser Schritt ist auch in Abbildung 2.14 dargestellt. Hier ist die Datei auszuwählen, aus der die Daten hinzugefügt werden sollen (diese Datei bleibt bestehen, die Fälle werden lediglich kopiert). Am einfachsten geht es, wenn die Datei, aus der man Daten kopieren möchte, bereits geöffnet ist. Aber auch nicht geöffnete Datendateien können über „An external SPSS Statistics data file“ ausgewählt werden.

Nach Klicken auf *Continue* scheinen Variablen, die nicht in beiden Dateien vorkommen, im Feld „Unpaired Variables“ auf, siehe Abbildung 2.15 links. Wenn Sie diese Variablen ebenfalls in der neuen Datendatei haben möchten (sie werden dann lediglich für einige Fälle keine Variablenwerte aufweisen), markieren Sie sie, klicken auf den Pfeil zwischen den beiden Feldern und schließlich auf OK. Das linke Feld ist dann leer und im rechten Feld unter „Variables in New Active Dataset“ sind alle Variablen aufgelistet, die in der neuen Datendatei enthalten sein sollen, siehe Abbildung 2.15 rechts. Wenn in dem Feld „Unpaired Variables“ von Beginn an keine Variablen enthalten sind, bedeutet es nur, dass bereits in beiden Datendateien ausschließlich dieselben Variablen mit denselben Eigenschaften vorliegen. Kontrollieren Sie zum Schluss, ob alle Variablen vorhanden sind, die in den neuen Datensatz übernommen werden sollten.

Speichern Sie anschließend die resultierende Datendatei unter einem neuen Dateinamen ab! Das ist insbesondere deshalb wichtig, weil die neuen Fälle durch die oben beschriebene Prozedur in die bereits geöffnete Datendatei hinzugefügt werden. Wenn sie diese dann unter demselben Namen abspeichern, haben sie ihre originale Datendatei überschrieben und das Original verloren!

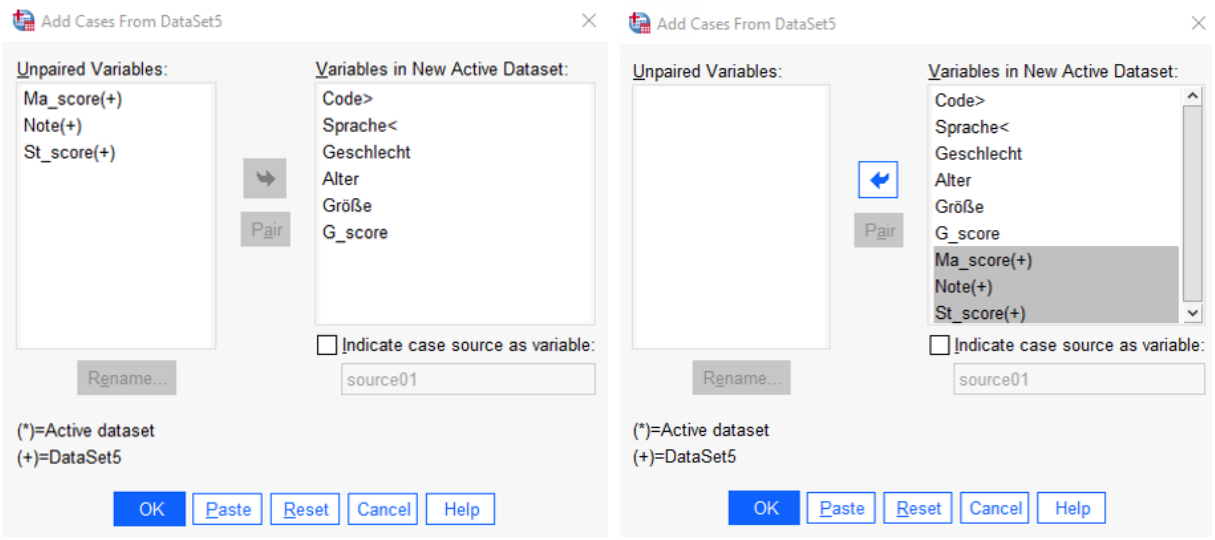


Abbildung 2.15. Übernehmen von „unpaired variables“.

Neue Variablen zu einem bestehenden Datensatz hinzufügen

Dies findet z.B. Einsatz, wenn Sie bei denselben Personen zu zwei Zeitpunkten Variablen erhoben haben (entweder unterschiedliche Variablen oder als klassische Messwiederholung dieselben Variablen, die dann aber in den beiden Datendateien unterschiedliche Variablennamen haben müssen). Wenn zum Beispiel bei einer Studie einmal „vor Ort“ Daten erhoben wurden und danach zusätzlich ein Onlinefragebogen auszufüllen war, sodass zwei Datendateien mit unterschiedlichen Variablen, aber für dieselben Personen resultieren.

Damit das Zusammenfügen der beiden Dateien funktioniert, müssen beide eine Variable beinhalten, durch die Variablenwerte aus beiden Dateien eindeutig jeweils derselben einzelnen Person zugewiesen werden können. Dabei kann es sich um jede Art eines eindeutigen persönlichen Codes handeln, der in SPSS als Schlüsselvariable bzw. Key-Variable bezeichnet wird. Diese Variable muss in beiden Datendateien exakt dieselben Eigenschaften und für jede einzelne Person dieselbe Ausprägung haben.

Außerdem müssen beide Datendateien nach dieser Schlüsselvariable aufsteigend sortiert sein. Dieses Sortieren kann entweder als erster Schritt manuell erfolgen (muss es aber nicht, siehe unten), indem in beiden Datendateien in der Datenansicht die Spaltenüberschrift mit der rechten Maustaste angeklickt und im Kontextmenü *Sort Ascending* (=aufsteigend sortieren) gewählt wird. Nach dem

Umsortieren sollten beide Datendateien gespeichert werden (sonst findet das anschließende Zusammenfügen teilweise mit den unsortierten Daten vor der Speicherung statt).

Über *Data >> Merge Files >> Add Variables* gelangen Sie zu dem in Abbildung 2.16 gezeigten Fenster zur Definition der Methode zum Zusammenfügen der Dateien. Voreingestellt ist hier „One-to-one merge based on key values“ und diese Option ist fast immer die beste Wahl. Außerdem ist die Option „Sort files by key values before merging“ auch voreingestellt. Dadurch werden die Fälle in beiden Datendateien vor dem Zusammenfügen nach der Schlüsselvariable sortiert. In dem in Abbildung 2.16 gezeigten Beispiel stehen im Feld mit der Überschrift „Key Variables“ mehrere Variablen. Dies sind alle Variablen, die in beiden Datendateien vorkommen. An dieser Stelle muss die Variable gewählt werden, die als Schlüsselvariable für das Zusammenfügen herangezogen werden soll. In diesem Falle ist es die Variable „Code“. Nach dem Zusammenfügen kontrollieren Sie die Variablen nochmals, um sicherzugehen, dass alles richtig funktioniert hat und speichern die resultierende Datei wieder unter einem neuen Dateinamen ab.

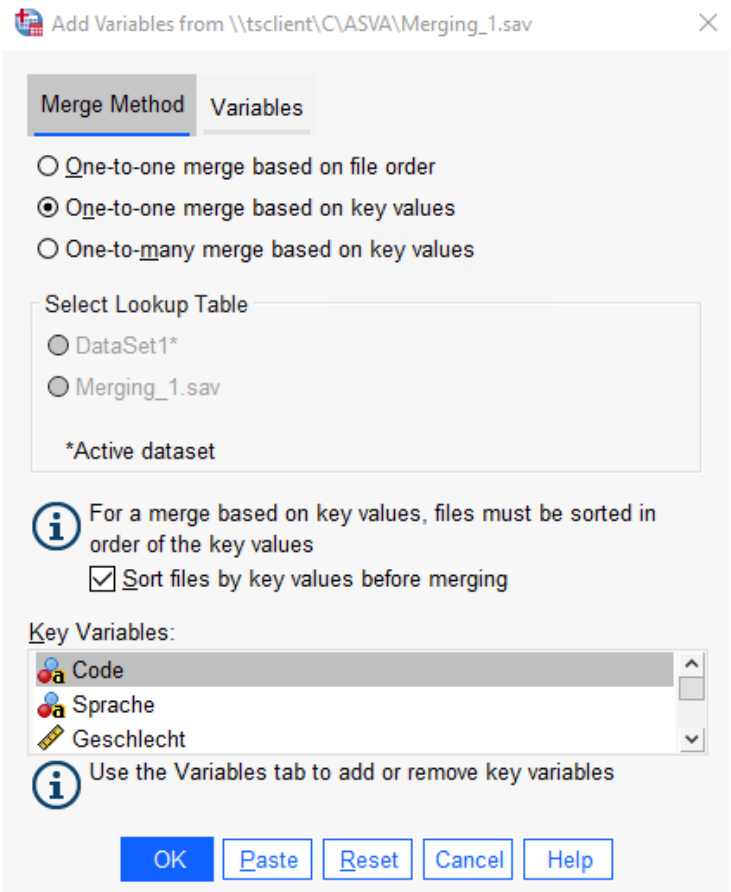


Abbildung 2.16. Hinzufügen neuer Variablen zu einem bestehenden Datensatz.

Übungsaufgaben

Wir werden nun all die in diesem Kapitel beschriebenen Inhalte an einer Reihe von Übungsbeispielen illustrieren.

Beispiel 2.1

Stellen Sie einen Fernzugriff zu SPSS her. Stellen Sie die Sprache auf Englisch um. Ändern Sie das Dezimaltrennzeichen in Ein- und Ausgabe von einem Komma auf einen Punkt mit einer entsprechenden Syntaxdatei. Speichern Sie die Syntaxdatei anschließend lokal auf Ihrem Rechner ab.

Beispiel 2.2

Welche der folgenden gehören zu den grundlegenden Programmfenstern in SPSS?

(a) Code-Editor

(b) Dateneditor

(c) Builder

(d) Syntax-Editor

Beispiel 2.3

Was stimmt für Kommentare in der SPSS-eigenen Programmiersprache SPSS Syntax?

(a) Kommentare müssen mit einem „*“ beginnen.

(b) Kommentare müssen mit einem „.“ aufhören.

(c) Kommentare können auch mit einer nachfolgenden Leerzeile aufhören.

(d) Kommentare beginnen immer mit einem „%“.

Beispiel 2.4

Welche der folgenden Aussagen ist/sind richtig/falsch?

Nr.	**Aussage**	**R/F**
1)	Daten müssen händisch in SPSS eingetippt werden.	
2)	SPSS verfügt über keine Funktionalität um MS Excel Dateien einzulesen.	
3)	Statistische Analysen können in SPSS ausschließlich mittels Point&Click durchgeführt werden.	

Beispiel 2.5

Laden Sie den Ordner mit elektronischem Zusatzmaterial (Engl.: „Electronic supplementary material“) für dieses Übungsbuch unter dem entsprechenden Link auf https://osf.io/9tcx3/ herunter. Entpacken Sie den Ordner in ein Verzeichnis Ihrer Wahl auf Ihrem Computer und öffnen Sie die SPSS-Datendatei „test.sav“ mit SPSS. Beantworten Sie folgende Fragen:

(a) Wie viele Variablen sind in dem Datensatz definiert?

(b) Wie viele Variablen liegen (mindestens) auf Intervallskalenniveau vor?

(c) Wie viele Fälle (Personen) liegen in dem Datensatz vor?

(d) Wie alt ist die Person mit Personencode „wryz0893_z0“? Wie groß ist die Person? Ist die Person verheiratet?

(e) In welcher Einheit wird die Körpergröße in dem Datensatz angegeben? In welcher Einheit wird das Gewicht angegeben?

Beispiel 2.6

Erstellen Sie eine neue Datendatei in SPSS und definieren Sie Variablen entsprechend Abbildung 2.10. Ergänzen Sie die Felder in der Spalte „Label“ und „Values“ sinngemäß. Wechseln Sie anschließend in die Datenansicht und fügen Sie der Datendatei drei Fälle (Personen) hinzu. Wählen Sie als Codes für diese drei Fälle: „ID1“, „ID2“, „ID3“ (ohne die Anführungszeichen). Erfinden Sie für die übrigen Messwerte einfach plausible Werte. Speichern Sie die Datei schließlich unter dem Dateinamen „Kap2UE6.sav“ ab.

Beispiel 2.7

Erstellen Sie mit MS Excel eine Datei mit dem in Abbildung 2.17 dargestellten Inhalt, lesen Sie diese anschließend in SPSS ein, und speichern Sie sie mit Dateinamen „Kap2UE7.sav“ ab.

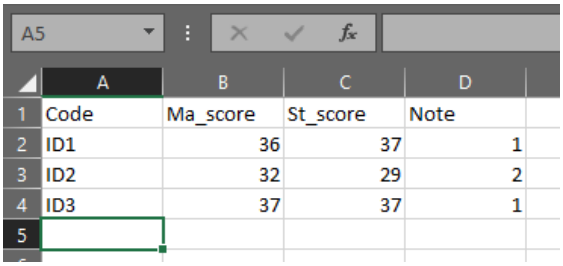


Abbildung 2.17. Eine mögliche MS Excel Datendatei.

Beispiel 2.8

Fügen Sie die Variablen „Ma_score“, „St_score“, „Note“ aus dem Datensatz, den Sie in Beispiel 2.7 erzeugt haben zu den Variablen aus Beispiel 2.6 für die drei Personen mit den Codes „ID1“, „ID2“ und „ID3“ hinzu. Kontrollieren Sie den resultierenden Datensatz und ergänzen Sie die Definition der Variablen gemäß Abbildung 2.13 und speichern Sie den neuen Datensatz als neue Datendatei mit Dateinamen „Kap2UE8.sav“ ab.

Beispiel 2.9

Sie erhalten die folgenden Daten für zwei weitere Personen mit den Codes „ID4“ und „ID5“:

Code	Sprache	Geschlecht	Alter	Größe	G_score	Ma_score	St_score	Note
ID4	1	1	19	164	92	25	21	3
ID5	2	2	20	183	95	20	15	3

Kopieren Sie die Tabelle in eine leere MS Excel Datei und speichern Sie die Datei als CSV-Datei ab. Öffnen Sie die Datei daraufhin mit einem einfachen Texteditor (unter Windows geben Sie einfach Editor in die Suchleiste ein und öffnen Sie die Datei damit). Sehen Sie sich den Inhalt der Datei an und versuchen Sie dann die Datei in SPSS einzulesen. Speichern Sie die resultierende Datendatei unter dem Dateinamen „Kap2UE9.sav“ ab.

Beispiel 2.10

Generieren Sie einen Syntax-Code, der die CSV-Datei aus Beispiel 2.9 einliest und speichern Sie die resultierende Syntax-Datei unter dem Dateinamen „Kap2UE10.sps“ ab.

Beispiel 2.11

Fügen Sie die Fälle aus dem resultierenden Datensatz aus Beispiel 2.8 mit jenen aus dem resultierenden Datensatz aus Beispiel 2.9 zusammen und speichern Sie den resultierenden Gesamtdatensatz unter dem Dateinamen „Kap2UE11.sav“ ab.

Beispiel 2.12

Stellen Sie sich vor, Sie hätten den in Abbildung 2.18 gezeigten Datensatz handschriftlich auf Papier erhalten und sollen nun eine geeignete Datendatei mit SPSS erstellen. Erstellen und definieren Sie entsprechende Variablen und tragen Sie anschließend die Daten ein. Bei der Variable Geschlecht soll die Ziffer 1 für „weiblich“ und die Ziffer 2 für „männlich“ stehen. Beachten Sie ferner, dass Werte von 999 bedeuten, dass der jeweilige Messwert fehlt. Speichern Sie die resultierende Datendatei unter dem Namen „Kap2UE12.sav“ ab.

Code	Alter	Geschlecht	Groesse	Mathe_score
FK14	33	1	155	67
JH23	23	1	162	92
EL06	18	2	170	999
DS12	42	2	176	58
NT08	29	1	180	88
BA17	56		185	98
ST26	37	2	999	39
JP19	25	2	168	69
SR05	64	1	165	75
HS12	999	2	175	84

Abbildung 2.18. Ein Datensatz mit fehlenden Werten.

Beispiel 2.13

Wenn Sie wirklich alle Übungsbeispiele dieses Kapitels bis hierher gemacht haben, haben Sie vermutlich viele Datendateien gleichzeitig in SPSS geöffnet. Woran erkennen Sie, welche Datendatei aktuell gerade aktiv ist?

Kapitel 3

Datenmanagement und deskriptive Statistiken

Stefan E. Huber

Im vorhergehenden Kapitel haben wir uns mit einigenden grundlegenden Funktionen von SPSS befasst: Wie kann SPSS (von zu Hause aus) genutzt werden? Wie können Datendateien unterschiedlicher Formate eingelesen werden? Wie sehen Datendateien aus? Wie können sie mit SPSS erstellt werden? Mit den Daten selbst haben wir allerdings noch nicht viel angestellt. Diesem Aspekt wenden wir uns nun in diesem Kapitel zu. Zuerst werden wir einige typische Verarbeitungsschritte von Daten in SPSS betrachten, die man immer wieder brauchen kann. Danach werden wir uns mit deskriptiven Statistiken zur Charakterisierung erhobener Stichproben befassen. Die einzelnen Arbeitsschritte werden wir der Einfachheit halber an einem Beispieldatensatz illustrieren. Mit den Übungsbeispielen am Ende des Kapitels können Sie dann die einzelnen Arbeitsschritte noch einmal an einem anderen Beispieldatensatz wiederholen. Beide Datensätze finden Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument, das Sie unter https://osf.io/9tcx3/ herunterladen können. Den Datensatz, mit dem wir in den direkt folgenden Abschnitten arbeiten werden, finden Sie in der Datei „Kap3daten.sav“.

Umkodieren von Variablen

Wenn man mit Fragebögen arbeitet, ist es häufig nötig, einzelne Items umzukodieren. Was bedeutet das? Betrachten wir dafür die folgenden beiden Items (Variablennamen: *politik_politik1* und *politik_politik2*), für die Sie die Messwerte für 51 Personen in dem Datensatz in der Datei „Kap3daten.sav“ finden:

1. Im Großen und Ganzen sehe ich mich selbst als einen politisch interessierten Menschen. (Variable: *politik_politik1*)
2. Politik ödet mich an. (Variable: *politik_politik2*)

Beide Items sind auf einer fünfstufigen Likert-Skala von „trifft überhaupt nicht zu“ (1) bis „trifft völlig zu“ (5) zu beantworten (überzeugen Sie sich davon, indem Sie in der Spalte „Values“ im SPSS-

Datensatz nachsehen). Beide Items zielen ebenfalls darauf ab zu quantifizieren, wie gerne sich jemand mit Politik beschäftigt. Das heißt, wir könnten auch sagen, das Merkmal, dessen Ausprägung mit diesen Items gemessen werden soll, ist die Politikaffinität einer Person. Eine sehr politikaffine Person wird das erste der beiden Items eher mit hohen Werten beantworten, das zweite eher mit niedrigen. Bei einer wenig politikaffinen Person wäre es gerade umgekehrt.

Wenn wir also nun die Politikaffinität von Personen mit diesen Items erfassen wollen, wäre es vorteilhaft, wenn für beide Items gelten würde, dass hohe Werte hohe Politikaffinität und niedrige Werte niedrige Politikaffinität bedeuten. Wäre das der Fall, könnten wir die Werte der beiden Items einfach zusammenzählen oder ihren Mittelwert bilden und hätten in beiden Fällen eine Zahl, die die Politikaffinität der befragten Person erfasst (zumindest bis auf einen Messfehler).

Um genau das zu erreichen, können wir die zweite der beiden Variablen umkodieren oder genauer: umpolen. Das heißt, wir drehen sozusagen die Skala um: niedrige Werte sollen hohe Werte bedeuten und hohe Werte niedrige. Wieso? Weil eine Person, die einen niedrigen Wert beim Item „Politik ödet mich an“ auswählt, eine hohe Politikaffinität hat und umgekehrt.

Wie können wir das in SPSS machen? Dazu wählen wir im Menü „Transform“ die Option „Recode into Different Variables…“ aus wie in Abbildung 3.1 gezeigt.

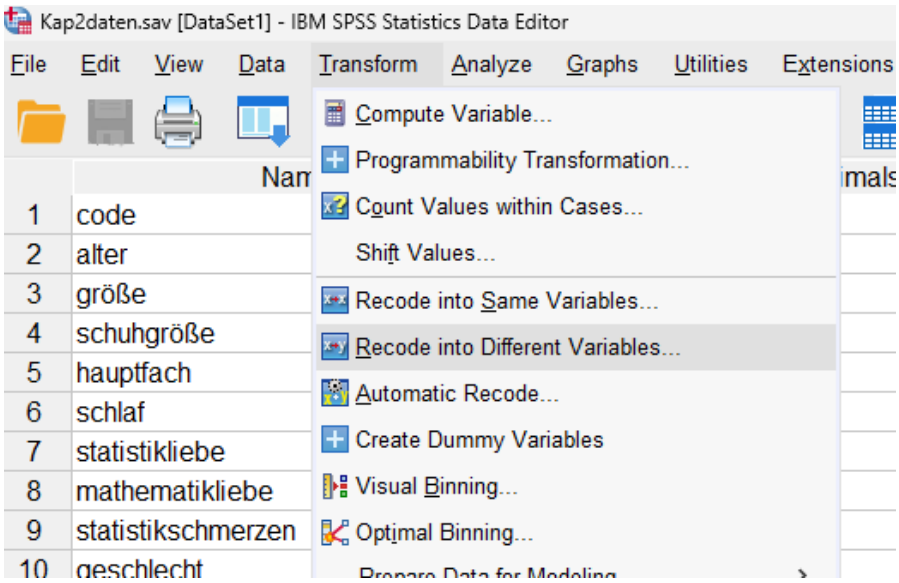


Abbildung 3.1. Um eine Variable umzukodieren benötigen wir im Menü „Transform“ die Option „Recode into Different Variables…“.

Im sich dadurch öffnenden Fenster wird nun links eine Liste mit all den Variablen des aktiven Datensatzes angezeigt. Allerdings werden diese Variablen standardmäßig durch die Angabe ihrer Labels angezeigt. Das ist oft recht unübersichtlich. Durch einen Rechtsklick irgendwo in das linke Fenster öffnet sich ein Kontextmenü. Wenn wir dort „Display Variable Names“ auswählen, werden uns die Variablennamen anstelle der Labels angezeigt. Dies ist allerdings nur dann ein Vorteil, wenn man sich gute, prägnante Bezeichnungen für die Variablen überlegt hat (dieser Aufwand macht sich also bei der Definition der Variablen eines Datensatzes häufig bezahlt).

Wählen wir nun die Variable *politik_politik2* aus (da wir diese Variable umkodieren wollen) und klicken auf den kleinen Pfeil, der nach rechts zeigt, wird diese Variable in das Fenster „Input Variable -> Output Variable“ verschoben. Alternativ können wir die Variable auch in dieses Fenster hineinziehen (mittels Linksklick und Halten bzw. Drag-and-drop). Unter „Output Variable“ können wir der neuen Variable, die wir aus der alten erzeugen werden, einen Namen und ein Label geben (diese werden dann in der Variablenübersicht auch genauso angezeigt werden). Dort tragen wir nun bei „Name“ den Text „politik_politik2_umk“ ein und bei „Label“ den Text „Umkodierung des Items "Politik ödet mich an."“. Durch Klicken auf „Change“ werden diese Bezeichnungen bestätigt, was wir daran erkennen, dass nun im zentralen Fenster der Variablenname der neuen Variable eingefügt wird.

Unter „Old and New Values…“ können wir die Regeln für die Umkodierung festlegen. Wenn wir auf diese Schaltfläche klicken, öffnet sich ein weiteres Fenster. In diesem Fenster können wir nun festlegen wie die Werte unserer ursprünglichen Variable (*politik_politik2*) auf die Werte unserer neuen Variable (*politik_politik2_umk*) abgebildet werden sollen. Hier geben wir nun bei „Old Value“ unter „Value“ die Zahl 5 ein, und anschließend bei „New Value“ unter „Value“ die Zahl 1. Dann klicken wir rechts in der Mitte des Fensters auf „Add“ und sehen daraufhin einen neuen Eintrag in dem Feld, der mit „Old --> New“ überschrieben ist. Dort steht jetzt „5 --> 1“. Das bedeutet, der hohe Wert 5 der alten Variable wird auf den Wert 1 der neuen Variable abgebildet; das heißt, hohe Werte für „Politik ödet mich an“ werden auf niedrige Werte des entsprechend umgepolten Items (mit der gegenteiligen Bedeutung, also etwa im Sinne von „Politik fasziniert mich“) abgebildet. Ganz analog verfahren wir nun mit den übrigen Werten 4, 3, 2, und 1 und bilden diese auf die neuen Werte 2, 3, 4, 5 ab. Wenn wir damit fertig sind, klicken wir auf „Continue“, woraufhin sich dieses Fenster schließt.

Damit sind wir eigentlich schon fertig. Wir könnten nun auf „OK“ klicken und uns an der neu eingefügten Variable in der Variablenansicht (und an den entsprechenden Messwerten in der Datenansicht) erfreuen. Allerdings werden wir die Gelegenheit gleich nutzen, um uns etwas in gute Praxis der Datenanalyse einzuarbeiten. Daher klicken wir nicht auf „OK“, sondern stattdessen auf „Paste“.

Dadurch öffnet sich eine Syntaxdatei, in der nun bereits einige Zeilen eingefügt sind. War bereits eine Syntaxdatei geöffnet, wurden diese Zeilen in dieser am Ende hinzugefügt. Diese Zeilen mit den entsprechenden Kommandos entsprechen nun genau dem Code, den SPSS ausgeführt hätte, wenn wir vorhin auf „OK“ geklickt hätten. Das ist also der Code, der unsere alte Variable in eine neue umkodiert. Theoretisch hätten wir diesen Code auch in eine Syntaxdatei eintippen können und dann ausführen und wir hätten dasselbe Ergebnis wie mit dem Klick auf „OK“ erhalten. Wir haben allerdings diesen Code noch nicht ausgeführt. Das machen wir jetzt, indem wir die Codezeilen markieren und dann auf das grüne „Abspielen“-Symbol klicken.

In der Variablenansicht sollte nun eine neue Variable mit dem Namen „politik_politik2_umk“ und dem Label, das wir vorhin definiert haben, hinzugekommen sein. Alle anderen Einstellungen können wir jetzt noch vornehmen. Das heißt, wir ändern die Anzahl der Dezimalstellen auf 0, fügen die Werte wie für die beiden Items *politik_politik1* und *politik_politik2* ein (das geht einfach mittels Copy & Paste) und ändern noch das Skalenniveau und die Rolle der Variablen entsprechend (selbstverständlich können wir auch die übrigen kosmetischen Einstellungen noch anpassen). In der Datenansicht können wir uns nun davon überzeugen, dass Personen, die niedrige Werte bei der Variablen *politik_politik2* hatten, hohe Werte bei der neuen Variable *politik_politik2_umk* aufweisen und umgekehrt. Wir sind also jetzt mit dem Umkodieren der Variable wirklich fertig, juhu!

Aber warum haben wir oben bloß diesen Umweg über die Syntaxdatei gemacht? Wie gesagt, das hat mit guter Praxis der Datenanalyse zu tun. Bei einer Datenanalyse können unter Umständen sehr viele Arbeitsschritte erfolgen bis man beim Endergebnis angelangt ist. Möchte man nun später wissen, was man selbst oder jemand anders bei einer Datenanalyse genau gemacht hat, kann man das sehr schlecht, wenn überhaupt nachvollziehen, wenn sich diejenige Person bei der Analyse einfach nur durch

alle Schritte „hindurchgeklickt“ hat und einem am Ende die fertige Ausgabedatei übergibt (oder in unserem aktuellen Fall den Datensatz mit der neuen Variable). Fügt man allerdings alle Arbeitsschritte in eine Syntaxdatei ein und speichert diese ab, so sind dort bereits alle Arbeitsschritte dokumentiert und man kann diese sogar noch mit Kommentaren versehen, um einzelne Schritte zu erläutern bzw. zu erklären (es ist nämlich überhaupt nicht immer selbstverständlich, weshalb man was genau an welcher Stelle macht und ganz selbsterklärend ist der SPSS Syntax Code auch nicht immer). Deshalb wechseln auch wir noch einmal in unser Syntaxfenster zurück und fügen über den Zeilen mit dem gerade ausgeführten Code noch die beiden Kommentarzeilen „* Kapitel 3: Datensatz Kap3daten.sav.“ und „* Umkodieren der Variable politik_politik2.“ hinzu. Danach speichern wir die Datei unter einem beliebigen Namen, z.B. „Kap3dokumentation.sps“, ab.

Mit allen weiteren Arbeitsschritten werden wir nun ebenso verfahren. Das heißt, wir werden sie immer erst in diese Syntaxdatei einfügen, die entsprechenden Kommandozeilen dort ausführen und zwischendurch abspeichern. Am Ende des Kapitels haben wir dann eine Dokumentation für alle Arbeitsschritte, die hier behandelt werden. Die entsprechende Dokumentation finden Sie ebenfalls in dem elektronischen Ergänzungsmaterial zu diesem Dokument, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Abbildung 3.2, Abbildung 3.3 und Abbildung 3.4 fassen noch einmal sämtliche Arbeitsschritte zum Umkodieren einer Variablen, die gerade im Detail erläutert wurden, zusammen. In den übrigen Abschnitten werden nicht alle Schritte verbal ausformuliert, sondern manchmal bloß auf entsprechende Abbildungen verwiesen. Je mehr Sie mit SPSS arbeiten, desto klarer sollte auch werden, dass man die meisten Abläufe nicht auswendig wissen muss, sondern mit ein bisschen Verständnis für das, was man eigentlich vorhat und der Fähigkeit sich umzuschauen bzw. sinnerfassend zu lesen, recht schnell an den einzelnen Bezeichnungen der Menüs und Optionen ablesen kann, wie man durchführen kann, was immer man eben vorhat. Am Anfang ist das natürlich noch sehr verwirrend, aber die eigene Übersicht ändert sich erfahrungsgemäß sehr rasch mit fortschreitender Übung.

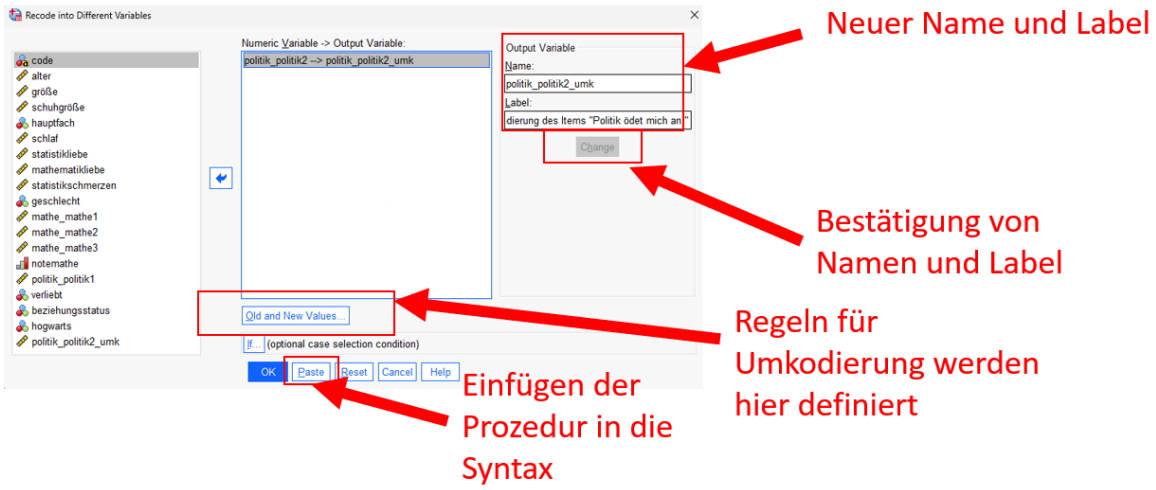


Abbildung 3.2. Fenster „Recode Into Different Variables…“.

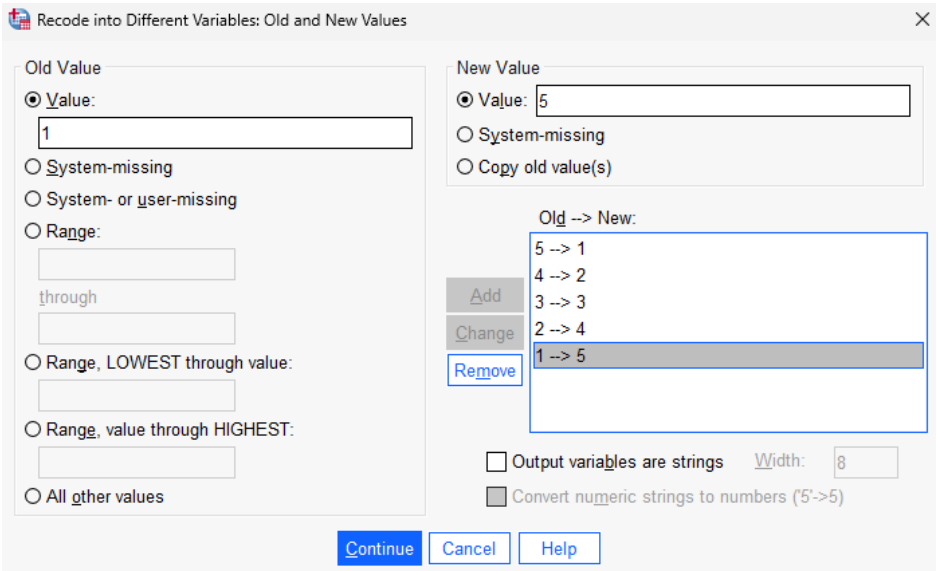


Abbildung 3.3. Fenster, das sich nach Klicken auf „Old and New Values…“ öffnet. Beachte, dass alle fünf Zuordnungen jeweils durch eine Angabe eines alten und eines neuen Werts und anschließendes Klicken auf „Add“ getätigt werden müssen.

Vielleicht ist auch aufgefallen, dass es neben der Option „Recode into Different Variables…“ auch die Möglichkeit gegeben hätte „Recode into Same Variables…“ auszuwählen. Damit hätten wir in der Tat, das Item *politik_politik2* mit seiner umkodierten Version direkt überschreiben können. Das mag angesichts der Tatsache, dass wir für die Erfassung der Politikaffinität von Personen das ursprünglich „verkehrt“ herum kodierte Item ja gar nicht mehr brauchen, durchaus sinnvoll erscheinen. Aus Sicht

guter Analysepraxis ist es aber keine gute Idee das zu machen, weil dadurch das ursprüngliche Item verlorengeht bzw. der ursprüngliche Datensatz verändert wird. Will man dann zu einem späteren Zeitpunkt noch einmal ganz neu anfangen, weil man sich irgendwo festgefahren hat, kann man u.U. nicht mehr zum originalen Datensatz zurück (mit einer guten Dokumentation ließe sich das zwar umkehren, aber benötigt trotzdem zusätzliche Arbeitsschritte). Es wird also empfohlen, niemals den ursprünglichen Datensatz direkt zu verändern, sondern in solchen Fällen immer neue Variablen zu generieren.

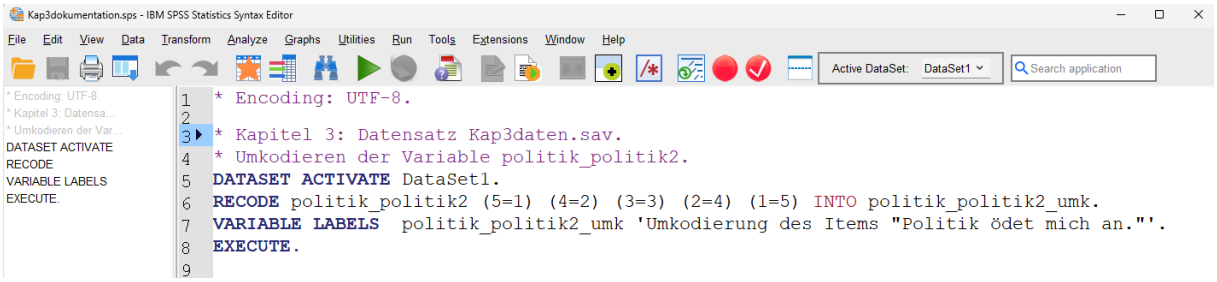


Abbildung 3.4. Syntaxdatei mit der Dokumentation der Prozedur für das Umkodieren der Variablen *politik_politik2* in die Variable *politik_politik2_umk*.

Index- oder Skalenbildung

Oben wurde bereits davon gesprochen, die beiden Items *politik_politik1* und *politik_politik2* bzw. *politik_politik2_umk* zu kombinieren, um einen Zahlenwert zu erhalten, der die Politikaffinität einer Person insgesamt beschreiben soll. Um einen solchen Index oder eine solche Skala zu bilden, können z.B. einfach die Summe oder der Mittelwert der beiden Items für jede Person berechnet werden. In diesen Fällen spricht man dann von einer Summen- oder einer Mittelwertskala (oder -index). Auch das lässt sich einfach mit SPSS machen.

Dafür wählen wir wieder das Menü „Transform“ und dort „Compute Variable…“ oder in Kurzform (wie es im Folgenden häufig geschrieben werden wird): *Transform >> Compute Variable…*.

Im sich öffnenden Fenster geben wir unter „Target Variable“ einen Namen für unsere neue Variable an. Für den vorliegenden Fall schreiben wir hier „Politikaffinität_Summe“. Danach können wir das Item *politik_politik1* in das Feld „Numeric Expression“ ziehen, die Schaltfläche mit dem „+“ anklicken und dann noch das Item *politik_politik2_umk* in das Feld „Numeric Expression“ ziehen.

Daraufhin klicken wir wieder auf „Paste“, woraufhin zwei neue Zeilen in unsere Syntaxdatei eingefügt werden. Zu diesen fügen wir noch einen entsprechenden Kommentar hinzu. Dann sollte alles aussehen wie in Abbildung 3.5.

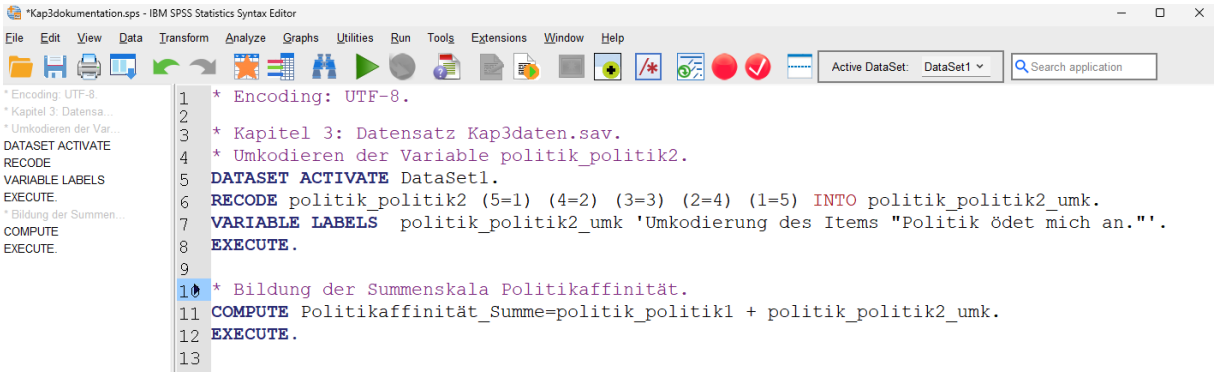


Abbildung 3.5. Bildung einer Summenskala für Politikaffinität aus den beiden Items *politik_politik1* *politik_politik2_umk*.

Ausführen der beiden letzten, gerade hinzugefügten Kommandozeilen erzeugt schließlich unsere Summenskala, die wir daraufhin in der Variablen- und Datenansicht bewundern können, und noch die nötigen Eigenschaften für sie definieren bzw. nachtragen sollten.

Für eine einfache Summenskala ist oft der Umweg über *Transform* >> *Compute Variable*... gar nicht notwendig. Wenn man das häufig macht, ist es bequemer einfach gleich die entsprechenden Zeilen in der Syntax einzutragen.

Ganz analog kann eine Mittelwertskala gebildet werden. Unter *Transform* >> *Compute Variable*... ist dafür erstmal ein neuer Name für diese Variable einzugeben, z.B. „Politikaffinität_Mittelwert“. Danach kann im Feld „Function group“ der Begriff „Statistical“ ausgewählt werden. Unter „Functions and Special Variables“ kann dann „Mean“ ausgewählt werden. Im Feld „Numeric Expression“ sind dann schließlich noch die beiden „?“ durch die beiden Items *politik_politik1* und *politik_politik2_umk* zu ersetzen (z.B. per Drag-and-drop oder per Doppelklick auf den Variablennamen). Mittelwerte können auch für mehr als zwei Items gebildet werden. Dafür muss innerhalb der Klammern schlichtweg ein weiteres Komma und dann eine weitere entsprechende Variable (und dies eventuell wiederholt) eingegeben werden.

Durch Klicken auf „Paste“ werden wieder die entsprechenden Kommandos der Syntaxdatei hinzugefügt. Wenn man das Kommando für die Berechnung des Mittelwerts kennt, kann man alternativ die entsprechende Zeile natürlich auch einfach gleich händisch in die Syntaxdatei eintragen. In jedem Fall sind die beiden Zeilen danach noch auszuführen (und zu kommentieren; gute Analysepraxis!). Die Syntaxdatei sollte dann wie in Abbildung 3.6 aussehen.

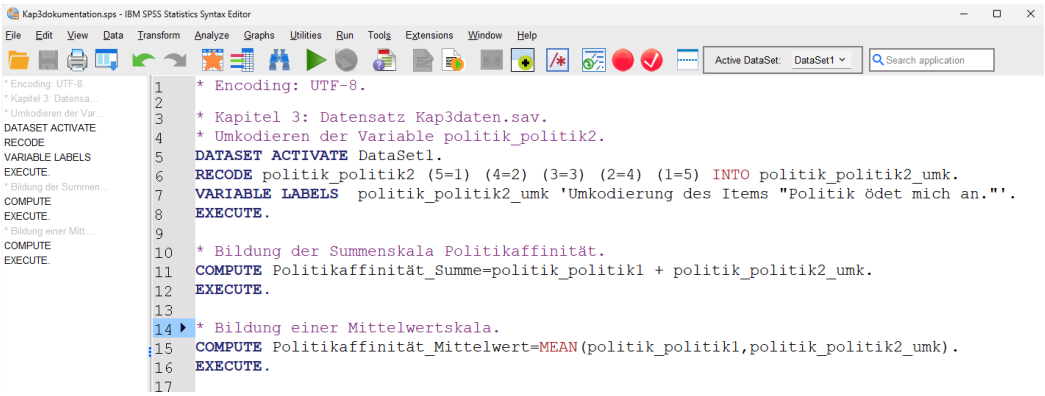


Abbildung 3.6. Bildung einer Mittelwertskala für Politikaffinität aus den beiden Items *politik_politik1* *politik_politik2_umk*.

Nach dem Ausführen der beiden soeben hinzugefügten Kommandozeilen gibt es wieder eine neue Variable in unserem Datensatz.

Kategorienbildung

Eine weitere Sache, die manchmal ganz nützlich sein kann, ist die Bildung von Kategorien. Beispielsweise kann es sein, dass man (z.B. zum Zwecke der Anonymisierung) nicht das genaue Alter von Personen angeben will, sondern lediglich Alterskategorien.

Wir sehen uns die Bildung von Kategorien allerdings am Beispiel der Schuhgrößen an. Die entsprechende Variable sollte sich im Datensatz relativ einfach ausfindig machen lassen. Unter *Transform >> Recode into Different Variables…* entfernen wir nun zuerst den bestehenden Eintrag im Feld „Input Variable -> Output Variable“ (der noch da ist, weil wir vorhin eine Variable umkodiert haben) und ziehen dann die Variable *schuhgröße* in dieses Feld. Als Namen für unsere neue Variable wählen wir „Schuhgrößenkategorie“ und als Label „Schuhgröße in den Kategorien klein (0) und groß (1)“. Dann klicken wir wieder auf „Change“ und das Ergebnis sollte wie Abbildung 3.7 aussehen.

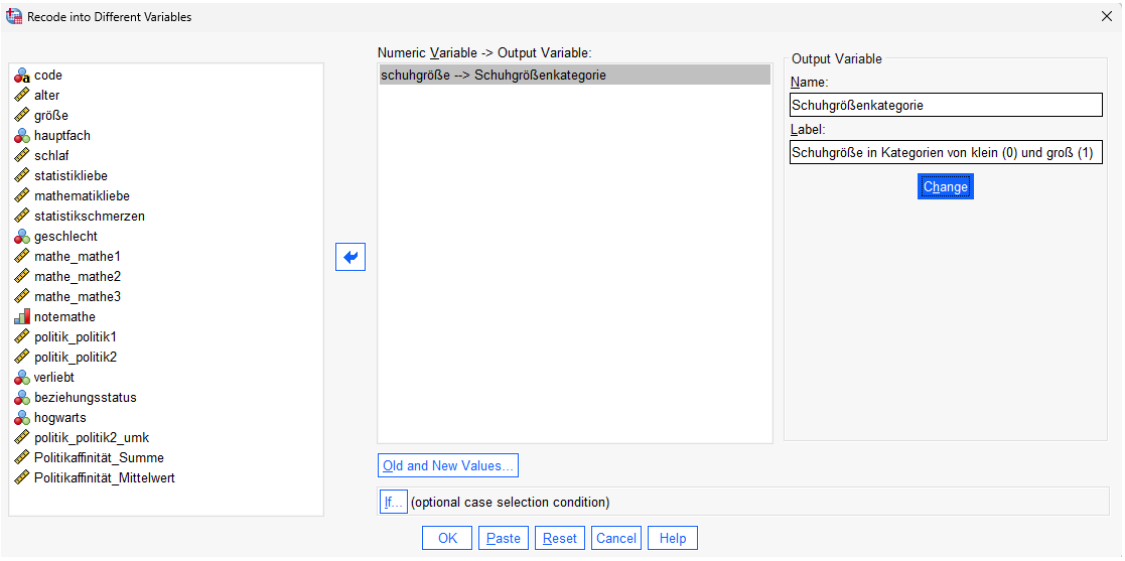


Abbildung 3.7. Bildung von Kategorien durch Umkodieren einer Variablen.

Im Untermenü „Old and New Values…“ entfernen wir zuerst alle alten Einträge aus dem Feld „Old --> New“. Dann wählen wir links „Range, LOWEST through value“ aus und tragen in dem Feld direkt darunter die Zahl 39 ein. In dem Feld unter „New Value“ tragen wir die Zahl 0 ein. Anschließend klicken wir auf „Add“. Danach wählen wir links „Range, value through HIGHEST“ aus und tragen in dem Feld direkt darunter die Zahl 40 ein. In dem Feld unter „New Value“ tragen wir die Zahl 1 ein. Anschließend klicken wir auf „Add“. Was wir gerade gemacht haben, bedeutet folgendes: Alle Schuhgrößen von der kleinsten bis zur Größe 39 werden der Kategorie 0, alle Schuhgrößen ab 40 bis zur größten der Kategorie 1 zugeordnet. Das Ganze sollte aussehen wie in Abbildung 3.8. Abschließend klicken wir auf „Continue“, dann auf „Paste“, kommentieren die neu hinzugekommenen Zeilen in der Syntaxdatei und führen diese aus. In der Variablenansicht sollten wir die neu hinzugekommen Variable noch mindestens um die Werte 0 mit dem Label „klein“ und 1 mit dem Label „groß“ in der Spalte „Values“ ergänzen (in die Zelle klicken und mit dem +-Symbol dann die entsprechenden zwei Zeilen ergänzen). In der Datenansicht können wir uns schließlich noch davon überzeugen, dass nun tatsächlich jede Person mit einer Schuhgröße von 39 oder kleiner in der Kategorie 0 und jede Person mit einer Schuhgröße von 40 oder größer in der Kategorie 1 gelandet ist. Mit der Schaltfläche „Value Labels“, siehe Abbildung 3.9, können wir auch zwischen den Zahlenwerten und den Kategorienbezeichnungen für kategoriale Variablen hin und her schalten (sofern diese definiert wurden, was wir aber aus diesem Grund für die Variable Schuhgrößenkategorie gerade getan haben).

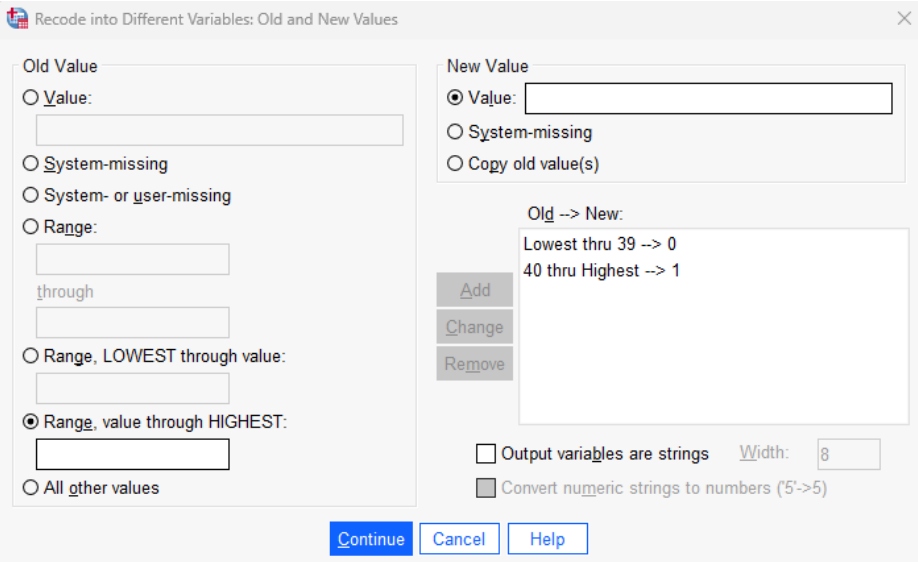
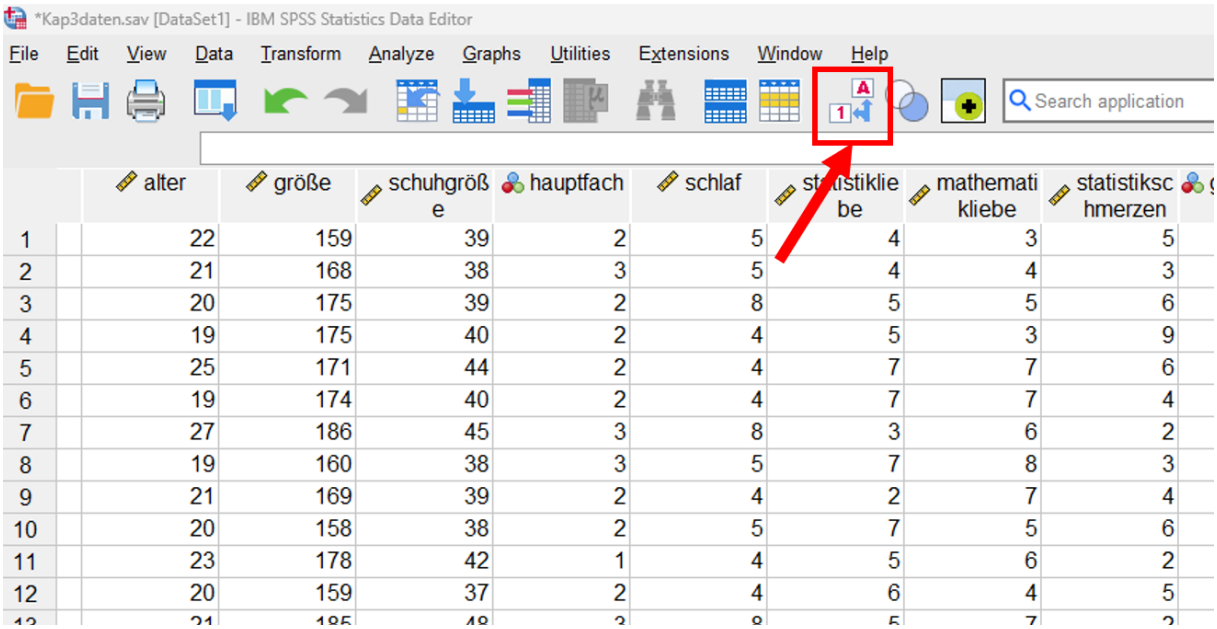


Abbildung 3.8. Kategorienbildung für Schuhgrößen.



	alter	größe	schuhgröße	hauptfach	schlaf	statistikliebe	mathematikliebe	statistikschmerzen
1	22	159	39	2	5	4	3	5
2	21	168	38	3	5	4	4	3
3	20	175	39	2	8	5	5	6
4	19	175	40	2	4	5	3	9
5	25	171	44	2	4	7	7	6
6	19	174	40	2	4	7	7	4
7	27	186	45	3	8	3	6	2
8	19	160	38	3	5	7	8	3
9	21	169	39	2	4	2	7	4
10	20	158	38	2	5	7	5	6
11	23	178	42	1	4	5	6	2
12	20	159	37	2	4	6	4	5

Abbildung 3.9. Schaltfläche „Value Labels“ in der Datenansicht.

Deskriptive Statistiken

Wie es der Name schon andeutet, dienen deskriptive Statistiken der Beschreibung. In unserem Fall dienen sie, genauer gesagt, der Beschreibung unserer Stichprobe(n). Es handelt sich also um Zahlen, die dabei helfen sollen, Stichproben zu charakterisieren. Zum Beispiel könnten wir uns fragen: Wie viele Männer und Frauen gibt es in unserer Stichprobe? Wie viele Personen hatten ein „Sehr gut" als Abschlussnote in Mathematik? Wie alt sind die Personen der Stichprobe im Mittel? Wie stark variiert das Alter der Personen in der Stichprobe? Für all diese Fragen gibt es Zahlen oder Größen, die zur Beantwortung herangezogen werden können. Damit werden wir uns in diesem Abschnitt befassen.

Häufigkeiten

Eine der einfachsten Möglichkeiten sich einen Überblick über die Ausprägungen einer bestimmten Variablen in einer Stichprobe zu verschaffen sind Häufigkeitstabellen. Diese können in SPSS über *Analyze >> Descriptive Statistics >> Frequencies…* generiert werden. Im linken Feld können wieder die Variablen ausgewählt werden, für die wir Häufigkeitstabellen generieren wollen. Wählen wir hier der Übersichtlichkeit halber erstmal nur drei aus: *alter*, *geschlecht*, *notemathe* (durch Halten der Taste „Strg" können mittels Linksklick gleich mehrere Variablen ausgewählt werden; mittels Rechtsklick irgendwo im linken Feld können wieder die Variablennamen anstelle der Labels angezeigt werden). Unter „Charts…" können zusätzlich noch einige Grafiken wie z.B. Balkendiagramme oder Histogramme ausgewählt werden. Wir wählen hier zusätzlich noch Balkendiagramme (d.h. „Bar Charts") aus. Wir bestätigen die Auswahl mit Klick auf „Continue" und klicken dann wieder auf „Paste". In der Syntaxdatei führen wir die soeben eingefügten Kommandozeilen wieder aus (nachdem wir sie entsprechend kommentiert haben; z.B. mit „*Häufigkeitstabellen und Balkendiagramme für Alter, Geschlecht, und Abschlussnote in Mathematik.").

Durch Ausführen der entsprechenden Kommandozeilen in der Syntaxdatei erhalten wir nun zum ersten Mal eine Ausgabe (sowohl numerisch als auch grafisch). Dafür öffnet sich das Ausgabefenster, das wir nun auch endlich aus gegebenem Anlass etwas eingehender untersuchen können. Zuallererst bietet es sich aber an, auch die Ausgabe gleich einmal zwischenzuspeichern, z.B. unter dem Dateinamen „Kap3ausgabe.spv". Sie finden eine Ausgabedatei zur Illustration für dieses Kapitel auch in dem

elektronischen Ergänzungsmaterial zu diesem Dokument, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Auf der linken Seite des Ausgabefensters sehen wir ein hierarchisches Inhaltsverzeichnis, das wir zum schnellen Manövrieren in der Ausgabe verwenden können. Jeder einzelne Teil der Ausgabe ist dort angeführt. Klicken wir zum Beispiel auf „Alter in Jahren“ unter „Bar Chart“ springen wir in der eigentlichen Ausgabe im rechten Feld sofort zum Balkendiagramm für unsere Altersvariable. Gerade bei sehr umfangreichen Ergebnissen kann es sehr praktisch sein, schnell zu einzelnen Abschnitten wechseln zu können.

Für den Moment ist unsere Ausgabe aber noch sehr übersichtlich. Zuallererst finden wir eine Angabe, auf welche Datendatei sich die gelisteten Ergebnisse überhaupt beziehen (diese erscheint u.U. nur, wenn mehrere Datendateien gleichzeitig geöffnet sind). Hier kann schnell erkannt werden, falls eine Analyse irrtümlich für eine falsche Datendatei durchgeführt wurde. Das kann schnell einmal passieren, wenn man viele Datendateien gleichzeitig geöffnet hat. Daher wird gerade für den Beginn der Arbeit mit (und Einübung in) SPSS empfohlen bei jeder Analyse nur einen einzigen Datensatz geöffnet zu haben.

Direkt im Anschluss finden wir eine Tabelle mit der Überschrift „Statistics“. Diese Tabelle zeigt uns lediglich an, dass für alle drei ausgewählten Variablen jeweils 51 Messwerte vorliegen und insbesondere für keine Variable fehlende Werte vorliegen, siehe Abbildung 3.10.

Statistics

		Alter in Jahren	Bitte geben Sie Ihr Geschlecht an.	Geben Sie Ihre Schulabschlussnote in Mathematik an.
N	Valid	51	51	51
	Missing	0	0	0

Abbildung 3.10. Unsere erste mit SPSS erzeugte Tabelle, wie schön!

Unter der Überschrift “Frequency Table“ finden wir anschließend drei Häufigkeitstabellen, je eine für unsere drei Variablen. In jeder Häufigkeitstabelle sind die einzelnen Messwertausprägungen der Größe nach geordnet (bei nominalen Variablen werden dafür, sofern vorhanden, die definierten Zahlenwerte für die einzelnen Kategorien verwendet) und sowohl absolute als auch relative Häufigkeiten (in Prozent) sowie kumulierte relative Häufigkeiten angegeben. Die Spalte mit der Bezeichnung „Valid Percent“ beinhaltet die relative Häufigkeit nach Bereinigung für fehlende Werte. Da in unserem Fall keine Werte fehlen, beinhalten die Spalten „Percent“ und „Valid Percent“ dieselben Werte.

An den Häufigkeitstabellen lassen sich nun bereits viele Eigenschaften über die Verteilung der drei Variablen in der Stichprobe ablesen. Beispielsweise erkennen wir an der Häufigkeitstabelle für das Alter, dass beinahe die Hälfte der Stichprobe 20 Jahre oder jünger ist, dass alle Personen volljährig sind, dass es nur eine Person über 30 gibt etc., siehe Abbildung 3.11. Zudem sehen wir an der Häufigkeitstabelle für das Geschlecht, dass fast drei Viertel der Stichprobe weiblich sind. An der Häufigkeitstabelle für die Abschlussnote in Mathematik sehen wir, dass mehr als die Hälfte mit „Sehr Gut“ oder „Gut“ abgeschlossen hat und knapp 10% mit „Genügend“, siehe Abbildung 3.12.

Alter in Jahren

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	5	9.8	9.8	9.8
	19	9	17.6	17.6	27.5
	20	10	19.6	19.6	47.1
	21	9	17.6	17.6	64.7
	22	4	7.8	7.8	72.5
	23	3	5.9	5.9	78.4
	24	2	3.9	3.9	82.4
	25	3	5.9	5.9	88.2
	26	2	3.9	3.9	92.2
	27	1	2.0	2.0	94.1
	28	1	2.0	2.0	96.1
	29	1	2.0	2.0	98.0
	37	1	2.0	2.0	100.0
	Total	51	100.0	100.0	

Abbildung 3.11. Die von SPSS generierte Häufigkeitstabelle für die Variable *alter*, mit der das Alter der Personen in der Stichprobe erfasst wurde.

Bitte geben Sie Ihr Geschlecht an.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Weiblich	37	72.5	72.5	72.5
	männlich	14	27.5	27.5	100.0
	Total	51	100.0	100.0	

Geben Sie Ihre Schulabschlussnote in Mathematik an.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	10	19.6	19.6	19.6
	2	21	41.2	41.2	60.8
	3	15	29.4	29.4	90.2
	4	5	9.8	9.8	100.0
	Total	51	100.0	100.0	

Abbildung 3.12. Die beiden Häufigkeitstabellen für das Geschlecht und die Schulabschlussnote in Mathematik für die Personen in der Stichprobe.

Auch wenn hier zu Illustrationszwecken die von SPSS generierten Tabellen in Form von Bildern gezeigt werden, wird hier bereits darauf hingewiesen, dass das Einfügen von Tabellen in Form von Abbildungen in Ergebnisberichten keine gute Analysepraxis ist (da sich dann u.a. einzelne Zahlen, Zeilen oder Spalten für etwaige Weiterverarbeitung der Daten durch Dritte nicht aus den Berichten extrahieren lassen; bzw. nicht auf vergleichsweise einfachem Weg). Für Ergebnisberichte sollten also im entsprechenden Dokument wirklich auch Tabellen erstellt werden, wenn tabellarische Ergebnisse zu berichten sind oder der Verständlichkeit der Ergebnisdarstellung dienlich sind.

Die angeforderten Balkendiagramme geben uns schließlich noch visuell Aufschluss über die Verteilung der drei Variablen in der Stichprobe. Das resultierende Balkendiagramm für die Mathematik-Abschlussnoten ist in Abbildung 3.13 (oberes Panel) dargestellt. Für die Balkendiagramme sind absolute Häufigkeiten auf der Ordinate abgetragen, im Kontextmenü unter „Charts…“ unter *Analyze >> Descriptive Statistics >> Frequencies…* können hierfür aber auch relative Häufigkeiten ausgewählt werden, indem „Percentages“ unter „Chart Values“ ausgewählt wird.

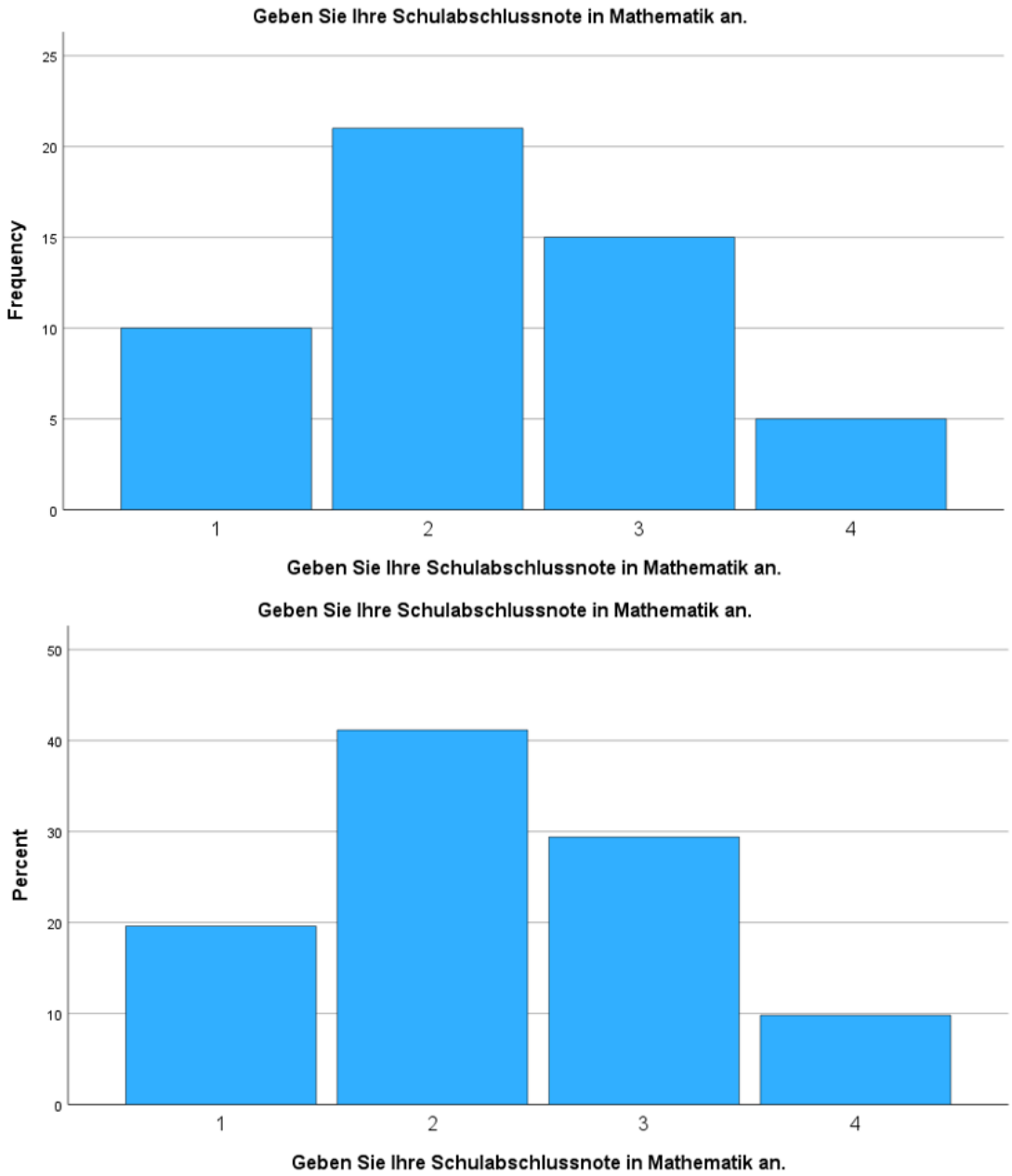


Abbildung 3.13. Oben: Balkendiagramm für die Mathematik-Abschlussnoten in absoluten Häufigkeiten. Unten: Balkendiagramm für die Mathematik-Abschlussnoten in relativen Häufigkeiten (in Prozent).

Aus den Häufigkeitstabellen lässt sich auch bereits ablesen, welche Ausprägung der jeweiligen Variable am häufigsten vorkommt. Zum Beispiel ist das häufigste Alter 20 Jahre, das häufigste Geschlecht „weiblich“ und die häufigste Abschlussnote „Gut“. Bei diesen Werten handelt es sich also um die sogenannten Modalwerte für diese Variablen. Der Modalwert ist bereits ein Beispiel für Maßzahlen zur Beschreibung von Stichproben. Diesen wenden wir uns jetzt zu.

Maßzahlen

Gerade für metrische Variablen mit sehr vielen einzelnen Ausprägungen sind oft Häufigkeitstabellen weniger informativ. Um die Verteilung von metrischen Variablen über die Stichprobe zu charakterisieren, bietet es sich stattdessen eher an, Maßzahlen wie Mittelwert, Median, Standardabweichung, Schiefe (Engl.: Skewness) oder Wölbung (auch: Kurtosis) zu ermitteln sowie auf grafische Darstellungen wie Boxplots oder Histogramme zurückzugreifen.

In SPSS gibt es viele verschiedene Möglichkeiten sich Maßzahlen ausgeben zu lassen. Im Folgenden sind einige Möglichkeiten wiederum für das Beispiel der Variable *alter* angeführt.

Eine Möglichkeit besteht beispielsweise in der Auswahl der gewünschten Maßzahlen unter *Analyze >> Descriptive Statistics >> Frequencies…* und dort unter „Statistics…“, siehe Abbildung 3.14. In diesem Fall haben wir den Mittelwert, den Median, den Modalwert als Lagemaße, die Standardabweichung, das Minimum und Maximum sowie die Spannweite (= Maximum – Minimum) als Streuungsmaße, sowie Schiefe und Wölbung zur Charakterisierung der Form der Verteilung des Alters in der Stichprobe ausgewählt.

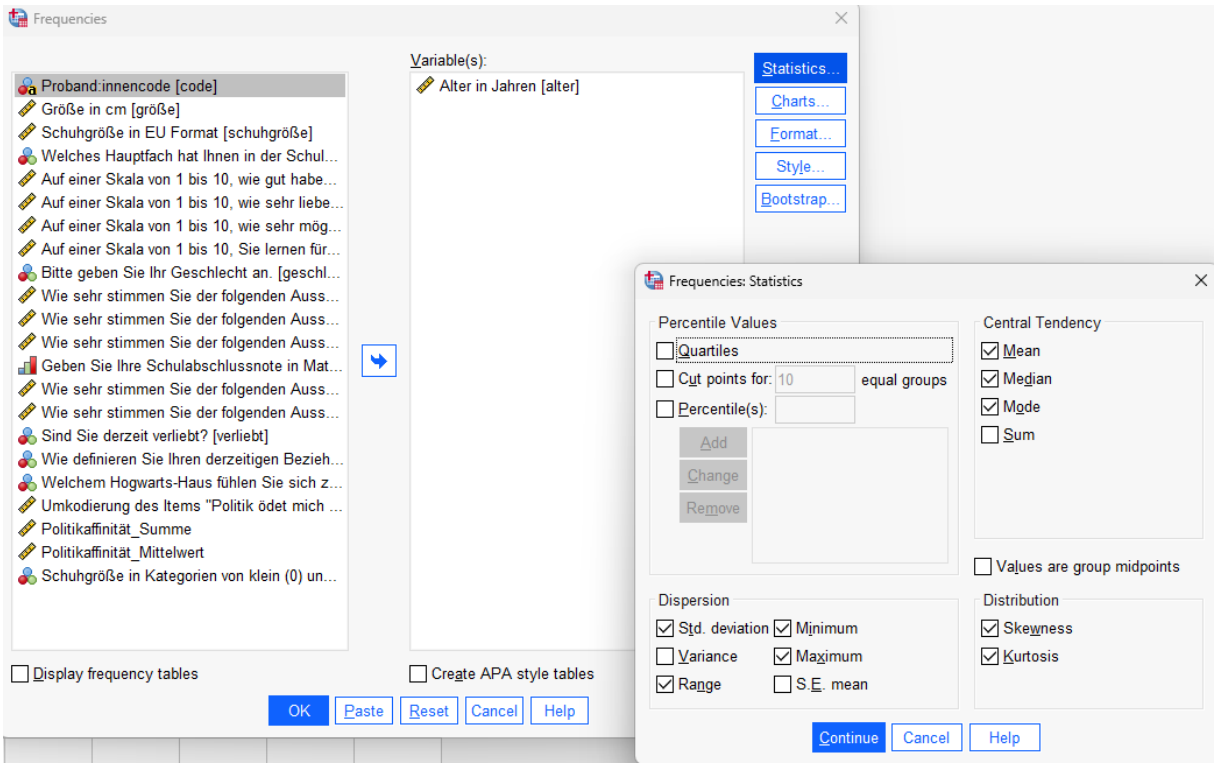


Abbildung 3.14. Auswahl einiger deskriptiver Statistiken für die Variable *alter*.

Das Ergebnis ist in Abbildung 3.15 dargestellt. Wir sehen wiederum, dass 51 Messwerte vorliegen und für keine Person das Alter fehlt. Das mittlere Alter in der Stichprobe beträgt 21.63 Jahre, der Median liegt bei 21 Jahren. Das hätten wir auch schon an der Häufigkeitstabelle oben ablesen können, da der Median ja gerade jene Variablenausprägung ist, die die Reihe der Größe nach geordneter Messwerte in zwei gleich große Hälften teilt. Auch der Modalwert von 20 Jahren war uns schon bekannt (dieser ist aber bei metrischen Variablen kaum je interessant). Die Standardabweichung beträgt ferner 3.49 Jahre. Hierbei ist wichtig zu beachten, dass es sich dabei nicht um die empirische Standardabweichung

$$s_{emp} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

handelt, sondern um den Schätzwert der Populationsstandardabweichung auf Basis der Stichprobe mittels der (erwartungstreuen) Schätzfunktion für die Populationsvarianz σ^2, d.h.

$$s = \sqrt{\hat{\sigma}^2_{Wert}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Statistics

Alter in Jahren

N	Valid	51
	Missing	0
Mean		21.63
Median		21.00
Mode		20
Std. Deviation		3.487
Skewness		2.120
Std. Error of Skewness		.333
Kurtosis		6.550
Std. Error of Kurtosis		.656
Range		19
Minimum		18
Maximum		37

Abbildung 3.15. Ergebnistabelle für die angeforderten Maßzahlen für die Verteilung der Altersvariable in der Stichprobe.

Am positiven Wert der Schiefe erkennen wir, dass es sich um eine rechtsschiefe bzw. linkssteile Verteilung handelt, siehe Abbildung 3.16. Das heißt, dass sich die Verteilung von einer symmetrischen Verteilung nach links hin weg neigt, was auch daran erkennbar ist, dass Median und Modalwert jeweils kleiner als der Mittelwert sind. Die positive Wölbung (Kurtosis) zeigt an, dass die Flanken der Altersverteilung zudem ausgeprägter sind als bei einer Normalverteilung, dass es also mehr extreme Ausreißer geben könnte als aufgrund einer Normalverteilung zu erwarten wären. Hier ist wichtig zu bedenken, dass SPSS nicht die eigentliche Wölbung, sondern den Exzess (Wölbung minus 3) gegenüber einer Normalverteilung ausgibt (eine Normalverteilung hat eine Wölbung von 3). Der Vergleich der Werte für Schiefe und Wölbung mit ihren jeweiligen Standardfehlern zeigt zudem an, dass beide Maßzahlen deutlich von denjenigen abweichen, die man für die Ziehung einer einfachen Zufallsstichprobe aus einer normalverteilten Grundgesamtheit erwarten könnte. In beiden Fällen weist ein Verhältnis der jeweiligen Maßzahl zu ihrem Standardfehler von mehr als 2 auf eine deutliche Abweichung hin. Zur Bedeutung dieser Maßzahlen für die Einschätzung, ob eine Variable durch eine normalverteilte Zufallsvariable approximiert werden kann, werden wir aber in späteren Kapiteln noch kommen. Es macht also nichts, falls diese Informationen jetzt noch sehr abstrakt klingen.

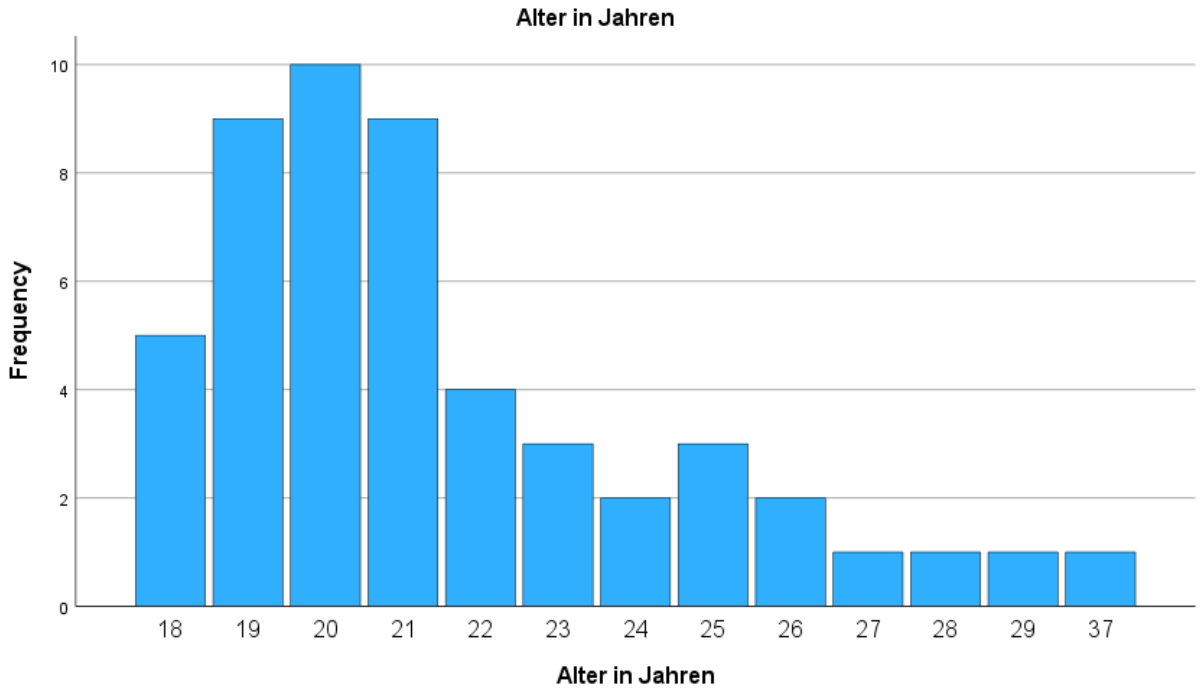


Abbildung 3.16. Rechtsschiefe bzw. linkssteile Altersverteilung.

Die Spannweite gibt schließlich an, dass das Alter in der Stichprobe einen Bereich von 19 Jahren abdeckt. Dies lässt sich auch leicht an Maximum und Minimum ablesen: die älteste Person war 37 Jahre alt, die jüngste 18.

Eine weitere Möglichkeit sich einen grafischen Überblick über die Verteilung des Alters in der Stichprobe zu verschaffen, besteht in einem sogenannten Boxplot. Dieses kann über *Graphs >> Chart Builder…* generiert werden. Falls wir den „Chart Builder“ noch nie verwendet haben, erscheint nun eine Meldung, die uns erklärt, dass es für ein angemessenes Funktionieren des „Chart Builder“-Assistenten wichtig ist, dass die Skalenniveaus aller Variablen korrekt angegeben werden. Sollten wir daran noch Zweifel haben, können wir über die Schaltfläche „Define Variable Properties…“ zu einem Assistenten gelangen, der uns dabei helfen kann, passende Skalenniveaus für unsere Variablen zu definieren. Wenn wir uns allerdings wie hier sicher sind, dass wir bei der Definition der Variablen alles richtig gemacht haben, können wir auch einfach auf die Schaltfläche „OK“ klicken, um zum eigentlichen „Chart Builder“-Assistenten zu gelangen. Durch Anwählen der Option „Don’t show this dialog again“ können wir das Erscheinen dieses Dialogfensters bei der nächsten Auswahl des „Chart Builder“-Assistenten auch verhindern.

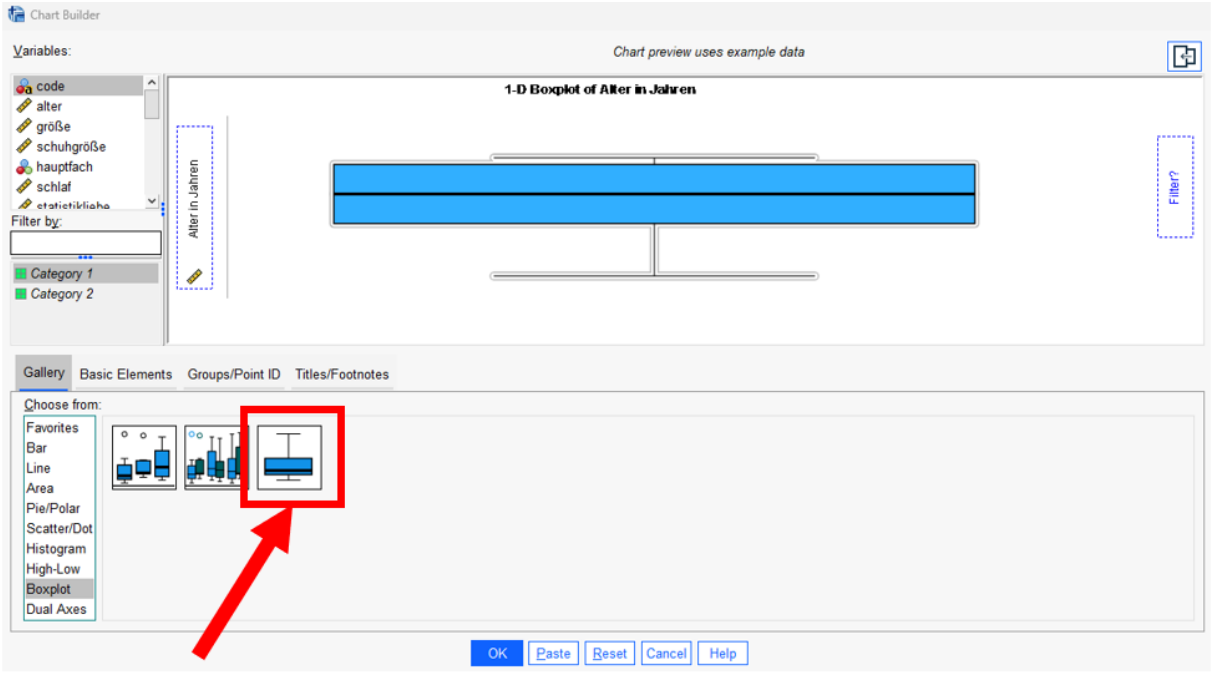


Abbildung 3.17. Auswahl eines einfachen Boxplots unter *Graphs >> Chart Builder…*.

Im Chart Builder kann jedenfalls anschließend aus dem Menü links unten „Boxplot“ und in der dortigen Auswahl durch Doppelklick dann das einfache Boxplot (dritte Möglichkeit ganz rechts) ausgewählt werden, siehe Abbildung 3.17. Die Variable *alter* kann dann auf die y-Achse (Ordinate; in SPSS vor Einfügen der Variablen allerdings als x-Achse bezeichnet) gezogen werden. Mittels „Paste“ kann der nötige Code in die Syntaxdatei eingefügt werden. Die Ausführung des Codes erzeugt dann das Boxplot, das in Abbildung 3.18 gezeigt ist.

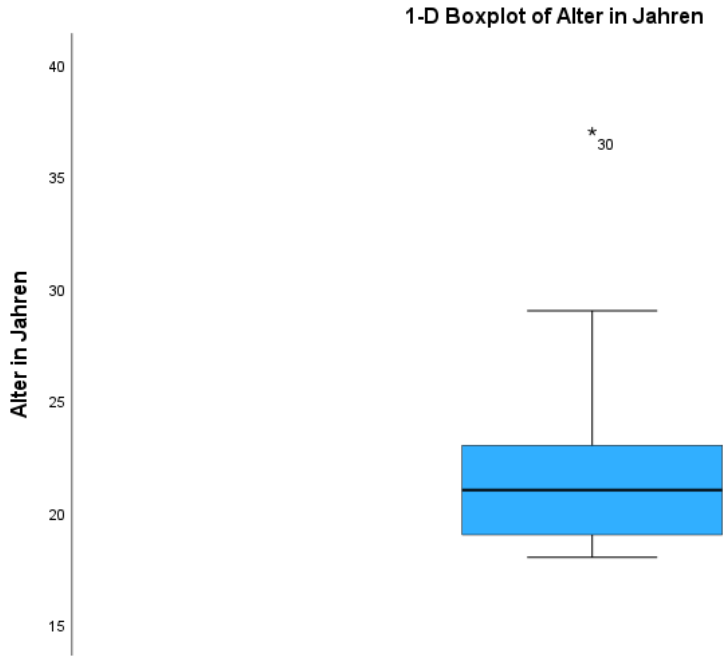


Abbildung 3.18. Boxplot der Altersverteilung.

Das resultierende Boxplot illustriert ebenfalls die deutliche Schiefe der Altersverteilung. Die mittleren 50% der Variable Alter drängen sich am unteren Ende der Verteilung, was man an der Lage der blauen Box in Abbildung 3.18 erkannt. Die obere Kante dieser Box entspricht dem 3. Quartil der Altersverteilung, d.h. jenem Alter unterhalb dessen 75% aller Messwerte liegen. Die untere Kante entspricht dem ersten Quartil, d.h. jenem Alter unterhalb dessen 25% aller Messwerte liegen. D.h. zwischen der unteren und der oberen Kante, d.h. innerhalb der Box liegen die Hälfte der Messwerte. Nach unten erstrecken sich die unteren 25% nicht sehr weit (da das Mindestalter in der Stichprobe 18 Jahre beträgt). Nach oben hin (also zu höherem Alter) erstrecken sich die Messwerte also noch recht weit. Mit einem 37-jährigen liegt sogar ein extremer Ausreißer (erkennbar durch den Stern in der Darstellung) vor. Extreme Ausreißer sind durch mehr als 3 Interquartilsabstände von der

nächstliegenden Kante der Box charakterisiert. Gewöhnliche Ausreißer sind durch mehr als 1.5 Interquartilsabstände von der nächstliegenden Kante der Box charakterisiert und würden durch einen Kreis dargestellt werden (hier haben wir allerdings keine gewöhnlichen Ausreißer). Messwerte zwischen dem Minimum und dem ersten Quartil bzw. zwischen dem dritten Quartil und dem Maximum werden durch die sogenannten Whiskers (die T-förmigen Ausformungen in Abbildung 3.18) dargestellt.

Wie oben bereits angekündigt, gibt es in SPSS aber noch jede Menge andere Möglichkeiten sich deskriptive Statistiken (und gegebenenfalls noch weitere grafische Veranschaulichungen) ausgeben zu lassen. Eine weitere Möglichkeit besteht unter *Analyze >> Descriptive Statistics >> Descriptives…*, wo unter „Options…" dann eine Auswahl an Maßzahlen getroffen werden kann (allerdings kann dort weder Median noch Modalwert ausgegeben werden).

Unter *Analyze >> Descriptive Statistics >> Descriptives…* kann zudem auf einfache Weise eine z-Transformation von Variablen durchgeführt werden. Dazu muss nur die Option „Save standardized values as variables" angewählt werden, siehe Abbildung 3.19. Die z-Transformation einer Variablen ergibt eine neue Variable mit Mittelwert 0 und Standardabweichung 1 und ist gegeben durch:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Die Verwendung z-transformierter Variablen kann im Rahmen linearer Regressionsmodelle (Kapitel 9-12) für die Interpretation von Ergebnissen nützlich sein.

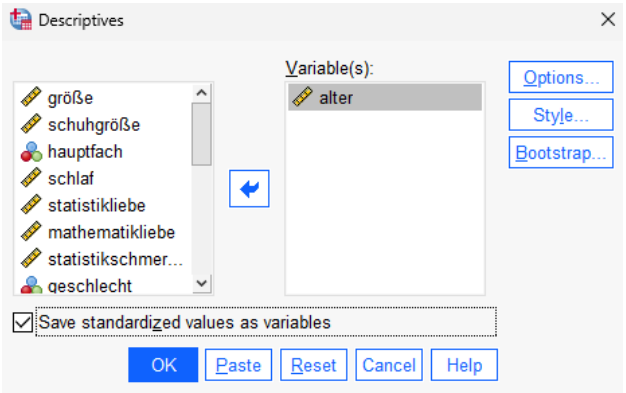


Abbildung 3.19. Das Menü „Descriptives" unter *Analyze >> Descriptive Statistics* bietet eine einfache Möglichkeit eine z-Transformation von Variablen durchzuführen.

Kreuztabellen

Kreuztabellen sind eine nützliche Möglichkeit um die Verteilung einer kategorialen Variablen über mehrere Kategorien einer anderen kategorialen Variablen hinweg darzustellen. Kreuztabellen können unter *Analyze* >> *Descriptive Statistics* >> *Crosstabs* generiert werden. Unter „Cells…“ können dort zudem relative Häufigkeiten für Zeilen, Spalten oder alle Zellen der sich ergebenden Kreuztabelle angefordert werden. Abbildung 3.20 zeigt die nötigen Eingaben für die Erstellung einer Kreuztabelle für die beiden Variablen *geschlecht* und *hauptfach*. Abbildung 3.21 zeigt die in der Ausgabe resultierende Kreuztabelle.

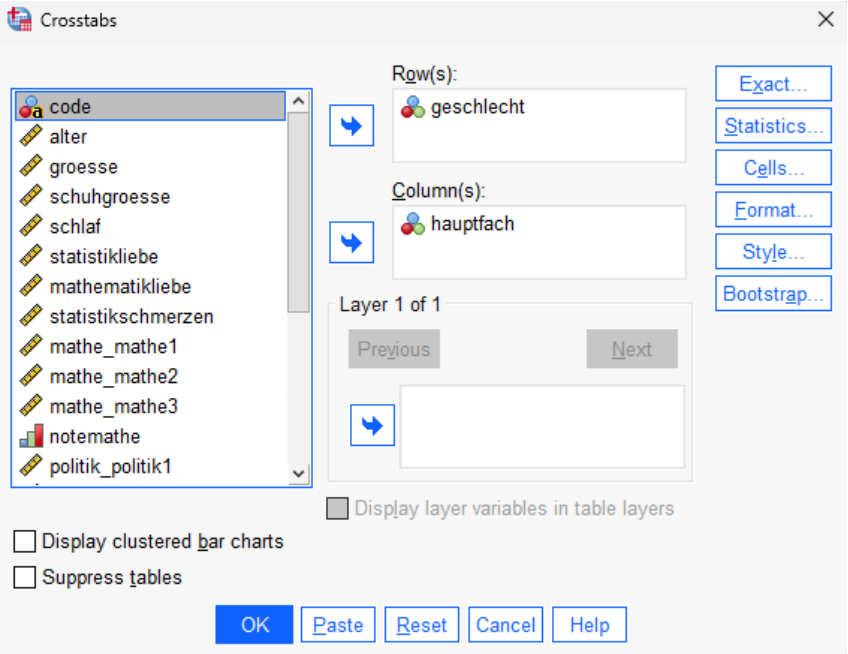


Abbildung 3.20. Erstellung einer Kreuztabelle (SPSS Version 29; in SPSS Version 30 ist in diesem Menü anstelle von „Row(s)“ die Bezeichnung „TargetList“ zu finden; lassen Sie sich davon nicht beirren, wie alles im Leben, verändern sich auch Programme und Software mit der Zeit bzw. mit fortschreitender Versionsnummer).

Bitte geben Sie Ihr Geschlecht an. * Welches Hauptfach hat Ihnen in der Schule am meisten Freude bereitet? Crosstabulation

Count

		Welches Hauptfach hat Ihnen in der Schule am meisten Freude bereitet?			
		Deutsch	Englisch	Mathematik	Total
Bitte geben Sie Ihr Geschlecht an.	Weiblich	8	22	7	37
	männlich	3	3	8	14
Total		11	25	15	51

Abbildung 3.21. Resultierende Kreuztabelle für die Variablen *geschlecht* und *hauptfach* in der Schule.

Maße des Zusammenhangs zwischen metrischen Variablen – Korrelationen

Bisher haben wir hauptsächlich deskriptive Statistiken betrachtet, die zur Charakterisierung der Lage (z.B. Mittelwert oder Median) oder der Streuung (z.B. Standardabweichung) oder anderer Charakteristika einzelner metrischer (oder auch kategorialer) Variablen verwendet werden. Mit den Kreuztabellen im vorhergehenden Abschnitt haben wir eine einfache Möglichkeit kennengelernt, eventuelle Zusammenhänge zwischen kategorialen Variablen zu veranschaulichen. Manchmal sind wir allerdings auch an Zusammenhängen zwischen metrischen Variablen interessiert. Um Maßzahlen zur Beschreibung solcher Zusammenhänge geht es in diesem Abschnitt. Die Eigenschaften und Unterschiede dieser Maßzahlen – und auch wie sie mithilfe von SPSS berechnet werden können – werden allesamt an folgendem Beispiel illustriert.

Im Datensatz „Kap3daten2.sav“ sind Gewicht und Größe für 20 (sehr athletische) Männer und Frauen gegeben, siehe Abbildung 3.22. Im Falle einer durchwegs sehr athletischen Stichprobe ist es naheliegend, dass zwischen Größe und Gewicht ein enger Zusammenhang besteht. Um uns von diesem Zusammenhang im wahrsten Sinne des Wortes ein Bild zu machen, können wir, bevor wir versuchen den Zusammenhang zu quantifizieren (d.h., in eine Zahl zu fassen), in SPSS ein sogenanntes Streudiagramm anfordern. Dazu wählen wir in der geöffneten Datendatei unter „Graphs“ den ersten Punkt „Chart Builder…“ aus. Im „Chart Builder“-Assistenten wählen wir unter der Rubrik „Gallery“ links unten „Scatter/Dot“ aus, und unter den daraufhin erscheinenden Optionen durch Doppelklick „Scatter Plot“, siehe Abbildung 3.23. Schließlich ziehen wir die Variable Größe auf die x-Achse und die Variable Gewicht auf die y-Achse wie in Abbildung 3.23 illustriert. Anschließend klicken wir auf „Paste“ und führen die eingefügten Kommandozeilen in der Syntax aus.

	ID	Größe	Gewicht	BMI	Geschlecht	Größe_zentriert	Gewicht_zentriert
1	JoCe	1.85	114.00	33.31	2	.08	28.30
2	Rock	1.96	118.00	30.72	2	.19	32.30
3	RaOr	1.96	113.00	29.41	2	.19	27.30
4	UnTa	2.08	140.00	32.36	2	.31	54.30
5	TriH	1.93	116.00	31.14	2	.16	30.30
6	SeRo	1.85	102.00	29.80	2	.08	16.30
7	SaZa	1.85	96.00	28.05	2	.08	10.30
8	AJSt	1.80	100.00	30.86	2	.03	14.30
9	Rey	1.68	79.00	27.99	2	-.09	-6.70
10	Punk	1.88	99.00	28.01	2	.11	13.30
11	AlBl	1.55	46.00	19.15	1	-.22	-39.70
12	ZeVe	1.54	48.00	20.24	1	-.23	-37.70
13	BeLy	1.68	61.00	21.61	1	-.09	-24.70
14	ChGr	1.70	57.00	19.72	1	-.07	-28.70
15	NiJa	1.83	123.00	36.73	1	.06	37.30
16	BiBe	1.70	70.00	24.22	1	-.07	-15.70
17	LiMo	1.65	57.00	20.94	1	-.12	-28.70
18	lySk	1.56	54.00	22.19	1	-.21	-31.70
19	RhRi	1.70	62.00	21.45	1	-.07	-23.70
20	LyVa	1.67	59.00	21.16	1	-.10	-26.70

Abbildung 3.22. Datenansicht zum Datensatz „Kap3daten2.sav“.

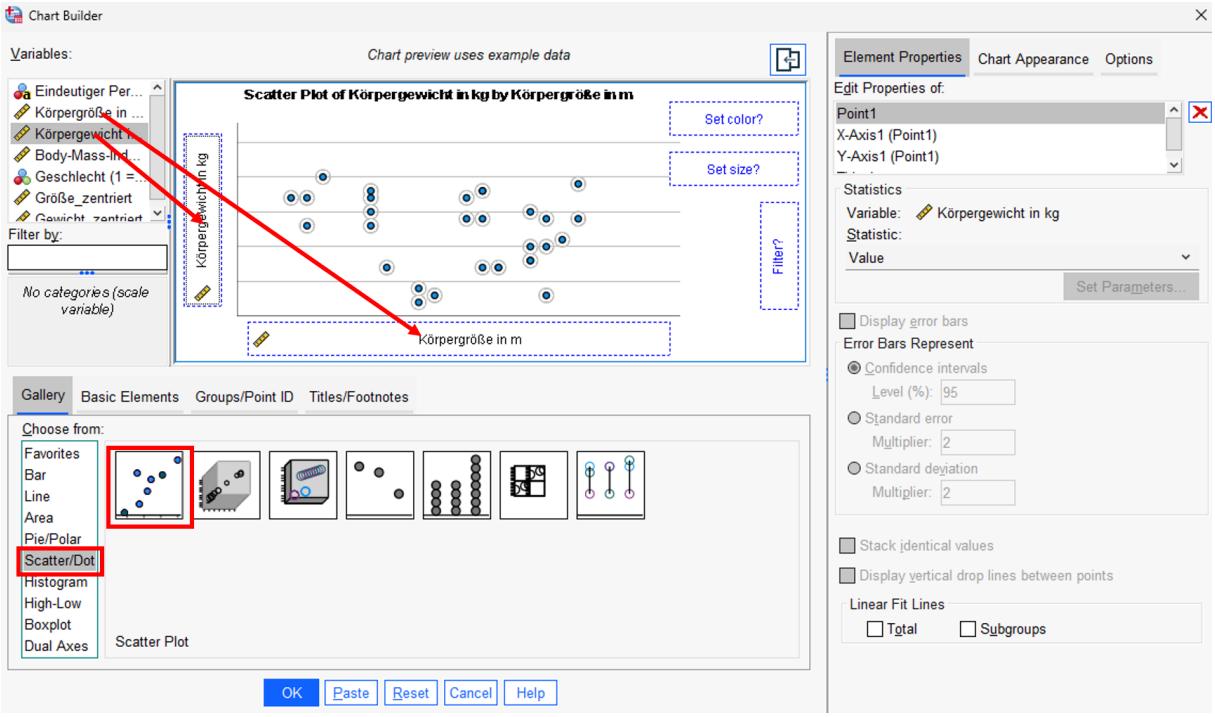


Abbildung 3.23. Anforderung eines Streudiagramms, in dem die Körpergröße auf der Abszisse (x-Achse) und das Körpergewicht auf der Ordinate (y-Achse) abgetragen wird.

Wird zum ersten Mal eine Abbildung in SPSS erstellt, kann es sein, dass die Erstellung etwas länger dauert. Aber mit ein wenig Geduld sollten wir schließlich in der Ausgabe die Grafik erhalten, die in Abbildung 3.24 dargestellt ist. Wir sehen, dass für unsere Stichprobe in der Tat ein sehr enger

Zusammenhang zwischen Körpergröße und -gewicht besteht: je größer eine Person, desto schwerer dürfte die Person in der Regel auch sein. Ausnahmen von dieser Regel sind im gegebenen Datensatz in der Tat Mangelware.

Eine grafische Inspektion mittels eines Streudiagramms ist insbesondere bei Vermutung eines Zusammenhangs zwischen zwei metrischen Variablen eigentlich immer zu empfehlen (Anscombe, 1973; siehe auch Übungsaufgabe 3.15). In diesem Fall überzeugt auch die graphische Darstellung bereits sehr deutlich vom tatsächlichen Bestehen eines solchen Zusammenhangs. Aber wie lässt sich dieser Zusammenhang nun auch in einer Maßzahl abbilden, d.h., quantifizieren?

Um dies zu erläutern, denken wir kurz darüber nach, was wir eigentlich meinen, wenn wir sagen, dass zwischen zwei metrischen Variablen ein Zusammenhang besteht. Wir sagen, dass zwischen zwei metrischen Variablen ein Zusammenhang besteht, üblicherweise dann, wenn eher große Werte der einen Variablen mit eher großen (oder eher kleinen) Werten der anderen Variablen einhergehen und umgekehrt. Im Falle des in Abbildung 3.24 dargestellten Beispiels sagen wir genau deshalb, es scheint eindeutig ein Zusammenhang zwischen Größe und Gewicht zu bestehen, weil Personen mit einem eher größeren Gewicht auch eher größere Personen sind und umgekehrt eher leichtere Personen auch eher kleinere Personen sind. Wir würden von einem Zusammenhang auch dann sprechen, wenn der Zusammenhang gerade umgekehrt wäre, also eher kleiner Ausprägungen einer Variablen mit eher größeren Ausprägungen der anderen Variablen assoziiert wären. Solche Zusammenhänge werden manchmal auch als „negative“ Zusammenhänge bezeichnet, aber ebenfalls als Zusammenhänge. Von keinem Zusammenhang würden wir nur dann sprechen, wenn die Ausprägungen einer Variablen in keiner erkennbaren Form mit den Ausprägungen der anderen Variablen assoziiert wären. In „keiner erkennbaren Form“ impliziert auch, dass Zusammenhänge nicht unbedingt monoton oder gar linear sein müssen. Es kann z.B. sein, dass kleine und große Werte einer Variablen jeweils mit großen Werten der anderen Variablen assoziiert sind, und mittlere Werte hingegen mit kleinen Werten. In diesem Fall könnte u.U. ein sog. quadratischer Zusammenhang zwischen den Variablen vorliegen.

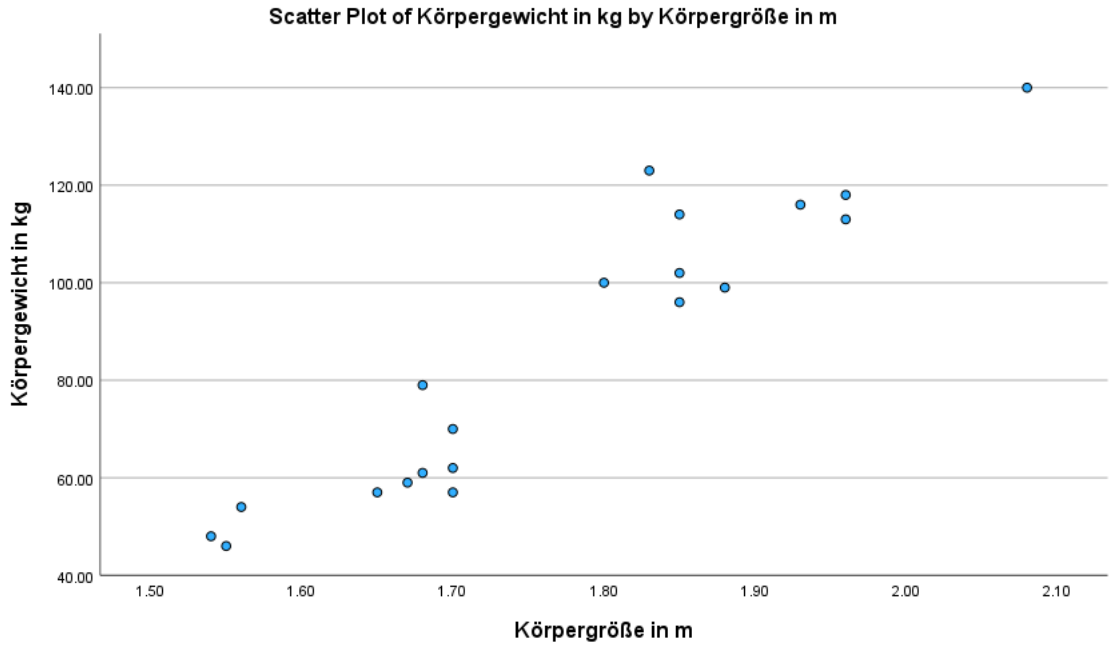


Abbildung 3.24. Streudiagramm für Körpergröße und -gewicht für den Datensatz „Kap3daten2.sav“.

Im Folgenden werden wir uns lediglich mit der Quantifizierung von monotonen und insbesondere linearen Zusammenhängen befassen. Monotone Zusammenhänge bezeichnen Zusammenhänge der Form „je, desto“: je größer eine Variable, desto größer auch die andere (ein sog. positiver Zusammenhang); oder je kleiner eine Variable, desto größer die andere (ein sog. negativer Zusammenhang). Mit einem linearen Zusammenhang ist ein Zusammenhang gemeint, bei dem im Mittel für jedes Paar von Werten der beiden betrachteten Variablen ein Unterschied zwischen dem Wert der einen Variablen und typischen Werten für diese Variable auf einer für diese Variable typischen Skala mit entsprechenden Unterschieden zwischen dem Wert der anderen Variablen und typischen Werten für die andere Variablen auf einer für die andere Variable typischen Skala einhergeht. Da das äußerst kompliziert klingt, ist es vermutlich eine gute Idee, diese Erläuterung noch einmal schrittweise am Beispiel der Körpergrößen und -gewichte zu erläutern.

Abbildung 3.24 zeigt uns, dass Körpergrößen für die untersuchte Stichprobe zwischen ca. 1.50 m und 2.10 m variieren. Körpergewichte variieren hingegen auf einer ganz anderen Skala: erstens variieren sie nicht in Metern sondern in Kilogramm, und auch zahlenmäßig in einem völlig anderen Intervall: in etwa von 40 kg bis 140 kg. Auch bei typischen Körpergrößen und -gewichten handelt es sich um völlig unterschiedliche Größen. Wählen wir als Maß für eine typische Körpergröße den

Mittelwert, so erhalten wir M = 1.77 m (mit einer Standardabweichung von SD = 0.15 m), für das typische Gewicht, ebenfalls in Form des Mittelwerts, erhalten wir M = 85.70 kg (mit einer Standardabweichung von SD = 29.40 kg). Die Standardabweichungen zeigen wiederum an, dass die beiden Größen typischerweise über völlig andere Bereiche variieren. Allerdings sprechen wir von einem linearen Zusammenhang zwischen den beiden Variablen genau dann, wenn eine Variation der einen Variablen auf der für sie typischen Skala mit einer proportionalen Variation der anderen Variablen auf der für die andere Variable typischen Skala einhergeht, und das über den gesamten Bereich beider Variablen hinweg. Abbildung 3.24 zeigt, dass Personen, die in etwa 10 cm größer sind als andere Personen, in etwa 20 kg schwerer sind als andere Personen. D.h. ein Größenunterschied von 1 cm geht in etwa mit einem Gewichtsunterschied von 2 kg einher und das in guter Näherung über die ganze Bandbreite an Größen von 1.50 m bis 2.10 m. Genau das ist gemeint, wenn von einem linearen Zusammenhang zwischen zwei Variablen die Rede ist: ändert sich eine Variable um einen bestimmten Wert, so verändert sich die andere um einen dazu proportionalen Wert, unabhängig davon, wo die Variablen in ihrer Bandbreite an möglichen Werten liegen. Dass dieser Zusammenhang nicht für jedes Wertepaar in Abbildung 3.24 exakt gilt, sondern lediglich approximativ und „im Großen und Ganzen“, hat damit zu tun, dass die Größe für das Gewicht zwar sicherlich ein bestimmender Faktor ist, aber nicht der einzige, sondern es noch andere Faktoren gibt, die dafür eine Rolle spielen, über die wir allerdings keine Informationen haben. Diese zusätzlichen, unbekannten Faktoren können den Zusammenhang in beide Richtungen beeinflussen (d.h., zu etwas geringerem oder größeren Gewicht bei gleicher Größe führen) und sorgen daher für eine Schwankung um den im Mittel sehr deutlichen linearen Zusammenhang.

Zur Quantifizierung dieses linearen Zusammenhangs müssen wir nun nur noch das oben Gesagte in uns bereits bekannten statistischen Größen zum Ausdruck bringen. Zuerst möchten wir wissen, ob eher große Abweichungen von der typischen Körpergröße auch mit eher großen Abweichungen vom typischen Gewicht einhergehen. Dazu können wir uns in einem ersten Schritt einfach einmal alle Abweichungen für beide Variablen von deren typischen Ausprägungen berechnen. Dies können wir tun, indem wir von jeder einzelnen Variablenausprägung den Mittelwert der Variablen abziehen. Dies wird auch als Zentrierung bezeichnet und könnte in SPSS z.B. unter *Transform* >>

Compute Variable… durchgeführt werden. Für diesen Datensatz wurde diese Zentrierung bereits vorgenommen, siehe die beiden Variablen *Größe_zentriert* und *Gewicht_zentriert* (siehe auch die beiden Spalten ganz rechts in Abbildung 3.22). An den beiden in Abbildung 3.22 dargestellten Spalten mit den Ausprägungen für diese beiden Variablen erkennen wir sehr gut den Zusammenhang zwischen diesen beiden Abweichungen: weicht die Größe sehr weit (auf der Skala für die Größe) nach oben hin vom typischen Wert ab (z.B. der Wert 0.31 in Zeile 4), so weicht auch das Gewicht sehr weit nach oben (auf der Skala für das Gewicht) vom typischen Gewicht ab (der Wert 54.30 in Zeile 4). Weicht um gekehrt der Wert für die Größe sehr weit nach unten von der typischen Größe ab (z.B. der Wert -0.23 in Zeile 12), so weicht auch das Gewicht sehr weit nach unten ab (der Wert -37.70 in derselben Zeile). Dies gilt „im Großen und Ganzen“ für alle Werte und jeweils proportional zur Größe der Abweichung vom jeweils typischen Wert. Um diesen Zusammenhang in eine einzelne Zahl zu fassen, könnte man nun die Produkte der einzelnen Zahlenpaare bilden, aufsummieren, und schließlich durch die Anzahl der Zahlenpaare dividieren. Dies wäre in der Tat eine Quantifikation des mittleren Zusammenhangs zwischen den beiden Variablen. Wird die Körpergröße der i-ten Person mit x_i und das Gewicht mit y_i bezeichnet, so würde die Formel für die soeben beschriebene Größe lauten:

$$cov_{emp}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Diese Größe wird als empirische Kovarianz zwischen den Variablen x und y bezeichnet. Der Zusatz „empirisch“ bezieht sich darauf, dass es sich dabei um die Kovarianz zwischen den konkreten Werten der beiden Variablen in der Stichprobe handelt (und nicht um die Schätzung der Kovarianz in der Population, aus der diese Stichprobe gezogen wurde). In der Tat handelt es sich bei der Kovarianz um eine Größe, die den linearen Zusammenhang zwischen zwei Variablen erfasst: je größer der lineare Zusammenhang, desto größer die Kovarianz. Das Vorzeichen der Kovarianz erfasst auch die Richtung des Zusammenhangs: gehen Abweichungen von typischen Werten nach oben in x typischerweise mit Abweichungen von typischen Werten nach oben in y einher, so sind die einzelnen Summanden in der Formel für die Kovarianz vorwiegend positiv und die Kovarianz insgesamt typischerweise eher positiv; gehen umgekehrt Abweichungen von typischen Werten nach oben in x mit Abweichungen von

typischen Werten nach unten in y einher, so sind die einzelnen Summanden in der Formel für die Kovarianz vorwiegend negativ und die Kovarianz insgesamt typischerweise eher negativ.

Allerdings hat die Kovarianz für die Quantifizierung des linearen Zusammenhangs den gravierenden Nachteil, dass sie von den Einheiten abhängt, mit denen die beiden Variablen erfasst wurden. Wird die Körpergröße in unserem Beispiel etwa in cm statt in m angegeben, so sieht man an der Formel oben, dass sich die Kovarianz schlagartig um den Faktor 100 verändern würde. Dies wäre aber nicht der Fall, weil der Zusammenhang zwischen Größe und Gewicht etwa größer geworden wäre; im Gegenteil, der Zusammenhang ist nach wie vor derselbe, nur die Maßeinheit (für eine der beiden Variablen) hat sich verändert. D.h., für ein brauchbares Maß des Zusammenhangs zwischen zwei Variablen verlangen wir zudem, dass es unabhängig von den Einheiten ist, mit welchen diese Variablen erfasst werden.

Diesen Aspekt haben wir allerdings oben beim Versuch zu beschreiben, was wir mit einem linearen Zusammenhang überhaupt meinen, bereits mit der Änderung einer Variablen auf ihrer jeweiligen Skala bereits charakterisiert. D.h., wir betrachten nicht nur die Abweichungen der Variablen von ihren jeweiligen typischen Werten an sich, sondern diese Abweichungen auf der für die Variable typischen Skala. Mit letzterer ist der Bereich an Werten gemeint, in dem Ausprägungen der jeweiligen Variablen typischerweise liegen. Eine Größe um diesen Bereich zu quantifizieren haben wir mit der Standardabweichung auch bereits kennengelernt. D.h., wenn wir die oben berechneten Abweichungen jeweils durch die Standardabweichung der jeweiligen Variablen dividieren, dann die Produkte bilden, und schließlich deren Mittelwert berechnen, bekommen wir eine Maßzahl, die den Zusammenhang zwischen beiden Variablen einheitenunabhängig quantifiziert. Der mathematische Ausdruck für diese Maßzahl lautet

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{s_{x,emp}}\frac{(y_i - \bar{y})}{s_{y,emp}} = \frac{1}{n}\sum_{i=1}^{n} z_{x,i} z_{y,i}.$$

Diese Maßzahl wird als Pearsons Produkt-Moment-Korrelationskoeffizient oder einfach als Pearson Korrelationskoeffizient (manchmal auch nur als Pearson Korrelation) bezeichnet. Mit $s_{x,emp}$ und $s_{y,emp}$ werden dabei die empirischen Standardabweichungen bezeichnet, mit $z_{x,i}$ und $z_{y,i}$ die z-

transformierten Variablen x und y. D.h., der Pearson Korrelationskoeffizient für die Variablen x und y entspricht der Kovarianz der beiden z-transformierten Variablen. In der Tat werden durch die z-Transformation (vgl. den entsprechenden Abschnitt oben) beide Operationen durchgeführt, auf die es uns für die Quantifizierung des linearen Zusammenhangs ankommt: zuerst die Zentrierung am Mittelwert und anschließend die Standardisierung (oder Normierung) an der Skala der Variablen (d.h. am für die Variable typischen Variationsbereich). Durch die Standardisierung wird schließlich auch der Wertebereich des Korrelationskoeffizienten auf das Intervall von -1 bis 1 beschränkt. Besteht zwischen zwei Variablen ein exakter positiver linearer Zusammenhang ist der Koeffizient gleich 1, ist der Zusammenhang exakt negativ linear ist der Koeffizient -1. Besteht überhaupt kein linearer Zusammenhang ist der Koeffizient gleich 0. Alle anderen Fälle liegen zwischen diesen Werten.

Im Vergleich zu allem bisherigen ist die Berechnung des Pearson Korrelationskoeffizienten mit SPSS äußerst einfach. Dazu schieben wir im sich öffnenden Menü nach Auswahl von *Analyze* >> *Correlate* >> *Bivariate…* einfach alle Variablen, zwischen denen wir den Pearson Korrelationskoeffizienten berechnen möchten, in das Feld „Variables“, siehe Abbildung 3.25. Wie wir in der Abbildung sehen, ist der Pearson Korrelationskoeffizient bereits vorab ausgewählt. In diesem Menü könnten wir auch einen der beiden anderen Korrelationskoeffizienten berechnen lassen, die im Folgenden noch kurz erläutert werden. Anschließend klicken wir wieder auf „Paste“ und führen die neuen Kommandozeilen in der Syntaxdatei aus. Die daraufhin erzeugte Ausgabe ist in Abbildung 3.26 dargestellt.

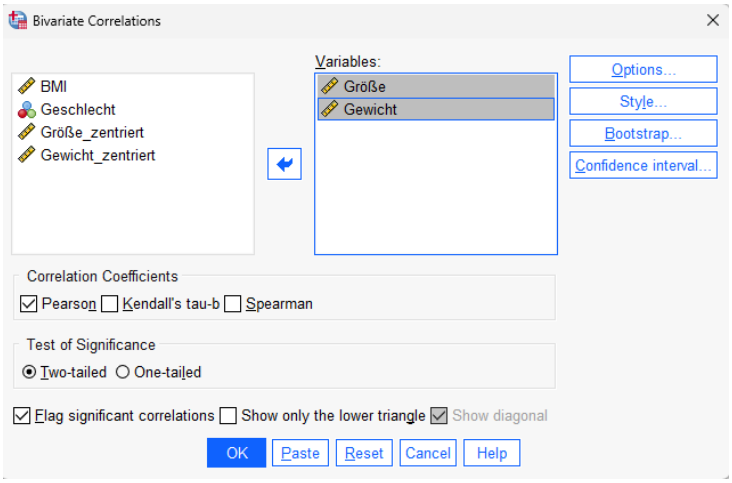


Abbildung 3.25. Berechnung des Pearson Korrelationskoeffizienten mit SPSS.

Correlations

		Körpergröße in m	Körpergewicht in kg
Körpergröße in m	Pearson Correlation	1	.943**
	Sig. (2-tailed)		<.001
	N	20	20
Körpergewicht in kg	Pearson Correlation	.943**	1
	Sig. (2-tailed)	<.001	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Abbildung 3.26. Ausgabe für den Pearson-Korrelationskoeffizienten zwischen der Körpergröße und dem Körpergewicht.

Wir sehen, dass zwischen Körpergröße und -gewicht in der gegebenen Stichprobe in der Tat ein sehr starker linearer Zusammenhang besteht, der Pearson Korrelationskoeffizient ist fast maximal, r = .94. In der Ausgabe sehen wir zusätzlich noch das Ergebnis eines Signifikanztests für die Nullhypothese, dass der Korrelationskoeffizient in der Population gleich Null ist. Mit Signifikanztests werden wir uns ab dem nächsten Kapitel im Detail befassen. Hier sei nur erwähnt, dass es sich beim hier durchgeführten Test um einen t-Test mit $n - 2$ Freiheitsgraden handelt, wobei n = 20 hier den Stichprobenumfang bezeichnet. Bestünde die Forschungsfrage darin, ob sich der Pearson Korrelationskoeffizient für die beiden Variablen in der Population von Null unterscheidet, könnte das Ergebnis des Signifikanztests in diesem Fall (unter der Annahme eines Signifikanzniveaus von .005) wie folgt berichtet werden: „Der Pearson Korrelationskoeffizient für den linearen Zusammenhang zwischen Körpergröße und -gewicht unterscheidet sich (mit α = .005) signifikant von Null, $r(18)$ = .94, p < .001. Gemäß Cohen (1988) handelt es sich um einen großen Effekt.“ Es ist üblich beim Bericht des Pearson Korrelationskoeffizienten die führende Null (d.h., die „0“ vor dem Dezimaltrennzeichen) wegzulassen, da sein Wertebereich zwischen -1 und 1 liegt. Zudem wurde im Ergebnisbericht auf die Heuristiken für Effektstärken von Cohen (1988) Bezug genommen. Effektstärken werden wir in den folgenden Kapiteln noch im Detail besprechen. Für den Moment genügt es festzuhalten, dass es sich auch beim Pearson Korrelationskoeffizienten um eine sog. Effektstärke handelt. Gemäß Cohens Heuristiken (1988) werden Korrelationskoeffizienten ab 0.1 als kleine, ab 0.3 als mittlere, und ab 0.5 als große Effekte bezeichnet. Dies gilt auch für einige andere Korrelationskoeffizienten, von welchen wir unten noch Spearmans Rangkorrelationskoeffizienten und Kendalls tau erläutern.

Bezüglich Pearsons Korrelationskoeffizienten ist es wichtig, sich noch einmal einige wesentliche Einschränkungen vor Augen zu führen. Der Koeffizient ist, wie oben im Detail erläutert, eine Maßzahl für den linearen Zusammenhang zwischen zwei Variablen. Nichtlineare Zusammenhänge (z.B. quadratische Zusammenhänge) können damit nicht beschrieben werden. Der Koeffizient ist zudem sehr empfindlich für sog. Ausreißer, d.h. Datenpunkte, die im Streudiagramm weit abseits der Punktewolke aller anderen Datenpunkte liegen. Diese Einschränkung kann für andere Korrelationskoeffizienten (wie z.B. die beiden unten beschriebenen) deutlich geringer ausfallen, allerdings ist dabei zu beachten, dass diese anderen Koeffizienten nicht dasselbe erfassen wie Pearsons Korrelationskoeffizient (d.h. nicht unbedingt den linearen Zusammenhang zwischen zwei Variablen, siehe die Erläuterung im Zusammenhang mit Kendalls tau unten).

Für alle Korrelationskoeffizienten gilt, dass sie keine Aussagen über einzelne Personen erlauben. Sie beschreiben lediglich im Mittel Zusammenhänge zwischen zwei Variablen; d.h. „im Großen und Ganzen“ und nicht „im Einzelnen und Partikulären“. Einzelne Fälle (vgl. auch wiederum mit der Thematik Ausreißer) können mitunter weit von mittleren Zusammenhängen abweichen.

Ebenfalls gilt für alle Korrelationskoeffizienten, dass sich aus den von ihnen beschriebenen Zusammenhängen keine kausalen Aussagen ableiten lassen. Handelt es sich nicht um einen rigoros kontrollierten, experimentellen Versuchsaufbau, gibt es grundsätzlich meist viele Erklärungen für das Zustandekommen eines Zusammenhangs zwischen zwei Variablen (Bühner et al., 2025). So kann es etwa in der Tat sein, dass Änderungen in der Variable x Änderungen in der Variablen y verursachen. Die beiden Variablen hängen aber auch miteinander zusammen (= kovariieren), wenn umgekehrt Änderungen in y Änderungen in x verursachen. Es kann auch beides gleichzeitig (zu unterschiedlichen Anteilen) der Fall sein, d.h., Variablen können sich gegenseitig ursächlich beeinflussen (z.B. depressive Stimmung und Schlafmangel, siehe Bühner et al., 2025). Zudem kann eine unbekannte Drittvariable sowohl x als auch y verursachen, was wiederum in einem Zusammenhang zwischen den beiden Variablen resultiert. Schließlich kann auch eine Reihe unbekannter Variablen ursächlich mit den beiden Variablen in Verbindung stehen und einem Zusammenhang zwischen x und y zugrunde liegen (Bühner et al., 2025).

Abschließend seien nun noch zwei weitere Korrelationskoeffizienten zur Beschreibung von Zusammenhängen zwischen zwei Variablen kurz erläutert. Beide Korrelationskoeffizienten erfassen lediglich monotone Zusammenhänge (nicht unbedingt aber lineare Zusammenhänge); nicht-monotone (z.B. quadratische) Zusammenhänge werden durch sie nicht erfasst.

Bei Spearmans Rangkorrelationskoeffizienten handelt es sich um einen Korrelationskoeffizienten, der auch dann verwendet werden kann, wenn beide Variablen oder eine von beiden Variablen lediglich auf Ordinalskalenniveau vorliegen. Betrachten wir beispielsweise eine Umfrage zum Thema Soziale Medien, in der Personen unterschiedlichen Alters nach der Häufigkeit des Konsums sozialer Medien befragt werden. Das Alter wird dabei mittels einer metrischen Variablen erhoben, die Häufigkeit des Konsums sozialer Medien allerdings nur mit einer Skala mit den Abstufungen „1 = so gut wie nie“, „2 = einmal pro Monat“, „3 = einmal pro Woche“, „4 = täglich“, „5 = öfters täglich, aber weniger als eine Stunde“, „6 = mehrere Stunden täglich“. Bei letzterer Skala handelt es sich offensichtlich nicht um eine Intervallskala und dementsprechend auch nicht um eine metrische Variable. Allerdings kann es dennoch interessant sein, ob zwischen dem Alter der befragten Personen und der ordinalen Variable Häufigkeit (des Konsums sozialer Medien) ein monotoner Zusammenhang besteht, d.h., ob Personen desto häufiger/seltener soziale Medien konsumieren, je jünger/älter sie sind. Um diese Frage zu erhellen, könnte in diesem Fall Spearmans Rangkorrelationskoeffizient verwendet werden.

Zur Berechnung von Spearmans Rangkorrelationskoeffizienten werden zuerst die Werte beider Variablen jeweils für jede der beiden Variablen in Ränge umgerechnet. Für eine intervallskalierte Variable (etwa das Alter) funktioniert die Umrechnung in Ränge wie folgt. Angenommen die Stichprobe umfasse lediglich die folgenden sieben Personen mit einem jeweiligen Alter von 18, 21, 19, 36, 25, 67, und 53 Jahren. Diese Altersangaben werden dann in Ränge umgerechnet, indem dem geringsten Alter der Wert 1, dem zweitgeringsten Alter der Wert 2 usw. vergeben wird. D.h., die zu den oben angegebenen Alterswerten gehörigen Ränge wären 1, 3, 2, 5, 4, 7, 6. Liegen sog. Rangbindungen vor, d.h. haben etwa im Beispiel mehrere Personen dasselbe Alter, z.B. 18, 21, 18, 36, 25, 25, 25, wird ihnen der durchschnittliche Rang der Positionen zugewiesen, die sie eingenommen hätten. D.h., in letzterem Fall wären die resultierenden Rangwerte für die gegebenen Alterswerte 1.5, 3, 1.5, 7, 5, 5, 5. Im Fall der

ordinalskalierten Variablen liegen zwar bereits so etwas wie Rangwerte vor, allerdings weisen in diesem Fall viele Fälle (Personen) dieselbe Variablenausprägung, d.h. Rangbindungen, auf.

Wurden beide Variablen in Rangwerte umgerechnet, kann schlichtweg der Pearson Korrelationskoeffizient für die in Rangwerte umgerechneten Variablen berechnet werden. Der resultierende Wert entspricht dann dem Spearman Rangkorrelationskoeffizienten, der üblicherweise mit dem Symbol r_S bezeichnet wird. Spearmans Rangkorrelationskoeffizient erfasst die Monotonie von Zusammenhängen insofern, dass er abbildet, dass positive Änderungen der einen Variablen mit positiven Änderungen der anderen Variablen einhergehen oder positive Änderungen der einen Variablen negativen Änderungen der anderen Variablen. Wie Pearsons Korrelationskoeffizient hat auch Spearmans Rangkorrelationskoeffizient im ersten Fall ein positives, im zweiten Fall ein negatives Vorzeichen. Genauso wie Pearsons Korrelationskoeffizient ist Spearmans Rangkorrelationskoeffizient auf den Bereich -1 bis 1 beschränkt. Im Gegensatz zu Pearsons Korrelationskoeffizient ist Spearmans Rangkorrelationskoeffizient allerdings weniger empfindlich auf Ausreißer. Allerdings können einige wenige ungewöhnliche Datenpunkte Spearmans Rangkorrelationskoeffizienten immer noch stark beeinflussen (siehe insbesondere Übungsaufgabe 3.16).

Noch etwas robuster gegenüber Ausreißern und gleichzeitig ein präziseres Maß für die Monotonie von Zusammenhängen zwischen zwei Variablen ist Kendalls tau (Wilcox, 2017). Für Kendalls tau, üblicherweise auch bezeichnet mit dem gleichnamigen griechischen Buchstaben τ, werden aus allen Datenpunkten alle möglichen Paare an Datenpunkten gebildet. Ein Paar von Datenpunkten, bestehend aus den Datenpunkten (x_i, y_i) und (x_j, y_j), wird genau dann als konkordant bezeichnet, wenn für $x_i < x_j$ auch gilt, dass $y_i < y_j$, bzw. für $x_i > x_j$ auch gilt, dass $y_i > y_j$ (d.h., wenn sich x nach oben/unten ändert, ändert sich y in die gleiche Richtung). Hingegen wird das Datenpunktepaar als diskordant bezeichnet, wenn für $x_i < x_j$ gilt, dass $y_i > y_j$, bzw. für $x_i > x_j$ gilt, dass $y_i < y_j$ (d.h., wenn sich x nach oben/unten ändert, ändert sich y in die jeweils andere Richtung). Anschließend wird zwischen konkordanten und diskordanten Paaren die Differenz gebildet und durch die Anzahl aller Datenpunktepaare dividiert. Sind alle Paare konkordant, so ist Kendalls tau gleich 1, sind alle Paare diskordant, so ist Kendalls tau gleich -1. Gibt es gleich viele diskordante wie konkordante Paare, dann

ist Kendalls tau gleich Null. Alle anderen Fälle liegen zwischen diesen Werten. Damit ist Kendalls tau ein sehr anschauliches Maß für die Monotonie des Zusammenhangs zwischen zwei Variablen.

Sind Rangbindungen vorhanden, sind manche Datenpunktepaare entsprechend der obigen Definitionen weder konkordant noch diskordant. Diesem Umstand muss dann bei der Berechnung von Kendalls tau Rechnung getragen werden (Kendall, 1945). Der Korrelationskoeffizient für diesen Fall wird als Kendalls tau-b mit dem Symbol τ_b bezeichnet.

In SPSS können sowohl Spearmans Rangkorrelationskoeffizient als auch Kendalls tau-b in demselben Menü angefordert werden, das wir schon zur Berechnung von Pearsons Korrelationskoeffizienten verwendet haben, siehe Abbildung 3.25. Wählen wir dort für unsere Beispielfragestellung zu Körpergröße und -gewicht für die im Datensatz „Kap3daten2.sav“ gegebene Stichprobe beide Korrelationskoeffizienten aus, fügen die entsprechenden Kommandozeilen in die Syntax ein und führen diese aus, so erhalten wir die in Abbildung 3.27 dargestellte Ausgabe. Wir sehen, dass auch diese beiden Korrelationskoeffizienten klar auf einen Zusammenhang zwischen den beiden Variablen hinweisen. Am Unterschied untereinander sowie zu den Werten für Pearsons Korrelationskoeffizienten (Abbildung 3.26) erkennen wir aber auch, dass die beiden Koeffizienten nicht dieselbe Art von Zusammenhang erfassen.

Correlations

			Körpergröße in m	Körpergewicht in kg
Kendall's tau_b	Körpergröße in m	Correlation Coefficient	1.000	.771**
		Sig. (2-tailed)	.	<.001
		N	20	20
	Körpergewicht in kg	Correlation Coefficient	.771**	1.000
		Sig. (2-tailed)	<.001	.
		N	20	20
Spearman's rho	Körpergröße in m	Correlation Coefficient	1.000	.901**
		Sig. (2-tailed)	.	<.001
		N	20	20
	Körpergewicht in kg	Correlation Coefficient	.901**	1.000
		Sig. (2-tailed)	<.001	.
		N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Abbildung 3.27. Ausgabe für Spearmans Rangkorrelationskoeffizienten sowie Kendalls tau-b.

Für weitere Korrelationskoeffizienten, die in manchen Datensituationen bzw. für manche Fragestellungen durchaus angemessener sein können (z.B. Zusammenhang zwischen einer künstlich

dichotomen und einer metrischen Variablen oder Zusammenhänge für typische Datenpunkte), wird auf entsprechende Fachliteratur verwiesen (siehe z.B. Bühner & Ziegler, 2017; Wilcox, 2017, 2022).

Bericht deskriptiver Statistiken

Für kategoriale Variablen werden im Rahmen der Charakterisierung von Stichproben meist schlicht Häufigkeiten berichtet (eventuell auch mittels Balkendiagrammen dargestellt). Bei großen Stichproben verschaffen oft relative Häufigkeiten einen besseren Überblick über die Verteilung mehrerer Kategorien über die Stichprobe hinweg. Bei kleinen Stichproben kann manchmal die Angabe absoluter Häufigkeiten ein treffenderes Bild der Verhältnisse geben (z.B. 3 von 10 befragten Personen anstelle von 30% der befragten Personen). Unter Umständen kann auch der Modalwert bei kategorialen Variablen von Interesse sein (z.B. hinsichtlich der Frage „Was war das beliebteste der drei Schulfächer?“ in unserer Stichprobe). Hier ist auch wichtig zu bedenken, dass es bei wenigen Kategorien auch sein kann, dass es mehrere Modalwerte gibt.

Zur Charakterisierung metrischer Variablen sollte zumindest eine Maßzahl der Lage (z.B. Mittelwert) und eine Maßzahl der Streuung (typischerweise Standardabweichung) angegeben werden. Wenn empirische Verteilungen ausgeprägte Schiefe oder Wölbung aufweisen ist es zusätzlich ratsam, diese etwa mittels geeigneter Maßzahlen zu charakterisieren. Oft sagt aber auch ein Bild mehr als tausend Worte (oder Zahlen). Bei außergewöhnlichen Verteilungen (z.B. mit zwei ausgeprägten Maxima, d.h. im Falle einer sog. bimodalen Verteilung) hilft häufig eine geeignete grafische Darstellung wie ein Histogramm (weshalb ist in diesem Fall ein Boxplot keine geeignete Wahl?) deutlich mehr als eine Liste von Maßzahlen (geeignete Maßzahlen können aber dennoch den visuellen Eindruck einer grafischen Darstellung unterstützen bzw. ein Verstehen des Dargestellten erleichtern/ermöglichen). Histogramme können wie Balkendiagramme über *Analyze >> Descriptive Statistics >> Frequencies…* und dort unter „Charts…“ angefordert werden.

Zur Charakterisierung von Zusammenhängen zwischen Variablen können Tabellen mit Korrelationskoeffizienten erstellt werden. Dabei ist allerdings zu betonen, dass für die Beurteilung des Zusammenhangs zwischen zwei Variablen häufig eine Abbildung deutlich erhellender sein kann als eine einzelne Maßzahl und auf die Inspektion entsprechender paarweiser Abbildungen deshalb keinesfalls

verzichtet werden sollte. Ein dahingehend illustratives Beispiel ist unten in Übungsaufgabe 3.15 gegeben und beruht auf einer Publikation zur grundsätzlichen Bedeutung graphischer Darstellungen für statistische Analysen überhaupt (Anscombe, 1973).

Format berichteter Maßzahlen

Für Ergebnisberichte und im Allgemeinen für den Bericht statistischer Kennwerte oder Maßzahlen haben sich in der Psychologie verschiedene Standards etabliert. Ein weit verbreiteter Standard ist derjenige der American Psychological Association (APA), der im Rahmen dieser Übungen gleich mitbehandelt werden soll.

Die wesentlichen Merkmale dieses Standards, die für uns im Rahmen dieser Übungen eine Rolle spielen werden, sind in Tabelle 3.1 illustriert. Es macht nichts, wenn Sie mit vielen der dort aufgeführten Symbole und Abkürzungen noch nichts anfangen können. Das wird sich im Lauf der verbleibenden Kapitel noch ganz von selbst (bzw. durch Üben, Üben und nochmal Üben) ändern. Im Wesentlichen ist zu beachten, dass statistische Kenngrößen kursiv zu setzen sind, Zahlen aber nicht, und, dass mit Ausnahme von p-Werten grundsätzlich auf zwei Nachkommastellen zu runden ist. Für p-Werte sind es drei Nachkommastellen und da sich dann ein Wert größer 0.999 oder kleiner 0.001 nicht mehr exakt angeben lässt, wird stattdessen $p > .999$ bzw. $p < .001$ geschrieben. Das ist eindeutig, da p-Werte bekanntlich nach unten durch 0 und nach oben durch 1 begrenzt sind. Bei Zahlen, bei denen das so ist (gilt z.B. auch für Korrelationskoeffizienten) wird zudem die führende 0 weggelassen, d.h. statt $p = 0.321$ wird $p = .321$ geschrieben. Auch Tabellen selbst folgen einem bestimmten Format nach dem APA-Standard, der mit der Form von Tabelle 3.1 illustriert wird.

Tabelle 3.1

Kennwerte nach APA-Richtlinien (American Psychological Association, 2019)

Kennwert	Zeichen	Darstellung	Beispiel
Mittelwert	M	*M* = x.xx	*M* = 13.68
Median	Mdn	*Mdn* = x.xx	*Mdn* = 14.57
Modus	Mo	*Mo* = x.xx	*Mo* = 13.5
Standardabweichung	SD	*SD* = x.xx	*SD* = 2.48
Standardfehler	SE	*SE* = x.xx	*SE* = 1.50
Freiheitsgrade	df	x (Ausnahme: x.xx)	27 (24.45)
t-Wert	t	*t*(df) = x.xx	*t*(38) = 2.89
Cohens‘ d	d	*d* = x.xx	*d* = 0.68
F-Wert	F	*F*(df1,df2) = x.xx	*F*(1, 121) = 37.46
Partielles Eta-Quadrat	η_p^2	η_p^2 = .xx	η_p^2 = .06
p-Wert	p	*p* = .xxx	*p* > .999 *p* = .567 *p* = .032 *p* < .001
Pearson Korrelationen	r	*r*(df) = .xx	*r*(120) = .35
Spearman Korrelationen	r_s	r_s(df) = .xx	r_s(120) = .79
Kendalls tau-b	τ_b	τ_b(df) = .xx	τ_b(120) = -.71
Partialkorrelationen	r_{part}	r_{part}(df/N-3) = .xx	r_{part}(119) = -.17
Chi-quadrat	Chi^2; χ^2	χ^2(df) = x.xx	χ^2(1) = 6.42
Z-Werte	z	*z* = x.xx	*z* = 1.54

Anmerkung. Dies ist eine Tabelle, die nach APA-Richtlinien (American Psychological Association, 2019) formatiert ist! Gemäß APA sollen neben der Durchnummerierung der Tabellen und der Vergabe eines Titels in kursiver Schrift auch nur horizontale und niemals vertikale Linien verwendet werden.

Übungsaufgaben

Für die Beispiele 3.1-10 kann durchwegs mit der Datendatei „Kap3daten_bearbeitet.sav“ gearbeitet werden die Sie wiederum in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden können, das Sie unter https://osf.io/9tcx3/ herunterladen können. Vergessen Sie nicht alle Arbeitsschritte in entsprechenden Syntaxdateien zu dokumentieren. Speichern Sie Syntax- und Ausgabedateien (sofern vorhanden) regelmäßig ab.

Beispiel 3.1

Stellen Sie sich vor, Sie möchten aus den Items *mathe_mathe1*, *mathe_mathe2*, und *mathe_mathe3* eine Skala generieren, die das Merkmal „Affinität zu Mathematik und Statistik“ operationalisieren soll. Das heißt, bei dieser Skala handelt es sich um eine Variable, die dieses Merkmal quantifizieren soll: Personen mit niedriger Affinität zu Mathematik und Statistik sollen eine niedrige Zahl bei dieser Variablen haben, Leute mit einer hohen Affinität eine hohe Zahl. Schauen Sie sich die Labels der Items in der Datendatei genau an. Welche der Items werden Sie für eine Bildung so einer Skala umkodieren bzw. umpolen müssen? Erzeugen Sie dann für diese Items neue Variablen, die den umkodierten Items entsprechen.

Beispiel 3.2

Berechnen Sie eine Summenskala, die das Merkmal „Affinität zu Mathematik und Statistik“ numerisch abbilden soll. Verwenden Sie dazu die Items *mathe_mathe1*, *mathe_mathe2*, *mathe_mathe3* bzw. entsprechend umkodierte Items.

Beispiel 3.3

Lassen Sie sich angemessene deskriptive Statistiken und grafische Darstellungen für die in Beispiel 3.2 erstellte Skala ausgeben.

Beispiel 3.4

Polen/Kodieren Sie das Item *statistikschmerzen* um.

Beispiel 3.5

Berechnen Sie eine Mittelwertskala, die (wie schon in Beispiel 3.2) das Merkmal „Affinität zu Mathematik und Statistik“ numerisch abbilden soll. Verwenden Sie dazu die Items *statistikliebe*, *mathematikliebe* und *statistikschmerzen* bzw. entsprechend umkodierte Items.

Beispiel 3.6

Lassen Sie sich angemessene deskriptive Statistiken und grafische Darstellungen für die in Beispiel 3.5 erstellte Skala ausgeben.

Beispiel 3.7

Wie sind die Vorlieben für die Hauptfächer Deutsch, Englisch und Mathematik unter den befragten Personen verteilt? Lassen Sie sich dazu eine entsprechende Häufigkeitstabelle sowie eine Balkengrafik ausgeben.

Beispiel 3.8

Angenommen, Sie glauben, dass Personen, die in einer Beziehung sind, auch verliebt sind. Lassen Sie sich eine Kreuztabelle für die Variablen *verliebt* und *beziehungsstatus* ausgeben, um zu überprüfen, ob diese Vermutung rein deskriptiv für die erhobene Stichprobe erfüllt wird.

Beispiel 3.9

Lassen Sie sich für Alter, Körpergröße und Schuhgröße angemessene deskriptive Statistiken ausgeben. Wie groß sind Mittelwerte und Standardabweichungen für die drei Variablen? Sehen Sie ein mögliches Problem für die Interpretation der mittleren Körper- und Schuhgröße für diese Stichprobe?

Beispiel 3.10

Lassen Sie sich eine Kreuztabelle für die Schulabschlussnote in Mathematik und dem Hogwarts-Haus, dem sich die Befragten am ehesten zugehörig fühlen, ausgeben.

Beispiel 3.11

In Österreich, Deutschland und der Schweiz wurde eine (fiktive) Befragung zum Thema Vereinbarkeit von Familie und Beruf durchgeführt. Unglücklicherweise sind die Daten der drei Länder alle jeweils separat in einer SPSS Datei gespeichert. Fügen Sie die drei Datensätze „Österreich.sav“, „Deutschland.sav“ und „Schweiz.sav“, die Sie wiederum in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) unter https://osf.io/9tcx3/ herunterladen können, zu einem großen Datensatz zusammen und speichern Sie diesen unter dem neuen Namen „dach.sav“ ab. Geben Sie an, von wie vielen Personen Daten im Gesamtdatensatz vorliegen. Geben Sie ferner an, wie viele Personen jeweils in den drei Ländern befragt wurden.

Als ich im Jahr 2023 zum ersten Mal die Lehrveranstaltung „Anwendung statistischer Verfahren am Computer“ an der Universität Graz abhalten durfte, konnte ich dankenswerterweise auf die Lernmaterialien einiger meiner Vorgänger:innen zurückgreifen. Darunter befanden sich diese drei Datensätze, die ich seitdem zu Lehr- und Lernzwecken (u.a. für dieses und die folgenden Beispiele) in einigen Hinsichten verändert und adaptiert habe. Leider enthalten die Datensätze keinerlei Hinweis darauf, wer sie ursprünglich erstellt hat. Sollte jemals jemand dieser Information habhaft werden, wäre ich äußerst dankbar, falls sie mit mir geteilt werden könnte, da ich dann jener Person oder jenen Personen die zustehende Würdigung und den verdienten Dank hier endlich nachtragen könnte.

Beispiel 3.12

Arbeiten Sie in diesem Beispiel mit dem Gesamtdatensatz, den Sie in Beispiel 3.11 erstellt haben. Falls Sie sich bei Ihrer Lösung für Beispiel 3.11 unsicher sind, können Sie auch den Datensatz „dach.sav“ verwenden, den Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden können, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Um sich einen ersten Überblick über die Zusammensetzung der Gesamtstichprobe zu verschaffen, lassen Sie sich sinnvolle deskriptive Statistiken für die Variablen *geschlecht*, *alter* und *bildung* ausgeben. Wählen Sie dafür die Ihrer Meinung nach geeigneten statistischen Kennwerte. Charakterisieren Sie in Worten die Stichprobe hinsichtlich dieser Variablen.

Sehen Sie sich schließlich noch die Variable *m_dur* an. Worum handelt es sich bei dieser Variablen? Lassen Sie sich geeignete statistische Kennwerte ausgeben. Was fällt Ihnen auf?

Beispiel 3.13

Arbeiten Sie in diesem Beispiel mit dem Gesamtdatensatz, den Sie in Beispiel 3.11 erstellt haben. Falls Sie sich bei Ihrer Lösung für Beispiel 3.11 unsicher sind, können Sie auch den bereits verfügbaren Datensatz „dach.sav“ verwenden (siehe Beispiel 3.12).

Geben Sie an, wie viele Männer und Frauen pro Nation befragt worden sind.

Beispiel 3.14

Arbeiten Sie in diesem Beispiel mit dem Gesamtdatensatz, den Sie in Beispiel 3.11 erstellt haben. Falls Sie sich bei Ihrer Lösung für Beispiel 3.11 unsicher sind, können Sie auch den bereits verfügbaren Datensatz „dach.sav“ verwenden (siehe Beispiel 3.12).

Bei den Befragten wurde unter anderem erhoben, wie gerecht sie in ihrer Beziehung die Aufteilung verschiedener Aspekte der Hausarbeit einschätzen: die Aufteilung der Haushaltsarbeit (Kochen, Putzen, Wäsche waschen etc.) und die Aufteilung der Kinderbetreuung. Um die allgemeine Einschätzung der Gerechtigkeit in einer Beziehung abschätzen zu können, soll eine Skala aus den beiden entsprechenden Items gebildet werden. Bilden Sie hierfür sowohl den Mittelwert (nennen Sie die resultierende Skala *justice_mean*) als auch die Summe (nennen Sie die resultierende Skala *justice_sum*) aus den beiden Variablen und geben Sie für beide jeweils Mittelwert und Standardabweichung an.

Beispiel 3.15

Im Datensatz „anscombe.sav“ sind vier Paare von Variablen gegeben, die jeweils mit x_i und y_i mit $i = 1, \dots, 4$ bezeichnet sind. Erzeugen Sie für jedes Variablenpaar (x_i, y_i) ein Streudiagramm, in dem Sie x_i auf der x-Achse und y_i auf der y-Achse auftragen. Berechnen Sie zudem für jedes Variablenpaar den Pearson Korrelationskoeffizienten. Was fällt Ihnen auf? Diskutieren Sie Ihr Ergebnis. Für welches der vier Variablenpaare erscheint es Ihnen sinnvoll, den Zusammenhang zwischen den beiden Variablen mittels Pearsons Korrelationskoeffizienten zu charakterisieren?

Beispiel 3.16

Im Datensatz „outliers.sav“ sind drei Variablenpaare gegeben. Alle drei Variablenpaare beziehen sich auf dieselben Datenpunkte für das Variablenpaar (x, y). Für das Variablenpaar (xwo, ywo) für lediglich zwei Datenpunkte von den Datenpunkten für das Variablenpaar (x, y) entfernt. Für das Variablenpaar $(xwo2, ywo2)$ wurden schließlich noch zwei weitere Datenpunkte entfernt. Erstellen Sie drei Streudiagramme, um sich veranschaulichen welche Datenpunkte jeweils entfernt wurden. Ermitteln Sie dann für jedes der drei Variablenpaare sowohl den Pearson Korrelationskoeffizienten als auch Spearmans Rangkorrelationskoeffizienten und Kendalls tau-b. Vergleichen Sie die Ergebnisse. Wie wirkt sich das Entfernen einzelner Punkte jeweils auf die Koeffizienten aus?

Beispiel 3.17

In der Datei „sterne.sav“ sind die Logarithmen der Oberflächentemperatur und der Leuchtkraft von 47 Sternen gegeben. Zwischen dem Logarithmus der Oberflächentemperatur und dem Logarithmus der Leuchtkraft eines Sterns im Hauptreihenstadium besteht laut Theorie näherungsweise ein linearer Zusammenhang: mit steigender Oberflächentemperatur nimmt die Leuchtkraft zu. Für die folgenden Berechnungen können Sie von einer bivariaten Normalverteilung für die beiden metrischen Variablen ausgehen.

(a) Ermitteln Sie dien Pearson-Korrelationskoeffizienten zwischen den beiden Variablen und erstellen Sie einen entsprechenden Ergebnisbericht. Wie würden Sie das Resultat in Hinsicht auf die theoretische Vorhersage interpretieren?

(b) Bei der Inspektion eines Streudiagramms für die 47 Sterne stellt ein Astrophysiker fest, dass das Diagramm vier Sterne enthält, die sehr hohe Leuchtkraft (> 5.5) bei sehr geringer Oberflächentemperatur (< 3.6) aufweisen. Da es sich bei diesen Sternen vermutlich nicht um Hauptreihensterne, sondern um sogenannte Rote Riesen handelt, empfiehlt der Astrophysiker die Berechnung der Korrelation unter Ausschluss dieser vier Sterne zu wiederholen. Zu welchem Ergebnis kommen Sie in diesem Fall und was schließen Sie daraus für den theoretisch postulierten Zusammenhang zwischen den Logarithmen von Oberflächentemperatur und Leuchtkraft?

Kapitel 4

Parameterschätzung und Testen statistischer Hypothesen über Populationsmittelwerte

Stefan E. Huber

Bis hierher haben wir uns vorrangig mit der grundlegenden Bedienung der Software SPSS (Kapitel 2) und der Beschreibung gegebener Datensätze bzw. Stichproben (Kapitel 3) befasst. In diesem Kapitel werden wir uns erstmals sogenannter inferenzstatistischer Fragestellungen annehmen. Das heißt, wir wollen auf der Basis einer (begrenzten) Stichprobe Aussagen über die Population treffen, aus der die Stichprobe gezogen wurde.

Z.B. möchten wir aufgrund unserer Stichprobe abschätzen wie hoch die mittlere Ausprägung einer Variablen in der Population ist. Könnten wir eine Messung dieser Variablen an jedem Fall der Population vornehmen, dann könnten wir schlichtweg den Mittelwert berechnen und hätten unsere Antwort. In der Realität ist es aber meistens nicht möglich eine gesamte Population zu vermessen. Man stelle sich beispielsweise vor, bei der Population handele es sich um alle erstsemestrigen Studierenden und bei dem interessierenden Merkmal um das Interesse am jeweiligen Studium. Letzteres soll mittels eines Fragebogens erfasst werden, d.h. die Skala, die aus den Items des Fragebogens generiert wird, soll das Merkmal „Interesse am Studium“ in einer Zahl abbilden. Bei der Ausprägung auf dieser Skala handelt es sich also um die Variable, mit der das interessierende Merkmal erfasst werden soll. Uns interessiert nun, wie hoch das Interesse am jeweils eigenen Studium unter Studienanfänger:innen (Erstsemestrigen) im Mittel ist, d.h. wir interessieren uns für den Populationsmittelwert. Diesen können wir nun aus zwei Gründen nicht durch Messung aller Erstsemestrigen erfassen. Der erste Grund ist ein rein praktischer: selbst wenn wir uns auf ein einziges Erhebungsjahr beschränken würden, gibt es (weltweit) sehr viele Erstsemestrige, was die Erhebung praktisch unmöglich macht (prinzipiell, d.h. denkbar, ist sie natürlich möglich). Der zweite Grund ist allerdings ein prinzipieller: unsere Fragestellung war ja nicht zeitlich beschränkt. Das heißt, wir wollen das mittlere Studieninteresse nicht nur bei Erstsemestrigen eines bestimmten Jahrgangs wissen, sondern bei Erstsemestrigen überhaupt. Diese zeitliche Unbeschränktheit der Fragestellung macht die Erhebung der gesamten Population

prinzipiell unmöglich, da wir ja in einem bestimmten Zeitraum keine Erhebungen an vergangenen und zukünftigen Erstsemestrigen durchführen können.

Allerdings erlaubt uns die Erhebung des Studieninteresses zu einem bestimmten Zeitpunkt an einer hinreichend großen Anzahl von Erstsemestrigen immer noch Aussagen über die mögliche Ausprägung des Populationsmittelwerts. Diese Aussagen sind dann allerdings aufgrund der Endlichkeit der Stichprobe schon rein statistisch mit Unsicherheiten behaftet. Dazu kommen natürlich noch andere Einschränkungen wie Kontexteffekte, kulturelle Unterschiede etc., die die Generalisierbarkeit der Aussagen über bloße statistische Unsicherheiten hinaus einschränken. Diese Limitationen klammern wir aber der Einfachheit halber für den weiteren Verlauf dieser Übungen zu statistischen Anwendungen erst einmal aus. In diesem Kapitel befassen wir uns also nur mit jenen rein statistischen Auswirkungen auf Aussagen, die wir aufgrund einer endlichen Stichprobe über den Populationsmittelwert treffen können.

Dafür werden wir uns zwei Fälle genauer ansehen: (1) die Schätzung des Populationsmittelwerts, (2) die Testung von Hypothesen über den Populationsmittelwert, jeweils auf der Basis einer endlichen Stichprobe. Für beide Fälle werden wir immer davon ausgehen, dass es sich bei unserer Stichprobe um eine einfache Zufallsstichprobe handelt. Das heißt, wir ziehen Personen oder im Allgemeinen Merkmalsträger:innen (wir werden noch sehen, dass es sich dabei nicht unbedingt um Personen handeln muss) zufällig aus der Population. Das bedeutet jede:r Merkmalsträger:in hat dieselbe Wahrscheinlichkeit aus der Population gezogen zu werden wie jede:r andere Merkmalsträger:in. Zudem sind die Ziehungen einzelner Merkmalsträger:innen unabhängig voneinander, d.h. die Ziehung einer beliebigen Person aus einer Population hängt nicht davon ab, welche anderen Personen bereits gezogen wurden oder noch gezogen werden sollen.

Im Folgenden werden wir uns ansehen, wie wir SPSS verwenden können, um die beiden inferenzstatistischen Fragestellungen (1) und (2) zu beantworten. Der Einfachheit halber werden wir dafür den Beispieldatensatz „Kap4daten.sav“ verwenden, den Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden können, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Der Datensatz enthält IQ-Werte von 240 fiktiven Studienanfänger:innen im Studiengang Psychologie. In einer zweiten Variablen (*IQsub*) sind IQ-Werte einer kleineren Menge an Studierenden aus der Gesamtmenge gegeben, mithilfe derer wir uns ein Bild davon machen können, wie sich der Stichprobenumfang auf unsere Ergebnisse auswirken kann. Die Fragestellungen, die wir mittels dieses Datensatzes mit SPSS beantworten möchten, lauten:

(1) Wie hoch ist der mittlere IQ von Studienanfänger:innen im Studiengang Psychologie?

(2) Unterscheidet sich der mittlere IQ von Studienanfänger:innen im Studiengang Psychologie vom mittleren IQ von 100 der Population aller jungen Erwachsenen?

Für beide Fragestellungen nehmen wir an, dass unsere fiktive Stichprobe repräsentativ für die Population von Studienanfänger:innen im Studiengang Psychologie ist.

In SPSS können beide Fragestellungen im Rahmen der Durchführung eines Einstichproben-t-Tests beantwortet werden. Prinzipiell handelt es sich aber beim Einstichproben-t-Test um einen statistischen Test, der verwendet wird, um den Unterschied eines Populationsmittelwerts von einem vorgegebenen Wert zu testen, d.h. hier insbesondere um Fragestellung (2) zu beantworten. Auch wenn es sehr angenehm ist, beide Antworten in SPSS gleich auf einmal zu bekommen, ist es trotzdem wichtig zu verstehen, dass es sich bei den beiden Fragestellungen prinzipiell um konzeptuell verschiedene Fragestellungen handelt. Aus diesem didaktischen Grund werden sie im Folgenden auch getrennt voneinander behandelt.

Zusätzlich zur Beantwortung der oben genannten Fragestellungen werden wir uns in diesem Kapitel auch noch ansehen, wie wir die Effektstärke für einen Einstichproben-t-Test mit SPSS ermitteln können und wie Ergebnisberichte für die durchgeführten Analysen gemäß APA-Format zu berichten sind. Schließlich werden wir uns noch ansehen wie eine Stichprobenplanung für einen Einstichproben-t-Test mit der frei verfügbaren Software G*Power durchgeführt werden kann.

Punkt- und Intervallschätzung eines Populationsmittelwerts

Aus der Theorie wissen wir (Bühner et al., 2025), dass es sich beim Stichprobenmittelwert um einen *erwartungstreuen*, *effizienten* und *konsistenten* Schätzer für den Populationsmittelwert handelt. Streng genommen gilt dies zwar nur, wenn es sich bei der Variablen, deren Mittelwert geschätzt werden soll, um eine identisch und unabhängig *normalverteilte* Zufallsvariable handelt, allerdings kann aufgrund des zentralen Grenzwerttheorems davon ausgegangen werden, dass sich für hinreichend große Stichproben, die Stichprobenkennwerteverteilung auch hinreichend gut durch eine Normalverteilung approximieren lässt und diese drei Eigenschaften der Schätzfunktion des Populationsmittelwerts in guter Näherung auch für anders verteilte Zufallsvariablen erhalten bleiben. In der Psychologie hat sich dafür die Konvention eingebürgert, dass Stichproben zumindest einen Umfang von 30 aufweisen sollten (ob dies im Einzelfall auch genügt, um pauschal von einer hinreichend guten Näherung auszugehen, ist zu bezweifeln, siehe z.B. Wilcox, 2022; für diese Übungen werden wir allerdings aus rein pragmatischen Gründen erst einmal davon ausgehen).

Das heißt, für eine Punktschätzung des Populationsmittelwerts wären wir mit der Ermittlung des Stichprobenmittelwerts bereits fertig. Wie im letzten Kapitel besprochen, könnten wir etwa über *Analyze >> Descriptive Statistics >> Frequencies…* unter „Statistics“ den Mittelwert (Engl.: Mean) anfordern und würden für unsere Stichprobe den Wert von 106.75 erhalten. In der Tat wäre das unsere Punktschätzung für den Populationsmittelwert auf Basis unserer Stichprobe.

Allerdings wurde oben bereits erwähnt, dass diese Schätzung aufgrund der Endlichkeit der Stichprobe mit einer Unsicherheit verbunden ist. Aus der Theorie wissen wir (Bühner et al., 2025), dass sich auch diese statistische Unsicherheit quantifizieren lässt. Eine Möglichkeit ist die Berechnung des Standardfehlers $SE = SD/\sqrt{n}$ mit SD dem Schätzwert der Populationsstandardabweichung (wie ihn uns SPSS praktischer gleich ausgibt) und n dem Stichprobenumfang. Der Standardfehler kann ebenfalls gleich unter *Analyze >> Descriptive Statistics >> Frequencies…* und dort unter „Statistics“ angefordert werden und muss nicht selbst berechnet werden. In unserem Fall ergibt sich für den Standardfehler des Mittelwerts $SE = 0.74$. Lassen wir uns zusätzlich die Standardabweichung ausgeben, lässt sich leicht überprüfen, dass der Ausdruck $\frac{SD}{\sqrt{n}} = \frac{11.42}{\sqrt{240}} = 0.74$ tatsächlich dem Wert des Standardfehlers entspricht.

Je geringer der Standardfehler, desto präziser unsere Punktschätzung, d.h., desto zuversichtlicher sind wir, uns mit unserer Punktschätzung auch in der Nähe des tatsächlichen Populationsmittelwerts zu befinden. Dies wird auch bei der sogenannten Intervallschätzung noch einmal deutlich, für die der Standardfehler auch eine wesentliche Rolle spielt.

Aus der Theorie wissen wir schließlich (Bühner et al., 2025), dass für die normalverteilte Schätzfunktion $\bar{X}$ des Mittelwerts die folgende Teststatistik

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

einer zentralen t-Verteilung mit $\nu = n - 1$ Freiheitsgraden folgt (Student, 1908). Hier bezeichnet S^2 die Schätzfunktion der Populationsvarianz, n ist wiederum der Stichprobenumfang, μ bezeichnet den (unbekannten) Populationsmittelwert.

Für eine gegebene Stichprobengröße können die Quantile $t_{\alpha/2}$ und $t_{1-\alpha/2}$ einer t-Verteilung berechnet werden (z.B. mit dem Online-Tool unter www.stattrek.com), so dass

$$P\left(t_{\frac{\alpha}{2}} \leq T \leq t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

D.h. es können jene Grenzen des Bereichs unterhalb der t-Verteilung berechnet werden, außerhalb derer sich jeweils genau der $\alpha/2$-Anteil der Fläche unterhalb der t-Verteilung befinden. Bekanntlich ist die Wahrscheinlichkeit, mit der sich eine Zufallsvariable in einem bestimmten Bereich unterhalb ihrer Wahrscheinlichkeitsdichteverteilung realisiert, proportional zur Fläche dieses Bereichs. Die Wahrscheinlichkeit, dass sich die Größe T im Bereich $t_{\frac{\alpha}{2}} \leq T \leq t_{1-\frac{\alpha}{2}}$ realisiert, ist also genau durch $1 - \alpha$ gegeben.

Einsetzen des Ausdrucks für T oben führt auf

$$P\left(t_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Die linke der beiden Ungleichungen im Argument der Wahrscheinlichkeitsfunktion lässt sich wie folgt nach μ auflösen:

$$t_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \quad \Rightarrow \quad t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}} \leq \bar{X}-\mu \quad \Rightarrow \quad \mu \leq \bar{X} - t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}} \quad \Rightarrow \quad \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}},$$

wobei hier im letzten Schritt die Symmetrie der zentralen t-Verteilung $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$ verwendet wurde.

Analog lässt sich die zweite der beiden Ungleichungen wie folgt nach μ auflösen:

$$t_{1-\frac{\alpha}{2}} \geq \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \quad \Rightarrow \quad t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}} \geq \bar{X}-\mu \quad \Rightarrow \quad \mu \geq \bar{X} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}.$$

Damit ergibt sich schließlich die aus der Theorie bekannte Gleichung (Bühner et al., 2025)

$$P\left(\bar{X} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}\right) = 1-\alpha.$$

Kann also $\bar{X}$ als normalverteilte Zufallsvariable approximiert werden, so realisiert sich diese Zufallsvariable im $(1-\alpha)$-Anteil aller möglichen Realisationen in einem Bereich, so dass der unbekannte Populationsmittelwert von den Grenzen dieses Bereiches eingeschlossen wird, also innerhalb des folgenden Intervalls liegt:

$$\left[\bar{X} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}\right] = [U, O]$$

mit U und O der unteren bzw. oberen Grenze des (zufälligen) Intervalls. Hier ist wichtig, noch einmal in Ruhe über die exakte Bedeutung dieses Intervalls zu reflektieren: Im Anteil (1-α), dem sog. *Konfidenzniveau*, aller möglichen Realisationen der Zufallszahl $\bar{X}$ befindet sich μ tatsächlich irgendwo zwischen den beiden Grenzen dieses Intervalls. Lediglich im α-Anteil aller möglichen Realisationen von $\bar{X}$ realisiert sich $\bar{X}$ so weit entfernt vom Populationsmittelwert, dass das auf diese Weise gebildete Intervall den Populationsmittelwert nicht enthält. Wird daher ein kleiner Wert für α gewählt, kann man sehr zuversichtlich sein (das „sehr“ lässt sich hier exakt spezifizieren: man kann mit $(1-\alpha) * 100\%$ zuversichtlich sein), dass das konkrete $(1-\alpha)$-Konfidenzintervall (häufig abgekürzt als KI), das sich

durch Einsetzen der jeweiligen Schätzwerte für die jeweiligen Schätzfunktionen, d.h. Einsetzen des Stichprobenmittelwerts $\bar{x}$ für $\bar{X}$ und des Standardfehlers $SE = \sqrt{s^2/n}$ für $\sqrt{S^2/n}$, ergibt

$$\left[\bar{x} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}, \bar{x} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n}}\right] = \left[\bar{x} - t_{1-\frac{\alpha}{2}} \cdot SE, \bar{x} + t_{1-\frac{\alpha}{2}} \cdot SE\right] = [u, o]$$

den Populationsmittelwert μ auch tatsächlich enthält.

Für unseren konkreten Datensatz kann dieses Konfidenzintervall über *Analyze >> Compare Mean and Proportions >> One-Sample T Test…* angefordert werden. Unter Options kann der Anteil $1 - \alpha$ (d.h., das Konfidenzniveau) in Prozent festgelegt werden, siehe Abbildung 4.1. Wählen wir z.B. $\alpha = 0.05$, geben wir dort 95% ein, was ohnehin der Voreinstellung entspricht. Wir erhalten dann ein sog. 95%-Konfidenzintervall (typischerweise abgekürzt zu 95%-KI).

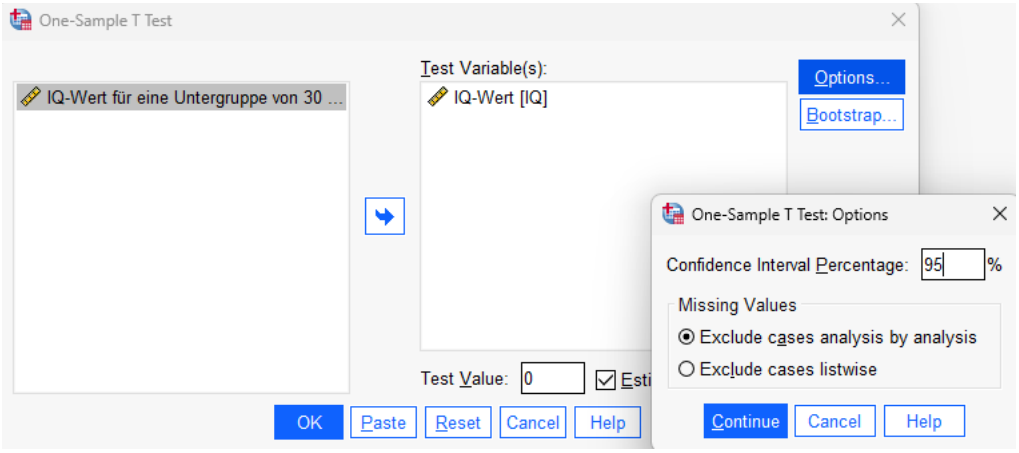


Abbildung 4.1. Anfordern eines 95%-Konfidenzintervalls für den Mittelwert auf Basis unserer Stichprobe.

Einfügen der entsprechenden Kommandos durch Klicken auf „Paste“ und Ausführen der Kommandozeilen in der Syntax (natürlich erst, nachdem wir die Syntax hinreichend dokumentiert haben) erzeugt die in Abbildung 4.2. gezeigte Ausgabe. In der Tabelle unter der Überschrift „One-Sample Statistics“ finden wir nochmals den Stichprobenumfang, die Punktschätzung des Populationsmittelwerts (= der Stichprobenmittelwert), die Schätzung der Populationsstandardabweichung sowie den sich aus letzterer und dem Stichprobenumfang ergebenden Standardfehler des

Mittelwerts. Eine separate Berechnung dieser Größen, wie oben rein zur Illustration durchgeführt, ist also nicht notwendig, wir bekommen all diese Informationen und noch mehr ohnehin auf diese Weise.

Das Konfidenzintervall für den Mittelwert können wir in der Tabelle „One-Sample Test“ ganz rechts unter der Überschrift „95% Confidence Interval of the Difference“ ablesen. Unter der Bezeichnung „Lower“ finden wir die untere Grenze, unter der Bezeichnung „Upper“ die obere Grenze.

Das Ergebnis unserer Intervallschätzung für den Populationsmittelwert könnten wir auf Basis dieser Ergebnisse wie folgt berichten: „Auf Basis unserer Stichprobe sind die Werte in dem 95%-Konfidenzintervall [105.29, 108.20] plausible Werte für den Mittelwert des IQ in der Population.“

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
IQ-Wert	240	106.75	11.417	.737

One-Sample Test

Test Value = 0

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
IQ-Wert	144.849	239	<.001	<.001	106.746	105.29	108.20

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
IQ-Wert	Cohen's d	11.417	9.350	8.502	10.196
	Hedges' correction	11.453	9.321	8.475	10.164

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

Abbildung 4.2. Ausgabe für einen Einstichproben-t-Test, die hier erstmal nur zur Ermittlung des 95%-Konfidenzintervalls für den Mittelwert herangezogen wird.

Hypothesentest für eine ungerichtete statistische Hypothese über den Populationsmittelwert

Die zweite Fragestellung, die wir mit Hilfe unseres Datensatzes beantworten wollten, lautete: Unterscheidet sich der mittlere IQ von Studienanfänger:innen im Studiengang Psychologie vom mittleren IQ von 100 der Population aller jungen Erwachsenen?

Da wir hier lediglich nach einem Unterschied fragen, handelt es sich hierbei (wenn auch als Frage formuliert) um eine ungerichtete Hypothese. Wir hätten ja auch vermuten können, dass der IQ von Studienanfänger:innen im Studiengang Psychologie größer oder kleiner als jener der Allgemeinpopulation junger Erwachsener ist. Dann hätte es sich um gerichtete Hypothesen gehandelt. Wie wir diese mit SPSS testen können, werden wir uns auch gleich im Anschluss an die Beantwortung der gerichteten Fragestellung ansehen.

Zurück zur vorliegenden, ungerichteten Hypothese, die sich wie folgt spezifizieren lässt:

$$H_0: \mu = \mu_0 = 100, H_1: \mu \neq \mu_0 = 100,$$

d.h. die Nullhypothese wäre „der Populationsmittelwert ist gleich demjenigen der Allgemeinpopulation von 100“ und die Alternativhypothese wäre „die beiden Populationsmittelwerte unterscheiden sich“. Hier haben wir bereits angenommen, dass die Stichprobenkennwerteverteilung des Mittelwerts durch eine Normalverteilung approximiert werden kann und daher die Schätzfunktion des Mittelwerts die nötigen Gütekriterien erfüllt, um den Populationsmittelwert $\bar{x}_{pop}$ zu schätzen.

Wie oben bereits erläutert, wissen wir, dass die Teststatistik

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

einer zentralen t-Verteilung mit $\nu = n - 1$ Freiheitsgraden folgt (Student, 1908). Hier bezeichnet S^2 die Schätzfunktion der Populationsvarianz, n ist wiederum der Stichprobenumfang, μ bezeichnet den (unbekannten) Populationsmittelwert.

Unter der Gültigkeit der Nullhypothese, d.h., wenn $\mu = \mu_0$, gilt dann entsprechend auch, dass

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}}$$

einer zentralen t-Verteilung mit $\nu = n - 1$ Freiheitsgraden folgt. Wenn dem so ist, und wir uns zudem vorstellen, sehr viele einfache Zufallsstichproben zu ziehen und jeweils die jeweilige Realisation dieser Teststatistik zu berechnen, so würde nur in $\alpha/2$ aller Fälle eine Teststatistik kleiner als das $\alpha/2$-Quantil der t-Verteilung resultieren. Genauso würde nur in $\alpha/2$ aller Fälle eine Teststatistik größer als das $(1 - \alpha/2)$-Quantil der t-Verteilung resultieren.

Wenn wir nun einen sehr kleinen Wert für α, das sog. *Signifikanzniveau*, wählen würden, könnten wir schlichtweg die Teststatistik für unsere konkrete Stichprobe berechnen, und würde das einen Wert kleiner als das $\alpha/2$-Quantil der t-Verteilung (mit $\nu = n - 1$ Freiheitsgraden) oder größer als das $(1 - \alpha/2)$-Quantil der t-Verteilung (mit $\nu = n - 1$ Freiheitsgraden) ergeben, könnten wir schlussfolgern, dass die Realisierung eines so extremen Wertes nur sehr selten der Fall wäre, wenn die Nullhypothese zuträfe. Daraus könnten wir dann den Umkehrschluss ziehen, dass die Annahme der Gültigkeit der Nullhypothese unplausibel erscheint. Auf dieser Grundlage könnten wir schließlich die Nullhypothese mit der sog. *Irrtumswahrscheinlichkeit* α (d.h., die Irrtumswahrscheinlichkeit entspricht gerade dem gewählten Signifikanzniveau) *verwerfen* oder *ablehnen*. Der Begriff Irrtumswahrscheinlichkeit bezieht sich hierbei auf die Tatsache, dass sich entsprechend extreme Werte für die Teststatistik ja tatsächlich selten, aber eben doch unter Gültigkeit der Nullhypothese ergeben würden. In diesen seltenen Fällen würden wir also die Nullhypothese mittels des oben beschriebenen Vorgehens ablehnen, obwohl sie zuträfe, d.h. wir würden uns in unserer Entscheidung irren. Das geht letztlich einfach darauf zurück, dass wir auf Basis einer endlichen Stichprobe keine sichere Entscheidung über Populationseigenschaften treffen können, es bleibt immer eine Unsicherheit.

Alternativ – aber, was die Testentscheidung anbelangt, völlig äquivalent zu dem eben erläuterten Vorgehen – können wir einen sogenannten p-Wert berechnen. Der p-Wert ist die maximale Wahrscheinlichkeit unter der Gültigkeit der Nullhypothese (und aller nötigen Annahmen für die t-Verteilung der Teststatistik, siehe oben bzw. auch die Erläuterungen unten zu Testannahmen) dafür, dass sich die Teststatistik in der beobachteten Realisation oder einer extremeren Realisation in Richtung der Alternativhypothese realisiert. Ist dieser p-Wert kleiner dem gewählten Signifikanzniveau α, dann liegt die Teststatistik auch im kritischen Bereich (d.h. hier: sie ist kleiner als das $\alpha/2$-Quantil der t-

Verteilung oder größer als das $(1 - \alpha/2)$-Quantil der t-Verteilung) und umgekehrt. Dies gilt auch für alle weiteren Hypothesentests, die wir im Rahmen dieser Übungen noch besprechen werden.

Wie können wir nun einen solchen Hypothesentest für unseren Datensatz mit SPSS durchführen? Dazu wählen wir wieder *Analyze >> Compare Mean and Proportions >> One-Sample T Test…* und im sich öffnenden Fenster geben wir nun unter „Test Value“ die Zahl 100 ein, siehe Abbildung 4.3. Das bedeutet, wir wollen einen ungerichteten Einstichproben t-Test durchführen, der die Gleichheit des auf Basis unserer Stichprobe geschätzten Populationsmittelwerts mit dem Wert 100 prüft. Alle anderen Einstellungen können wir belassen wie sie sind und dann auf „Paste“ klicken um wieder die entsprechenden Kommandozeilen in die Syntax einzufügen. Ausführen dieser Zeilen ergibt die Ausgabe, die wir in Abbildung 4.4 bewundern können. In der Tabelle mit der Überschrift „One-Sample Test“ finden wir unseren p-Wert unter „Two-Sided p“, der kleiner als 0.001 ausfällt, daher wird in der Ausgabe (ganz konform mit den APA-Richtlinien) lediglich „< .001“ ausgegeben. Möchten wir den Wert aber exakt wissen, dann können wir in der Ausgabe die Tabelle doppelt anklicken und dort nochmals auf den entsprechenden Wert doppelt klicken, um den exakten Wert einzusehen.

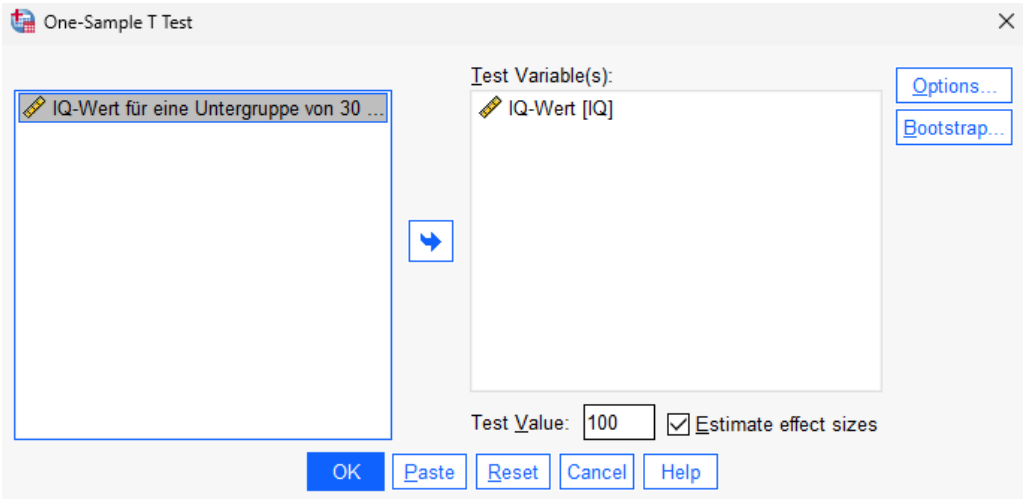


Abbildung 4.3. Anforderung eines ungerichteten Einstichproben t-Tests um die Gleichheit des Populationsmittelwerts mit dem Wert 100 zu prüfen.

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
IQ-Wert	240	106.75	11.417	.737

One-Sample Test

Test Value = 100

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
IQ-Wert	9.154	239	<.001	<.001	6.746	5.29	8.20

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
IQ-Wert	Cohen's d	11.417	.591	.453	.727
	Hedges' correction	11.453	.589	.452	.725

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

Abbildung 4.4. Ausgabe für den soeben angeforderten Einstichproben t-Test.

An der Ausgabe sehen wir auch, dass sich unser 95%-Konfidenzintervall geändert hat. Anstelle des Intervalls [105.29, 108.20] wird uns das Intervall [5.29, 8.20] angegeben. Das ist aber nur ein scheinbarer Unterschied, da sich dieses Intervall auf die Differenz zwischen dem geschätzten Populationsmittelwert und dem gewählten Vergleichswert für den Mittelwert bezieht. Dasselbe gilt für die Punktschätzung des Mittelwerts. In der Tabelle „One-Sample Test“ wird uns der Wert 6.75 für die mittlere Differenz angezeigt anstelle des Werts 106.75 in Abbildung 4.2. Dort war aber der Wert „gegen den wir getestet haben“ die Voreinstellung von null (vgl. Abbildung 4.1 mit Abbildung 4.3). Wir können also prinzipiell alles auf einmal, d.h. Punktschätzung, Intervallschätzung und einen Hypothesentest für einen bestimmten Vergleichswert in SPSS ausführen. Wir dürfen dabei dann nur nicht vergessen, den Wert, auf dessen Gleichheit wir testen, wieder zu den Grenzen für das Konfidenzintervall, das wir berechnen, bzw. für die Punktschätzung zur mittleren Differenz, die wir erhalten, zu addieren, wenn wir auch an Punkt- und Intervallschätzungen für den Populationsmittelwert interessiert sind. Bei der Punktschätzung besteht allerdings geringere Gefahr, da diese ohnehin auch zusätzlich in der Tabelle „One-Sample Statistics“ abzulesen ist.

Schließlich sehen wir in der Ausgabe auch einen Wert für „One-Sided p“, den wir heranziehen müssten, wenn wir ursprünglich eine gerichtete Hypothese über den Populationsmittelwert gehabt hätten. Diesen p-Wert könnten wir allerdings auch sehr leicht selbst aus dem p-Wert für ungerichtete Hypothesen (in der Tabelle unter „Two-Sided p“) ermitteln, da es sich dabei schlichtweg um die Hälfte des Werts für Letzteren handelt. In dieser Ausgabe kann das allerdings wiederum nur durch Doppelklick auf die Tabelle und anschließende Inspektion der exakten Werte sichtbar gemacht werden.

Effektstärke für den Einstichproben t-Test

In der Theorie haben wir Effektstärken als nützliche, einheitsunabhängige Maße kennengelernt, die als die Größe eines Unterschieds oder die Stärke eines Zusammenhangs interpretiert werden können (Bühner et al., 2025). Im Falle eines Einstichproben t-Tests können in SPSS zwei solcher Maße unter *Analyze >> Compare Mean and Proportions >> One-Sample T Test…* angefordert werden, indem „Estimate effect sizes“ ausgewählt wird (als Voreinstellung ist dies bereits grundsätzlich ausgewählt), siehe Abbildung 4.1 oder Abbildung 4.3.

Die entsprechende Ausgabe dieser beiden Effektstärken findet sich dann in der Tabelle „One-Sample Effect Sizes“ und bezieht sich auf die Größe des Unterschieds zwischen dem geschätzten Populationsmittelwert und dem Wert gegen den mittels des Einstichproben t-Tests getestet wurde. Die Effektstärken selbst sind in dieser Tabelle in der Spalte „Point Estimate“ zu finden, da es sich dabei wiederum um Punktschätzungen einer prinzipiell unbekannten Effektstärke in der Population handelt. Die Effektstärke Cohens d entspricht dem Verhältnis der Differenz zwischen Schätzwert für den Populationsmittelwert und dem Testwert und der geschätzten Standardabweichung SD, d.h.

$$d = \frac{\bar{x} - \mu_0}{SD}.$$

In unserem konkreten Fall lässt sich damit die Ausgabe in Abbildung 4.4 sehr leicht nachvollziehen:

$$d = \frac{\bar{x} - \mu_0}{SD} = \frac{106.75 - 100}{11.417} = \frac{6.75}{11.417} = 0.591.$$

Wie an der SPSS Ausgabe in Abbildung 4.4 abzulesen handelt es sich bei der zweiten Effektstärke um eine Effektstärke, bei der im Nenner für Cohens d zusätzlich ein Korrekturfaktor verwendet wird. Häufig sind die sich ergebenden Effektstärken aber, wie auch in unserem Beispiel, sehr ähnlich. Wir werden im Rahmen dieser Übungen grundsätzlich immer Cohens d verwenden. Auch für beide Effektstärken erhalten wir jeweils ein Konfidenzintervall.

Für Cohens d gibt es Heuristiken nach Cohen (1988), mit deren Hilfe eine einfache, schnelle Einschätzung der Größe der Effektstärke durchgeführt werden kann. Gemäß Cohen (1988) wird Cohens d im Bereich 0.2 bis 0.5 als klein, im Bereich 0.5 bis 0.8 als mittel, und für Werte größer als 0.8 als groß bezeichnet. Bei Werten unterhalb von 0.2 spricht man manchmal auch von vernachlässigbarer Effektstärke. Bei der Effektstärke ist schließlich meist nur der Betrag interessant, da sich die Abweichung des geschätzten Populationsmittelwerts vom Testwert ohnehin an der mittleren Differenz bzw. am Wert des geschätzten Populationsmittelwerts zeigt.

Ergebnisbericht

Die Ergebnisse, die bei der Schätzung des Populationsmittelwerts bzw. des Tests einer statistischen Hypothese über den Populationsmittelwert, erhalten werden, können grundsätzlich wie folgt berichtet bzw. interpretiert werden.

Im Falle des vorliegenden Beispiels würde ein Ergebnisbericht etwa wie folgt aussehen: „Im Mittel ist der IQ der getesteten Studienanfänger:innen um 6.75 IQ-Punkte höher als der Vergleichswert von 100 (n = 240, M = 106.75, 95%-KI [105.29, 108.20], SD = 11.42). Der mittlere IQ unterscheidet sich (mit α = .005) signifikant vom Vergleichswert, $t(239) = 9.15$, $p < .001$ (ungerichtet), Cohens d = 0.59, 95%-KI [0.45, 0.73]. Gemäß Cohens Heuristik (1988) handelt es sich um einen mittleren Effekt."

Sehen wir uns die einzelnen Bestandteile dieses Ergebnisberichts noch einmal im Detail an. Im ersten Satz werden schlichtweg deskriptive Statistiken sowie Ergebnisse der Punkt- und Intervallschätzung berichtet. Der zweite Satz bezieht sich dann auf den durchgeführten Hypothesentest. Es wird mitgeteilt, dass sich der mittlere IQ signifikant vom Vergleichswert unterscheidet, d.h., der p-Wert ist kleiner als das (vorab) gewählte Signifikanzniveau α. Zudem werden die ermittelte Teststatistik (der t-Wert), die Anzahl der Freiheitsgrade, der p-Wert und die Effektstärke inkl. Konfidenzintervall

berichtet. Die Punktschätzung der Effektstärke wird schließlich mittels der Heuristik nach Cohen (1988) interpretiert. Lateinische Buchstaben, die statistische Größen kennzeichnen, sind gemäß APA-Richtlinien kursiv gesetzt, Dezimalzahlen werden auf zwei Nachkommastellen gerundet, mit Ausnahme des p-Werts, der auf drei Nachkommastellen genau angegeben wird; und p-Werte kleiner 0.001 werden mit „< .001“, p-Werte größer als 0.999 mit „> .999“ gekennzeichnet (siehe auch die entsprechenden Erläuterungen in Kapitel 3).

Voraussetzungen für den Einstichproben t-Test und die Ermittlung von Konfidenzintervallen auf Grundlage der t-Verteilung

Für die Gültigkeit der eben beschriebenen Verfahren zur Intervallschätzung bzw. zur Testung der entsprechenden statistischen Hypothesen müssen einige Voraussetzungen gelten, die teilweise immer wieder erwähnt wurden und teilweise impliziert waren, die aber wichtig genug sind, um noch einmal explizit angeführt zu werden. Diese Voraussetzungen sind:

- Die Varianz der Population, aus der die Stichprobe gezogen wurde, ist nicht bekannt und muss mittels der Stichprobendaten geschätzt werden. Ist diese Varianz bekannt, kann stattdessen ein z-Test zur Hypothesentestung bzw. die Standardnormalverteilung zur Konstruktion von Konfidenzintervallen verwendet werden (siehe z.B. Bühner & Ziegler, 2017, S. 267-268).
- Die Messwerte sind mindestens intervallskaliert. Nur dann ist die Bildung eines Mittelwerts und sein numerischer Vergleich mit einem bestimmten Vergleichswert durch eine Differenz auch sinnvoll.
- Die abhängige Variable ist normalverteilt oder es liegt eine hinreichend große Stichprobe vor, dass von einer guten Näherung der Stichprobenkennwerteverteilung des Mittelwerts durch eine Normalverteilung aufgrund des zentralen Grenzwerttheorems ausgegangen werden kann.

Sind diese Voraussetzungen nicht erfüllt, dann sind die Argumente, die oben verwendet wurden, um die Testentscheidungen plausibel zu machen, nicht mehr gültig. Liegt beispielsweise keine (näherungsweise) Normalverteilung der Stichprobenkennwerteverteilung des Mittelwerts vor, dann ist die Verteilung der Teststatistik T, die oben betrachtet wurde, im Allgemeinen nicht bekannt, die kritischen Bereiche können nicht ermittelt werden, und es ist nicht klar, wie p-Werte auf der Grundlage

fälschlich angenommener t-Verteilungen zu interpretieren sind. In diesen Fällen kann auf Verfahren zurückgegriffen werden, für die diese Voraussetzung(en) nicht gelten müssen, um belastbare Entscheidungsgrundlagen zu liefern.

Eine Möglichkeit bietet hierzu das sog. Bootstrap-Verfahren, das hier zwar nicht im Detail besprochen wird, aber auf dessen Durchführbarkeit mit SPSS wenigstens hingewiesen werden soll. Entsprechende 95%-Bootstrap-Konfidenzintervalle können unter *Analyze >> Compare Means and Proportions >> One-Sample T Test…* und dort im Untermenü „Bootstrap…“ durch Auswählen der Option „Perform bootstrapping“ angefordert werden. Für weitere Details wird an dieser Stelle aber auf spezialisierte Literatur bzw. Lernmaterialien verwiesen (siehe z.B. Bühner & Ziegler, 2017; Field, 2024; Wilcox, 2022).

Teststärke und Stichprobenplanung

Die Irrtumswahrscheinlichkeit (bzw. die Wahrscheinlichkeit für sog. *Fehler 1. Art*) wurde oben bereits kurz erläutert. Zur Wiederholung: Es handelte sich dabei um die Häufigkeit bei wiederholter Ziehung einfacher Zufallsstichproben *unter Geltung der Nullhypothese* die Nullhypothese mit dem Signifikanzniveau α zu verwerfen, d.h. die Nullhypothese fälschlicherweise zu verwerfen. Der sozusagen umgekehrte Irrtum, nämlich die Nullhypothese nicht zu verwerfen, obwohl die Alternativhypothese zutrifft (d.h. im hier betrachteten Fall, dass auch wirklich ein Unterschied zwischen dem Populationsmittelwert und einem gegebenen Testwert besteht), wird als Fehler 2. Art (oder β-Fehler) bezeichnet. In diesem Kapitel bezieht er sich auf die Frage: Wie oft verwerfen wir die Nullhypothese nicht mit einem Einstichproben t-Test, obwohl sie nicht gilt? Die Komplementärhäufigkeit bzw. -wahrscheinlichkeit, mit der bei wiederholter Ziehung einer einfachen Zufallsstichprobe die Nullhypothese verworfen wird, wenn die Alternativhypothese zutrifft, steht in direktem Zusammenhang zum β-Fehler und wird als Teststärke (Engl.: power) bezeichnet: Ist die Wahrscheinlichkeit für einen Fehler 2. Art gleich β, so ist die Teststärke gleich $1 - \beta$ und umgekehrt.

Die Teststärke eines Verfahrens hängt dabei von drei Größen ab: der Effektstärke, dem Signifikanzniveau α und dem Stichprobenumfang n. Zur Ermittlung der Teststärke muss also vorab bekannt sein, wie groß die Effektstärke eines bestimmten Mittelwertsunterschieds ist, um dann für ein

gegebenes Signifikanzniveau und einen gegebenen Stichprobenumfang berechnen zu können, wie häufig bei wiederholter Ziehung einer einfachen Zufallsstichprobe mit einem p-Wert kleiner α zu rechnen ist. Umgekehrt kann dieser Zusammenhang zwischen den vier Größen (Teststärke, Effektstärke, Signifikanzniveau, Stichprobenumfang) aber genutzt werden, um bei einer fundierten Vermutung für eine Mindesteffektstärke, den Stichprobenumfang so planen zu können, dass bei einer geringen Fehlerwahrscheinlichkeit für einen Fehler 1. Art gleichzeitig auch eine hohe Wahrscheinlichkeit für die Verwerfung der Nullhypothese (d.h. auch eine geringe Wahrscheinlichkeit für einen Fehler 2. Art) besteht.

Wenn zum Beispiel die begründete Vermutung besteht, dass sich ein Populationsmittelwert mindestens um eine Effektstärke von Cohens $d = 0.2$ von einem bestimmten Testwert unterscheiden sollte und die Fehlerwahrscheinlichkeit 1. Art mit einem α von 0.005 klein gehalten werden soll (man möchte die Nullhypothese also nicht *fälschlicherweise* verwerfen), dann kann β so berechnet werden, dass in beispielsweise 95% aller Einstichproben t-Tests für die jeweilige einfache Zufallsstichprobe ein $p < \alpha$ resultiert, sofern tatsächlich Cohens $d = 0.2$ gilt.

Eine Stichprobenplanung dieser Art kann beispielsweise mit dem unter https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower frei verfügbaren Programm G*Power durchgeführt werden. Nach dem Herunterladen und Öffnen der aktuellsten Version kann unter „Test family“ die Option „t tests“ ausgewählt werden. Unter „Statistical test“ ist dann für das vorliegende Beispiel „Means: Difference from constant (one sample case)“ auszuwählen. Unter „Type of power analysis“ kann die Voreinstellung „A priori: Compute required sample size – given α, power, and effect size“ so belassen werden. Nun sind die „Input Parameters“ zu wählen. Wir haben eine ungerichtete Hypothese („ein Populationsmittelwert *unterscheidet sich* von einem bestimmten Testwert“), d.h. wir wählen unter „Tail(s)“ die Option „Two“, da sich bei einer ungerichteten Hypothese der kritische Bereich unter der t-Verteilung in zwei Bereiche (einer für positive, einer für negative Werte der Teststatistik) gliedert. Diese beiden Bereiche unterhalb der t-Verteilung werden als Flanken bzw. Ränder der Verteilung (Engl.: Tails) bezeichnet. Ferner möchten wir mindestens eine Effektstärke von 0.2 (gemessen in Einheiten von Cohens *d*) mit hoher Wahrscheinlichkeit detektieren (d.h. ein signifikantes Ergebnis

erhalten, d.h. $p < \alpha$), d.h. wir geben im Feld „Effect size d“ die Zahl 0.2 ein. Wir möchten allerdings die Nullhypothese nicht bzw. nur mit geringer Wahrscheinlichkeit verwerfen, wenn sie zutrifft, daher geben wir im Feld „α err prob“ den Wert 0.005 ein. Schließlich möchten wir – wie oben bereits gesagt – den vermuteten Effekt mit hoher Wahrscheinlichkeit detektieren, daher geben wir im Feld „Power (1-β) err prob“ (das ist also die Teststärke, die wir gerne hätten) den Wert 0.95 ein (der netterweise schon voreingestellt ist). Diese Auswahlen und Eingaben sind in Abbildung 4.5 illustriert.

Durch einen Klick auf die Schaltfläche „Calculate“ wird die Berechnung des benötigten Stichprobenumfangs durchgeführt. Neben dem unserem Signifikanzniveau entsprechenden kritischen t-Wert und den Freiheitsgraden der entsprechenden t-Verteilung erhalten wir auch den uns hauptsächlich interessierenden Stichprobenumfang von $n = 500$. Wenn wir also einen so kleinen Mittelwertsunterschied verlässlich (beide Fehlerarten betreffend) detektieren wollen, müssen wir eine recht umfangreiche Stichprobe erheben. Zusätzlich erhalten wir in der Ausgabe noch einen „Noncentrality parameter δ“, der uns im Rahmen dieser Übungen nicht interessieren muss, und die eigentliche Teststärke für die gegebene Effektstärke, das gegebene Signifikanzniveau und die erhaltene Stichprobengröße. Der Grund dafür, dass die eigentliche Teststärke nicht genau der von uns geforderten entspricht, besteht schlichtweg darin, dass der Stichprobenumfang nur natürliche Zahlen annehmen kann und daher bei fixierter Effektstärke und fixiertem Signifikanzniveau eine kleine Abweichung von der geforderten Teststärke in Kauf genommen werden muss, damit sich für den Stichprobenumfang genau eine ganze Anzahl von Fällen ausgeht. Die erhaltene Ausgabe für die soeben durchgeführte Stichprobenplanung ist in Abbildung 4.6 gezeigt.

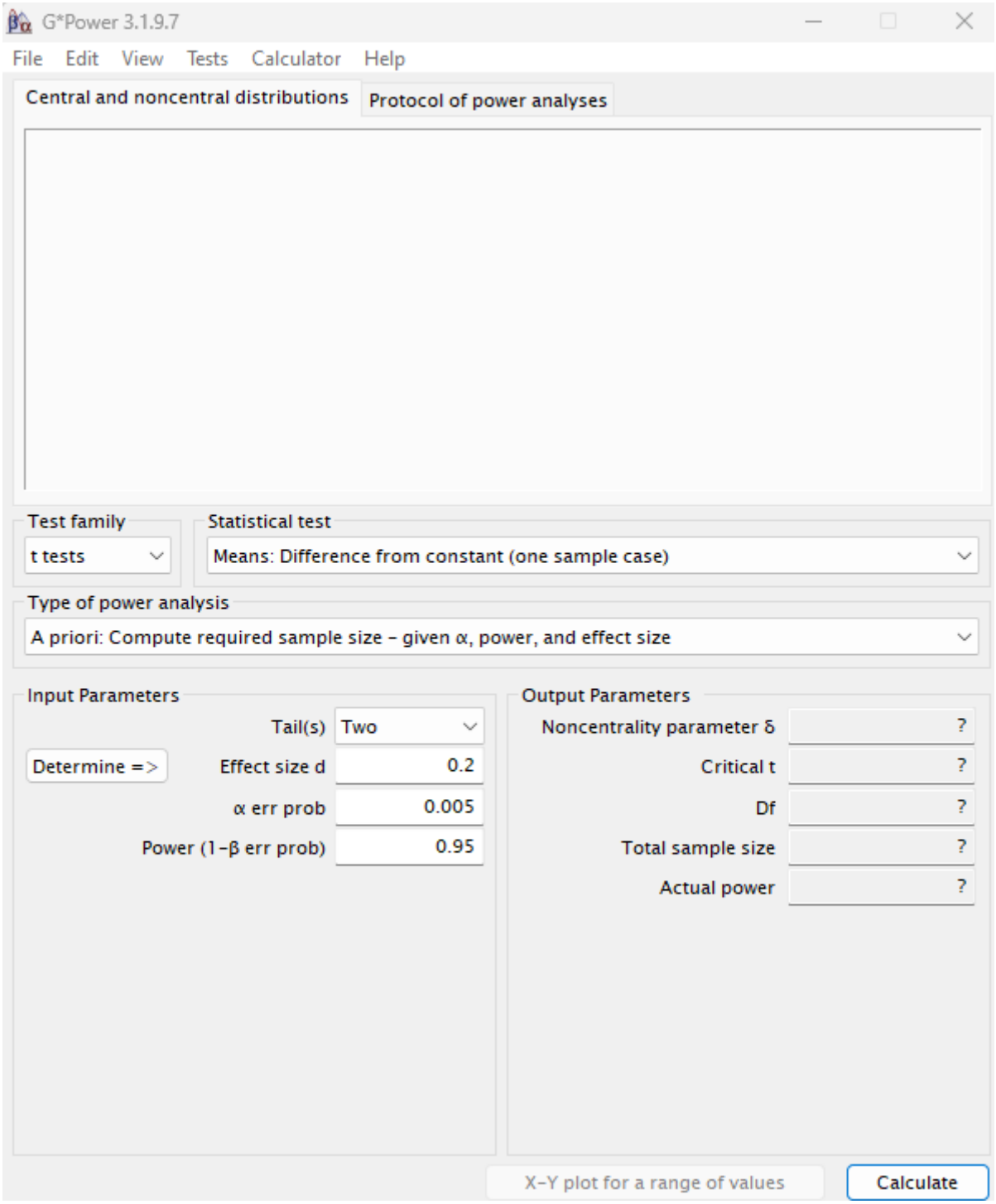


Abbildung 4.5. Eingaben für eine Stichprobenplanung für einen Einstichproben t-Test mit dem Programm G*Power.

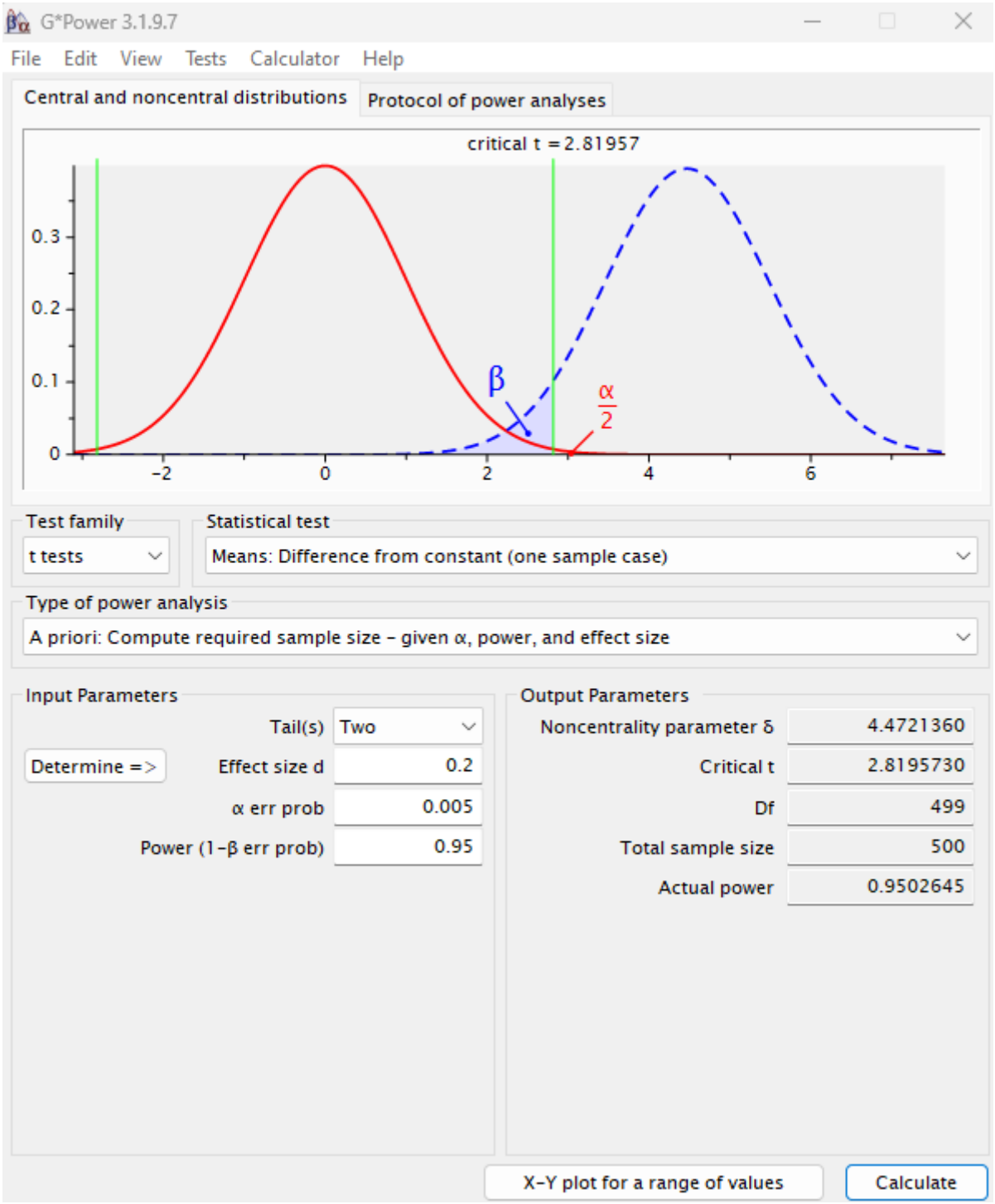


Abbildung 4.6. Ausgabe für die soeben durchgeführte Stichprobenplanung.

Übungsaufgaben

Beispiel 4.1

Was gehört zur Definition einer einfachen Zufallsstichprobe?

(a) Alle Merkmalsträger:innen einer Population haben dieselbe Wahrscheinlichkeit in die Stichprobe gezogen zu werden.

(b) Die einzelnen Ziehungen müssen unabhängig voneinander sein.

(c) Merkmalsträger:innen müssen Personen sein.

(d) Der Stichprobenumfang muss größer 30 sein.

Beispiel 4.2

Welche Aussage(n) zum p-Wert trifft(treffen) zu?

(a) Der p-Wert ist die Wahrscheinlichkeit, dass die Nullhypothese wahr ist.

(b) Aus dem p-Wert kann man die Wahrscheinlichkeit ableiten, dass die Alternativhypothese wahr ist.

(c) Der p-Wert ist die Wahrscheinlichkeit dafür sich bei der Verwerfung der Nullhypothese zu irren.

(d) Würde man das Experiment sehr oft wiederholen so würde man in $(1-p)\cdot 100\%$ aller Fälle ein signifikantes Ergebnis erhalten.

Beispiel 4.3

Welche der folgenden Aussagen ist/sind richtig/falsch?

Nr.	**Aussage**	**R/F**
1)	Es kann sein, dass der p-Wert kleiner als das Signifikanzniveau α ist, aber die Teststatistik T nicht im Ablehnungsbereich der Nullhypothese liegt.	
2)	Für eine ungerichtete Hypothese ist der p-Wert die Wahrscheinlichkeit unter Annahme der Gültigkeit der Nullhypothese dafür, dass sich die Teststatistik in der beobachteten Realisation oder einer extremeren Realisation in Richtung der Alternativhypothese realisiert.	
3)	Ist der p-Wert klein, dann liegt der wahre Populationsmittelwert weit weg vom Testwert.	
4)	Ist der p-Wert klein, dann hat man einen Effekt mit großer Effektstärke detektiert.	

Beispiel 4.4

Welche der folgenden Aussagen ist/sind richtig/falsch?

Nr.	**Aussage**	**R/F**
1)	Ein p-Wert größer als das gewählte Signifikanzniveau bedeutet, dass es keinen Unterschied zwischen dem Populationsmittelwert und dem Testwert gibt.	
2)	Ein p-Wert größer als das gewählte Signifikanzniveau bedeutet, dass die Nullhypothese stimmt.	
3)	Ein p-Wert größer als das gewählte Signifikanzniveau bedeutet, dass die Nullhypothese eher stimmt als die Alternativhypothese.	
4)	Ein p-Wert kleiner als das gewählte Signifikanzniveau bedeutet, dass die Alternativhypothese zutrifft.	

Beispiel 4.5

Welche Aussage(n) trifft(treffen) zu?

(a) Um eine Stichprobenplanung in G*Power durchzuführen muss man wissen, wie viele Personen man insgesamt in einer Studie testen wird.

(b) Die Teststärke (power) hängt von der Effektstärke, dem Signifikanzniveau und dem Stichprobenumfang ab.

(c) Bei Cohens *d* handelt es sich um ein einheitsunabhängiges Maß für die Teststärke (power).

(d) Ab einem Cohens $d > 2$ spricht man gemäß Cohens Heuristik (1988) von einem großen Effekt.

Beispiel 4.6

Was gehört zu den Voraussetzungen für einen Einstichproben t-Test?

(a) Die Varianz der Population ist bekannt.

(b) Die abhängige Variable ist normalverteilt oder es liegt eine hinreichend große Stichprobe vor.

(c) Die abhängige Variable muss mindestens ordinalskaliert sein.

(d) Die Varianz der Population ist nicht bekannt.

Beispiel 4.7

In Österreich beträgt das Durchschnittsalter von Studierenden 27.1 Jahre (jedenfalls gemäß https://www.studium.at/oesterreichische-studenten-sind-im-schnitt-aelter-und-arbeiten-haeufiger). In der Datei „Kap3daten.sav“ finden Sie das Alter von 51 (fiktiven) Studierenden zu einem Zeitpunkt, an dem sie am Kurs „Anwendung statistischer Verfahren am Computer“ teilgenommen hatten. Verwenden Sie einen Einstichproben t-Test, um zu einer Entscheidung zu kommen, ob sich das Alter der Kursteilnehmer:innen vom oben angegebenen Durchschnittsalter von Studierenden unterscheidet. Verwenden Sie ein Signifikanzniveau von 0.5%. Berichten Sie Ihre Resultate gemäß APA-Richtlinien und geben Sie auch eine Intervallschätzung für den Populationsmittelwert der Kursteilnehmer:innen an. Sind die Voraussetzungen für einen Einstichproben t-Test erfüllt?

Beispiel 4.8

Ein Medikament werde als die Verkehrstüchtigkeit einschränkend bezeichnet, sobald es die Reaktionszeit im Mittel um mehr als 50 ms verringere.

Um die Wirkung eines neuen Medikaments auf die Verkehrstüchtigkeit zu testen, lässt daher ein (fiktives) Pharmaunternehmen die Wirkung des Medikaments auf die Reaktionszeitverzögerung (RZV) bei 42 Versuchspersonen prüfen. In der Datendatei „Kap4UE8.sav“, die Sie in dem elektronischen Ergänzungsmaterial finden, das Sie unter https://osf.io/9tcx3/ herunterladen können, finden Sie den dazugehörigen Datensatz. Verwenden Sie ein angemessenes statistisches Verfahren, um zu testen, ob die Reaktionszeitverzögerung in der Population durch Einnahme des Medikaments mehr als 50 ms im Mittel beträgt. Wählen Sie dafür ein Signifikanzniveau von 0.05. Berichten Sie Ihre Ergebnisse gemäß APA-Richtlinien.

Beispiel 4.9

Eine Forscherin zweifelt an der Sinnhaftigkeit der Überprüfung des Medikaments aus dem vorhergehenden Beispiel durch das Pharmaunternehmen. Ihre Argumentation lautet wie folgt.

Aufgrund der Bedeutung der Überprüfung (Verkehrstüchtigkeit) sollten schon sehr kleine Änderungen über den Schwellenwert von 50 ms hinaus zu einer entsprechenden Kennzeichnung des Medikaments führen. Die Forscherin setzt dafür ein Cohens $d = 0.1$ als Limit an und fordert, dass eine

ernstzunehmende Untersuchung einen solchen Unterschied mit hoher Teststärke von mindestens 95% detektieren können sollte. Sie argumentiert weiter, dass das Signifikanzniveau dafür durchaus auf $\alpha = 0.1$ erhöht werden könne, da ein Fehler 1. Art in diesem Fall das weitaus geringere Übel darstelle.

(a) Verwenden Sie G*Power, um für die angegebenen Werte den nötigen Stichprobenumfang einer entsprechenden Untersuchung zu ermitteln.

(b) Verwenden Sie G*Power, um zu ermitteln, welche Teststärke die Untersuchung des vorhergehenden Beispiels aufwies, um einen Effekt der Stärke Cohens $d = 0.1$ zu detektieren. *Hinweis*: Sehen Sie sich dafür die Option „Post hoc: Compute achieved power…“ unter „Type of power analysis“ an.

Beispiel 4.10

Die historische Entwicklung der Statistik hat bekanntlich viel mit der Bierbrauerei zu tun. So arbeitete beispielsweise William Gosset, der den sog. Studentschen t-Test entwickelt hat und hinter dem Pseudonym Student in der entsprechenden Arbeit aus dem Jahr 1908 steckt, in der Guinness Brauerei in Dublin. Noch heute kann man dort eine Plakette bewundern, die ihm zu Ehren dort angebracht wurde, siehe z.B. https://en.wikipedia.org/wiki/William_Sealy_Gosset.

Seine Arbeit (vermutlich in der Qualitätssicherung) in der Brauerei stelle ich mir gerne wie folgt vor: Ein Bauer bringt einige Proben der aktuellen Hopfenernte vorbei, weil er neuer Hauptlieferant der ehrwürdigen Brauerei (und entsprechend gut bezahlt) werden möchte. William Gosset unterzieht die Proben einigen Tests, aus denen schließlich für jede Probe ein bestimmter Qualitätskennwert resultiert, z.B. diese Liste an Zahlen: 53, 77, 44, 62, 57, 48, 75, 71, 65, 65. Aus langjähriger Erfahrung weiß Gosset, dass es sich um qualitativ hochwertigen Hopfen handelt, wenn dieser im Mittel einen Qualitätskennwert von mindestens 50 übersteigt.

Verwenden Sie einen geeigneten statistischen Test, um zu testen, ob die Ernte des Bauern diesem Qualitätsanspruch potentiell gerecht werden kann. Sie können für dieses Beispiel davon ausgehen, dass die Qualitätskennwerte durch eine identisch und unabhängig normalverteilte Zufallsvariable approximiert werden können. Erstellen Sie einen Ergebnisbericht gemäß APA-Richtlinien.

Beispiel 4.11

Wie viele Personen muss eine Stichprobe umfassen, damit ein Unterschied eines Populationsmittelwerts der Stärke Cohens $d = 0.25$ von einem gegebenen, konstanten Wert von 100 für ein Signifikanzniveau $\alpha = .005$ mit einer Teststärke (= power) von 80% detektiert werden kann?

Beispiel 4.12

Tun Sie sich für diese Übungsaufgabe mit einem:einer Kolleg:in zusammen. Erstellen Sie jeweils unabhängig voneinander einen Ergebnisbericht für Beispiel 4.7. Überprüfen Sie danach gegenseitig Ihre Ergebnisberichte mit der am Ende dieses Dokuments bereitgestellten Lösung. Markieren und korrigieren Sie etwaige Fehler und seien Sie dabei ruhig möglichst streng. Diskutieren Sie anschließend Ihre gegenseitigen Korrekturen und klären Sie gemeinsam Fragen, die sich dabei ergeben.

Beispiel 4.13

Tun Sie sich für diese Übungsaufgabe mit einem:einer Kolleg:in zusammen. Erstellen Sie jeweils unabhängig voneinander einen Ergebnisbericht für Beispiel 4.8. Überprüfen Sie danach jeweils selbst die Korrektheit Ihres Ergebnisberichts mit der am Ende dieses Dokuments bereitgestellten Lösung. Fügen Sie anschließend in Ihren Ergebnisbericht 5 Fehler ein, ohne sie Ihrem:Ihrer Kolleg:in mitzuteilen (und es dürfen durchaus Fehler sein, die nur schwer zu entdecken sind). Tauschen Sie anschließend Ihre fehlerhaften Ergebnisberichte aus. Versuchen Sie nun jeweils die fünf Fehler zu identifizieren und zu korrigieren, indem Sie ausschließlich die Angabe von Beispiel 4.8, den entsprechenden Datensatz und SPSS verwenden, d.h. insbesondere, ohne dabei die Musterlösung zu verwenden. Für die korrekte Identifikation eines Fehlers gibt es einen Punkt, für die korrekte Korrektur eines Fehlers einen weiteren Punkt. D.h. Sie können jeweils 10 Punkte erreichen. Wer mehr Punkte erreicht gewinnt! Bei einem Unentschieden spielen Sie einfach noch eine Runde.

Beispiel 4.14

Reflektieren Sie schriftlich: Welche Voraussetzungen müssen für einen Einstichproben-t-Test erfüllt sein? Wie können Sie überprüfen, ob diese Voraussetzungen erfüllt sind? Was sind die Konsequenzen, wenn diese Voraussetzungen nicht erfüllt sind? Welche Alternativen haben Sie, falls die Voraussetzungen nicht erfüllt sind? Machen Sie vor allem zur Beantwortung der letzten Frage Gebrauch von entsprechender Literatur. Sie können auch von generativer Künstlicher Intelligenz Gebrauch machen, aber überprüfen Sie die erhaltenen Antworten jedenfalls mit einschlägiger Literatur.

Kapitel 5

Schätzung und Testung von Mittelwertsunterschieden zwischen zwei Gruppen

Stefan E. Huber

Im vorhergehenden Kapitel haben wir sehr viele grundlegende theoretische Konzepte wiederholt. Dies haben wir nicht grundlos getan. In der Tat haben wir mit der Wiederholung dieser Grundkonzepte bereits alles vorbereitet, was wir in diesem Kapitel brauchen werden, um Mittelwertsunterschiede zwischen zwei Gruppen schätzen und gegen einen vorgegebenen Testwert prüfen zu können. Wesentliche Aspekte dieser Grundkonzepte werden wir auch in den kommenden Kapiteln immer wieder brauchen können. D.h. auch, dass wir ab jetzt den *Durchführungsaspekten* der unterschiedlichen statistischen Verfahren mehr und mehr Platz einräumen werden, da wir für konzeptuelle Betrachtungen oder Wiederholungen weitgehend auf diese Grundkonzepte zurückgreifen können und diese nur stellenweise ergänzen werden müssen.

In diesem Kapitel sehen wir uns das einmal für die Schätzung und Testung von Mittelwertsunterschieden zwischen zwei Gruppen an. Aus der Theorie wissen wir (Bühner et al., 2025), dass wir dafür zwei Fälle unterscheiden müssen (strenggenommen gibt es noch einen dritten Fall, mit dem wir uns aber nicht befassen werden, siehe z.B. Wilcox, 2022, S. 203-210). Im ersten Fall ziehen wir zwei einfache Zufallsstichproben aus zwei unterschiedlichen Populationen und möchten den Unterschied zwischen den beiden Mittelwerten dieser Populationen schätzen bzw. gegen einen Testwert prüfen. In diesem Fall spricht man von zwei unabhängigen Stichproben. Bezüglich der Testung ist dabei häufig der Spezialfall der Gleichheit bzw. Ungleichheit der beiden Populationsmittelwerte interessant, was der Testung des Mittelwertsunterschieds gegen den Wert Null entspricht. Im zweiten Fall ziehen wir grundsätzlich nur eine einfache Zufallsstichprobe, aber erfassen für jeden Fall zwei aufeinander bezogene Variablen. Allerdings spricht man auch in diesem Fall von zwei Stichproben, allerdings nun von zwei abhängigen Stichproben. Dabei kann es sich etwa um die Erfassung derselben Variablen zu zwei verschiedenen Zeitpunkten handeln (etwa die Reaktionszeit vor und nach Einnahme eines bestimmten Medikaments) oder um zwei Variablen, die aber eindeutig miteinander zusammenhängen (etwa der systolische Blutdruck, einmal gemessen am linken Arm und einmal gemessen am rechten Arm

jeweils derselben Person). Aus Gründen, die sehr bald klarer sein werden, beginnen wir mit dem zweiten der beiden Fälle: der Schätzung und Testung der Mittelwertsunterschiede für zwei abhängige Stichproben.

Schätzung und Testung der Mittelwertsunterschiede für zwei abhängige Stichproben

Im Falle zweier abhängiger Stichproben lässt sich leicht einsehen, dass wir zur Schätzung und Testung des Unterschieds zwischen den Populationsmittelwerten auch die mittlere Differenz zwischen den beiden abhängigen Variablen schätzen bzw. testen können, da

$$\bar{X}_1 - \bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{1i} - \frac{1}{n}\sum_{i=1}^{n} X_{2i} = \frac{1}{n}\sum_{i=1}^{n}(X_{1i} - X_{2i}) = \frac{1}{n}\sum_{i=1}^{n} X_{Di} = \bar{X}_D$$

mit X_{1i} der ersten der beiden Zufallsvariablen für Fall (Person) i, X_{2i} der zweiten der beiden Zufallsvariablen für Fall (Person) i, X_{Di} der Differenz der beiden Zufallsvariablen für Fall (Person) i, $\bar{X}_1, \bar{X}_2, \bar{X}_D$ den entsprechenden Mittelwerten, und n dem Stichprobenumfang. Hier wurde bereits angenommen, dass sich die interessierenden Variablen durch entsprechende Zufallsvariablen approximieren lassen.

Sofern sich nun X_{1i} und X_{2i} insbesondere als identisch und unabhängig *normalverteilte* Zufallsvariablen mit Erwartungswerten μ_1 und μ_2 und Varianzen σ_1^2 und σ_2^2 approximieren lassen, wissen wir ebenfalls, dass sich X_{Di} als identisch und unabhängig *normalverteilte* Zufallsvariable mit Erwartungswert $\mu_1 - \mu_2$ und Varianz σ_D approximieren lässt (für hinreichend große Stichproben würde uns aber auch hier wieder das zentrale Grenzwerttheorem zu Hilfe kommen). Die Varianz σ_D ist dabei nicht bekannt und muss aus der Stichprobe geschätzt werden.

Alles, was wir benötigen, um den Erwartungswert einer solchen Zufallsvariable zu schätzen und zu testen, haben wir allerdings schon im vorhergehenden Kapitel besprochen! Denn obwohl es sich ursprünglich um zwei abhängige Variablen handelte, handelt es sich bei der Differenz der beiden Variablen nur um eine Variable für eine einfache Zufallsstichprobe. Können wir deren Wert schätzen, dann haben wir damit auch die Differenz der Mittelwerte der beiden ursprünglichen Variablen geschätzt. Testen wir deren Wert gegen eine bestimmte Konstante so haben wir die Differenz der beiden Variablen gegen einen bestimmten Wert geprüft.

D.h. wir können dazu in SPSS auch prinzipiell völlig gleich verfahren. Sehen wir uns das an einem Beispiel an, für den Sie den entsprechenden Datensatz in der Datei „Kap5daten.sav“ finden, die Sie wieder in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden können, das Sie unter https://osf.io/9tcx3/ herunterladen können. Es handelt sich dabei wiederum um fiktive Daten von 200 (ebenso fiktiven) Studierenden, deren Statistikwissen vor (Variable *Punkte_vorher*) und nach (Variable *Punkte_nachher*) dem Besuch eines Tutoriums mit einem entsprechenden Test gemessen wurde, der für jede:n Studierende:n einen Testwert zwischen 0 und 100 Punkten ergibt.

Um nun die Differenz der Populationsmittelwerte mittels der Mittelwertdifferenz der beiden Variablen *Punkte_vorher* und *Punkte_nachher* zu schätzen und gleich auch gegen einen Testwert von 0 zu testen, können wir eine neue Variable unter *Transform* >> *Compute Variable*... berechnen, die wir z.B. einfach *Diff* (für Differenz) nennen, siehe Abbildung 5.1. Durch Einfügen in eine Syntaxdatei (zur Dokumentation) und Ausführen der entsprechenden Kommandozeilen wird die neue Differenzvariable erzeugt.

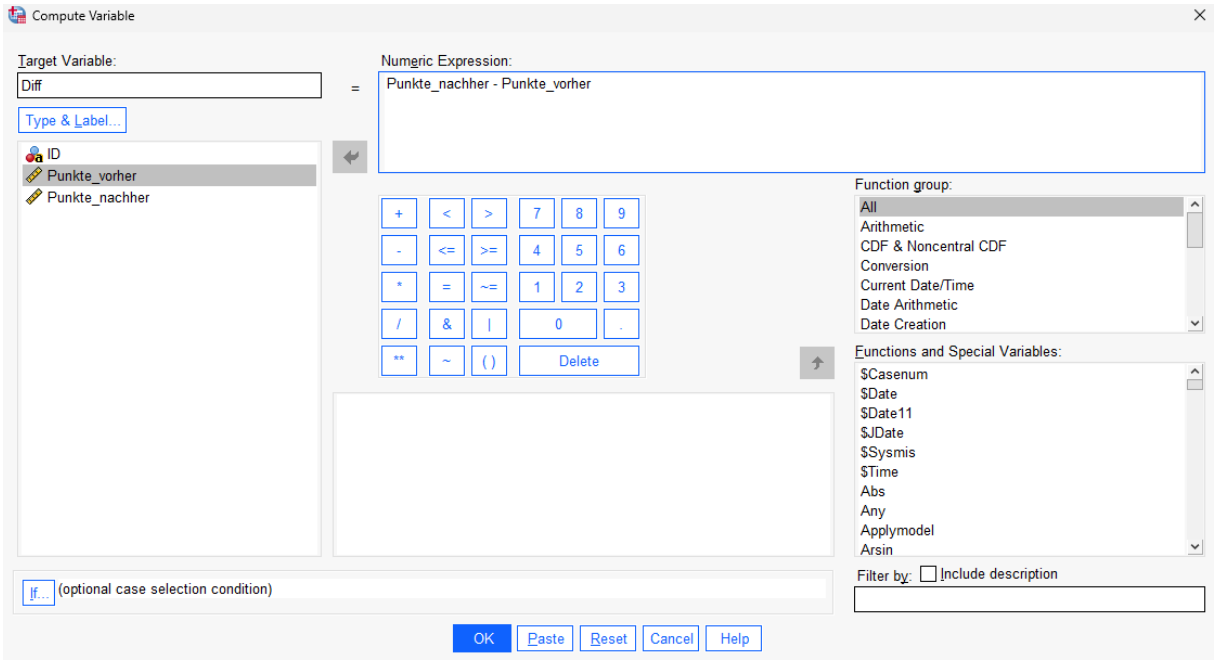


Abbildung 5.1. Bildung einer neuen Differenzvariable.

Für diese Differenzvariable können wir nun unter *Analyze >> Compare Means and Proportions >> One-Sample T Test…* einen Einstichproben t-Test gegen den Testwert 0 berechnen lassen und uns auch ein 95%-Konfidenzintervall (KI) sowie Effektstärken ausgeben lassen, siehe Abbildung 5.2.

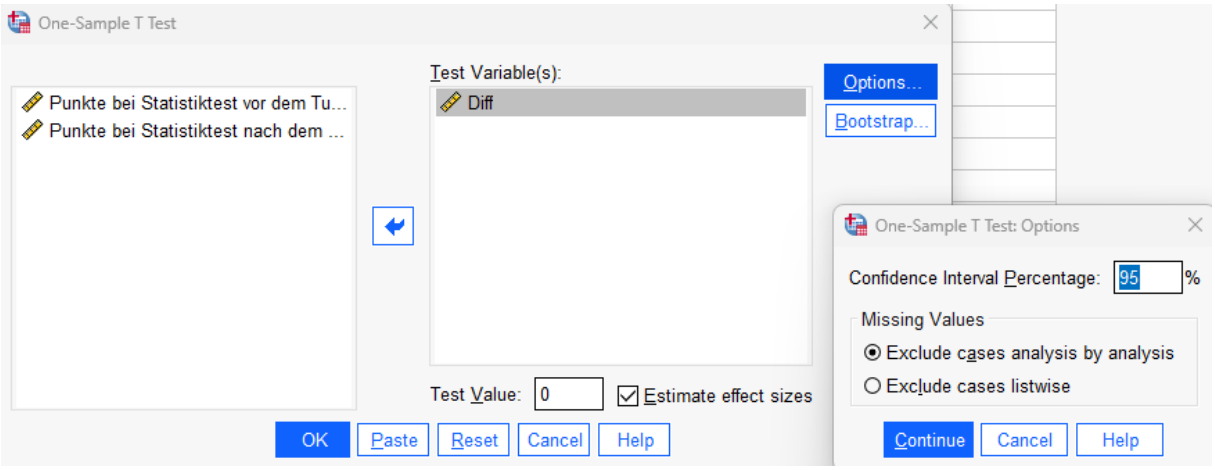


Abbildung 5.2. Anforderung eines Einstichproben t-Tests für unsere Differenzvariable.

Am Ergebnis unseres Einstichproben t-Tests für unsere Differenzvariable erkennen wir, siehe Abbildung 5.3, dass die Punktschätzung für die Differenz zwischen den Populationsmittelwerten $\bar{x}_D = \bar{x}_2 - \bar{x}_1 =$ 9.5 Testpunkte beträgt, mit einem 95%-KI von [6.84, 12.16]. Zudem weicht die Mittelwertdifferenz (mit α = .05) signifikant vom Wert Null ab, $t(199) = 7.05$, $p < .001$, Cohens d = 0.50 mit 95%-KI [0.35, 0.65]. Gemäß Cohens Heuristik (1988) liegt also ein kleiner bis mittlerer Effekt vor. Es sieht also ganz danach aus, als hätte das Tutorium (zumindest im Mittel) auch etwas für das Statistikwissen gebracht!

Allerdings gibt es in SPSS noch einen bequemeren Weg dieselben Ergebnisse zu erhalten, den wir uns jetzt ansehen werden. Wir haben die Berechnung bisher nur deshalb etwas umständlicher durchgeführt, um uns davon zu überzeugen, dass (mit einem kleinen Unterschied) auf die bequemere Art und Weise genau dasselbe einfach im Hintergrund von SPSS durchgeführt wird. Für die bequemere Durchführung wählen wir unter *Analyze >> Compare Means and Proportions* nun nicht „One-Sample T Test“, sondern stattdessen „Paired-Samples T Test“ aus. Im sich öffnenden Fenster fügen wir die Variable *Punkte_vorher* unter „Variable1“ und *Punkte_nachher* unter „Variable2“ ein; alle übrigen Voreinstellungen lassen wir genauso wie sie sind, siehe Abbildung 5.4. Das Ergebnis ist in Abbildung 5.5 dargestellt.

T-Test

[DataSet1] C:\Users\hubestef\Documents\UniGraz2023\Lehre\SoSe2025\ASVA_dieses_semester\resources\Kap5daten.sav

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Diff	200	9.5000	19.04953	1.34700

One-Sample Test

Test Value = 0

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Diff	7.053	199	<.001	<.001	9.50000	6.8438	12.1562

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Diff	Cohen's d	19.04953	.499	.351	.645
	Hedges' correction	19.12170	.497	.350	.643

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

Abbildung 5.3. Ergebnis des Einstichproben t-Tests für unsere Differenzvariable.

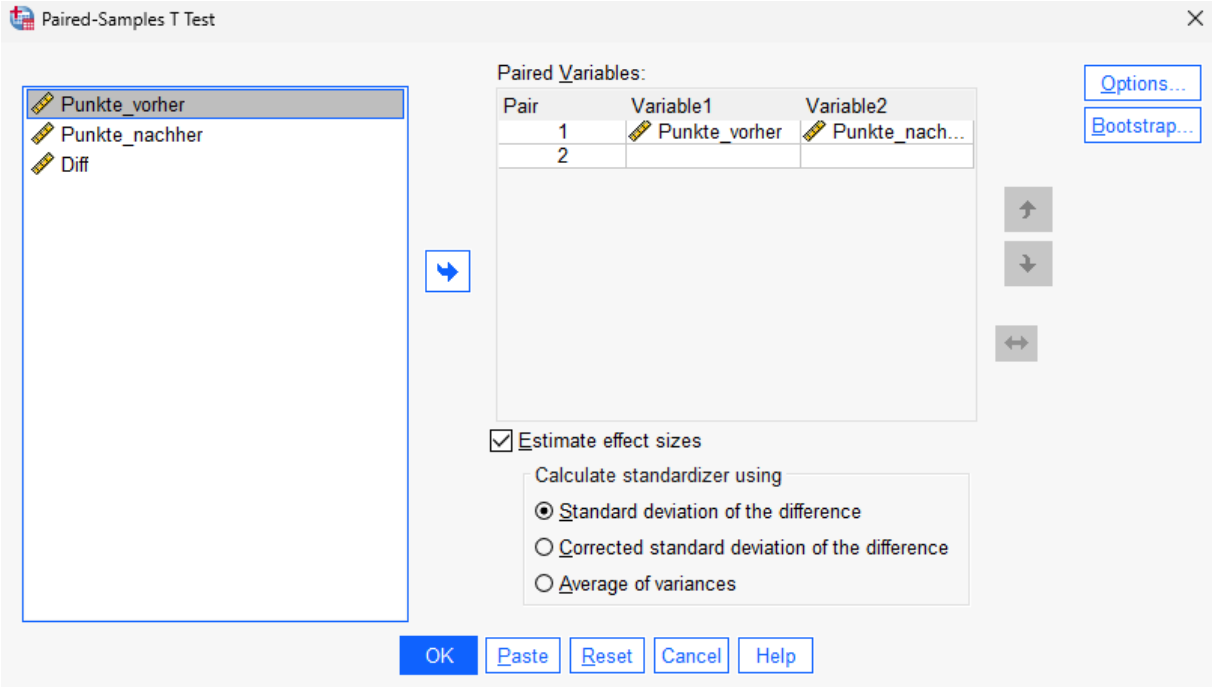


Abbildung 5.4. Eine bequemere Art in SPSS eine Schätzung und Testung (gegen 0) einer Mittelwertdifferenz für zwei abhängige Stichproben durchzuführen.

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Punkte bei Statistiktest vor dem Tutorium	30.97	200	19.918	1.408
	Punkte bei Statistiktest nach dem Tutorium	40.47	200	25.783	1.823

Paired Samples Correlations

		N	Correlation	Significance One-Sided p	Significance Two-Sided p
Pair 1	Punkte bei Statistiktest vor dem Tutorium & Punkte bei Statistiktest nach dem Tutorium	200	.680	<.001	<.001

Paired Samples Test

		Paired Differences					t	df	Significance One-Sided p	Significance Two-Sided p
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper				
Pair 1	Punkte bei Statistiktest vor dem Tutorium - Punkte bei Statistiktest nach dem Tutorium	-9.500	19.050	1.347	-12.156	-6.844	-7.053	199	<.001	<.001

Paired Samples Effect Sizes

			Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Pair 1	Punkte bei Statistiktest vor dem Tutorium - Punkte bei Statistiktest nach dem Tutorium	Cohen's d	19.050	-.499	-.645	-.351
		Hedges' correction	19.122	-.497	-.643	-.350

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference.
Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

Abbildung 5.5. Ergebnisse der Durchführung des t-Tests für abhängige Stichproben in SPSS.

Das praktische an dieser Art der Durchführung eines sog. t-Tests für abhängige Stichproben in SPSS ist, dass wir zusätzlich zu den Informationen, die wir mit dem Einstichproben t-Test für die Differenzvariable erhalten haben, noch einige andere nützliche Informationen erhalten. Zum einen bekommen wir in der Tabelle „Paired Samples Statistics“ die für einen Ergebnisbericht (siehe unten) ohnehin benötigten deskriptiven Statistiken (d.h. Stichprobenumfang, Mittelwerte und Standardabweichungen jeweils für beide Variablen). In der Tabelle „Paired Samples Correlations“ bekommen wir zudem noch den Pearson-Korrelationskoeffizienten für die beiden Variablen r = .68, an dem wir erkennen, dass zwischen den beiden Variablen eine erhebliche Korrelation besteht. Gemäß Cohens Heuristiken (1988) gelten Korrelationskoeffizienten von 0.1-0.3 als klein, von 0.3-0.5 als mittel, und größer 0.5 als groß. Mit einer erheblichen Korrelation zwischen unseren beiden Variablen ist im vorliegenden Fall zu rechnen, da wir ja davon ausgegangen sind, dass die beiden Variablen voneinander abhängen. Hätten wir an dieser Stelle gesehen, dass zwischen den Variablen kaum ein Zusammenhang besteht (zumindest wie er durch den linearen Pearson Korrelationskoeffizienten überhaupt zum

Ausdruck kommen kann), hätte das unsere Annahme zweier abhängiger Variablen durchaus in Zweifel gezogen.

In den Tabellen „Paired Samples Test“ und „Paired Samples Effect Sizes“ finden wir dieselben Werte, die wir bereits oben erhalten haben. Falls es verwundert, dass wir hier nun ein negatives Vorzeichen für die mittlere Differenz (und alle davon abgeleiteten Größen) bekommen, so liegt das schlichtweg daran, dass SPSS im Rahmen des t-Tests für abhängige Stichproben die Differenz Variable1-Variable2 schätzt bzw. gegen den Wert Null testet und wir oben bei der Bildung der Differenzvariablen die Differenz gerade umgekehrt (Variable2-Variable1) gebildet haben. Inhaltlich ist das Ergebnis aber ganz identisch: die Punkte im Test vor dem Tutorium sind im Mittel kleiner als im Test nach dem Tutorium. Deshalb ist es für die inhaltliche Interpretation stets wichtig, die deskriptiven Statistiken zu betrachten, da man an diesen gut erkennt, zu welchem Zeitpunkt der Mittelwert größer bzw. kleiner ist. Vom Betrag her sind alle Zahlen dieser beiden Tabellen identisch mit jenen aus Abbildung 5.3.

Ein kleiner Nachteil des t-Tests für abhängige Stichproben in SPSS ist, dass die Mittelwertdifferenz nur über einen Umweg gegen einen anderen Testwert als Null getestet werden kann. Für den Einstichproben t-Test kann hierfür hingegen ein beliebiger Testwert gewählt werden. Gegen einen anderen Wert als Null zu testen kann zum Beispiel gewünscht sein, wenn eine Fragestellung vorliegt, in der ein Unterschied zwischen zwei abhängigen Variablen einen Mindestbetrag über- oder unterschreiten soll, wie es etwa in Fällen der Qualitätssicherung der Fall sein kann. Will man in so einem Fall dennoch den t-Test für abhängige Stichproben in SPSS verwenden, kann man den Wert gegen den man testen möchte, schlichtweg zum Subtrahenden addieren, d.h. wenn man die Nullhypothese $H_0: \mu_1 - \mu_2 = 5$ testen möchte, kann man eine neue Variable $x'_{2i} = x_{2i} + 5$ und anschließend den t-Test für abhängige Stichproben verwenden, um die Nullhypothese $H'_0: \mu_1 - \mu'_2 = 0$ zu testen. Die Verwerfung dieser Nullhypothese ist dann äquivalent zur Verwerfung der Nullhypothese H_0.

Ergebnisbericht für Schätzung und Testung der Mittelwertsunterschiede für zwei abhängige Stichproben

Wie nach jeder statistischen Analyse sind auch in diesem Fall die Ergebnisse in einem entsprechenden Ergebnisbericht festzuhalten. Dieser sollte jedenfalls die von SPSS ausgegebenen deskriptiven Statistiken für die beiden Variablen enthalten (mit Ausnahme des Standardfehlers, der ohnehin aus den anderen Größen berechnet werden könnte). Anschließend sollte die Punktschätzung für die Mittelwertdifferenz zusammen mit dem Konfidenzintervall (inkl. des verwendeten Signifikanzniveaus) angegeben werden. Zudem sollten wieder die wesentlichen Bestandteile des Signifikanztests (Art und Wert der Teststatistik, Freiheitsgrade, p-Wert) und schließlich die Effektstärke mit dem entsprechenden Konfidenzintervall angegeben werden. Für die Effektstärke wählen wir wiederum Cohens d.

Für den vorliegenden Fall könnte ein adäquater Ergebnisbericht also wie folgt aussehen: „Der Mittelwert der Punkte beim Test vor dem Tutorium ($M = 30.97$, $SD = 19.92$) von $N = 200$ Studierenden war niedriger als der Mittelwert der Punkte beim Test nach dem Tutorium ($M = 40.47$, $SD = 25.78$). Die Punktschätzung für die mittlere Populationsdifferenz ergab sich entsprechend zu 9.5 mit einem 95%-KI [6.84, 12.16]. Diese Mittelwertdifferenz unterscheidet sich (mit $\alpha = .05$) signifikant von Null, $t(199) = 7.05$, $p < .001$, Cohens $d = 0.50$ mit 95%-KI [0.35, 0.65]. Gemäß Cohens Heuristik (1988) liegt also ein kleiner bis mittlerer Effekt vor.“ Das APA-Format ist selbstverständlich auch wieder zu beachten.

Teststärke und Stichprobenplanung

Auch für einen t-Test für abhängige Stichproben kann eine Stichprobenplanung mittels G*Power für eine gewünschte Teststärke bei gegebenem Signifikanzniveau und vermuteter Effektstärke vorgenommen werden. Unter „Test family“ ist dafür wiederum „t tests“ auszuwählen. Unter „Statistical test“ ist diesmal „Means: Difference between two dependent means (matched pairs)“ auszuwählen. Unter “Type of power analysis” ist wieder “A priori: Compute required sample size – given α, power, and effect size” auszuwählen. Bei den “Input Parameters” ist wieder anzugeben, ob eine gerichtete (“one tail“) oder eine ungerichtete („two tails“) statistische Hypothese getestet werden soll. Bei der Effektstärke ist wiederum die vermutete Effektstärke in Einheiten von Cohens d anzugeben. Das Signifikanzniveau ist wiederum unter „α err prob“ und die gewünschte Teststärke unter „Power (1-β err prob)“ anzugeben.

Angenommen, wir vermuteten einen Effekt der Stärke 0.3 für einen Unterschied zwischen den Mittelwerten zweier abhängiger Stichproben (d.h. wir hätten eine ungerichtete Hypothese), möchten uns mit $\alpha = .005$ wiederum stark gegen einen Fehler 1. Art absichern und einen Effekt der veranschlagten Stärke schon mit einer Wahrscheinlichkeit von mindestens 80% detektieren (d.h. ein Ergebnis „$p < \alpha$“ in 80% einer Serie unendlicher vieler hypothetischer, äquivalenter Replikationen erhalten), dann wären in G*Power die in Abbildung 5.6 gezeigten Einstellungen vorzunehmen. Abbildung 5.7 zeigt, dass sich daraus ein benötigter Stichprobenumfang von $n = 152$ ergibt.

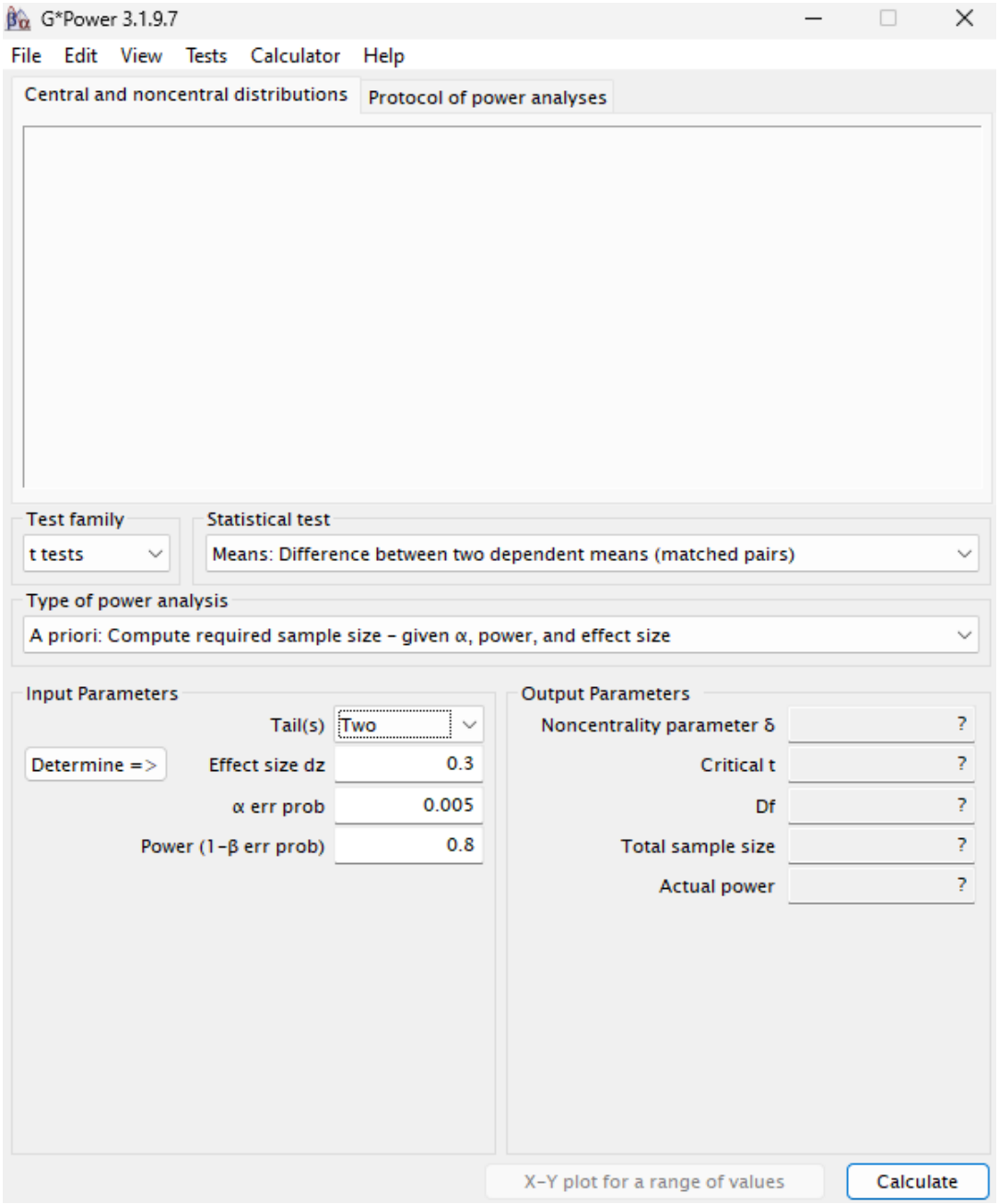


Abbildung 5.6. Einstellungen in G*Power für das im Text beschriebene Beispiel.

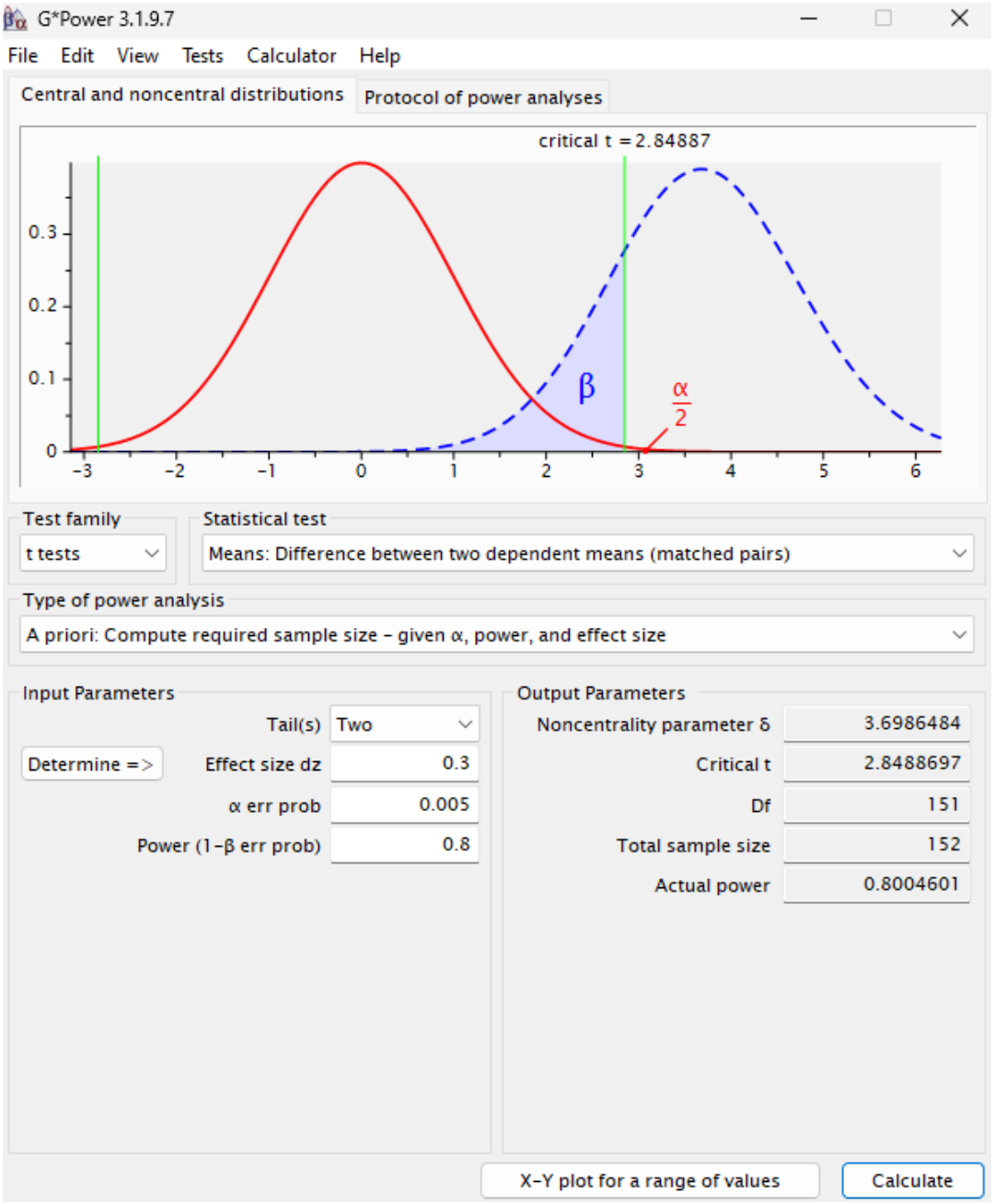


Abbildung 5.7. Ergebnisse der im Text beschriebenen Stichprobenplanung.

Voraussetzungen für den t-Test für abhängige Stichproben

Da es sich bei dem t-Test für abhängige Stichproben gemäß den vorhergehenden Erläuterungen um nichts anderes als einen Einstichproben t-Test in Verkleidung handelt, sind auch die Voraussetzungen ganz analog. D.h., es muss sich um mindestens intervallskalierte Variablen handeln, die Varianz der Differenz der Variablen ist nicht bekannt und muss aus den Daten geschätzt werden, und die Differenz der Variablen muss sich entweder selbst durch eine Normalverteilung approximieren lassen oder der Stichprobenumfang muss hinreichend groß sein.

Schätzung und Testung der Mittelwertsunterschiede für zwei unabhängige Stichproben

Auch für die Schätzung und Testung der Mittelwertsunterschiede für zwei unabhängige Stichproben haben wir die wesentlichen Grundkonzepte bereits im vorhergehenden Kapitel rekapituliert. Der einzig neu hinzukommende Aspekt ist der theoretische Befund, dass, wenn die Schätzfunktionen der beiden Populationsmittelwerte $\bar{X}_1$ und $\bar{X}_2$ jeweils durch identisch und unabhängig normalverteilte Zufallsvariablen approximiert werden können und die beiden Populationsvarianzen gleich (oder zumindest hinreichend gleich, siehe unten) sind, die Teststatistik

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{SE_{MD}}$$

einer t-Verteilung mit $\nu = n_1 + n_2 - 2$ Freiheitsgraden folgt. Hier werden mit μ_1 und μ_2 die beiden unbekannten Populationsmittelwerte, mit $SE_{MD} = \sqrt{S^2_{pool}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ die Schätzfunktion des Standardfehlers der Mittelwertdifferenz, wobei hier $S^2_{pool} = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ die Schätzfunktion der sogenannten gepoolten Varianz ist, und mit n_1 und n_2 die beiden Stichprobenumfänge bezeichnet (vgl. z.B. Wilcox, 2022, S. 171).

Für die Punktschätzung der Populationsmittelwertdifferenz sind dann wiederum lediglich die Schätzfunktionen durch die Schätzwerte aus den konkreten Stichproben zu ersetzen, d.h. der Schätzwert für $\mu_1 - \mu_2$ ist schlichtweg $\bar{x}_1 - \bar{x}_2$. Analog kann mittels den prinzipiell berechenbaren (nicht von uns aber z.B. von SPSS) Quantilen der t-Verteilung das konkrete $(1-\alpha)$-Konfidenzintervall zur Intervallschätzung der Populationsmittelwertdifferenz auf Basis der gegebenen Stichproben zu

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{1-\frac{\alpha}{2}} \cdot SE_{MD}, (\bar{x}_1 - \bar{x}_2) + t_{1-\frac{\alpha}{2}} \cdot SE_{MD}\right] = [u, o]$$

ermittelt werden.

Schließlich kann mittels der Realisation der Teststatistik T für die gegebenen Stichproben die Wahrscheinlichkeit (= p-Wert) berechnet werden (wiederum nicht von uns, sondern z.B. von SPSS), eine so extreme oder extremere Teststatistik unter der Annahme der Gültigkeit der Nullhypothese (d.h. für einen bestimmten Wert von $\mu_1 - \mu_2$, z.B. für $\mu_1 - \mu_2 = 0$) zu erhalten. Ist dieser p-Wert kleiner

oder gleich als das vorab festgelegte Signifikanzniveau, kann wiederum aufgrund desselben Plausibilitätsarguments wie im vorhergehenden Kapitel die Nullhypothese verworfen werden, da dann die Teststatistik mit Sicherheit in dem auf der Basis des Signifikanzniveaus berechneten kritischen Bereich liegt.

Um all die nötigen Berechnungen kümmert sich netterweise SPSS, solange wir nur die richtigen Eingaben tätigen und die Ausgabe richtig lesen können. Schauen wir uns daher wiederum an einem Beispiel an, wie das geht. Dazu arbeiten wir wieder mit einem fiktiven Datensatz, in dem wir uns diesmal anschauen wollen, ob (ebenfalls fiktive) Psychologiestudierende im Mittel einen höheren IQ haben als (natürlich genauso fiktive) BWL-Studierende. Der Datensatz ist in der Datei „Kap5daten2.sav“ zu finden, die Sie wieder in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden können.

Wenn wir die Datendatei in SPSS geöffnet haben, können wir unter *Analyze >> Compare Means and Proportions* diesmal die Option „Independent-Samples T Test“ auswählen. Im sich öffnenden Fenster fügen wir die Variable *IQ* in das Feld „Test Variable(s)“ ein und die Variable *Gruppe* in das Feld „Grouping Variable“. Bei der Variable Gruppe handelt es sich um eine nominalskalierte Variable, die uns sagt, welchen Studiengang ein:e spezifische:r Studierende:r gewählt hat. Dabei ist der Studiengang Psychologie mit der Zahl 0 und der Studiengang BWL mit der Zahl 1 kodiert (wer sich schon beim Öffnen der Datei einen Überblick über die Variablen und deren Eigenschaften verschafft hat, ist jetzt klar im Vorteil). Diese Zuweisung der Gruppen-Codes zu den Gruppen, deren Mittelwerte verglichen werden sollen, müssen wir jetzt noch unter „Define Groups…“ vornehmen (wir könnten nämlich auch zwei Gruppen auf Basis einer Variablen mit 3, 4, oder beliebig vielen Kategorien vergleichen). Im Menü „Define Groups“ spezifizieren wir daher „Group 1“ mit dem Wert 0 (für die Psychologiestudierenden) und „Group 2“ mit dem Wert 1 (für die BWL-Studierenden), siehe Abbildung 5.8. Danach klicken wir auf „Continue“ und fordern unter „Options“ zur Abwechslung noch breitere 99%-Konfidenzintervalle für eine näherungsweise Kompatibilität mit einem strengen Signifikanzniveau von $\alpha = .005$ (für eine gerichtete Hypothese) an. Dann wählen wir (ab SPSS Version 31) noch die Option „Homogeneity of variance test“ im Hauptmenü aus, fügen alles in die Syntax ein und führen die neuen Kommandozeilen aus. Die daraufhin erzeugte Ausgabe ist in Abbildung 5.9 dargestellt.

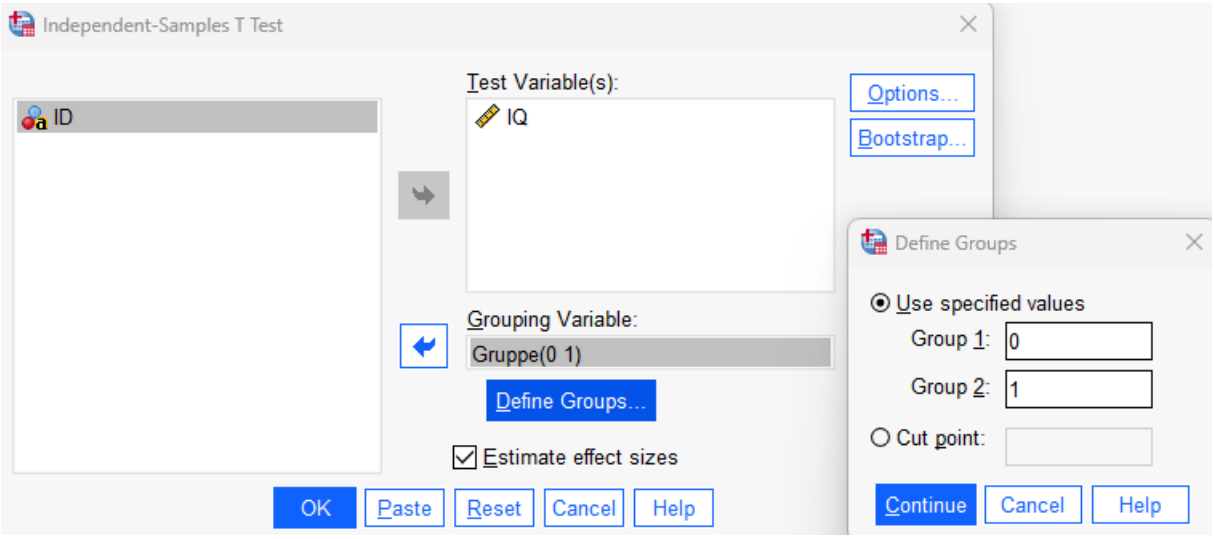


Abbildung 5.8. Spezifikation des t-Tests für unabhängige Stichproben.

In der Tabelle „Group Statistics“ finden wir dieses Mal unsere deskriptiven Statistiken. Wir sehen, dass in der Gruppe der Psychologiestudierenden der IQ von 27 Studierenden gemessen wurde und im Mittel 108.89 beträgt (mit einer Standardabweichung von 10.26). Der mittlere IQ in der Gruppe der BWL-Studierenden, die nur aus 19 Studierenden bestand, beträgt hingegen nur 99.95 (mit einer sehr ähnlichen Standardabweichung von 10.94).

In der Tabelle „Independent Samples Test“ finden wir die Ergebnisse der Punkt- und Intervallschätzung für unsere Mittelwertdifferenz sowie unseres Hypothesentests für die Nullhypothese $H_0: \mu_1 - \mu_2 = 0$. Auffällig ist dabei, dass wir in dieser Tabelle zwei Zeilen haben, in denen wir jeweils t-Werte, Freiheitsgrade, p-Werte etc. vorfinden. Dies liegt daran, dass eine oben bereits erwähnte Voraussetzung für die Berechnung der Konfidenzintervalle bzw. des Hypothesentests die Gleichheit der Populationsvarianzen war. In der Tat handelt es sich bei dem t-Test, für den diese Voraussetzung gelten muss, genauer gesagt um den sog. Student’schen t-Test (Student, 1908). Ist die Voraussetzung der Varianzgleichheit (oder Homoskedastizität) nicht erfüllt, kann weiterhin mit einer t-Verteilung gerechnet werden, solange die Freiheitsgrade entsprechend korrigiert werden. Bei dieser Art der Berechnung handelt es sich dann um den sog. t-Test nach Welch (oder kurz: Welch-Test). SPSS führt im Falle eines t-Tests für unabhängige Stichproben einfach immer beide Tests durch. Die Ergebnisse des Student’schen t-Tests befinden sich in der oberen Zeile der Tabelle „Independent Samples Test“, die Ergebnisse des Welch-Tests in der unteren. Ganz vorne in dieser Tabelle finden sich auch noch die

Ergebnisse eines sog. Levene-Tests. Dabei handelt es sich um einen statistischen Test, der die Gleichheit der Varianzen in den beiden Gruppen prüft (d.h. die Nullhypothese lautet: die Varianzen sind gleich). Ist dieser Test signifikant (üblicherweise mit α = .05), so bedeutet das, dass sich die Varianzen signifikant unterscheiden. In diesem Fall wäre dann jedenfalls die untere Zeile, also die Ergebnisse des Welch-Tests zu berichten. Ist der Levene-Test nicht signifikant, könnte man argumentieren, dass man Varianzhomogenität annehmen kann (zumindest kann man sie nicht mit Irrtumswahrscheinlichkeit α verwerfen) und daher die Ergebnisse des Student'schen t-Tests berichten. Allerdings ist es so, dass man durch die Korrektur der Freiheitsgrade in der Praxis kaum je deutlich an Teststärke verliert, sich aber gleichzeitig bezüglich Fehler 1.Art durch Berücksichtigung ungleicher Populationsvarianzen absichert. Daher (neben weiteren Gründen) argumentieren manche Autoren auch dafür, einfach immer die Ergebnisse des Welch-Tests zu berichten, unabhängig von den Ergebnissen des Levene-Tests (Ruxton, 2006; Zimmerman, 2004).

Hier in unserem Fall ist der Levene-Test nicht signifikant, F = 0.152, p = .698. Wir werden unten aber dennoch aus den oben genannten Gründen die Ergebnisse des Welch-Tests berichten. Auch Sie können daher im Rahmen der Übungen in dieser Tabelle durchwegs die Ergebnisse des Welch-Tests in der unteren Zeile verwenden. In dieser Zeile finden wir wieder einen t-Wert und die nach Welch korrigierte Anzahl an Freiheitsgraden in der Spalte „df" für „degrees of freedom". Die Anzahl der Freiheitsgrade entspricht keiner ganzen Zahl mehr, das kommt durch die Korrektur nach Welch zustande. In den nächsten beiden Spalten finden wir p-Werte für gerichtete und ungerichtete Hypothesen. In unserem Fall würden wir diesmal den p-Wert für gerichtete Hypothesen („One-Sided p") berichten, da wir untersuchen wollten, ob Psychologiestudierende einen *höheren* IQ als BWL-Studierende haben, d.h. es lag eine gerichtete Hypothese vor. Danach finden wir die Punktschätzung für die Populationsmittelwertdifferenz, deren geschätzten Standardfehler und das aus all diesen Größen konstruierte 99%-Konfidenzintervall.

In der Tabelle „Independent Samples Effect Sizes" finden wir wiederum Punktschätzungen und Konfidenzintervalle für drei verschiedene Effektstärken, von denen uns vorwiegend wiederum Cohens d interessiert. Für zusätzliche Erläuterungen zu den anderen beiden Effektstärken siehe z.B. Bühner und Ziegler (2017).

T-Test

Group Statistics

	Studiengang	N	Mean	Std. Deviation	Std. Error Mean
Intelligenzquotient	Psychologie	27	108.89	10.259	1.974
	BWL	19	99.95	10.937	2.509

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance		Mean	Std. Error	99% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Difference	Difference	Lower	Upper
Intelligenzquotient	Equal variances assumed	.152	.698	2.833	44	.003	.007	8.942	3.157	.443	17.440
	Equal variances not assumed			2.801	37.295	.004	.008	8.942	3.193	.276	17.607

Independent Samples Effect Sizes

				99% Confidence Interval	
		Standardizer[a]	Point Estimate	Lower	Upper
Intelligenzquotient	Cohen's d	10.542	.848	.039	1.650
	Hedges' correction	10.726	.834	.038	1.621
	Glass's delta	10.937	.818	-.036	1.656

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

Abbildung 5.9. Ausgabe für den t-Test für unabhängige Stichproben.

Ergebnisbericht für Schätzung und Testung der Mittelwertsunterschiede für zwei unabhängige Stichproben

Ein Bericht dieser Ergebnisse könnte wie folgt aussehen: „Die untersuchte Stichprobe von Psychologiestudierenden hatte im Mittel einen höheren IQ (M = 108.89, SD = 10.26, n = 27) als die untersuchte Stichprobe von BWL-Studierenden (M = 99.95, SD = 10.94, n = 19). Ein Welch-Test ergab, dass der Mittelwertsunterschied von 8.94 mit 99%-KI [0.28, 17.61] signifikant war (mit α = .005, gerichtet), $t(37.30) = 2.80$, p = .004, Cohens d = 0.85 mit 99%-KI [0.04, 1.65]. Gemäß Cohens Heuristiken (1988) entspricht die Punktschätzung einem großen Effekt.“ Die APA-Richtlinien für Berichte von statistischen Ergebnissen sind selbstverständlich wiederum einzuhalten.

Teststärke und Stichprobenplanung

Auch für einen t-Test für unabhängige Stichproben kann eine Stichprobenplanung mittels G*Power für eine gewünschte Teststärke bei gegebenem Signifikanzniveau und vermuteter Effektstärke vorgenommen werden. Unter „Test family“ ist dafür wiederum „t tests“ auszuwählen. Unter „Statistical test“ ist diesmal „Means: Difference between two independent means (two groups)“ auszuwählen. Unter “Type of power analysis” ist wieder “A priori: Compute required sample size – given α, power, and effect size” auszuwählen. Bei den “Input Parameters” ist wieder anzugeben, ob eine gerichtete (“one tail“) oder eine ungerichtete („two tails“) statistische Hypothese getestet werden soll. Bei der

Effektstärke ist wiederum die vermutete Effektstärke in Einheiten von Cohens d anzugeben. Das Signifikanzniveau ist wiederum unter „α err prob“ und die gewünschte Teststärke unter „Power (1-β err prob)“ anzugeben. Unter „Allocation ratio N2/N1“ ist schließlich das Verhältnis der beiden Stichprobenumfänge zueinander anzugeben, da es ja im Fall unabhängiger Stichproben – wie wir auch im vorhergehenden Beispiel gesehen haben – sein kann, dass ungleich viele Messwerte in den beiden Gruppen erhoben werden.

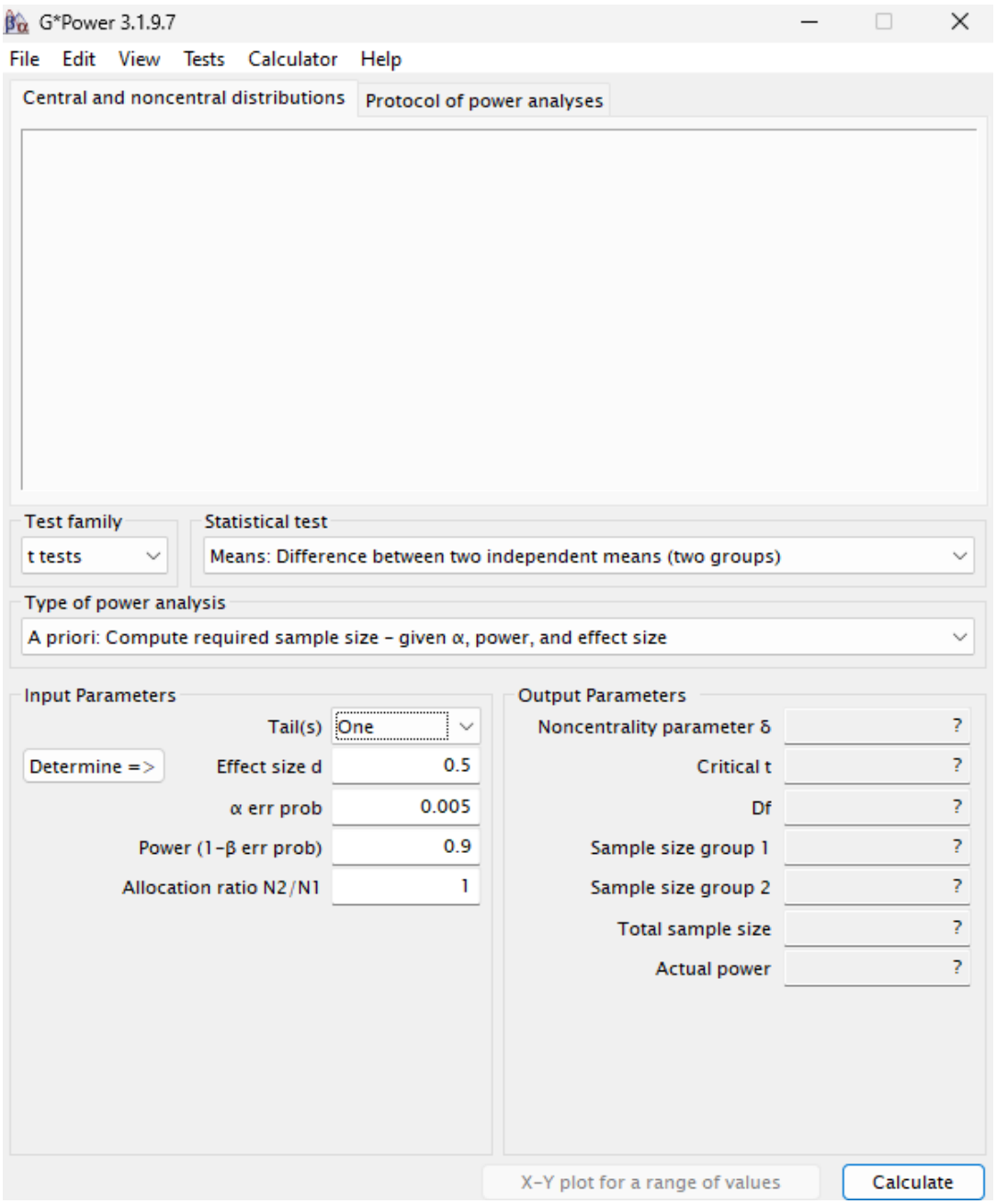


Abbildung 5.10. Für unser Beispiel nötige Eingaben in G*Power.

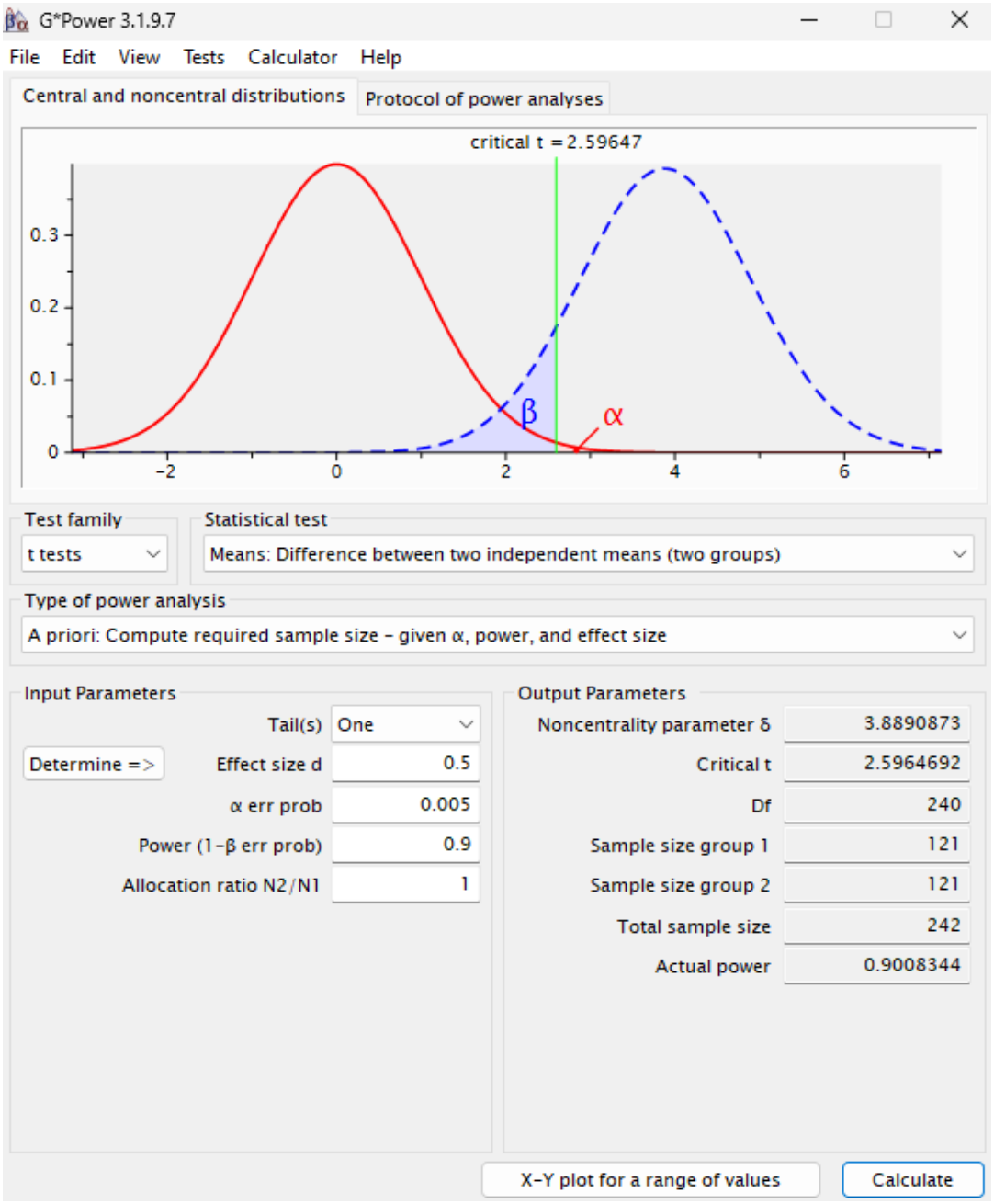


Abbildung 5.11. Nötiger Stichprobenumfang für unser (wieder einmal fiktives) Replikationsexperiment.

Zur Illustration der Stichprobenplanung nehmen wir an, dass wir das vorhergehende Beispiel replizieren wollen. Allerdings haben wir ein bisschen Zweifel an der Effektstärke, die wir vorhin erhalten haben. Die beiden Gruppen waren ja, was ihre Größe betrifft, recht überschaubar und es ist durchaus denkbar, dass wir zufällig einige sehr intelligente Leute unter den Psychologiestudierenden erwischt haben und zufällig ausgerechnet keine entsprechend intelligenten Leute unter den BWL-Studierenden. Daher möchten wir in unserem Replikationsexperiment auch eine kleinere Effektstärke von Cohens $d = 0.5$ immer noch verlässlich mit einer Teststärke von 90% detektieren können. Das

Signifikanzniveau belassen wir auf den strengen α = 0.5%. Wie in der vorhergehenden Untersuchung haben wir auch für die Replikation eine gerichtete Hypothese vorliegen. Der Einfachheit halber möchten wir allerdings Daten für gleich viele Psychologiestudierende wie für BWL-Studierende erheben, d.h. es soll $n_1/n_2 = 1$ gelten. Die diesen Vorgaben entsprechenden, für die Ermittlung des Stichprobenumfang nötigen Einstellungen in G*Power sind in Abbildung 5.10 gezeigt.

Die Ergebnisse der Berechnung sind in Abbildung 5.11 gezeigt. Für unser Replikationsexperiment benötigen wir 242 Personen (d.h. 121 in jeder Gruppe).

Voraussetzungen für den t-Test für unabhängige Stichproben

Wie oben bereits teilweise erläutert müssen auch für den t-Test für unabhängige Stichproben einige Voraussetzungen erfüllt sein, damit wir belastbare Ergebnisse erhalten. Diese Voraussetzungen sind:

- Die Varianzen beider Populationen, aus welchen die Stichproben gezogen wurden, sind nicht bekannt und müssen mittels der Stichprobendaten geschätzt werden.
- Es muss sich um unabhängige Zufallsstichproben handeln. D.h. insbesondere, es dürfen keine systematischen Abhängigkeiten in den Daten vorliegen (z.B. geclusterte Datenstruktur wie etwa Daten von Schüler:innen aus gewissen Klassen in einer Gruppe, Daten von Schüler:innen aus gewissen anderen Klassen in anderer Gruppe).
- Die Messwerte sind mindestens intervallskaliert.
- Die Messwerte sind in der jeweiligen Population normalverteilt oder es liegen hinreichend große Stichproben vor.
- (Die Varianzen in beiden Populationen sind gleich.)

Der letzte Punkt wurde eingeklammert, da dieser nur eine Voraussetzung für den Student'schen t-Test darstellt. Der Welch-Test berücksichtigt ungleiche Varianzen und wird in SPSS ohnehin immer durchgeführt. Diese Voraussetzung muss also nicht separat geprüft werden. Die Unkenntnis bezüglich der Varianzen in den Stichproben und das Intervallskalenniveau sind durch adäquate Fragestellungen bzw. Erhebungsinstrumente festgelegt, auch hier ist daher für uns nichts im Rahmen der Datenanalyse zu prüfen. Die Unabhängigkeit der Zufallsstichproben wird durch das Untersuchungsdesign festgelegt. Besteht diese Unabhängigkeit nicht, muss auf andere (hierarchische) Verfahren der Datenanalyse

zurückgegriffen werden, die allerdings hier nicht behandelt werden (für eine Einführung in solche Verfahren siehe z.B. Field, 2024).

Die vierte Voraussetzung, d.h. die Normalverteilung der Messwerte in der jeweiligen Population, ist allerdings im vorliegenden Fall von uns zu prüfen, da im vorliegenden Fall keine hinreichend großen Stichproben vorliegen. Wie man für diese Voraussetzungsprüfung verfahren kann, schauen wir uns im nächsten Abschnitt an.

Überprüfung der Normalverteilungsvoraussetzung

Für die Gültigkeit unserer Argumentation (= t-Verteilung der Teststatistik) war ausschlaggebend, dass, jedenfalls bei kleinen Stichproben, die Variablen X_{1i} und X_{2i} als identisch und unabhängig normalverteilte Zufallsvariablen modelliert (oder approximiert) werden können. D.h. diese Voraussetzung bezieht sich auf die Populationen, aus denen die jeweiligen Stichproben gezogen werden, nicht auf die Verteilung der Daten in den konkreten, vorliegenden Stichproben. D.h. auch die Überprüfung dieser Voraussetzung kann strenggenommen niemals ein definitives Ergebnis erbringen, sondern lediglich Indikatoren, die mehr oder weniger für eine Verletzung der Voraussetzung oder die Kompatibilität der vorliegenden Daten mit der Voraussetzung sprechen. Im Folgenden schauen wir uns mehrere dieser Indikatoren an.

Dazu wählen wir unter *Analyze >> Descriptive Statistics* die Option „Explore…". Dort fügen wir die Variable *IQ* im Feld „Dependent List" und die Variable *Gruppe* im Feld „Factor List" ein. Unter „Plots…" fordern wir „Normality plots with tests" an und wählen „Stem-and-Leaf" ab (per Voreinstellung leider eingestellt), siehe Abbildung 5.12.

In der resultierenden Ausgabe finden wir in der Tabelle „Tests of Normality" sog. Kolmogorov-Smirnov- sowie Shapiro-Wilk-Tests, die die Kompatibilität mit der Normalverteilungsvoraussetzung jeweils in beiden Gruppen prüfen, siehe Abbildung 5.13. Wie oben schon beim Levene-Test gilt auch hier: falls diese Tests signifikant sind (üblicherweise mit $\alpha = .05$), dann ist dies ein Hinweis auf eine Verletzung der Voraussetzung. Allerdings weisen die Tests gerade bei geringen Stichprobenumfängen wie viele statistische Verfahren keine große Teststärke auf (siehe z.B. Wilcox, 2022) und sind dadurch prädestiniert dafür gerade dann zu „versagen", wenn sie am ehesten benötigt werden (für große

Stichproben kommt uns ohnehin das zentrale Grenzwerttheorem „zu Hilfe“). Umgekehrt heißt das aber, dass wenn einer dieser Tests bei geringen Stichprobenumfängen eine signifikante Abweichung anzeigt, man wohl gut daran tut, diese Voraussetzung als verletzt zu betrachten. Sind die Tests jedoch nicht signifikant, wie hier in unserem Fall, gibt es noch weitere Indikatoren, die man inspizieren kann.

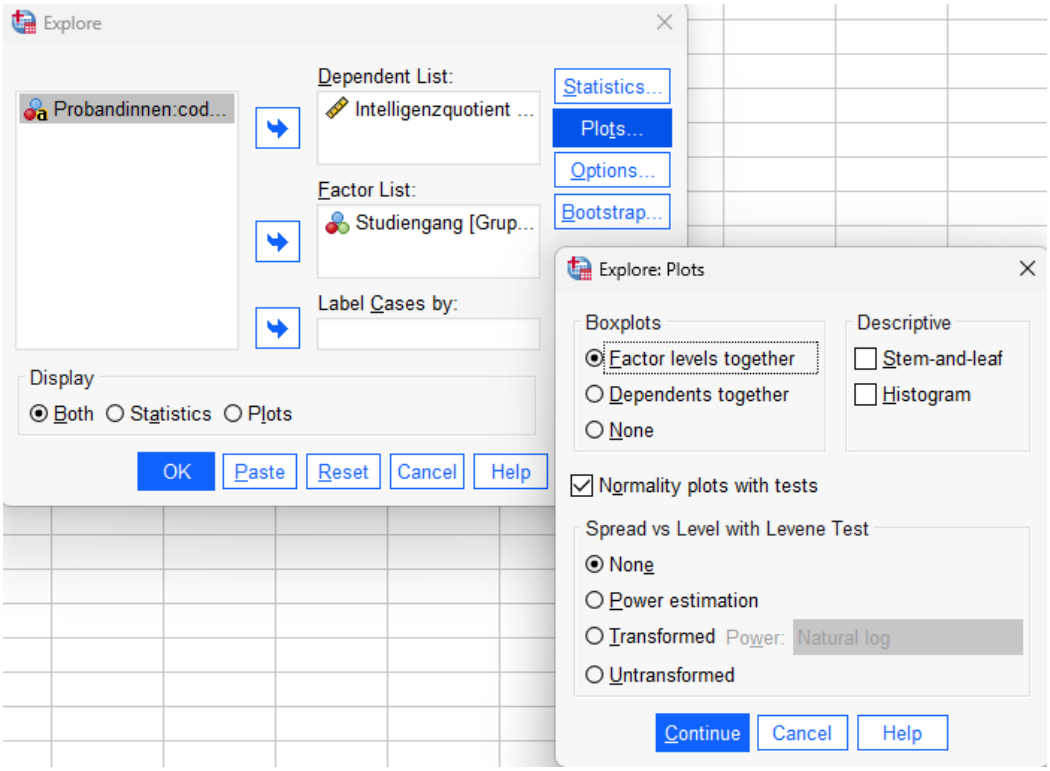


Abbildung 5.12. Anforderung von Tests auf Verträglichkeit mit Normalverteilungsvoraussetzung.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Studiengang	Statistic	df	Sig.	Statistic	df	Sig.
Intelligenzquotient	Psychologie	.129	27	.200*	.946	27	.172
	BWL	.138	19	.200*	.933	19	.195

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Abbildung 5.13. Kolmogorov-Smirnov- und Shapiro-Wilk-Tests zur Prüfung der Verträglichkeit mit der Normalverteilungsvoraussetzung.

In der Tabelle „Descriptives“ findet wir Angaben zu Schiefe (Skewness) und Wölbung (Kurtosis) sowie deren Standardfehler jeweils für beide Gruppen. In Kapitel 3 haben wir bereits kurz angedeutet, dass diese Kenngrößen auch zur Abschätzung der Verträglichkeit mit der

Normalverteilungsannahme herangezogen werden können. Dies beruht darauf, dass die Dichtefunktion der Normalverteilung sowohl eine Schiefe als auch eine Exzess-Wölbung (= was SPSS als Kurtosis ausgibt) von Null hat. Werden nun wiederholt einfache Zufallsstichproben einer normalverteilten Zufallsvariable gezogen, so sind die empirischen Schiefen und Wölbungen, die man für deren empirische Verteilungen erhält, nicht exakt gleich Null, aber mit höherer Wahrscheinlichkeit um Null herum verteilt als weit von Null entfernt. Etwas vereinfacht gesagt, kann man sagen, die meisten (präziser: etwa 95%) Werte von Schiefe und Kurtosis bei vielen solcher Zufallsstichproben, die tatsächlich aus einer Normalverteilung kommen, sollten innerhalb von zwei Standardfehlern um den Wert Null herum zu liegen kommen. Nur etwa 5% sollten weiter entfernt sein. Das kann nun verwendet werden, um grob abzuschätzen, ob die Schiefe und Kurtosis, die man im konkret vorliegenden Fall erhalten hat, einem Fall entspricht, den man nur äußerst selten erhalten würde, wenn die empirischen Verteilungen tatsächlich aus Zufallsziehungen einer normalverteilten Zufallsvariable generiert werden würden. Dazu betrachtet man einfach die Absolutwerte für Schiefe und Kurtosis für die beiden Gruppen und wenn einer der Werte mehr als zweimal so groß wie sein Standardfehler ist, dann ist das ein weiterer Indikator gegen die Verträglichkeit mit der Normalverteilungsvoraussetzung. In unserem Fall ist die Schiefe (bzw. ihr Betrag oder Absolutwert, d.h. der Wert ohne sein Vorzeichen) in der Stichprobe der Psychologie-Studierenden 0.19, was deutlich kleiner als der Standardfehler von 0.448 ausfällt. Der Wert für Kurtosis von 1.211 ist zwar größer als sein Standardfehler von 0.872, aber immer noch deutlich kleiner als das Zweifache dieses Werts. Die Werte für Schiefe (0.241) und Kurtosis (0.553) in der Stichprobe der BWL-Studierenden sind beide kleiner als ihre Standardfehler, deuten also ebenfalls nicht auf eine Verletzung der Normalverteilungsvoraussetzung hin.

Die beiden letzten Indikatoren (die wir hier zumindest kurz erläutern) sind in den beiden Grafiken mit der Überschrift „Normal Q-Q Plot of Intelligenzquotient“ zu finden. Solange sich die Punkte nahe um die schwarze Linie herum verteilen, spricht dies für die Kompatibilität mit der Normalverteilungsvoraussetzung, ansonsten dagegen. Auch hier deutet nichts auf grobe Verletzungen dieser Voraussetzung hin.

In unserem Fall scheint also die Berechnung eines t-Tests für unabhängige Stichproben in Form eines Welch-Tests zur Beantwortung unserer Fragestellung legitim. Was aber, wenn die Normalverteilungsvoraussetzung tatsächlich (grob) verletzt ist? In diesem Fall empfiehlt es sich auf Verfahren zurückzugreifen, die robuster gegenüber dieser Voraussetzung bzw. diese Voraussetzung nicht haben. Für eine Erläuterung solcher Verfahren wird an dieser Stelle allerdings auf weiterführende Literatur verwiesen (siehe z.B. Bühner & Ziegler, 2017; Wilcox, 2022; Field, 2024).

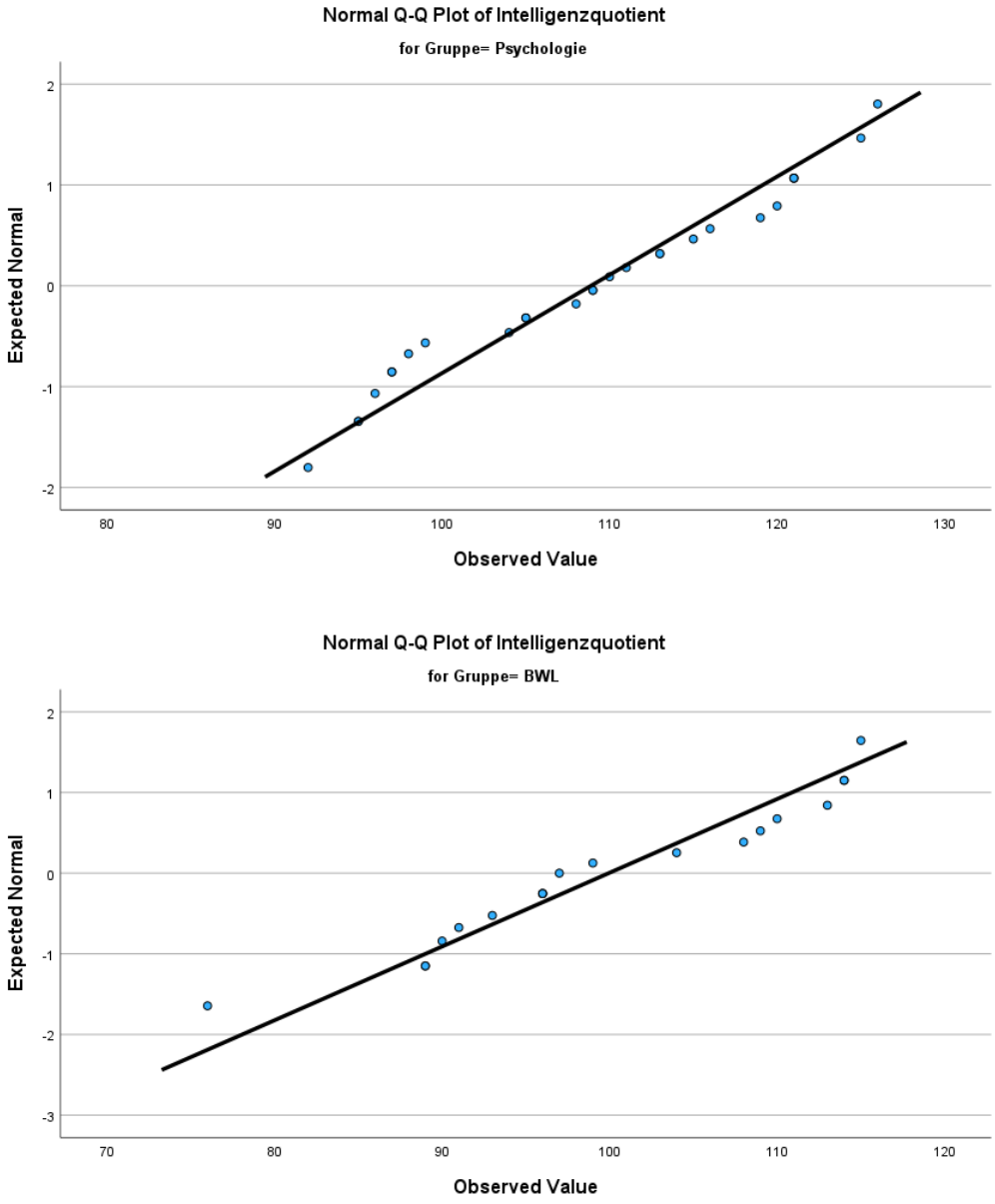


Abbildung 5.14. „Normal Q-Q Plots“ zur Einschätzung der Kompatibilität mit der Normalverteilungsvoraussetzung.

Übungsaufgaben

Die Datendateien, die Sie für manche der folgenden Übungsaufgaben benötigen, finden Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Beispiel 5.1

Welche Voraussetzungen müssen für einen t-Test für abhängige Stichproben erfüllt sein?

(a) Die gemessenen Variablen müssen mindestens intervallskaliert sein.

(b) Die Varianzen der gemessenen Variablen müssen bekannt sein.

(c) Die gemessenen Variablen müssen unabhängig voneinander sein.

(d) In kleinen Stichproben müssen die gemessenen Variablen jeweils normalverteilt sein bzw. durch normalverteilte Zufallsvariablen approximiert werden können.

Beispiel 5.2

Welche Voraussetzungen müssen für einen t-Test für unabhängige Stichproben erfüllt sein?

(a) Soll ein Student'scher t-Test gerechnet werden, müssen die Varianzen in beiden Populationen gleich sein.

(b) Soll ein t-Test nach Welch gerechnet werden, müssen die Varianzen in beiden Populationen gleich sein.

(c) Die gemessene Variable muss in beiden Gruppen normalverteilt sein bzw. durch eine normalverteilte Zufallsvariable approximiert werden können, sofern die Stichproben in beiden Gruppen nicht hinreichend groß sind.

(d) Die gemessene Variable muss mindestens nominalskaliert sein.

Beispiel 5.3

Welche Aussage(n) trifft(treffen) zu?

(a) Mit dem Levene-Test kann die Normalverteilungsvoraussetzung geprüft werden.

(b) Der Shapiro-Wilk-Test überprüft Homoskedastizität.

(c) Eine Normalverteilung hat eine Schiefe von 3.

(d) Ist ein Kolmogorov-Smirnov-Test signifikant, dann ist das ein Hinweis auf eine Verletzung der Normalverteilungsvoraussetzung.

Beispiel 5.4

In der Datei „Kap5UE4.sav“ finden Sie einen Datensatz für 67 (fiktive) psychiatrische Patient:innen, deren depressive Symptomatik vor und nach einer Behandlung mit Becks Depressionsinventar untersucht wurde, was für jede:n Patient:in je einen Messwert zwischen 0 und 63 ergibt. Die Variable *BDI_t1* beinhaltet die BDI-Werte zu Testzeitpunkt 1 und die Variable *BDI_t2* die BDI-Werte zu Testzeitpunkt 2. Es soll mit einem Signifikanzniveau von $\alpha = .005$ statistisch geprüft werden, ob die BDI-Werte der untersuchten Patient:innen von Zeitpunkt 1 zu Zeitpunkt 2 im Mittel abnehmen. Erstellen Sie für Ihre Ergebnisse einen mit den APA-Richtlinien konformen Ergebnisbericht.

Der Datensatz für dieses Beispiel beruht auf dem Datensatz für ein entsprechendes Beispiel bei Bühner und Ziegler (2017, S. 305-307). Aufgrund leicht abgeänderter Messwerte stimmen allerdings die sich ergebenden Resultate nicht exakt überein.

Beispiel 5.5

Uns interessiert, ob die Studierenden im Kurs „Anwendung statistischer Verfahren am Computer“ den Aussagen „Ich habe Angst vor der nächsten Statistikprüfung“ (Variable *mathe_mathe3*) und „Ich hasse Statistik“ (Variable *mathe_mathe2*) unterschiedlich stark zustimmen. Sie finden die entsprechenden Daten in der Datei „Kap3daten.sav“. Wählen Sie einen passenden statistischen Test, um die Fragestellung zu erhellen. Berichten Sie Ihre Ergebnisse gemäß APA Richtlinien.

Beispiel 5.6

Männer sind bekanntlich im Mittel größer als Frauen. Trifft dies auch auf Studierende im Kurs „Anwendung statistischer Verfahren am Computer“ zu? Sie finden die entsprechenden Daten in der Datei „Kap3daten.sav“. Wählen Sie einen passenden statistischen Test, um die Fragestellung zu erhellen. Berichten Sie Ihre Ergebnisse gemäß APA Richtlinien.

Beispiel 5.7

Männer haben bekanntlich größere Füße als Frauen. Trifft dies auch auf Studierende im Kurs „Anwendung statistischer Verfahren am Computer“ zu? Sie finden die entsprechenden Daten in der Datei „Kap3daten.sav“. Wählen Sie einen passenden statistischen Test und eine passende Variable, um die Fragestellung zu erhellen. Berichten Sie Ihre Ergebnisse gemäß APA Richtlinien.

Sie können für diese Übung davon ausgehen, dass die Voraussetzungen für einen t-Test für unabhängige Stichproben erfüllt sind.

Beispiel 5.8

Überprüfen Sie die Voraussetzungen für einen t-Test für unabhängige Stichproben für das vorhergehende Beispiel.

Beispiel 5.9

Ein Medikament werde als praktisch effektiver angesehen, wenn es eine um mindestens Cohens $d = 0.2$ höhere Wirkung als ein Plazebo hat. Aus Pilotstudien ist bekannt, dass die Varianzen für Experimental- und Plazebobedingungen sehr ähnlich sind. Das Verfahren, um die Wirksamkeit des Medikaments zu messen, liefert eine metrische Variable. Wie groß muss die Stichprobe sein, um bei einer Irrtumswahrscheinlichkeit von $\alpha = 0.005$ für einen Effekt der Stärke $d = 0.2$ in 80% der Fälle (= Teststärke) einen signifikanten Unterschied zwischen den Mittelwerten der Experimental- und Plazebogruppe zu finden? Nehmen Sie zur Beantwortung dieser Frage an, dass es sich bei Experimental- und Plazebogruppe jeweils um einfache, voneinander unabhängige Zufallsstichproben handelt.

Beispiel 5.10

Aus einer Pilotstudie ergibt sich eine Wirksamkeit für eine Therapie von Cohens $d = 0.95$. Dabei handelt es sich um die Verringerung depressiver Symptomatik durch die Therapie bei am Pilotexperiment

teilnehmenden Personen. Die depressive Symptomatik wurde jeweils vor und nach der Therapie erhoben. Mit einem Prüfexperiment soll dieses vielversprechende Ergebnis nun statistisch abgesichert werden.

Verwenden Sie G*Power, um die nötige Stichprobengröße für eine Teststärke von 80% und eine Irrtumswahrscheinlichkeit von 0.5% zu ermitteln.

Beispiel 5.11

Ein Forscher möchte wissen, ob es einen Unterschied im Angstniveau zwischen zwei bestimmten Personengruppen gibt. Daher rekrutiert er aus beiden betreffenden Populationen (im Folgenden schlicht als Gruppe 1 und 2 bezeichnet) Personen, die einen entsprechenden Fragebogen ausfüllen. Anschließend untersucht er die erhobenen Daten mit SPSS und erhält die in Abbildung 5.15 gezeigte Ausgabe.

Wie würden Sie diese Ergebnisse (in 2-3 Sätzen) für die Wahl eines Signifikanzniveaus von α = .05 berichten? Wie groß ist die Effektstärke gemäß Cohens Heuristik (1988)?

Group Statistics

	Gruppe (1 oder 2)	N	Mean	Std. Deviation	Std. Error Mean
Angstniveau (metrische Variable: Zahl von 0 bis 10 bzw. wenig bis viel Angst)	1	60	5.223	1.0588	.1367
	2	30	5.660	.9804	.1790

Independent Samples Test

		Levene's Test for Equality of Variances		t-test			
						Significance	
		F	Sig.	t	df	One-Sided p	Two-Sided p
Angstniveau (metrische Variable: Zahl von 0 bis 10 bzw. wenig bis viel Angst)	Equal variances assumed	.863	.356	-1.889	88	.031	.062
	Equal variances not assumed			-1.939	62.280	.029	.057

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval	
				Lower	Upper
Angstniveau (metrische Variable: Zahl von 0 bis 10 bzw. wenig bis viel Angst)	Cohen's d	1.0336	-.422	-.864	.021
	Hedges' correction	1.0426	-.419	-.857	.021
	Glass's delta	.9804	-.445	-.895	.011

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

Abbildung 5.15. Ausgabe für die Fragestellung aus Beispiel 5.11.

Beispiel 5.12

Ein Forscher möchte wissen, ob es einen Unterschied im Depressionsniveau zwischen zwei bestimmten Personengruppen gibt. Daher rekrutiert er aus beiden betreffenden Populationen (im Folgenden schlicht als Gruppe 1 und 2 bezeichnet) Personen, die einen entsprechenden Fragebogen ausfüllen. Anschließend untersucht er die erhobenen Daten mit SPSS und erhält die in Abbildung 5.16 gezeigte Ausgabe.

Wie würden Sie diese Ergebnisse (in 2-3 Sätzen) für die Wahl eines Signifikanzniveaus von $\alpha = .005$ berichten? Wie groß ist die Effektstärke gemäß Cohens Heuristik (1988)?

Group Statistics

	Gruppe (1 oder 2)	N	Mean	Std. Deviation	Std. Error Mean
Depressionsniveau (metrische Variable: Zahl von 0 bis 10 bzw. wenig bis viel Depression)	1	70	4.907	1.0465	.1251
	2	30	5.250	.7780	.1420

Independent Samples Test

		Levene's Test for Equality of Variances		t-test			
						Significance	
		F	Sig.	t	df	One-Sided p	Two-Sided p
Depressionsniveau (metrische Variable: Zahl von 0 bis 10 bzw. wenig bis viel Depression)	Equal variances assumed	4.440	.038	-1.612	98	.055	.110
	Equal variances not assumed			-1.811	72.968	.037	.074

Independent Samples Effect Sizes

				95% Confidence Interval	
		Standardizer[a]	Point Estimate	Lower	Upper
Depressionsniveau (metrische Variable: Zahl von 0 bis 10 bzw. wenig bis viel Depression)	Cohen's d	.9748	-.352	-.781	.080
	Hedges' correction	.9823	-.349	-.775	.079
	Glass's delta	.7780	-.441	-.879	.005

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

Abbildung 5.16. Ausgabe für die Fragestellung aus Beispiel 5.12.

Beispiel 5.13

Eine Forscherin möchte wissen, ob die Konzentrationsfähigkeit von Schüler:innen mit ADHS durch eine bestimmte Intervention erhöht werden kann. Dazu erhebt sie die Konzentrationsfähigkeit von 73 Schüler:innen mit ADHS vor und nach der Intervention. Anschließend untersucht sie die erhobenen Daten mit SPSS und erhält die in Abbildung 5.17 gezeigte Ausgabe.

Wie würden Sie diese Ergebnisse (in 2-3 Sätzen) für die Wahl eines Signifikanzniveaus von α = .04 berichten? Wie groß ist die Effektstärke gemäß Cohens Heuristik (1988)?

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Kozentrationsfähigkeit vor Intervention (Skala von 0 bis 100)	49.88	73	14.859	1.739
	Kozentrationsfähigkeit nach Intervention (Skala von 0 bis 100)	56.30	73	29.333	3.433

Paired Samples Test

		Paired Differences					t	df	Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			One-Sided p	Two-Sided p
Pair 1	Kozentrationsfähigkeit vor Intervention (Skala von 0 bis 100) - Kozentrationsfähigkeit nach Intervention (Skala von 0 bis 100)	-6.425	28.337	3.317	-13.036	.187	-1.937	72	.028	.057

Paired Samples Effect Sizes

			Standardizer[a]	Point Estimate	95% Confidence Interval	
					Lower	Upper
Pair 1	Kozentrationsfähigkeit vor Intervention (Skala von 0 bis 100) - Kozentrationsfähigkeit nach Intervention (Skala von 0 bis 100)	Cohen's d	28.337	-.227	-.458	.006
		Hedges' correction	28.636	-.224	-.454	.006

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference.
Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

Abbildung 5.17. Ausgabe für die Fragestellung aus Beispiel 5.13.

Beispiel 5.14

Ein Forscher möchte untersuchen, ob sich die wiederholte Beschäftigung mit Spielen, bei denen räumliche Rätsel gelöst werden müssen, auf die Fähigkeit der mentalen Rotation auswirkt. Für eine entsprechende Studie rekrutiert er daher 93 Schulkinder, die sich noch nie zuvor mit solchen Spielen befasst haben. Zu einem Zeitpunkt 1 erfasst er deren Fähigkeit zur mentalen Rotation auf einer Skala von 0 bis 100 mittels eines geeigneten Testverfahrens. Danach sollen sich die Kinder innerhalb von zwei Wochen insgesamt 4 Stunden mit dem Lösen räumlicher Rätselaufgaben befassen. Nach Ablauf der beiden Wochen, zu Zeitpunkt 2, wird wiederum die Fähigkeit zur mentalen Rotation erhoben. Die Daten zu diesem Experiment sind in der Datei *Kap5UE14.sav* zu finden. Ermitteln Sie mittels eines geeigneten statistischen Verfahrens, ob sich die Fähigkeit zur mentalen Rotation zwischen den beiden Zeitpunkten im Mittel unterscheidet und berichten Sie Ihre Resultate gemäß APA-Richtlinien.

Beispiel 5.15

Eine Forscherin möchte untersuchen, ob textuelle Informationen in digitalen Lernspielumgebungen leichter verarbeitet werden können, wenn diese schriftlich dargestellt oder gesprochen werden. Um diese Fragestellung zu untersuchen, rekrutiert die Forscherin 172 Versuchspersonen und weist diese randomisiert entweder der Gruppe „Schrift“ oder der Gruppe „Sprache“ zu. In der Gruppe „Schrift“ werden lernspielrelevante Texte am Bildschirm schriftlich dargestellt. In der Gruppe „Sprache“ werden dieselben Informationen von einem professionellen Sprecher eingesprochen und dann durch entsprechende Sprachaufzeichnungen im Lernspiel vermittelt. Einen Tag, nachdem sich die Versuchspersonen mit dem Lernspiel befasst haben, absolvieren sie einen Test zu den Inhalten des Lernspiels, bei dem Sie zwischen 0 und 100 Punkte erreichen können. Die Testergebnisse und Gruppenzugehörigkeiten sind in der Datei *Kap5UE15.sav* zu finden. Ermitteln Sie mittels eines geeigneten statistischen Verfahrens, ob sich die beiden Gruppen hinsichtlich der Testergebnisse im Mittel unterscheiden und berichten Sie Ihre Resultate.

Beispiel 5.16

Forscher:innen haben eine Studie in einer großen Firma durchgeführt. Sie untersuchten die Coping-Fähigkeiten der Angestellten dieser Firma, um herauszufinden, ob sich Personen mit und ohne Burnout hinsichtlich ihrer Coping-Fähigkeit unterscheiden und wie gestresst sich die Angestellten am Arbeitsplatz und in ihrem Privatleben fühlen. Untersuchen Sie anhand der Datendatei „Kap5UE16.sav“ die folgenden Fragestellungen mittels SPSS und berichten Sie die Ergebnisse Ihrer Berechnungen. Berichten Sie bei statistischen Ergebnissen immer alle relevanten Kennwerte (Mittelwerte und Standardabweichungen, Teststatistiken, Freiheitsgrade, p-Werte, Effektstärken). Das Signifikanzniveau soll für alle statistischen Tests zu 0.005 gewählten werden.

(a) Unterscheiden sich Personen mit und ohne Burnout (Variable: *burnout*) hinsichtlich ihres mittleren Stressempfindens am Arbeitsplatz (Variable: *stress_arbeit*)? Falls ja, bei welchen Personen ist das Stressempfinden höher?

(b) Unterscheidet sich das mittlere Ausmaß von Stress am Arbeitsplatz (Variable: *stress_arbeit*) von dem von Stress im Privatleben (Variable: *stress_privat*)? Falls ja, welcher Stress fällt höher aus?

Beispiel 5.17

Ein neues Schmerzmedikament soll erprobt werden. Dazu werden unterschiedliche Experimente durchgeführt. In einem Experiment wird das Schmerzmedikament an 200 Personen mit einer chronischen Schmerzerkrankung ausgegeben. Für jede Person wird die Schmerzintensität vor der Einnahme des Medikaments auf einer kontinuierlichen Skala von 0 bis 10 erfasst. Im Anschluss werden die Personen gebeten das Medikament eine Woche wie empfohlen einzunehmen. Danach wird wiederum die Schmerzintensität erfasst. Die Daten sind in der Datei „Kap5UE17.sav“ gegeben. Wählen Sie ein geeignetes statistisches Verfahren, um die Frage zu erhellen, ob die Einnahme des Medikaments im Mittel die Schmerzen der Personen lindert. Erstellen Sie anschließend einen entsprechenden Ergebnisbericht.

Beispiel 5.18

Sauer macht bekanntlich lustig. Allerdings geht diese Redewendung auf das altdeutsche Wort „gelüstig“ zurück und bezieht sich eher auf die geschmacksverstärkende Wirkung von Säure als auf Spaß und Gelächter.

Um die Redewendung im Zusammenhang mit dem Geschmack von Speisen zu untersuchen hat eine Forscherin daher 150 Versuchspersonen rekrutiert. Eine Hälfte der Personen bekam ein Glas Zitronensaft zu trinken, die andere Hälfte stattdessen Wasser. Danach wurde allen Personen dasselbe dreigängige Menü serviert, das schließlich von jeder Person auf einer kontinuierlichen Skala von 0 (= „schrecklich“) bis 10 (= „vorzüglich“) zu bewerten war. Die entsprechenden Daten sind in der Datei „Kap5UE18.sav“ zu finden.

Wählen Sie ein geeignetes statistisches Verfahren, um die Frage zu erhellen, ob die Personen, die den Zitronensaft zu trinken bekamen, das Menü im Mittel besser bewerteten als die Personen, die stattdessen Wasser zu trinken bekamen. Erstellen Sie anschließend einen entsprechenden Ergebnisbericht.

Beispiel 5.19

Eine Lehrerin möchte wissen, ob sich eine kurze Aktivierungsübung positiv auf die Lernmotivation im Mathematikunterricht auswirkt. Dazu erhebt sie die Lernmotivation ihrer Schüler:innen mit einem Kurzfragebogen während des Mathematikunterrichts jeweils vor und nach der Aktivierungsübung.

Der Kurzfragebogen umfasst drei Items, die jeweils auf einer Skala von 1 (= „trifft gar nicht zu“) bis 4 („trifft voll und ganz zu“) beantwortet werden. Höhere Werte bedeuten höhere Lernmotivation. Aus diesen drei Items ist schließlich eine Mittelwertskala zu bilden, um die Lernmotivation zu erfassen.

Die von der Lehrerin erhobenen Daten sind in der Datei „Kap5UE19.sav“ gegeben. Wählen Sie ein geeignetes statistisches Verfahren, um die Frage der Lehrerin zu erhellen, ob sich eine kurze Aktivierungsübung im Mittel positiv auf die Lernmotivation im Mathematikunterricht auswirkt. Erstellen Sie anschließend einen entsprechenden Ergebnisbericht.

Beispiel 5.20

Eine Therapeutin möchte wissen, ob sich eine kurze Atemübung positiv auf das allgemeine Entspannungsniveau von Klient:innen auswirkt. Dazu erhebt sie das allgemeine Entspannungsniveau von 60 ihrer Klient:innen mit einem Kurzfragebogen jeweils vor und nach der Atemübung.

Die von der Therapeutin erhobenen Daten sind in der Datei „Kap5UE20.sav“ gegeben. Wählen Sie ein geeignetes statistisches Verfahren, um die Frage der Therapeutin zu erhellen, ob sich eine kurze Atemübung im Mittel positiv auf das allgemeine Entspannungsniveau auswirkt. Erstellen Sie anschließend einen entsprechenden Ergebnisbericht.

Beispiel 5.21

Wie viele Personen muss eine Gesamtstichprobe umfassen, damit bei Aufteilung in zwei gleich große Gruppen und einem Signifikanzniveau $\alpha = .01$ ein Unterschied zwischen beiden Gruppenmittelwerten (unabhängige Stichproben) der Stärke Cohens $d = 0.5$ mit einer Teststärke (= power) von 90% detektiert werden kann? Fügen Sie für Ihre Antwort auch einen Screenshot Ihrer Berechnung des Stichprobenumfangs mit G*Power ein.

Beispiel 5.22

Wie viele Personen muss eine Gesamtstichprobe umfassen, damit bei Aufteilung in zwei gleich große Gruppen und einem Signifikanzniveau $\alpha = .005$ ein Unterschied zwischen beiden Gruppenmittelwerten (unabhängige Stichproben) der Stärke Cohens $d = 0.4$ mit einer Teststärke (= power) von 80% detektiert werden kann? Fügen Sie für Ihre Antwort auch einen Screenshot Ihrer Berechnung des Stichprobenumfangs mit G*Power ein.

Beispiel 5.23

Wie viele Personen muss eine Stichprobe umfassen, damit ein Unterschied zwischen zwei abhängigen Variablen der Stärke Cohens $d = 0.2$ für ein Signifikanzniveau $\alpha = .005$ mit einer Teststärke (= power) von 80% detektiert werden kann? Fügen Sie für Ihre Antwort auch einen Screenshot Ihrer Berechnung des Stichprobenumfangs mit G*Power ein.

Beispiel 5.24

Tun Sie sich für diese Übungsaufgabe mit einem:einer Kolleg:in zusammen. Erstellen Sie jeweils unabhängig voneinander jeweils drei Ergebnisberichte für drei beliebige aus den folgenden Übungsaufgaben: 5.4-5.6, 5.14-5.15, 5.17-5.20. Überprüfen Sie danach jeweils selbst die Korrektheit Ihrer Ergebnisberichte mit den am Ende dieses Dokuments bereitgestellten Lösungen. Fügen Sie anschließend in jeden Ihrer Ergebnisbericht 5 Fehler ein, ohne sie Ihrem:Ihrer Kolleg:in mitzuteilen (und es dürfen durchaus Fehler sein, die nur schwer zu entdecken sind). Tauschen Sie anschließend Ihre fehlerhaften Ergebnisberichte aus. Versuchen Sie nun jeweils die Fehler zu identifizieren und zu korrigieren, ohne dabei auf Musterlösungen zurückzugreifen. Für die korrekte Identifikation eines Fehlers gibt es einen Punkt, für die korrekte Korrektur eines Fehlers einen weiteren Punkt. D.h. Sie können beide jeweils maximal 30 Punkte erreichen. Wer mehr Punkte erreicht gewinnt!

Beispiel 5.25

Reflektieren Sie schriftlich: Welche Voraussetzungen müssen für einen t-Test für abhängige Stichproben erfüllt sein? Wie können Sie die Gültigkeit dieser Voraussetzungen prüfen? Welche Konsequenzen hat es, wenn die Voraussetzungen nicht erfüllt sind?

Kapitel 6

Einfaktorielle Varianzanalyse ohne Messwiederholung

Stefan E. Huber

In den verbleibenden Kapiteln dieses Dokuments werden wir uns mit varianz- und regressionsanalytischen Verfahren befassen. In beiden Fällen wird die methodische Sichtweise auf die entsprechenden Fragestellungen etwas anders gelagert sein als für die statistischen Verfahren, die wir bisher behandelt haben (wobei es grundsätzlich möglich ist, auch die Vergleiche von zwei Gruppenmittelwerten sowie eines Mittelwerts unter die allgemeine Gruppe linearer Verfahren zu subsumieren). Konkret heißt dies, das für die verbleibenden Kapitel die folgende (experimentelle- bzw. quasi-experimentelle wie etwa in dem im nächsten Absatz beschriebenen Beispiel) Perspektive auf die Problemstellungen eingenommen wird: ein:e Forscher:in variiert eine unabhängige Variable (UV) und registriert Veränderungen in einer abhängigen Variable (AV). Bei der unabhängigen Variablen kann es sich um eine diskrete oder eine kontinuierliche Variable handeln. Für diskrete UVn ist es nach wie vor häufig üblich, deren Auswirkungen auf die entsprechende AV mit varianzanalytischen Modellen zu untersuchen (auch wenn diese einen Spezialfall regressionsanalytischer Modelle darstellen). Da die varianzanalytischen Modelle aber einen relativ einfachen Einstieg in allgemeinere statistische Modelle erlauben, wird dieser Spezialfall hier zuerst behandelt. In den Kapiteln 9-11 werden wir uns schließlich mit dem allgemeineren regressionsanalytischen Zugang befassen und am Ende (genauer: in Kapitel 11) auf regressionsanalytische Modelle mit diskreten UVn (auch Prädiktoren genannt) zurückkommen.

Als Beispiel für dieses Kapitel werden wir die folgende Fragestellung betrachten: eine Forscherin erhebt das Merkmal Depression mit einem entsprechenden Fragebogen. Der Wert, die eine Person in diesem Fragebogen für das allgemeine Depressionsniveau erzielt, ist die Ausprägung der AV für diese Person. Die Forscherin erhebt das Depressionsniveau für drei unterschiedliche Gruppen: junge Erwachsene, Erwachsene mittleren Alters sowie ältere Erwachsene. D.h. die Forscherin zieht einfache Zufallsstichproben aus diesen mit entsprechenden Altersgrenzen begrenzten Populationen. Die jeweilige Altersgruppe entspricht also der UV, die die Forscherin dadurch systematisch variiert, indem sie systematisch aus diesen Zielgruppen einfache Zufallsstichproben zieht (sie variiert also nicht das

Alter systematisch, was schwer möglich wäre, sondern die Wahl der Altersgruppe). Ob das Depressionsniveau von der jeweiligen Altersgruppe abhängt, soll mithilfe der Gruppenmittelwerte beantwortet werden. Sind die Gruppenmittelwerte für das Depressionsniveau unterschiedlich, dann wird geschlossen, dass das Depressionsniveau von der Altersgruppe abhängt (d.h. nicht, dass ein Kausalzusammenhang zwischen den beiden Größen besteht, sondern nur, dass das typische Depressionsniveau eines älteren Erwachsenen typischerweise ein anderes ist als das eines jungen Erwachsenen).

Alle anderen möglichen Merkmale, die eine Rolle für das Depressionsniveau spielen könnten, werden erstmal ausgeklammert. Das heißt, es wird lediglich eine UV untersucht. Die UVn in varianzanalytischen Modellen werden auch als Faktoren bezeichnet. D.h., es handelt sich hier um eine einfaktorielle Varianzanalyse, da es genau *einen* Faktor gibt. Da drei Altersgruppen untersucht werden, handelt es sich um einen Faktor mit drei Stufen. D.h., im vorliegenden Fall kann man auch von einer einfaktoriellen dreistufigen Varianzanalyse sprechen. Der eine Faktor ist die Altersgruppe. Dessen Stufen sind: junge Erwachsene (Stufe 1), mittelalte Erwachsene (Stufe 2), ältere Erwachsene (Stufe 3).

Zum besseren Verständnis betrachten wir kurz noch einen zweiten Fall. Angenommen ein anderer Forscher würde die Auswirkung der Altersgruppe sowie vorliegender psychiatrischer Vorerkrankungen auf das Depressionsniveau untersuchen. Für die Altersgruppe würde er wieder dieselben Gruppen wie oben betrachten. Für das Vorliegen psychiatrischer Vorerkrankungen würde er zwischen „Person hat psychiatrische Vorerkrankungen“ und „Person hat keine psychiatrischen Vorerkrankungen“ unterscheiden. Die AV wäre in diesem Fall wieder das Depressionsniveau. Allerdings würden in diesem Fall nun zwei UVn bzw. Faktoren vorliegen, die Altersgruppe und das Vorliegen psychiatrischer Vorerkrankungen. Die Altersgruppe hätte wiederum drei Stufen. Das Vorliegen psychiatrischer Vorerkrankungen hätte zwei Stufen. In diesem Fall würde also eine *zwei*faktorielle Varianzanalyse mit einem 3 x 2 Design vorliegen. Die Angabe „3 x 2 Design“ bedeutet also, dass ein 3-stufiger und ein 2-stufiger Faktor vorliegt, insgesamt hat man es also mit 6 Populationen zu tun. Mit solchen mehrfaktoriellen Varianzanalysen werden wir uns im nächsten Kapitel befassen.

Zudem gibt es noch Varianzanalysen mit Messwiederholung. Bei diesen wird das interessierende Merkmal beispielsweise zu mehreren Messzeitpunkten erhoben (es können prinzipiell aber auch andere als zeitliche Abhängigkeiten zwischen den entsprechenden Variablen bestehen). Das Kennzeichen von Varianzanalysen mit Messwiederholung ist also (häufig) das wiederholte Vorliegen eines interessierenden Merkmals für jede Person bzw. jeden Fall. In den vorhergehenden Fällen tauchte jede Person nur einmal mit einem bestimmten Depressionsniveau auf. In einer Varianzanalyse mit Messwiederholung würde in einem entsprechenden Beispiel das Depressionsniveau für eine Person öfter (etwa zu Beginn und nach Ende einer Therapie) auftauchen. Mit Varianzanalysen mit Messwiederholung werden wir uns in Kapitel 8 befassen.

Schließlich kann ein Design einer Varianzanalyse balanciert oder unbalanciert sein. In einem balancierten Design ist der Umfang der Stichproben aus allen betreffenden Populationen gleich groß, in einem unbalancierten Design liegen unterschiedlich große Stichproben vor.

Varianzanalytisches Modell für eine einfaktorielle Varianzanalyse ohne Messwiederholung

Im varianzanalytischen Modell für eine einfaktorielle Varianzanalyse ohne Messwiederholung werden die Ausprägungen der AV für die einzelnen Personen in den unterschiedlichen Gruppen als Realisationen von identisch und unabhängig normalverteilten Zufallsvariablen aufgefasst. Diese Zufallsvariablen werden üblicherweise mit Y bezeichnet (anstatt wie für die bisherigen statistischen Verfahren mit X). Die zu diesen Zufallsvariablen gehörigen Normalverteilungen haben als Erwartungswert jeweils den einzelnen Gruppenmittelwert der AV, der um einen bestimmten Betrag $\Delta\mu_j$ vom Gesamtmittelwert μ (über alle Gruppen) abweicht (das Symbol Δ wird als „delta“ ausgesprochen und zeigt den Unterschied zwischen zwei Werten an, hier zwischen dem Gesamt- und dem jeweiligen Gruppenmittelwert), und alle dieselbe Varianz σ^2, d.h. das statistische Modell lautet

$$Y_{ij} \sim N\left(\mu + \Delta\mu_j, \sigma^2\right).$$

In Worten bedeutet dies: Die AV der i-ten Person in Gruppe j entspricht einer Zufallsvariable, die einer Normalverteilung mit Erwartungswert $\mu + \Delta\mu_j$ und Varianz σ^2 folgt. Da dies für alle Gruppen und alle Personen gleichermaßen gilt, haben insbesondere alle Zufallsvariablen dieselbe Varianz, d.h., es liegt Homoskedastizität vor. Zudem sind alle Zufallsvariablen normalverteilt. Noch dazu ergibt es nur Sinn

mindestens intervallskalierte Variablen durch solche Zufallsvariablen zu modellieren. Schließlich gilt dies für alle Gruppen und Personen unabhängig voneinander (sonst müsste die Abhängigkeit ja irgendwie im statistischen Modell ausgewiesen werden). Zusammengenommen bedeutet das, dass alle auf diesem Modell basierenden Analysen nur dann hinsichtlich der Irrtumswahrscheinlichkeit α interpretierbare Ergebnisse liefern werden, wenn diese Modellannahmen für die entsprechende Fragestellung als gültig angenommen werden können (oder zumindest in guter Näherung als gültig angenommen werden können). Darauf werden wir unten im Rahmen der Voraussetzungsprüfungen für einfaktorielle Varianzanalysen ohne Messwiederholung noch zurückkommen.

Für den Moment nehmen wir einmal an, alle diese Voraussetzungen seien erfüllt. Wie lässt sich dann prüfen, ob sich die Populationsmittelwerte der einzelnen Gruppen voneinander unterscheiden? Bis auf eine prinzipiell unhintergehbare statistische Unsicherheit (die wieder durch die Irrtumswahrscheinlichkeit zum Ausdruck kommen wird), lässt sich diese Frage mit einem sog. Omnibustest für die Gleichheit aller Populationsmittelwerte, d.h. $\Delta\mu_j = 0$ für alle $j = 1, \ldots, m$ mit m der Anzahl der betrachteten Gruppen bzw. Populationen, prüfen. Diesen Omnibustest rekapitulieren wir kurz im nächsten Abschnitt.

Omnibustest für ein einfaktorielles varianzanalytisches Modell

Vorweg: Bei diesem Omnibustest handelt es sich um die einfaktorielle Varianzanalyse ohne Messwiederholung. Letzteres ist also der Name für das Verfahren, das dem Omnibustest für diesen Fall entspricht. Der englische Ausdruck für Varianzanalyse ist „analysis of variance“, was meist mit dem Akronym ANOVA abgekürzt wird. Falls Sie also von einer einfaktoriellen ANOVA lesen, ist auch damit eine einfaktorielle Varianzanalyse ohne Messwiederholung gemeint (eine Varianzanalyse mit Messwiederholung wird zusätzlich mit dem Ausdruck „repeated measures“ ANOVA als solche qualifiziert).

Wie es der Begriff Varianzanalyse schon zum Ausdruck bringt, beruht dieses Verfahren auf der Analyse der Varianz. Gemeint ist damit die (unbekannte) Varianz der Zufallsvariablen des oben beschriebenen varianzanalytischen Modells. Bei Vorliegen eines entsprechenden Datensatzes kann diese grundsätzlich auf zwei Arten geschätzt werden. Eine der beiden Arten schätzt dabei immer die

unbekannte Varianz σ^2. Die andere Art schätzt diese Varianz nur, wenn die Nullhypothese der Gleichheit der Populationsmittelwerte auch tatsächlich gilt. Sonst überschätzt sie die Varianz (liefert also zu große Werte für sie). Durch Bildung des Verhältnisses aus diesen beiden Schätzungen kann dann abgeschätzt werden, wie plausibel das Zutreffen der Nullhypothese erscheint.

Nehmen wir zur Illustration dieses Vorgehens an, dass wir die Ausprägungen einer AV für insgesamt m Personengruppen vorliegen haben. Diese Personengruppen können grundsätzlich unterschiedliche Anzahlen an Personen beinhalten. Diese Anzahlen werden mit $n_1, n_2, \ldots, n_m$ bezeichnet, wobei $n = n_1 + n_2 + \cdots + n_m = \sum_{j=1}^{m} n_j$ den Umfang aller Stichproben zusammen bezeichnet.

Die erste Art, σ^2 zu schätzen, besteht nun darin, die Varianz von Y_{ij} jeweils in den einzelnen Populationen durch Ermittlung der Varianz der entsprechenden Schätzwerte y_{ij} für jede der Stichproben zu schätzen und diese m Schätzwerte zu einer sog. „gepoolten“ Schätzung der Varianz von Y_{ij} zu kombinieren. Den Spezialfall einer gepoolten Varianzschätzung aus zwei Stichproben haben wir bereits im vorhergehenden Kapitel bei der Ermittlung des Standardfehlers für die mittlere Differenz zweier Populationsmittelwerte kennen gelernt. Für den allgemeinen Fall von m Gruppen, kann die Schätzfunktion der gepoolten Varianz als

$$S_{pool}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_m - 1)S_m^2}{n_1 + n_2 + \cdots + n_m - m} = \frac{\sum_{j=1}^{m}(n_j - 1)S_j^2}{n - m}$$

mit Schätzwert

$$s_{pool}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_m - 1)s_m^2}{n_1 + n_2 + \cdots + n_m - m} = \frac{\sum_{j=1}^{m}(n_j - 1)s_j^2}{n - m}$$

geschrieben werden, wobei hier

$$s_j^2 = \frac{\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{(n_j - 1)}$$

den auf Basis der j-ten Stichprobe geschätzten Wert für die j-te Populationsvarianz darstellt. Da jedoch die Zufallsvariable in den m Populationen gemäß der oben diskutierten Voraussetzungen jeweils

dieselbe Varianz aufweisen sollte, sollten sich diese Schätzwerte nur aufgrund der zufälligen Ziehung von Zufallsvariablen zufällige Schwankungen unterscheiden und die gepoolte Varianz sollte insgesamt eine Schätzung von σ^2 darstellen. Theoretisch kann gezeigt werden, dass in der Tat

$$E\left(S_{pool}^2\right) = \sigma^2.$$

Die zweite Art, σ^2 zu schätzen, sofern die Nullhypothese der Gleichheit der Populationsmittelwerte zutrifft, besteht nun darin die einzelnen Gruppenmittelwerte für die m Stichproben zur Schätzung von σ^2 heranzuziehen. Mathematisch lässt sich zeigen, dass, wenn eine Zufallsvariable normalverteilt mit Erwartungswert μ und Varianz σ^2 ist, der Mittelwert von n_j solcher Zufallsvariablen wiederum eine normalverteilte Zufallsvariable mit Erwartungswert μ und Varianz σ^2/n_j ist. D.h. im vorliegenden Fall handelt es sich bei den Gruppenmittelwerten $\bar{Y}_j$ unter Geltung der Nullhypothese jeweils um Zufallsvariablen mit Erwartungswert μ und Varianz σ^2/n_j. Die um μ verschobene Zufallsvariable $(\bar{Y}_j - \mu)$ hat dann Erwartungswert Null und nach wie vor Varianz σ^2/n_j. Multiplikation (Skalierung) mit n_j führt also jeweils zurück auf eine Zufallsvariable mit Erwartungswert Null und Varianz σ^2. Die jeweilige Schätzung von μ durch $\bar{Y} = \frac{1}{m}\sum_{j=1}^{m} \bar{Y}_j$ sowie Division durch die Anzahl der Gruppen minus 1 (= $m - 1$) führt dann wiederum zu einer erwartungstreuen Schätzfunktion für die unbekannte Varianz σ^2 (siehe Theorie zur erwartungstreuen Schätzung der Varianz einer normalverteilten Zufallsvariable, z.B. bei Bühner et al., 2025), d.h.

$$S^2 = \frac{\sum_{j=1}^{m} n_j\left(\bar{Y}_j - \bar{Y}\right)^2}{m-1},$$

wobei hier $\bar{Y} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} Y_{ij}}{\sum_{j=1}^{m} n_j} = \frac{1}{m}\sum_{j=1}^{m} \bar{Y}_j$ für die Schätzfunktion des Gesamtmittelwerts gilt. Unter Geltung der Nullhypothese gilt wiederum, dass

$$E(S^2) = \sigma^2$$

und dass damit

$$s^2 = \frac{\sum_{j=1}^{m} n_j\left(\bar{y}_j - \bar{y}\right)^2}{m-1}$$

ebenfalls einem konkreten Schätzwert der unbekannten Populationsvarianz σ^2 entspricht.

Setzt man nun diese beiden Schätzwerte ins Verhältnis, d.h. bildet man den Quotienten s^2/s_{pool}^2, so würde man unter Geltung der Nullhypothese einen Wert nahe 1 für dieses Verhältnis erwarten. Gilt die Nullhypothese allerdings nicht, so hängt die zweite Art der Schätzung von σ^2 zusätzlich zu den zufälligen Schwankungen der Zufallsvariablen um ihren Erwartungswert auch noch vom realen Unterschied zwischen den jeweiligen Gruppenmittelwerten und dem Gesamtmittelwert ab. Theoretisch kann gezeigt werden, dass für den allgemeinen Fall $\Delta\mu_j \neq 0$

$$E(S^2) = \sigma^2 + \frac{\sum_{j=1}^{m} n_j\left(\Delta\mu_j - \mu\right)^2}{m-1},$$

d.h. insbesondere, da der zweite Summand ≥ 0 sein muss, dass im Falle, dass die Nullhypothese nicht gilt, größere Werte für das Verhältnis s^2/s_{pool}^2 erwartet werden können. Je größer also dieses Verhältnis, desto unplausibler die Nullhypothese.

Insbesondere kann schließlich gezeigt werden, dass unter Geltung der Nullhypothese die Teststatistik

$$F = \frac{\frac{\sum_{j=1}^{m} n_j\left(\bar{Y}_j - \bar{Y}\right)^2}{m-1}}{\frac{\sum_{j=1}^{m}\left(n_j - 1\right)S_j^2}{n-m}}$$

einer sog. F-Verteilung mit den Freiheitsgraden $\nu_1 = m-1$ und $\nu_2 = n-m$ folgt. Die Ausdrücke $\sum_{j=1}^{m} n_j\left(\bar{Y}_j - \bar{Y}\right)^2$ und $\sum_{j=1}^{m}\left(n_j - 1\right)S_j^2$ werden in der Theorie auch häufig als „Quadratsumme zwischen“ und „Quadratsumme innerhalb“ bezeichnet (Bühner et al., 2025), die Freiheitsgrade als Zähler- (ν_1; da im Zähler des Ausdrucks für F) und als Nennerfreiheitsgrade (ν_2; da im Nenner des Ausdrucks für F). Aus dem Verhältnis der beiden Quadratsummen (ohne Division durch den jeweiligen Freiheitsgrad) lässt sich eine Effektstärke für die einfaktorielle Varianzanalyse ermitteln (siehe unten).

Die Realisation dieser Teststatistik in der konkreten Datensituation ist dann gegeben durch

$$f = \frac{\frac{\sum_{j=1}^{m} n_j\left(\bar{y}_j - \bar{y}\right)^2}{m-1}}{\frac{\sum_{j=1}^{m}\left(n_j - 1\right)s_j^2}{n-m}}.$$

Für eine gegebene Datensituation kann dann jeweils die Realisation dieser Teststatistik berechnet werden sowie der p-Wert dafür unter Geltung der Nullhypothese eine so große oder extremere Teststatistik zu erhalten (durch Integration der F-Verteilung über den Bereich $[f, +\infty)$). Ergibt sich daraus ein p-Wert kleiner oder gleich dem vorab festgelegten Signifikanzniveau, kann wiederum geschlossen werden, dass ein solches Ergebnis unter Geltung der Nullhypothese bei wiederholter Ziehung einfacher Zufallsstichproben so selten wäre, dass es unplausibel erscheint, dass die Nullhypothese gilt und sie daher abgelehnt wird.

Durchführung der einfaktoriellen Varianzanalyse mit SPSS

Zur Illustration der Durchführung der einfaktoriellen Varianzanalyse mit SPSS sei wieder auf unser einleitendes Beispiel mit dem Depressionsniveau für die drei verschiedenen Altersgruppen junger Erwachsener, Erwachsener mittleren Alters und älterer Erwachsener zurückgegriffen. Ein entsprechender (wiederum fiktiver) Datensatz ist in der Datei „Kap6daten.sav“ zu finden, die Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Um nun eine einfaktorielle Varianzanalyse ohne Messwiederholung an diesem Datensatz zur Testung der Nullhypothese der Gleichheit aller drei Populationsmittelwerte (d.h. die Alternativhypothese lautet, dass sich mindestens zwei der drei Populationsmittelwerte unterscheiden) in SPSS durchzuführen, wählen wir *Analyze >> General Linear Model >> Univariate…*. Im sich öffnenden Fenster ziehen wir unsere AV, d.h. die Variable *Depressionsniveau*, in das Feld „Dependent Variable“ und unsere UV, d.h. die Variable Altersgruppe, in das Feld „Fixed Factor(s)“, siehe Abbildung 6.1. Unter „Options…“ wählen wir noch „Descriptive statistics“, „Homogeneity tests“ und „Estimate of effect size“ sowie unser Signifikanzniveau aus (hier belassen wir es einfach einmal bei der Voreinstellung von .05), siehe Abbildung 6.2. Danach fügen wir alles wieder in eine Syntaxdatei ein, dokumentieren diese und führen die eingefügten Kommandozeilen aus.

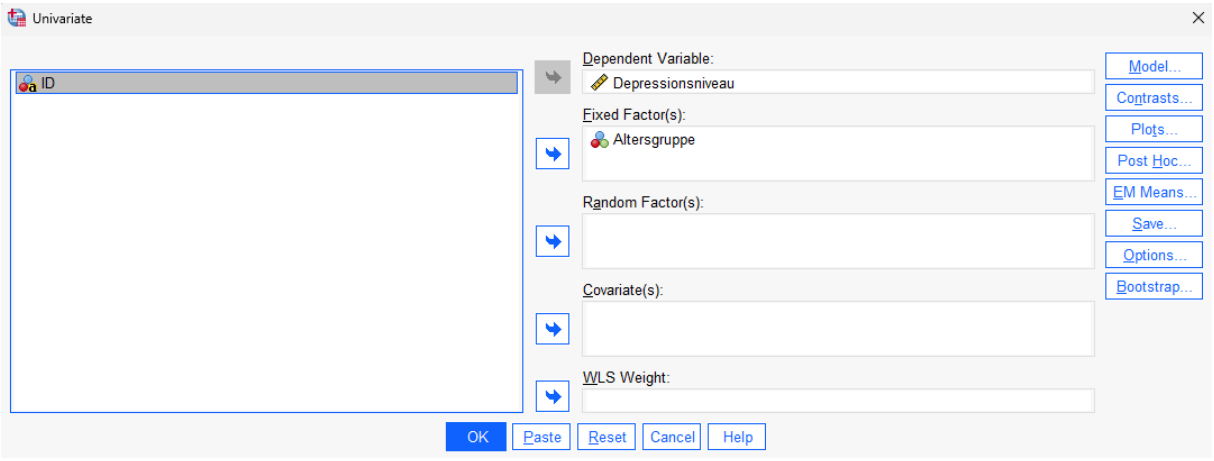


Abbildung 6.1. Auswahl einer einfaktoriellen Varianzanalyse in SPSS.

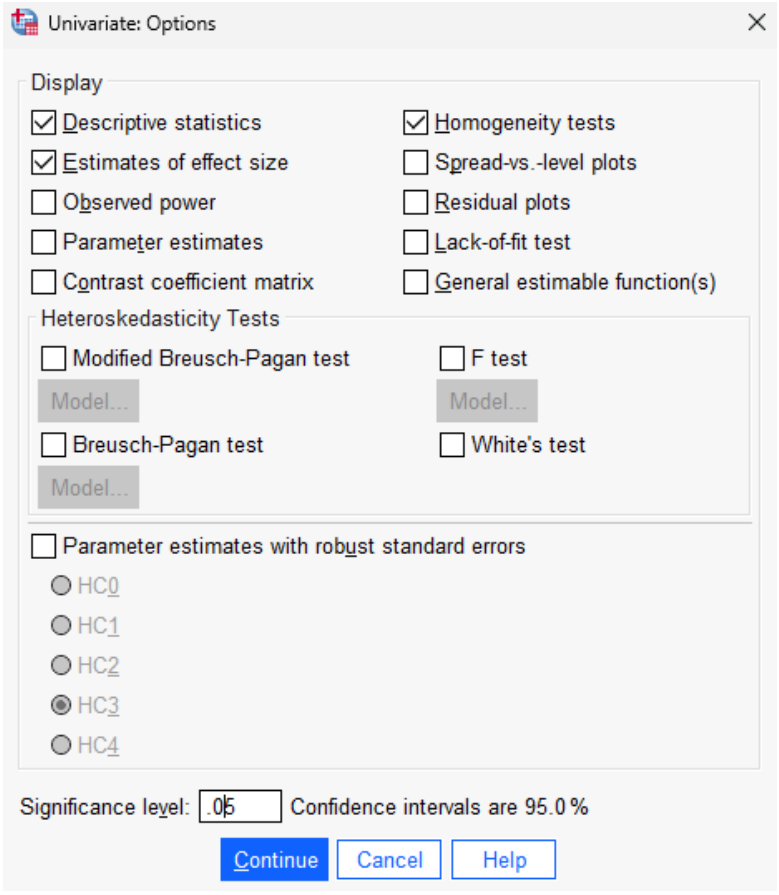


Abbildung 6.2. Auswahl der Optionen für eine einfaktorielle Varianzanalyse ohne Messwiederholung.

Die Ausgabe besteht aus insgesamt vier Tabellen. Die erste der Tabellen mit Überschrift „Between-Subjects Factors“ führt die drei Stufen unseres Faktors mitsamt den entsprechenden Labels sowie der Anzahl der Personen pro Stufe auf. Hier sehen wir, dass es sich hier um ein balanciertes Design handelt, da in jeder Stufe bzw. Stichprobe 60 Personen vorliegen. An dieser Stelle sei darauf

hingewiesen, dass Faktoren, die zwischen unabhängigen Gruppen unterscheiden, auch als Zwischensubjektfaktoren (Engl.: Between-subjects factors) bezeichnet werden, da es sich bei den Personen in den verschiedenen Gruppen auch immer um verschiedene Personen handelt. Bei Varianzanalysen mit Messwiederholung werden wir sog. Innersubjektfaktoren (Engl.: Within-subjects factors) kennenlernen, da dort mehrere Messwerte derselben AV für ein und dieselbe Person vorliegen werden, d.h. die Stufen des Faktors liegen dort jeweils innerhalb (Engl.: within) ein und derselben Person.

In der Tabelle „Descriptive Statistics“ finden wir deskriptive Statistiken sowohl getrennt für alle drei Stichproben als auch für die Gesamtstichprobe. Diese Werte werden wir im Ergebnisbericht brauchen (siehe unten). Allerdings werden sie dort noch dem APA-Format entsprechend anzupassen sein (z.B. Runden auf zwei Nachkommastellen).

In der nächsten Tabelle finden wir Ergebnisse für Levenes Tests, die sich jeweils darin unterscheiden bezüglich welchen Referenzwerts die Varianzen in den verschiedenen Stichproben verglichen werden (Mittelwerte, Mediane, Mediane mit Freiheitsgradanpassung, getrimmte Mittelwerte). Da uns die Prüfung der Varianzgleichheit bezogen auf die einzelnen Populationsmittelwerte interessiert, schauen wir uns hier die erste Zeile an und sehen, dass der Levenes Test nicht signifikant ist, $p = .405$, siehe Abbildung 6.3. Da Varianzhomogenität (= Gleichheit der Varianzen oder auch Homoskedastizität) für Varianzanalysen eine sehr wichtige Voraussetzung ist, wählen wir für diese Überprüfung üblicherweise das Signifikanzniveau $\alpha = .05$. D.h. Levenes Test wäre signifikant für $p < .05$ (unabhängig und eventuell verschieden von unserem Signifikanzniveau für unseren eigentlichen Hypothesentest) und wäre dem so, würden wir anstelle der einfaktoriellen Varianzanalyse eine Varianzanalyse nach Welch durchführen, die ungleiche Populationsvarianzen berücksichtigen kann (eine Beschreibung der Durchführung einer solchen Varianzanalyse folgt unten). In diesem Fall ist aber Levenes Test nicht signifikant und wir können uns endlich den eigentlichen Ergebnissen der Varianzanalyse zuwenden.

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)	Based on Mean	.907	2	177	.405
	Based on Median	.895	2	177	.410
	Based on Median and with adjusted df	.895	2	176.714	.410
	Based on trimmed mean	.894	2	177	.411

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)

b. Design: Intercept + Altersgruppe

Abbildung 6.3. Überprüfung der Varianzgleichheit.

Diese finden wir in der Tabelle „Tests of Between-Subjects Effects“, die zur Illustration auch noch einmal in Abbildung 6.4 dargestellt ist. Die relevanten Zeilen bzw. Zellen dieser Ausgabe sind in Abbildung 6.4 rot markiert. In der Zeile „Altersgruppe“ sehen wir, dass unsere Varianzanalyse mit p = .005 bei einem Signifikanzniveau von α = .05 signifikant ist (bei α = .005 wäre dem nicht so; dazu müssten wir im SPSS Ausgabe auf die entsprechende Tabelle doppelt links klicken und dann im sich öffnenden Fenster noch einmal doppelt auf den p-Wert, um den exakten Wert angezeigt zu bekommen). Wir sehen auch den Schätzwert für unsere Teststatistik in der Spalte „F“ (da es sich unter Geltung der Nullhypothese um eine F-verteilte Teststatistik handelt) sowie die Zählerfreiheitsgrade $\nu_1 = 2$ (= Anzahl der Gruppen minus 1, siehe oben) und die Nennerfreiheitsgrade $\nu_2 = 177$ (= gesamter Stichprobenumfang minus Anzahl der Gruppen, siehe oben). All diese Werte werden wir im Ergebnisbericht brauchen.

Was wir schließlich auch noch für den Ergebnisbericht und in späterer Folge für eine Stichprobenplanung (etwa eines Replikationsexperiments) brauchen werden, ist eine Effektstärke. Diese finden wir in der letzten Spalte mit der Überschrift „Partial Eta Squared“. In dieser lesen wir einen Wert für unsere Effektstärke $\eta^2 = 0.06$ ab. Gemäß Theorie entspricht diese Effektstärke dem Verhältnis der Varianz in der AV, die durch die Gruppenzugehörigkeit aufgeklärt werden kann (= Quadratsumme zwischen), zu der Varianz in der AV insgesamt (= totale Quadratsumme; siehe z.B. Bühner et al., 2025). Diese beiden Varianzen sind durch die Quadratsummen in der Zeile „Altersgruppe“, d.h. 2209.544, und in der Zeile „Corrected Total“, d.h. 38340.061, gegeben. Man kann sich leicht überzeugen, dass Division dieser beiden Zahlen die entsprechende Zahl in der letzten Spalte ergibt.

Tests of Between-Subjects Effects

Dependent Variable: Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2209.544[a]	2	1104.772	5.412	.005	.058
Intercept	171556.939	1	171556.939	840.441	<.001	.826
Altersgruppe	2209.544	2	1104.772	5.412	.005	.058
Error	36130.517	177	204.127			
Total	209897.000	180				
Corrected Total	38340.061	179				

a. R Squared = .058 (Adjusted R Squared = .047)

Abbildung 6.4. Die Ergebnisse unserer ersten einfaktoriellen Varianzanalyse ohne Messwiederholung mit SPSS.

Der Zusatz „partial“ in der letzten Spalte (für die Effektstärke) geht darauf zurück, dass im Falle mehrfaktorieller Varianzanalysen für jeden Faktor eine eigene Effektstärke berechnet werden kann, worauf wir im nächsten Kapitel zu sprechen kommen werden. Im Falle einer einfaktoriellen Varianzanalyse entspricht das partielle eta-Quadrat aber schlichtweg dem „gesamten“ eta-Quadrat, d.h. hier $\eta_p^2 = \eta^2$. Für diese Ausgabe könnte die sich ergebende Effektstärke wie folgt interpretiert werden: „Der Schätzwert für den Anteil an der Gesamtvarianz des Depressionsniveaus in der Population, der durch Zugehörigkeit zu einer der drei Altersgruppen erklärt werden kann, beträgt 6%.“

Auch für die Effektstärke η^2 gibt es Heuristiken nach Cohen (1988) dafür, wie groß diese Effektstärken einzuschätzen sind. Dementsprechend werden Effektstärken im Bereich 0.01-0.06 als klein, im Bereich 0.06-0.14 als mittel, und ab 0.14 als groß bezeichnet.

Ergebnisbericht für eine einfaktorielle Varianzanalyse ohne Messwiederholung

Ein Ergebnisbericht für dieses Beispiel könnte wie folgt aussehen: „Deskriptive Statistiken für die Depressionsschwere in den betrachteten drei Altersgruppen sind in Tabelle 6.1 angegeben. Die Mittelwerte der drei Altersgruppen unterscheiden sich (mit α = .05) signifikant, $F(2, 177) = 5.41$, p = .005, η^2 = .06, d.h. der Schätzwert für den Anteil an der Gesamtvarianz des Depressionsniveaus in der Population, der durch Zugehörigkeit zu einer der drei Altersgruppen erklärt werden kann, beträgt 6%. Gemäß Cohens Heuristik (1988) entspräche dies gerade einem mittleren Effekt für den auf zwei

Nachkommastellen gerundeten Schätzwert für η^2. Bezieht man sich auf den numerisch genaueren Wert von 0.058 handelt es sich nach der Heuristik gerade noch um einen kleinen Effekt."

Hier ist zu beachten, dass die führende Null bei der Effektstärke gemäß APA-Richtlinien wegzulassen ist, da es sich bei dieser Effektstärke um eine Zahl zwischen Null und Eins handelt. Auch das APA-Format der Tabelle für die deskriptiven Statistiken ist zu beachten.

Tabelle 6.1

Deskriptive Statistiken

Altersgruppe	*M*	*SD*	*n*
Junge Erwachsene	26.38	15.37	60
Erwachsene mittleren Alters	31.30	13.58	60
Ältere Erwachsene	34.93	13.86	60

Stichprobenplanung für eine einfaktorielle Varianzanalyse ohne Messwiederholung

Auch für diesen Fall kann wieder eine Stichprobenplanung mit G*Power durchgeführt werden. Dafür ist unter „Test family" die Option „F tests" auszuwählen, unter „Statistical test" die Option „ANOVA: Fixed effects, omnibus, one-way" und unter „Type of power analysis" wiederum „A priori: Compute required sample size – given α, power, and effect size". Unter „Input Parameters" sind dann die gewünschten Werte für die Effektstärke, das Signifikanzniveau, die Teststärke sowie die Anzahl der Gruppen einzutragen. Bei der Effektstärke ist dabei zu beachten, dass G*Power hier die Effektstärke in Form der Größe f benötigt (nicht zu verwechseln mit der Realisation der Teststatistik oben), die sich aus η^2 wie folgt berechnen lässt:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}.$$

Diese Berechnung kann allerdings gleich in G*Power durchgeführt werden, indem man auf die Schaltfläche „Determine =>" klickt und im sich öffnenden Menü den gewünschten Wert für η^2 unter „Partial η^2" eingibt (da für die einfaktorielle Varianzanalyse ohne Messwiederholung $\eta^2 = \eta_p^2$ gilt).

Angenommen wir wollten ein Replikationsexperiment für das obige Beispiel durchführen und den Stichprobenumfang für α = .005, eine Teststärke von 90% und die Effektstärke η^2 = .06 ermitteln. Dann würden wir in G*Power die in Abbildung 6.5 gezeigten Eingaben tätigen und einen benötigten Stichprobenumfang von n = 312, d.h. 104 Personen pro Gruppe, erhalten.

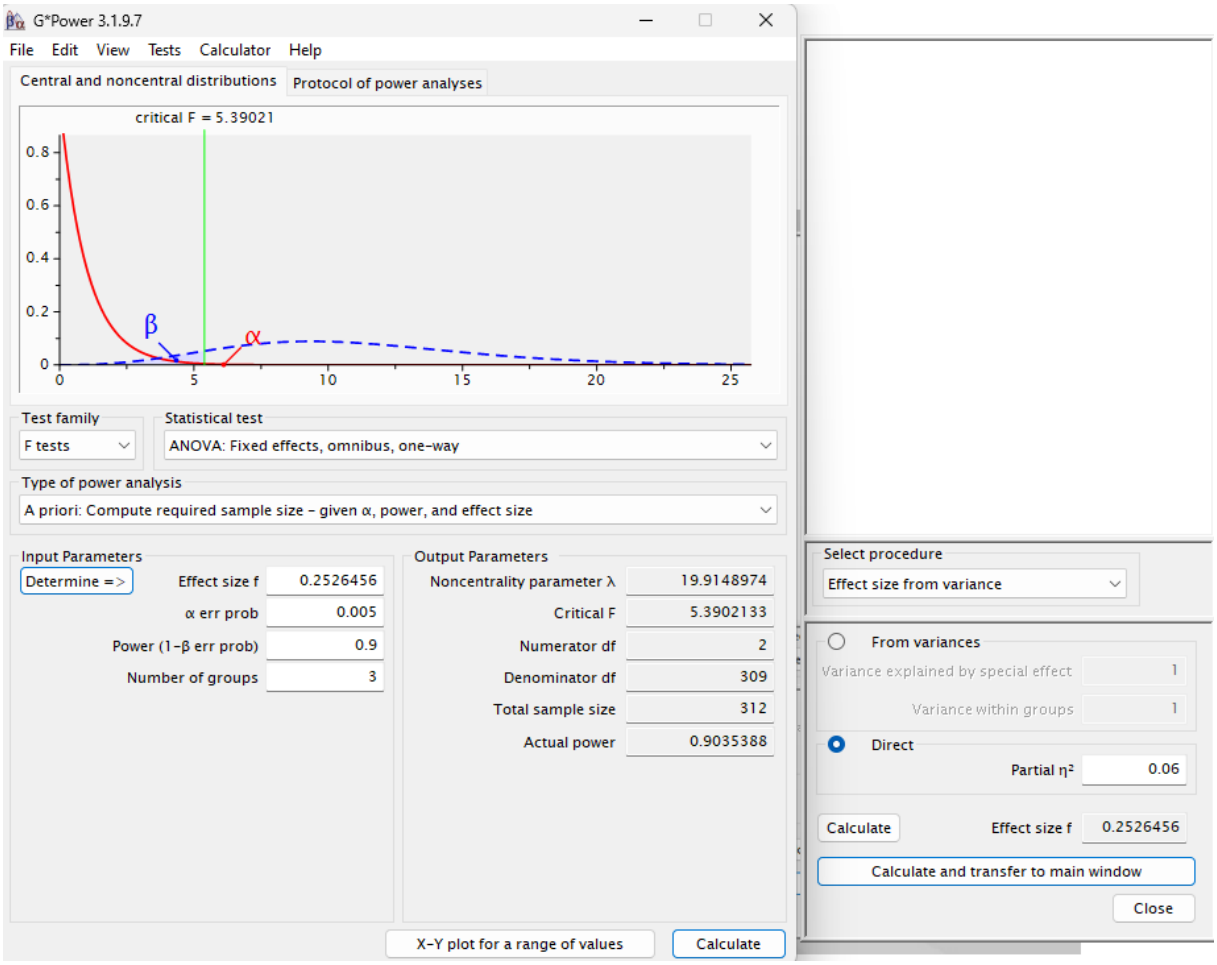


Abbildung 6.5. Stichprobenplanung für eine einfaktorielle Varianzanalyse ohne Messwiederholung mit G*Power.

Paarweise post-hoc Vergleiche

Am oben gegebenen Ergebnisbericht können wir einen unangenehmen Aspekt der Durchführung des Omnibustests für Gleichheit aller Populationsmittelwerte erkennen. Zwar können wir die Unterschiedlichkeit der Populationsmittelwerte mit einer Irrtumswahrscheinlichkeit von 5% feststellen, aber wir können keine statistische Aussage darüber machen, welcher Populationsmittelwert sich von welchem anderen unterscheidet. Deskriptiv können wir zwar aus den Ergebnissen ablesen, dass der

Stichprobenmittelwert am größten für ältere Erwachsene ist, am zweitgrößten für Erwachsene mittleren Alters, und am kleinsten für junge Erwachsene. Aussagen über die statistische Signifikanz paarweiser Populationsmittelwertsunterschiede können wir allerdings auf Basis des Omnibustests alleine nicht treffen. Deshalb ist auch der Omnibustest alleine in der Praxis selten hilfreich (Bühner et al., 2025).

Sollen deshalb im Anschluss an einen Omnibustest noch paarweise Mittelwertvergleiche gemacht werden, kann dies in SPSS im Rahmen sog. post-hoc Vergleiche angefordert werden. Sofern tatsächlich im Vorhinein keinerlei Hypothesen darüber bestanden wie sich die Populationsmittelwerte unterscheiden könnten, ist jedenfalls im Rahmen solcher post-hoc Vergleiche für multiple Vergleiche zu korrigieren, da dies der Testung einer zusammengesetzten Hypothese mit der Verknüpfung „oder" entspricht (d.h., wir würden sagen, dass sich die Populationsmittelwerte voneinander unterscheiden, wenn mindestens einer der paarweisen Vergleiche einen signifikanten Unterschied ergibt). Alternativ könnte man allerdings von vornherein an allen drei paarweisen Unterschieden interessiert sein; dann wäre es allerdings unnötig vorab einen Omnibustest durchzuführen; man könnte stattdessen einfach die drei paarweisen Vergleiche (allerdings mit der gepoolten Standardabweichung, da dies die Teststärke erhöht) ohne Korrektur der p-Werte zur Kontrolle der FWER (= family-wise error rate), sondern stattdessen mit einem geringen α = .005 und einer hohen Teststärke (etwa 80%) arbeiten, um die FDR (= false discovery rate) zu kontrollieren. Eine Angabe der Teststärke setzt aber gewisses Vorwissen voraus; d.h. sofern dieses nicht besteht (etwa aufgrund der Ergebnisse eines Explorationsexperiments), bleibt nur der Weg über die post-hoc Vergleiche.

Für letztere ist unter *Analyze >> General Linear Model >> Univariate…* im Menü „Post Hoc…" zuerst die Variable *Altersgruppe* in das Feld „Post Hoc Tests for" zu ziehen. In der Mitte links, unter „Equal Variances Assumed" können eine Vielzahl von post-hoc Vergleichen mit entsprechenden Korrekturen für p-Werte ausgewählt werden. Hier wählen wir zu Illustrationszwecken gleich die folgenden drei aus: „LSD", „Bonferroni" und „Tukey". Daraufhin fügen wir alles wieder in die Syntax ein und führen die Kommandozeilen aus.

Zusätzlich zu den bereits bekannten Ergebnissen erhalten wir dadurch auch eine Tabelle mit der Überschrift „Multiple Comparisons". In dieser Tabelle finden wir Vergleiche für alle möglichen Paare unserer drei Altersgruppen für jede ausgewählte Methode für die Korrektur der p-Werte. Für jeden

paarweisen Vergleich ist die Punktschätzung der Mittelwertdifferenz aufgeführt, sowie deren Standardfehler (basierend auf der gepoolten Varianz aller drei Gruppen, weshalb alle Standardfehler auch den gleichen Wert für alle Vergleiche haben), den p-Wert (in der Spalte „Sig“), und ein Konfidenzintervall mit den plausiblen Werten für den paarweisen Vergleich.

Wir sehen, dass für alle drei Methoden jeweils der Unterschied zwischen jungen und älteren Erwachsenen (mit $\alpha = .05$) signifikant ist ($p < .05$), während alle anderen Vergleiche nicht signifikant sind. Wir sehen zudem auch wie sich die einzelnen Korrekturen auf die p-Werte auswirken. Bei der LSD-Methode (LSD steht für Fishers Least-Significant-Difference Test) wird außer der Verwendung der gepoolten Varianz zur Berechnung des Standardfehlers keine Korrektur der p-Werte an sich vorgenommen. Allerdings kontrolliert die Methode die FWER im Fall von genau drei Gruppen exakt (Meier, 2006; Marcus et al., 1976), weshalb es sich in diesem Fall um die Methode mit der höchsten Teststärke handelt. Die Bonferroni-Methode hingegen multipliziert jeden p-Wert mit dem Faktor 3, da hier drei paarweise Mittelwertvergleiche durchgeführt werden (man sieht dies z.B. gut am p-Wert für den Vergleich junger Erwachsener mit Erwachsenen mittleren Alters: für die LSD-Methode ergibt sich $p = .061$, was genau einem Drittel von $p = .183$ bei der Bonferroni-Methode entspricht). Diese Korrektur des p-Werts ist allerdings sehr konservativ, was auf Kosten der Teststärke geht. Dahingehend ist für jede Anzahl von Gruppen Tukeys HSD-Test (HSD für „honestly significant difference“) der Bonferroni-Methode vorzuziehen. In Tukeys Methode werden die Abhängigkeiten der einzelnen paarweisen Vergleiche untereinander direkt berücksichtigt, was eine exaktere Korrektur der p-Werte und daher höhere Teststärke ermöglicht (man sieht das daran, dass die p-Werte in der Tabelle für die HSD-Methode etwas geringer ausfallen als für die Bonferroni-Methode), ohne auf die Kontrolle der FWER zu verzichten.

Zusammengefasst lässt sich also folgendes festhalten. Sollten genau drei Gruppen vorliegen, sollte für post-hoc Vergleiche die LSD-Methode gewählt werden, da sie bei exakter Kontrolle der FWER die höchste Teststärke aufweist. Für mehr als drei Gruppen sollte Tukeys HSD-Methode gewählt werden. Die Bonferroni-Methode sollte im Rahmen der einfaktoriellen Varianzanalyse ohne Messwiederholung nie gewählt werden. In den nächsten Kapiteln werden wir allerdings auf die

Bonferroni-Methode zurückgreifen müssen. Die Durchführung von post-hoc Vergleichen ist natürlich auch für den Ergebnisbericht zu berücksichtigen, den wir uns im nächsten Abschnitt ansehen.

Ergebnisbericht für eine einfaktorielle Varianzanalyse ohne Messwiederholung mit paarweisen post-hoc Vergleichen

Ein Ergebnisbericht für dieses Beispiel unter Einbeziehung der Ergebnisse der post-hoc Vergleiche könnte wie folgt aussehen:

„Deskriptive Statistiken für die Depressionsschwere in den betrachteten drei Altersgruppen sind in Tabelle 6.1 angegeben. Die Mittelwerte der drei Altersgruppen unterscheiden sich (mit α = .05) signifikant, $F(2, 177) = 5.41$, $p = .005$, $\eta^2 = .06$, d.h. der Schätzwert für den Anteil an der Gesamtvarianz des Depressionsniveaus in der Population, der durch Zugehörigkeit zu einer der drei Altersgruppen erklärt werden kann, beträgt 6%. Gemäß Cohens Heuristik (1988) entspricht dies einem mittleren Effekt.

Paarweise Vergleiche mittels Fishers LSD-Test ergeben einen signifikanten Unterschied zwischen den Mittelwerten für junge Erwachsene und ältere Erwachsene, $p = .001$. Die verbleibenden beiden paarweisen Unterschiede sind nicht statistisch signifikant ($p = .061$ für den Vergleich zwischen jungen Erwachsenen und Erwachsenen mittleren Alters, $p = .165$ für den Vergleich zwischen älteren Erwachsenen und Erwachsenen mittleren Alters)."

Voraussetzungen für eine einfaktorielle Varianzanalyse ohne Messwiederholung

Die Voraussetzungen für eine einfaktorielle Varianzanalyse ohne Messwiederholung wurden oben bereits erläutert und sind hier noch einmal zusammengefasst:

- Normalverteilung der AV in den einzelnen Populationen,
- Varianzgleichheit (auch als Varianzhomogenität bzw. Homoskedastizität bezeichnet),
- Unabhängigkeit der Beobachtungen bzw. Messungen,
- Intervallskalenniveau der AV.

Die Normalverteilungsvoraussetzung kann wie im vorhergehenden Kapitel beschrieben überprüft werden. Allerdings erwies sich die Varianzanalyse gegenüber der Verletzung der Normalverteilungsvoraussetzung in Simulationsstudien als relativ robust. Insbesondere bei balancierten Designs sind

Signifikanzniveau und Teststärke kaum von dieser Voraussetzungsverletzung betroffen (Bühner & Ziegler, 2017; Bühner et al., 2025).

Umso wichtiger ist allerdings im Rahmen von Varianzanalysen die Voraussetzung der Varianzgleichheit, insbesondere da darauf die Berücksichtigung aller Gruppen für die Schätzung der Varianz σ^2 durch die gepoolte Varianz S^2_{pool} beruht. Aus diesem Grund ist die Varianzanalyse noch empfindlicher auf die Verletzung dieser Voraussetzung als der t-Test im vorhergehenden Kapitel. Die Voraussetzung der Varianzhomogenität sollte daher grundsätzlich immer überprüft werden.

Um die Varianzhomogenität statistisch zu prüfen, kann der Levene-Test wie oben beschrieben verwendet werden. Ist dieser signifikant (mit α = .05) so kann von einer Verletzung der Voraussetzung ausgegangen werden und es sollte statt der einfaktoriellen Varianzanalyse eine Varianzanalyse nach Welch durchgeführt werden (siehe nächster Abschnitt).

Die Unabhängigkeit der Messungen und das Intervallskalenniveau der AV wird durch das experimentelle Design festgelegt und kann im Rahmen der Datenanalyse als gegeben vorausgesetzt werden bzw. an dieser Stelle nicht mehr überprüft werden.

Durchführung einer Varianzanalyse nach Welch

Die Durchführung ist wiederum am Beispiel des Depressionsniveaus für die untersuchten drei Altersgruppen erklärt. Hier kann zwar von einer Verträglichkeit mit der Voraussetzung der Varianzgleichheit ausgegangen werden, aber das Beispiel dient einerseits lediglich zur Illustration des Vorgehens und andererseits sollte sich so auch zeigen, dass im Falle der Erfüllung der Vorrausetzung keine stark unterschiedlichen Ergebnisse zu erwarten sind. Um eine Varianzanalyse nach Welch in SPSS durchzuführen ist unter *Analyze >> Compare Means and Proportions >> One-Way ANOVA…* erst die Variable *Depressionsniveau* in das Feld „Dependent List“ und die Variable *Altersgruppe* in das Feld „Factor“ zu verschieben, siehe Abbildung 6.6. Unter „Options“ ist anschließend „Welch test“ auszuwählen.

In der resultierenden Ausgabe, siehe Abbildung 6.7, ist in der Tabelle „ANOVA“ zuerst wieder das Ergebnis einer gewöhnlichen Varianzanalyse angeführt (zumindest die wesentlichen Ergebnisse) und in der Tabelle „ANOVA Effect Sizes“ unterschiedliche Maße für Effektstärken (inkl. der uns

wohlbekannten Effektstärke η^2 in der ersten Zeile) mit dem netten Zusatz eines 95%-Konfidenzintervalls für die Effektstärken.

In der letzten Tabelle ist schließlich das Ergebnis der Varianzanalyse nach Welch mit dem Wert der Teststatistik, den beiden Freiheitsgraden und dem p-Wert, angeführt. In einem Ergebnisbericht könnte man dieses Ergebnis etwa wie folgt berichten: „Aufgrund der Verletzung der Voraussetzung der Varianzgleichheit (getestet mit Levenes Test, $p < .05$) wurde eine Varianzanalyse nach Welch durchgeführt. Die Varianzanalyse ergab, dass sich die Mittelwerte (mit $\alpha = .05$) signifikant voneinander unterscheiden, $F(2, 117.68) = 5.10$, $p = .008$."

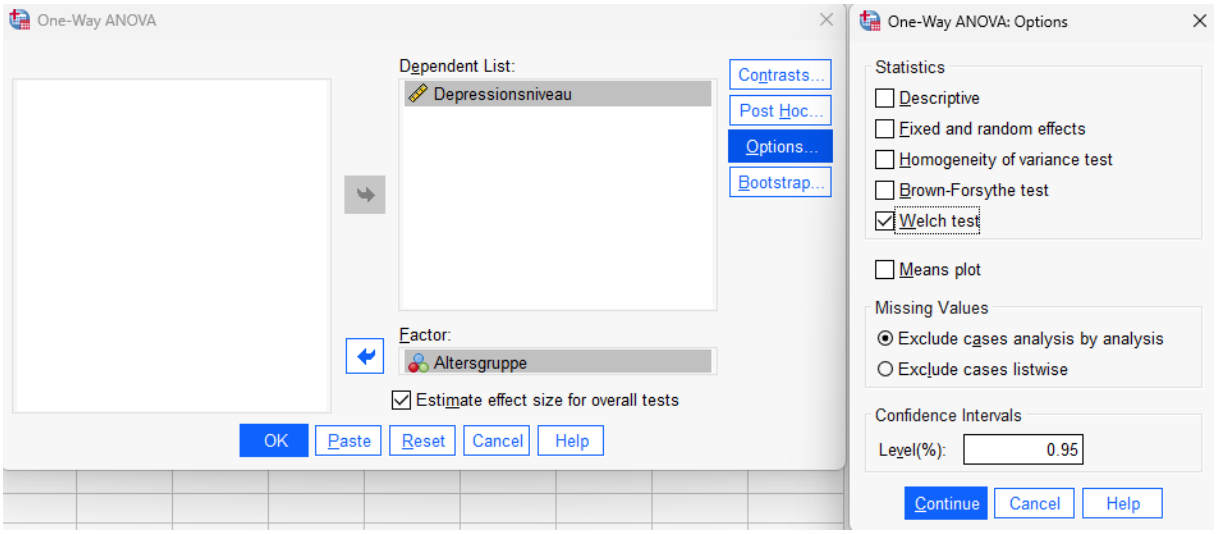


Abbildung 6.6. Auswahl einer Varianzanalyse nach Welch.

A-priori Vergleiche

Ein Omnibus-Test mit anschließenden post-hoc Vergleichen hat aufgrund der geringeren effektiven Signifikanzniveaus wegen der nötigen Korrektur der p-Werte für die einzelnen paarweisen Vergleiche häufig den Nachteil reduzierter Teststärke. Bei vorab formulierten Hypothesen für spezifische Mittelwertsunterschiede ist es daher meist von Vorteil stattdessen sog. a-priori Vergleiche durchzuführen. Vorteile derselben sind (siehe z.B. Bühner et al., 2025):

- Typischerweise höhere Teststärke.
- Das Ergebnis der statistischen Testung ist häufig informativer als ein Omnibus-Test.
- Es können gerichtete Hypothesen formuliert werden.
- Eine vorhergehende Durchführung eines Omnibus-Tests ist nicht notwendig.

Zudem ist es für eine Reihe spezifischer Vergleiche nicht notwendig eine Korrektur der p-Werte bezüglich der FWER durchzuführen (Bühner & Ziegler, 2017), allerdings ist eine Kontrolle der FDR durch ein geringes Signifikanzniveau und ausreichende Teststärke dennoch wünschenswert. Entsprechende a-priori Vergleiche sind allerdings vor der Datenerhebung zu formulieren.

Aufgrund dieser Vorteile soll die Durchführung von a-priori Vergleichen mit SPSS anhand des vorliegenden Datensatzes noch für die folgenden beiden Fragestellungen illustriert werden:

(a) Uns interessiert, ob das mittlere Depressionsniveau bei jungen Erwachsenen niedriger als bei Erwachsenen mittleren Alters ist und genauso, ob das mittlere Depressionsniveau bei Erwachsenen mittleren Alters niedriger ist als bei älteren Erwachsenen.

(b) Um wie viel ist das mittlere Depressionsniveau bei jungen Erwachsenen niedriger als der Mittelwert des Depressionsniveaus von Erwachsenen mittleren Alters und älteren Erwachsenen.

ANOVA

Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2209.544	2	1104.772	5.412	.005
Within Groups	36130.517	177	204.127		
Total	38340.061	179			

ANOVA Effect Sizes[a,b]

		Point Estimate	95% Confidence Interval Lower	Upper
Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)	Eta-squared	.058	.006	.129
	Epsilon-squared	.047	-.005	.119
	Omega-squared Fixed-effect	.047	-.005	.118
	Omega-squared Random-effect	.024	-.003	.063

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

b. Negative but less biased estimates are retained, not rounded to zero.

Robust Tests of Equality of Means

Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)

	Statistic[a]	df1	df2	Sig.
Welch	5.099	2	117.680	.008

a. Asymptotically F distributed.

Abbildung 6.7. Ausgabe für die unter *Analyze >> Compare Means and Proportions >> One-Way ANOVA…* angeforderte Varianzanalyse inklusive der robusten Variante nach Welch.

Zu Illustrationszwecken wählen wir für beide Fragestellungen ein Signifikanzniveau $\alpha = .05$. Zur Beantwortung der ersten Fragestellung wählen wir unter *Analyze >> Compare Means and Proportions >> One-Way ANOVA…* das Menü „Contrasts…". Dort geben wir im Feld „Coefficients" die Zahl 1 ein und klicken anschließend auf „Add". Danach geben wir die Zahl -1 ein und klicken wieder auf „Add". Schließlich geben wir die Zahl 0 ein und klicken wieder auf „Add". Dadurch haben wir den ersten Kontrast definiert: wir möchten die Differenz der Mittelwerte der ersten Stufe (junge Erwachsene) und der zweiten Stufe (Erwachsene mittleren Alters) unseres Faktors gegen Null testen, während der Mittelwert der dritten Stufe unberücksichtigt bleibt (das bedeuten hier die drei Zahlen in der Reihenfolge 1, -1, 0). Für die Eingabe des zweiten Vergleichs klicken wir zunächst auf „Next" und geben anschließend die Zahl 0, gefolgt von Klicken auf „Add", dann die Zahl 1, gefolgt von Klicken auf „Add", und schließlich die Zahl -1, gefolgt von Klicken auf „Add", ein. Hier wollen wir also die Stufe 2 unseres Faktors (Erwachsene mittleren Alters) mit Stufe 3 (ältere Erwachsene) vergleichen. Haben wir beide a-priori Vergleiche definiert, klicken wir auf „Continue" und wählen unter „Options…" noch die Varianzanalyse nach Welch ab, falls sie von vorhin noch ausgewählt war. Danach führen wir die entsprechenden Kommandozeilen wieder in der Syntax aus, nachdem wir sie dort eingefügt und dokumentiert haben.

In den drei sich ergebenden Tabellen, siehe Abbildung 6.8, sind in der ersten Tabelle mit der Überschrift „Contrast Coefficients" noch einmal die von uns definierten Vergleiche angeführt. Hier können wir also sehen, ob wir überhaupt die richtigen Vergleiche für unsere Fragestellung durchgeführt haben. In der zweiten Tabelle mit der Überschrift „Contrast Tests" sind die t-Tests für die beiden paarweisen Vergleiche aufgeführt. Diese t-Tests schätzen aber den Standardfehler für den Mittelwertsunterschied auf Basis aller drei Gruppen, d.h. sie verfügen im Allgemeinen über höhere Teststärke als t-Tests, die nur die beiden zu vergleichenden Gruppen berücksichtigen würden. Zudem sind für beide paarweisen Vergleiche sowohl Student'sche als auch Welch t-Tests angegeben. Allerdings ist nur ein p-Wert für ungerichtete Hypothesen angegeben, der aber bei Vorliegen einer gerichteten Hypothese halbiert werden kann. Im vorliegenden Fall würden wir also für ein $\alpha = .05$ die Nullhypothese für die erste der beiden Hypothesen in Fragestellung (a) verwerfen ($p = .033$), aber für die zweite beibehalten ($p = .075$).

Contrast Coefficients

Contrast	Altersgruppe (1=18-35 Jahre, 2=36-60 Jahre, 3=älter als 60 Jahre): junge Erwachsene	Erwachsene mittleren Alters	ältere Erwachsene
1	1	-1	0
2	0	1	-1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	95% Confidence Interval Lower	95% Confidence Interval Upper
Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)	Assumes equal variances	1	-4.92	2.608	-1.885	177	.061	-10.06	.23
		2	-3.63	2.608	-1.393	177	.165	-8.78	1.51
	Does not assume equal variances	1	-4.92	2.647	-1.857	116.238	.066	-10.16	.33
		2	-3.63	2.504	-1.451	117.951	.149	-8.59	1.33

Contrast Effect Sizes

		Contrast	Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)	Cohen's d	1	14.287	-.344	-.703	.016
		2	14.287	-.254	-.613	.105
	Hedges' correction	1	14.348	-.343	-.700	.016
		2	14.348	-.253	-.610	.104

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation for all the groups.
Hedges' uses pooled standard deviation for all the groups, plus a correction factor.

Abbildung 6.8. Ausgabe für a-priori Vergleiche für Fragestellung (a).

Die Definition des Vergleichs im Menü „Contrasts…“ für Fragestellung (b) ist in Abbildung 6.9 gezeigt.

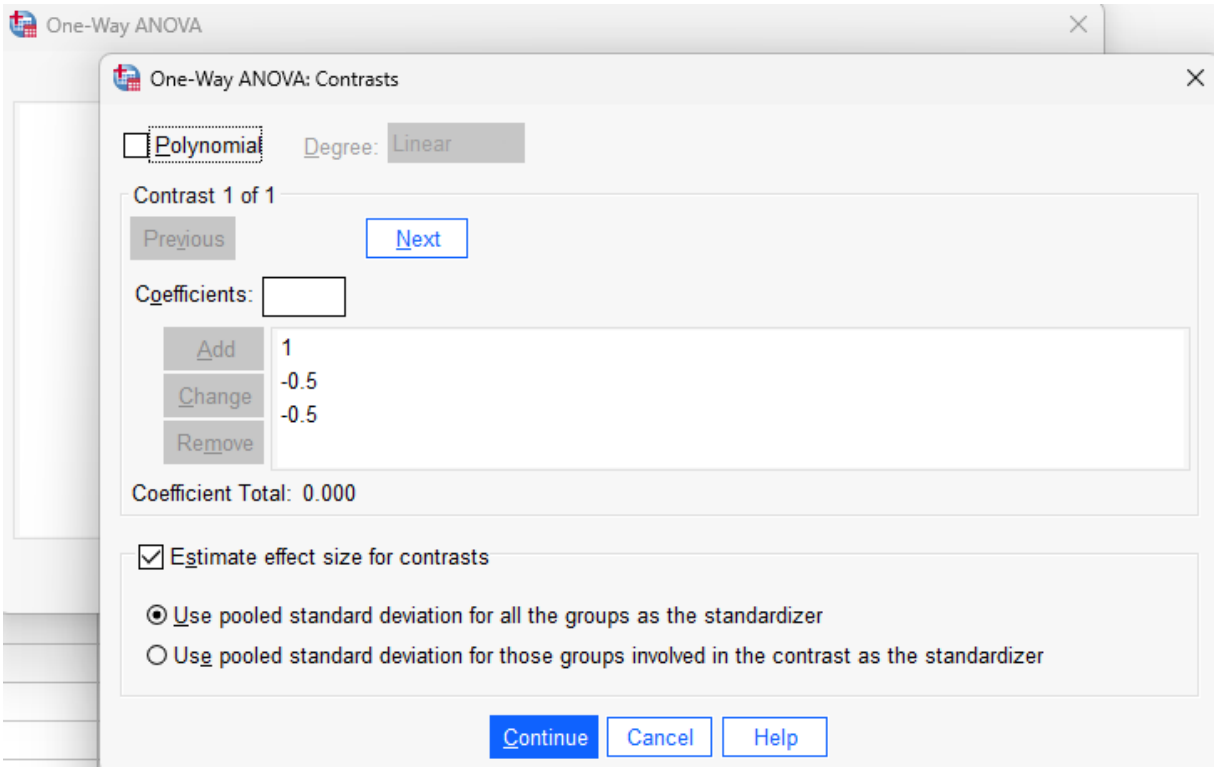


Abbildung 6.9. Vergleich eines Gruppenmittelwerts mit dem Mittelwert aus beiden anderen.

Die Ausgabe ist in Abbildung 6.10 gezeigt. Wir sehen, dass die plausiblen Werte gemäß des 95%-KIs für den Mittelwertsunterschied zwischen der Population junger Erwachsener und den beiden Populationen älterer Erwachsener und Erwachsener mittleren Alters im Bereich [-11.38, -2.08] liegen.

Contrast Coefficients

	Altersgruppe (1=18-35 Jahre, 2=36-60 Jahre, 3=älter als 60 Jahre)		
Contrast	junge Erwachsene	Erwachsene mittleren Alters	ältere Erwachsene
1	1	-.5	-.5

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	95% Confidence Interval Lower	95% Confidence Interval Upper
Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)	Assumes equal variances	1	-6.73	2.259	-2.981	177	.003	-11.19	-2.28
	Does not assume equal variances	1	-6.73	2.346	-2.870	106.902	.005	-11.38	-2.08

Contrast Effect Sizes

		Contrast	Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Wert bei Becks Depressionsinventar (Zahl zwischen 0 und 63)	Cohen's d	1	14.287	-.471	-.784	-.157
	Hedges' correction	1	14.348	-.469	-.781	-.156

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation for all the groups.
Hedges' uses pooled standard deviation for all the groups, plus a correction factor.

Abbildung 6.10. Ausgabe für Fragestellung (b).

Ergebnisbericht für a-priori Kontraste

Für Fragestellung (a) aus dem vorhergehenden Abschnitt könnte ein Ergebnisbericht so aussehen: „Deskriptive Statistiken für das Depressionsniveau in den betrachteten drei Altersgruppen sind in Tabelle 6.1 angegeben. Das Depressionsniveau junger Erwachsener ist (mit $\alpha = .05$) signifikant niedriger als das von Erwachsenen mittleren Alters, $t(116.24) = 1.86$, $p = .033$, Cohens $d = 0.34$ mit 95%-KI [-0.02, 0.70]. Das Depressionsniveau von Erwachsenen mittleren Alters ist hingegen nicht signifikant niedriger als das von älteren Erwachsenen, $t(117.95) = 1.45$, $p = .075$, Cohens $d = 0.25$ mit 95%-KI [-0.11, 0.61]. Beide Unterschiede entsprechen gemäß Cohens Heuristik (1988) einem kleinen Effekt."

Für Fragestellung (b) könnte ein Ergebnisbericht wie folgt aussehen: „Die plausiblen Werte gemäß des 95%-KIs für den Betrag, um den das mittlere Depressionsniveau junger Erwachsener niedriger ist als der Durchschnitt des mittleren Depressionsniveaus in den Altersgruppen der Erwachsenen mittleren Alters und der älteren Erwachsenen, liegen zwischen 2.08 und 11.38."

Übungsaufgaben

Die Datendateien, die Sie für manche der folgenden Übungsaufgaben benötigen, finden Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Beispiel 6.1

Was gehört zu den Voraussetzungen der einfaktoriellen Varianzanalyse ohne Messwiederholung?

(a) Die UV muss mindestens intervallskaliert sein.

(b) Die AV muss mindestens intervallskaliert sein.

(c) Die Varianz der AV muss in allen Populationen dieselbe sein.

(d) Die einzelnen Messungen müssen abhängig voneinander sein.

Beispiel 6.2

Was gehört zu den Voraussetzungen der einfaktoriellen Varianzanalyse ohne Messwiederholung?

(a) Die AV muss in der Grundgesamtheit normalverteilt sein, kann aber in den einzelnen Populationen von einer Normalverteilung abweichen.

(b) Die einzelnen Messungen müssen unabhängig voneinander sein.

(c) Die Varianz der AV muss sich zwischen den Populationen unterscheiden.

(d) Es muss Homoskedastizität vorliegen.

Beispiel 6.3

Was gehört zu den Vorteilen von a-priori Vergleichen (gegenüber post-hoc Vergleichen)?

(a) Eine vorhergehende Durchführung eines Omnibus-Tests ist nicht notwendig.

(b) Es können ungerichtete Hypothesen formuliert werden.

(c) A-priori Vergleiche kontrollieren die FWER strenger.

(d) Das Ergebnis der statistischen Testung ist häufig informativer als ein Omnibus-Test.

Beispiel 6.4

Geben Sie für jede der folgenden Aussagen an, ob sie richtig oder falsch ist.

Nr.	**Aussage**	**R/F**
1)	Gemäß Cohens Heuristik (1988) wird ein $\eta^2 = 0.4$ als kleiner Effekt bezeichnet.	
2)	Eine Effektstärke für die einfaktorielle ANOVA heißt *f* und kann aus η^2 berechnet werden. Diese Berechnung kann auch in G*Power durchgeführt werden.	
3)	Fishers least-significant-difference (LSD) Test hat für den Fall einer einfaktoriellen Varianzanalyse ohne Messwiederholung für drei Gruppen eine höhere Teststärke als Tukeys honestly-significant-difference (HSD) Test und ist diesem daher vorzuziehen.	
4)	Falls die Voraussetzung der Varianzhomogenität nicht erfüllt ist, kann anstelle einer einfaktoriellen Varianzanalyse ohne Messwiederholung eine Varianzanalyse nach Welch gerechnet werden.	
5)	Die Voraussetzung der Normalverteilung der AV ist wichtiger als die Voraussetzung der Varianzgleichheit für einfaktorielle Varianzanalysen ohne Messwiederholung.	
6)	Bei η^2 zwischen 0.5 und 0.8 spricht man gemäß Cohens Heuristik (1988) von einem mittleren Effekt.	

Beispiel 6.5

Eine einfaktorielle Varianzanalyse ohne Messwiederholung wird üblicherweise durchgeführt, um die Frage zu erhellen, ob sich mehrere Gruppenmittelwerte voneinander unterscheiden. D.h. insbesondere, dass es sich auch nur um zwei Gruppenmittelwerte handeln kann. Wiederholen Sie Übungsaufgabe 5.6, verwenden Sie aber dieses Mal eine Varianzanalyse, um zu ermitteln, ob Männer im Kurs „Anwendung statistischer Verfahren am Computer“ signifikant größer sind als Frauen. Verwenden Sie ein Signifikanzniveau von $\alpha = .005$ und berichten Sie Ihre Ergebnisse gemäß APA-Richtlinien. Sie finden die entsprechenden Daten in der Datei „Kap3daten.sav“.

Beispiel 6.6

Eine Forscherin möchte untersuchen, ob textuelle Informationen in digitalen Lernspielumgebungen leichter verarbeitet werden können, wenn diese schriftlich dargestellt oder gesprochen werden. Um diese Fragestellung zu untersuchen, rekrutiert die Forscherin 172 Versuchspersonen und weist diese randomisiert entweder der Gruppe „Schrift“ oder der Gruppe „Sprache“ zu. In der Gruppe „Schrift“ werden lernspielrelevante Texte am Bildschirm schriftlich dargestellt. In der Gruppe „Sprache“ werden dieselben Informationen von einem professionellen Sprecher eingesprochen und dann durch entsprechende Sprachaufzeichnungen im Lernspiel vermittelt. Einen Tag, nachdem sich die Versuchspersonen mit dem Lernspiel befasst haben, absolvieren sie einen Test zu den Inhalten des Lernspiels, bei dem Sie zwischen 0 und 100 Punkte erreichen können. Die Testergebnisse und Gruppenzugehörigkeiten sind in der Datei „Kap5UE15.sav“ zu finden. Ermitteln Sie mittels einer Varianzanalyse ohne Messwiederholung, ob sich die beiden entsprechenden Populationen hinsichtlich der Testergebnisse im Mittel unterscheiden und berichten Sie Ihre Resultate gemäß APA-Richtlinien.

Beispiel 6.7

Verwenden Sie G*Power, um folgende Frage zu beantworten: Für wie viele Personen müssen bei gleicher Aufteilung auf drei Gruppen Daten erhoben werden, wenn mit einer Irrtumswahrscheinlichkeit von 0.5% und einer Teststärke von 80% ein gemäß Cohen (1988) kleiner Effekt von $\eta^2 = .01$ detektiert werden soll?

Beispiel 6.8

Verwenden Sie die Datendatei „Kap3daten.sav“ für diese Aufgabe, um die folgende Frage zu beantworten. Gibt es Unterschiede in der Abneigung gegenüber Statistikprüfungen (erfasst durch die Variable *statistikschmerzen*) in Abhängigkeit vom bevorzugten Schulfach (Variable *hauptfach*)? Falls ja, führen Sie geeignete post-hoc Vergleiche durch, um diese Unterschiede genauer zu charakterisieren. Verwenden Sie ein Signifikanzniveau von $\alpha = .05$ für diese Aufgabe und korrigieren Sie p-Werte für Ihre post-hoc Vergleiche entsprechend. Berichten Sie Ihre Ergebnisse entsprechend APA-Richtlinien.

Beispiel 6.9

Verwenden Sie wiederum die Datendatei aus der vorherigen Aufgabe, d.h. „Kap3daten.sav“. Prüfen Sie nun aber die folgenden beiden Hypothesen mit angemessen definierten a-priori Vergleichen:

(a) Der Mittelwert für die Abneigung gegenüber Statistikprüfungen (wieder Variable *statistikschmerzen*) von Mathematik-affinen Studierenden (d.h. Studierenden mit Lieblingshauptfach Mathematik) ist niedriger als der Mittelwert für die Abneigung gegenüber Statistikprüfungen von Sprach-affinen Studierenden (d.h. Lieblingshauptfach entweder Englisch oder Deutsch).

(b) Der Mittelwert für die Abneigung gegenüber Statistikprüfungen (wieder Variable *statistikschmerzen*) von Deutsch-affinen Studierenden (d.h. Studierenden mit Lieblingshauptfach Deutsch) ist niedriger als der Mittelwert für die Abneigung gegenüber Statistikprüfungen von Englisch-affinen Studierenden (d.h. Lieblingshauptfach Englisch).

Formulieren Sie einen geeigneten Ergebnisbericht gemäß APA-Richtlinien.

Beispiel 6.10

Verwenden Sie für diese Übung die Datei „Sales.sav“. Der Datensatz enthält u.a. die Verkaufszahlen (in tausenden von Alben; Variable *Sales*), die Häufigkeit, mit der die entsprechende Musik im Radio gespielt wird (Variable *Airplay*), sowie die Attraktivität (Variable *Attr_Group*) von 200 verschiedenen Bands. Die Attraktivität ist dabei in den Kategorien 1 = „ugly“, 2 = „average“ und 3 = „beautiful“ gegeben. Bei dem Datensatz handelt es sich um eine adaptierte Version eines (fiktiven) Datensatzes, der von Andy Field für sein berüchtigtes Statistiklehrbuch „Discovering Statistics Using IBM SPSS Statistics“ (Field, 2024) erstellt wurde. Den Originaldatensatz finden Sie in der Datei „Album Sales.sav“, die Sie von der Webseite für Fields Buch https://edge.sagepub.com/field5e/student-resources/datasets herunterladen können

Versuchen Sie mit einem entsprechenden statistischen Verfahren die Frage zu beantworten, ob Bands unterschiedlicher Attraktivität unterschiedlich viele Alben verkaufen. Für den Fall, dass sich ein signifikanter Unterschied für die Mittelwerte der drei Kategorien ergibt, prüfen Sie alle paarweisen Vergleiche auf statistische Signifikanz mittels Fishers LSD-Test.

Fassen Sie Ihre Resultate in einem entsprechenden Ergebnisbericht gemäß APA-Richtlinien zusammen.

Beispiel 6.11

Wiederholen Sie die vorhergehende Übung 6.10, aber verwenden Sie dieses Mal sowohl den Bonferroni-Test sowie Tukeys HSD-Test im Rahmen der post-hoc Vergleiche. Vergleichen Sie die Resultate und erläutern Sie Unterschiede zu Fishers LSD-Test im vorhergehenden Beispiel.

Beispiel 6.12

Für die unten angegebene Fragestellung hat ein Freund, der Sie um Hilfe bei einer Statistikaufgabe bittet, bereits eine entsprechende Analyse in SPSS durchgeführt und einen Ergebnisbericht erstellt. Ihre Aufgabe besteht darin, die erhaltenen Ergebnisse zu überprüfen und gegebenenfalls zu korrigieren.

Fragestellung: In einer Studie wurde überprüft, wie gut bestimmte Therapieformen bzw. Kontrollbedingungen zur Behandlung von bestimmten Essstörungen geeignet sind. Dazu wurde die über den Zeitraum der Therapie erzielte Gewichtszunahme (in kg) für 4 Therapieformen bzw. Kontrollbedingungen verglichen: Kognitive Verhaltenstherapie (KVT = Code 1), lösungsfokussierte Kurzzeittherapie (LKT = Code 2), sowie als Kontrollbedingungen ein sogenanntes treatment as usual (TAU = Code 3) und keine Behandlung (KB = Code 4). Die konkreten Fragestellungen lauteten:

1) Führen die beiden Therapieformen zu einer größeren Gewichtszunahme als die beiden Kontrollbedingungen?
2) Gibt es jeweils innerhalb der Therapieformen und innerhalb der Kontrollbedingungen Unterschiede in der erzielten Gewichtszunahme?

Um diese Fragestellung zu untersuchen hat Ihr Freund eine einfaktorielle Varianzanalyse für unabhängige Stichproben durchgeführt, geeignete a-priori Vergleiche definiert und inferenzstatistisch untersucht. Als Signifikanzniveau wurde $\alpha = .05$ für jeden der Vergleiche gewählt. Im Anschluss erstellte er den unten folgenden Ergebnisbericht.

Dieser Ergebnisbericht ist leider teilweise fehlerhaft. Markieren und korrigieren Sie die Fehler. Die Daten zur Aufgabe befinden sich in der Datei „Kap6UE12.sav“.

Ergebnisbericht: Die Stichprobe umfasste insgesamt 200 Personen. Der erste Kontrast verglich die beiden Therapien mit den beiden Kontrollbedingungen. Es zeigte sich, dass die beiden Therapien zu weniger Gewichtszunahme führten als die beiden Kontrollbedingungen ($t(928.74) = 7.13$, $p < .001$, $d = 1.01$; d.h. gemäß Cohen (1988) ein großer Effekt). Zwischen den beiden Therapieformen gab es keinen signifikanten Unterschied zwischen den mittleren Gewichtszunahmen für die KVT-Gruppe ($M = 5.90$, $SD = 2.66$) und die LKT-Gruppe ($M = 4.52$, $SD = 3.05$; $t(96.19) = 2.40$, $p = .018$, $d = 0.47$; d.h. gemäß Cohen (1988) ein großer Effekt). Auch innerhalb der Kontrollbedingungen fand sich ein signifikanter Unterschied zwischen den mittleren Gewichtszunahmen der TAU-Gruppe ($M = 3.20$, $SD = 2.73$) und der KB-Gruppe ($M = 1.37$, $SD = 3.15$; $t(96.06) = 3.12$, $p = .020$, $d = 0.63$; d.h., ein mittlerer Effekt gemäß Cohen(1988)).

Beispiel 6.13

Eine Forschungsgruppe untersucht die Wirksamkeit unterschiedlicher psychotherapeutischer Ansätze und vergleicht dafür Psychoanalyse, Verhaltenstherapie und eine Kontrollbedingung (tau = treatment as usual) bei einer bestimmten Form von Zwangsstörungen. Dazu werden 120 geeignete Versuchspersonen rekrutiert, die dann zufällig auf die drei Therapien aufgeteilt werden. Ein halbes Jahr nach Therapiebeginn wird bei jeder Person die Minderung der Zwangssymptomatik mit einem geeigneten psychometrischen Instrument auf einer Skala von -50 bis +50 erhoben.

Verwenden Sie die in der Datei „Kap6UE13.sav“ gegebenen Daten, um mit einem geeigneten statistischen Verfahren die folgenden beiden Hypothesen zu prüfen: (a) Die beiden therapeutischen Ansätze wirken im Mittel besser (d.h. reduzieren die Symptomatik stärker) als die Kontrollbedingung; (b) die Verhaltenstherapie wirkt im Mittel besser als die Psychoanalyse. Verfassen Sie anschließend einen entsprechenden Ergebnisbericht.

Beispiel 6.14

Eine Forschungsgruppe fragt sich, ob die Statistikangst von Studienanfänger:innen davon abhängt, welchen Schultyp diese besucht haben. Daher erhebt die Forschungsgruppe mit einem geeigneten psychometrischen Verfahren die Statistikangst von 225 Studienanfänger:innen und erhebt auch deren Schultyp. Die Schultypen werden in drei Kategorien eingeteilt: Schwerpunkt: Sprachen; Schwerpunkt: Naturwissenschaft und Technik; Schwerpunkt: Kunst & Design.

Verwenden Sie die in der Datei „Kap6UE14.sav“ gegebenen Daten, um mit einem geeigneten statistischen Verfahren die folgende Hypothese zu prüfen: Die mittlere Statistikangst von Absolvent:innen von Schulen mit Schwerpunkt Naturwissenschaft und Technik ist niedriger als die mittlere Statistikangst von Absolvent:innen der beiden anderen Schultypen. Verfassen Sie anschließend einen entsprechenden Ergebnisbericht.

Beispiel 6.15

Eine Forschungsgruppe fragt sich, ob die Statistikangst von Studienanfänger:innen davon abhängt, welchen Schultyp diese besucht haben. Daher erhebt die Forschungsgruppe mit einem geeigneten psychometrischen Verfahren die Statistikangst von 300 Studienanfänger:innen und erhebt auch deren Schultyp. Die Schultypen werden in vier Kategorien eingeteilt: Schwerpunkt: Sprachen; Schwerpunkt: Naturwissenschaft und Technik; Schwerpunkt: Kunst & Design; Schwerpunkt: Sport.

Verwenden Sie die in der Datei „Kap6UE15.sav“ gegebenen Daten, um mit einem geeigneten statistischen Verfahren die folgenden Hypothesen zu prüfen: (i) Die mittlere Statistikangst von Absolvent:innen von Schulen mit Schwerpunkt Naturwissenschaft und Technik ist niedriger als die mittlere Statistikangst von Absolvent:innen der Schultypen mit den Schwerpunkten Sprachen und Kunst & Design; (ii) die mittlere Statistikangst von Absolvent:innen von Schulen mit Schwerpunkt Sport unterscheidet sich von der mittleren Statistikangst von Absolvent:innen der Schultypen mit Schwerpunkten Sprachen und Kunst & Design. Verfassen Sie anschließend einen entsprechenden Ergebnisbericht.

Kapitel 7

Zweifaktorielle Varianzanalyse ohne Messwiederholung

Stefan E. Huber

Als ob es nicht schon kompliziert genug wäre, sich mit der einfaktoriellen Varianzanalyse ohne Messwiederholung zu befassen, werden wir uns in diesem Kapitel nun auch noch mit der *zwei*faktoriellen Varianzanalyse ohne Messwiederholung beschäftigen. Allerdings haben wir dafür, was den konzeptuellen Hintergrund betrifft, im letzten Kapitel das Gröbste schon gut vorbereitet. Daher werden wir in diesem Kapitel die Darstellung der Grundkonzepte darauf beschränken, noch einmal zu rekapitulieren wie das Vorgehen der einfaktoriellen Varianzanalyse auf mehrere Faktoren, insbesondere zwei, erweitert werden kann. Für den Hauptteil des Kapitels werden wir anschließend den Schwerpunkt auf Durchführungsaspekte in SPSS legen.

Zweifaktorielles varianzanalytisches Modell

Das zweifaktorielle varianzanalytische Modell ist durch die folgende Gleichung gegeben:

$$Y_{ijk} \sim N\left(\mu + \Delta\mu_j + \Delta\mu_k + \Delta\mu_{jk}, \sigma^2\right),$$

mit $i = 1, \ldots, n$, $j = 1, \ldots, m$ und $k = 1, \ldots, q$, wobei n wieder dem Umfang der gesamten Stichprobe entspricht, m der Anzahl der Stufen bzw. untersuchten Populationen des ersten Faktors, und q der Anzahl der Stufen des zweiten Faktors. Hat z.B. der erste Faktor zwei Stufen und der zweite Faktor drei Stufen, so würden insgesamt 2x3 = 6 Populationen untersucht werden. Zur Erinnerung (vorhergehendes Kapitel): Man würde in diesem Fall also von einer zweifaktoriellen Varianzanalyse ohne Messwiederholung mit einem 2 x 3 Design sprechen.

Aus der obigen Modellspezifikation geht ferner hervor, dass wiederum angenommen wird, dass die AV in jeder untersuchten Population durch eine identisch und unabhängig normalverteilte Zufallsvariable approximiert werden kann, mit eventuell je nach Stufe der beiden Faktoren unterschiedlichem Populationsmittelwert, aber jeweils derselben Varianz σ^2. D.h. insbesondere, dass die Messwerte in den jeweiligen Gruppen ausschließlich durch Populationsmittelwert und Varianz bestimmt sind, woraus die Bedingung folgt, dass zwischen den einzelnen Gruppen bzw. Populationen

keine Abhängigkeiten bestehen. Wenn also z.B. Messwerte für ein und dieselbe Person in mehreren der Populationen vorhanden wären, würde das diese Modellvoraussetzung verletzen. Dasselbe wäre der Fall, wenn z.B. in einer Gruppe die Messwerte jeweils aller Brüder und in einer anderen Gruppe die Messwerte jeweils aller Schwestern von Geschwisterpaaren vorliegen würden. Ein weiterer Fall, in dem das Modell nicht gültig wäre, wäre gegeben, wenn die Gruppen Messwerte zu ein und derselben Person zu verschiedenen Zeitpunkten enthalten würden. In all diesen Fällen würden abhängige (oder „verbundene“) Stichproben vorliegen und wir müssten mit einem varianzanalytischen Modell mit Messwiederholung arbeiten (siehe nächstes Kapitel).

Die Voraussetzungen für das zweifaktorielle varianzanalytische Modell sind demgemäß alle durch das oben angegebene Modell spezifiziert. Im Einzelnen werden sie im nächsten Abschnitt noch einmal zusammengefasst und es wird kurz wiederholt wie sie jeweils in SPSS überprüft werden können.

Voraussetzungen für das zweifaktorielle varianzanalytische Modell

Die Voraussetzungen für das zweifaktorielle varianzanalytische Modell lauten wie folgt:

- Intervallskalenniveau der AV.
- Unabhängigkeit der Messungen bzw. Beobachtungen, d.h. insbesondere keine Abhängigkeiten zwischen den Gruppen.
- Normalverteilung der AV in jeder Gruppe.
- Gleichheit der Varianzen der AV in allen Gruppen (auch bekannt als Varianzhomogenität oder Homoskedastizität).

Die ersten beiden dieser Voraussetzungen sind wiederum durch das experimentelle bzw. messtheoretische Design festgelegte und können im Rahmen der Datenanalyse nicht mehr überprüft werden. Die Normalverteilung in den unterschiedlichen Gruppen kann prinzipiell in SPSS über *Analyze >> Descriptive Statistics >> Explore…* überprüft werden, siehe Kapitel 5. Dafür ist es allerdings nötig, die Überprüfung für jede mögliche Kombination der Stufen beider Faktoren durchzuführen, was durch Aufteilung der Datendatei *Data >> Split File…* und dortiger Eingabe beider Faktorvariablen realisiert werden kann. Eine Anleitung dafür ist auch unter https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics-2.php zu finden. Wie schon bei der einfaktoriellen Varianzanalyse

erwähnt, ist die Varianzanalyse aber gegenüber der Verletzung der Normalverteilungsvoraussetzung relativ robust (zumindest, wenn nicht gleichzeitig Heteroskedastizität vorliegt). Kritischer ist hingegen die Prüfung auf Varianzgleichheit, welche wiederum im Rahmen der Durchführung der Varianzanalyse in SPSS angefordert werden kann (siehe unten). Sollte diese Voraussetzung verletzt sein, empfiehlt sich jedenfalls die Verwendung eines robusteren Verfahrens zur Hypothesenprüfung (siehe z.B. Mair & Wilcox, 2020). Darauf wird allerdings in diesen Übungen nicht weiter eingegangen.

Omnibustests im zweifaktoriellen varianzanalytischen Modell

Zur statistischen Testung von Unterschieden zwischen Populationsmittelwerten lassen sich für das zweifaktorielle varianzanalytische Modell drei verschiedene Omnibustests durchführen. Prinzipiell wird in jedem dieser Omnibustests wieder das Verhältnis zweier Varianzen gebildet, die unter Geltung der jeweiligen Nullhypothese wiederum beide Schätzungen der unbekannten Varianz σ^2 darstellen. D.h. unter Geltung der Nullhypothese ergibt sich ein Verhältnis der beiden Varianzschätzungen nahe 1. Insbesondere ist dieses Verhältnis unter Geltung der Nullhypothese wieder F-verteilt (wobei sich die beiden Freiheitsgrade wieder aus der Stichprobengröße und den Anzahlen der untersuchten Gruppen bzw. Faktorstufen ergeben), woraus wiederum folgt, dass es nur selten den Wert 1 sehr weit übersteigt (und nach unten mit Null begrenzt ist). Ergibt sich also ein Testwert, der sehr weit (nach oben) vom unter der Nullhypothese erwarteten Wert von 1 abweicht, kann die Nullhypothese aufgrund der üblichen Argumentation als unplausibel abgelehnt werden. Als Entscheidungskriterium kann dafür auch wieder ein p-Wert berechnet und mit einem vorab gewählten Signifikanzniveau verglichen werden. All das erledigt netterweise SPSS für uns und, was die Durchführung anbelangt, müssen wir lediglich wissen, wie wir eine entsprechende Analyse ausführen und wo wir die einzelnen Informationen, die wir zur Entscheidungsfindung benötigen, finden können.

Bevor wir uns diesem Vorgehen zuwenden, rekapitulieren wir aber noch einmal, um welche drei Hypothesentests es sich bei diesen Omnibustests denn nun spezifisch handelt. Der erste dieser Hypothesentests testet die Nullhypothese $H_0: \Delta\mu_j = 0 \; \forall \; j = 1, \dots, m$ wobei das Symbol „$\forall$“ als „für alle“ zu lesen ist. Die entsprechende Alternativhypothese lautet $H_1: \exists j: \Delta\mu_j \neq 0$ wobei hier „$\exists$“ als „es gibt“ zu lesen ist. Die Alternativhypothese bedeutet, dass es mindestens einen Unterschied zwischen

den Populationsmittelwerten für die Faktorstufen j des ersten Faktors gibt, während die Nullhypothese bedeutet, dass es keinen solchen Unterschied gibt. Wird die Nullhypothese verworfen, sagt man auch, dass ein Haupteffekt für den ersten Faktor vorliegt. Die Freiheitsgrade für die F-Verteilung sind für dieses Hypothesenpaar zu $\nu_1 = m - 1$ und $\nu_2 = n - m \cdot q$ gegeben.

Der zweite dieser Hypothesentests testet analog die Nullhypothese $H_0: \Delta\mu_k = 0 \; \forall \; k = 1, \dots, q$. Die entsprechende Alternativhypothese lautet $H_1: \exists k: \Delta\mu_k \neq 0$. Die Alternativhypothese bedeutet, dass es mindestens einen Unterschied zwischen den Populationsmittelwerten für die Faktorstufen k des zweiten Faktors gibt, während die Nullhypothese bedeutet, dass es keinen solchen Unterschied gibt. Wird die Nullhypothese verworfen, sagt man auch, dass ein Haupteffekt für den zweiten Faktor vorliegt. Die Freiheitsgrade für die F-Verteilung sind für dieses Hypothesenpaar zu $\nu_1 = q - 1$ und $\nu_2 = n - m \cdot q$ gegeben.

Sobald es sich um ein mindestens zweifaktorielles Untersuchungsdesign handelt, sind auch sogenannte Interaktionseffekte zu berücksichtigen. Bei einer Interaktion handelt es sich schlichtweg um den Fall, dass die Auswirkung eines Faktors auf die AV von der Ausprägung des anderen Faktors abhängt. Ein prägnantes Beispiel für eine Interaktion findet sich bei Oswald Huber (2019, S. 160, Hervorhebungen im Original): „Wenn beispielsweise ein Fremder in das persönliche Territorium eines Menschen eindringt (z.B. in dessen persönliches Büro), dann hängt die Reaktion dieses Menschen meist vom Verhalten des Eindringlings ab: Klopft der Eindringling vor dem Öffnen der Türe an, fragt er, ob er eintreten darf, grüßt er höflich, dann wird der *Territoriumsinhaber* in der Regel nicht unfreundlich reagieren. Tut der Eindringling all das nicht, muss er mit einem unfreundlichen Empfang rechnen. In diesem Fall wirken die beiden Variablen *Eindringen in fremdes Territorium* und *Beschwichtigendes Verhalten des Eindringlings* nicht unabhängig voneinander, sondern sie interagieren. Welche Wirkung eintritt, hängt von beiden Variablen gemeinsam ab."

Der Omnibustest für die Interaktion testet die Nullhypothese $H_0: \Delta\mu_{jk} = 0 \; \forall \; j = 1, \dots, m \; \wedge \; \forall \; k = 1, \dots, q,$ wobei hier „$\wedge$" die logische Operation UND bezeichnet. Die entsprechende Alternativhypothese lautet $H_1: \exists \; (j, k): \; \Delta\mu_{jk} \neq 0$. Inhaltlich bedeutet das Vorliegen einer Interaktion, dass ein einzelner Populationsmittelwert sich nicht additiv aus den Haupteffekten der beiden Faktoren

und dem Gesamtmittelwert zusammensetzt. Dies geht damit einher, dass eine Interaktion (oder auch Wechselwirkung genannt) dann vorliegt, wenn der Verlauf des Einflusses eines Faktors auf die AV sich über die Stufen des jeweils anderen Faktors hinweg unterscheidet. Die Freiheitsgrade für die F-Verteilung sind für die Testung der Interaktion zu $\nu_1 = (m-1)(q-1)$ und $\nu_2 = n - m \cdot q$ gegeben.

Liegt eine Interaktion vor, kann man die Einflüsse der beiden Faktoren im Allgemeinen nicht getrennt voneinander betrachten. So kann es z.B. sein, dass kein Haupteffekt eines Faktors vorliegt, man aber nicht sagen kann, dass dieser keinen Einfluss auf die AV hat, siehe Abbildung 7.1 links. Genauso kann es sein, dass ein Haupteffekt eines Faktors auf die AV vorliegt, dieser aber nicht für jede Faktorstufe des anderen Faktors einen Einfluss auf die AV hat, siehe Abbildung 7.1 rechts. Weitere solcher Fälle werden im Folgenden noch an praktischen Beispielen illustriert.

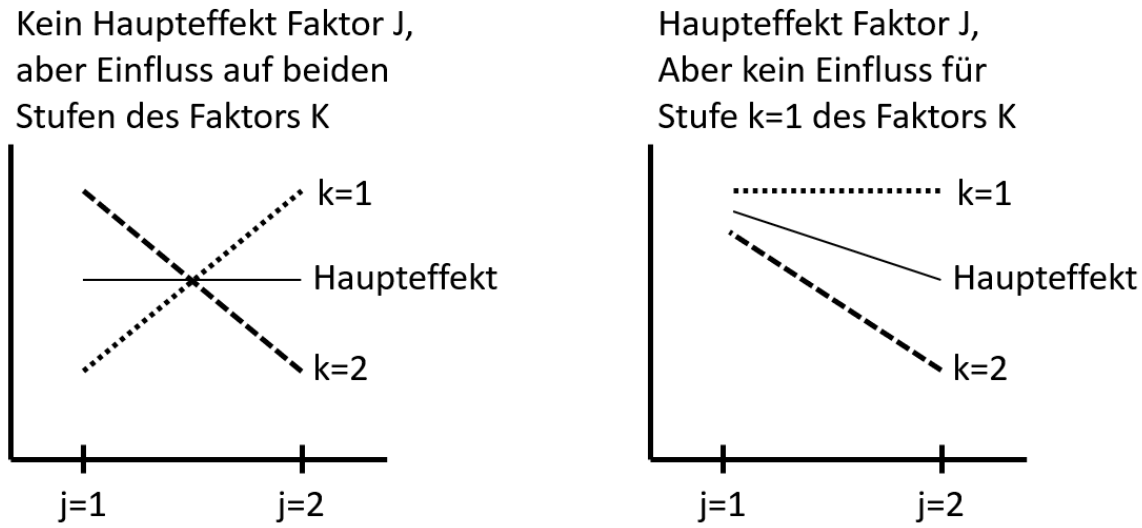


Abbildung 7.1. Illustrationen, weshalb Haupteffekte im Allgemeinen nicht mehr zu interpretieren sind, wenn eine Interaktion vorliegt.

Zweifaktorielle Varianzanalyse ohne Messwiederholung mit 2x2 Design in SPSS

Die Durchführung einer zweifaktoriellen Varianzanalyse ohne Messwiederholung mit einem 2x2 Design wird an dem Datensatz in der Datendatei „Kraft.sav“ illustriert, die Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Gegeben sind in diesem Datensatz Messungen der Körperkraft auf einer Skala von 0-100 für 109 männliche und weibliche Versuchspersonen aus zwei Altersgruppen (unter 50 und über 50). Die Fragestellung, für die diese (fiktiven) Daten (genauso fiktiv) erhoben wurden, lautete: Wie wirken sich Alter und Geschlecht auf die Körperkraft aus?

Um diese Fragestellung zu beantworten, wird eine zweifaktorielle Varianzanalyse ohne Messwiederholung mit den beiden Faktoren Geschlecht und Altersgruppe durchgeführt. Beide Faktoren haben jeweils 2 Stufen. Zur Durchführung in SPSS wählen wir zuerst *Analyze* >> *General Linear Model* >> *Univariate…* und ziehen anschließend die Variable Kraft in das Feld „Dependent Variable“ und die Variablen *Geschlecht* und *Alter_Gruppe* in das Feld „Fixed Factor(s)“, siehe Abbildung 7.2.

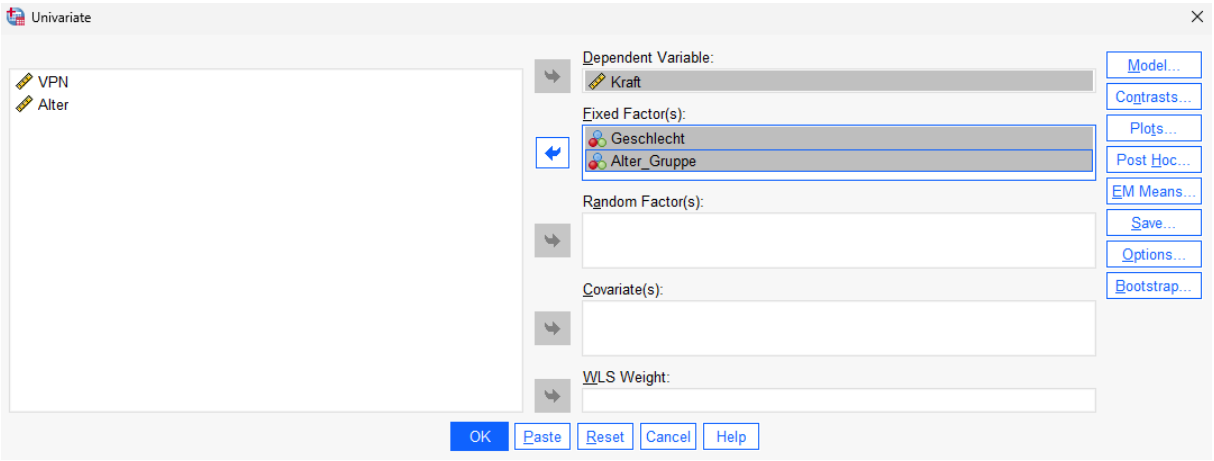


Abbildung 7.2. Durchführung einer zweifaktoriellen Varianzanalyse ohne Messwiederholung mit SPSS.

Anschließend öffnen wir das Menü „Plots…“ und ziehen dort die Variable *Geschlecht* in das Feld „Horizontal Axis“ und die Variable *Alter_Gruppe* in das Feld „Separate Lines“. Genauso könnten wir auch die Variable *Geschlecht* in das Feld „Separate Lines“ und die Variable *Alter_Gruppe* in das Feld „Horizontal Axis“ ziehen. Beide Darstellungsformen sind völlig äquivalent. Bei komplexeren

Designs (mit mehr Stufen) ist aber manchmal eine der beiden Formen einleuchtender, weshalb es oftmals bequem ist, sich einfach beide ausgeben zu lassen und dann hinterher herauszufinden, welche einfacher zu interpretieren ist. Zu Illustrationszwecken machen wir das auch in diesem Beispiel so (auch wenn es hier keinerlei Vorteile bringt). Zusätzlich wählen wir noch aus, dass uns 95%-KI für die Mittelwerte angezeigt werden sollen, siehe Abbildung 7.3. Danach klicken wir auf „Continue“.

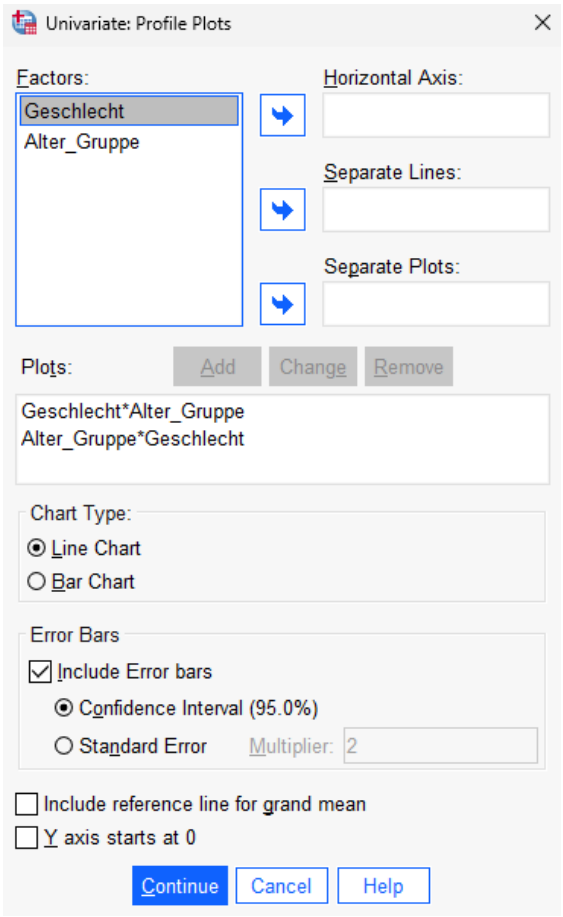


Abbildung 7.3. Ein Bild sagt oft mehr als tausend Worte.

Danach gehen wir noch ins Menü „Options…“ um hier jedenfalls „Descriptive statistics“, „Homogeneity tests“ und „Estimates of effect size “ anzufordern, siehe Abbildung 7.4. Danach klicken wir wieder auf „Continue“, dann auf „Paste“, dokumentieren die sich öffnende Syntaxdatei entsprechend und führen schließlich die gerade eingefügten Kommandozeilen aus. Das generiert eine relativ umfangreiche Ausgabe, die wir im Folgenden Schritt für Schritt bzw. Tabelle für Tabelle besprechen. Die Ausgabe ist hier nicht (vollständig) wiedergegeben. Das heißt, um die folgende Beschreibung nachvollziehen zu können, wird empfohlen die Analyse erst in SPSS auszuführen, um die entsprechenden Tabellen einsehen zu können.

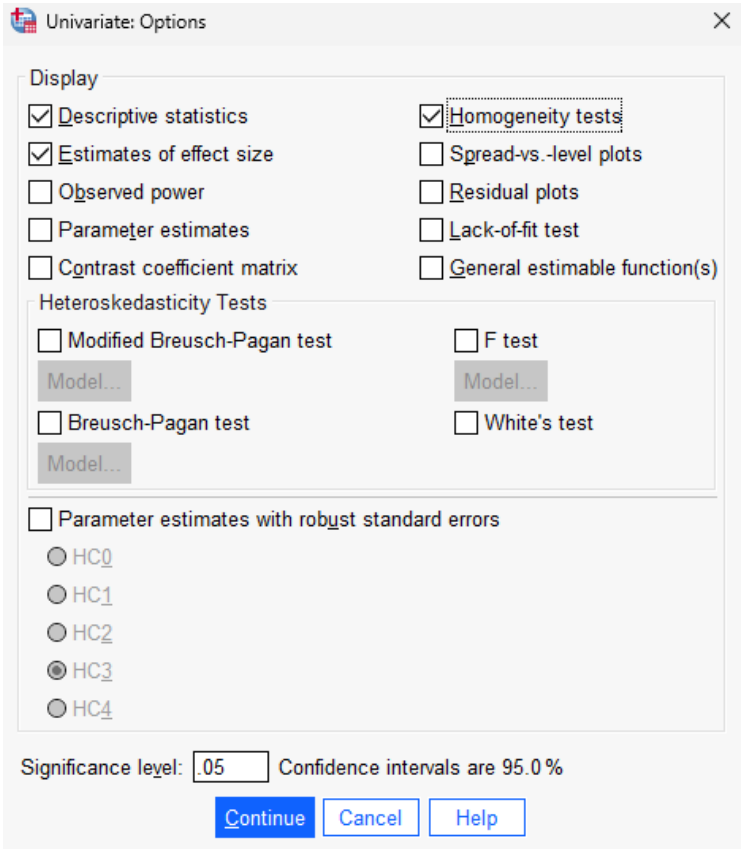


Abbildung 7.4. Optionen für unsere zweifaktorielle Varianzanalyse ohne Messwiederholung.

In der Tabelle „Between-Subjects Factors“ finden wir eine Übersicht zu den Stichprobenumfängen für die Stufen unserer beiden Faktoren. Wir sehen, dass wir Daten für 60 Männer und 49 Frauen vorliegen haben. Ferner sehen wir, dass insgesamt 52 Personen unter 50 Jahre alt waren, und 57 älter.

In der nächsten Tabelle mit der Überschrift „Descriptive Statistics“ sehen wir, wie sich diese Versuchspersonen genau in die 2x2 = 4 Stichproben aufteilen (letzte Spalte). Zudem sind wieder Mittelwerte und Standardabweichungen für jede Untergruppe, aber auch für alle übergeordneten Gruppen angegeben.

In der Tabelle „Levene’s Test of Equality of Error Variances“ finden wir wieder Ergebnisse für unterschiedliche Levenes Tests. Da wir uns wieder für Mittelwertsunterschiede interessieren, ist für uns hier besonders die erste Zeile „Based on Mean“ interessant. Dort sehen wir in der Spalte „Sig.“ den p-Wert für den Levenes Test, an dem wir erkennen, dass dieser nicht signifikant ist (d.h. > .05), $p = .256$. Wir entscheiden daher von Varianzhomogenität auszugehen.

In der Tabelle „Tests of Between-Subjects Effects“, siehe auch Abbildung 7.5, bekommen wir schließlich die Resultate unserer eigentlichen Varianzanalyse. In der Zeile „Geschlecht“ können wir die Freiheitsgrade $\nu_1 = 1$ in der Spalte „df“, den F-Wert 36.03, den p-Wert < .001, sowie die Effektstärke η_p^2 = .26 für unsere Variable *Geschlecht* ablesen. Es liegt also ein signifikanter Haupteffekt für die Variable *Geschlecht* vor. Bei der Effektstärke handelt es sich um das sogenannte partielle Eta-Quadrat. Jeder Faktor (sowie die Interaktion) kann einen gewissen Anteil der Varianz in der AV aufklären, während der restliche Teil unerklärt bleibt. Das partielle Eta-Quadrat entspricht dem Verhältnis des Anteils, der durch den betrachteten Faktor erklärt wird, und diesem Anteil zusammen mit dem unerklärten Varianzanteil. Das partielle Eta-Quadrat bemisst also wie groß der durch den Faktor erklärte Anteil relativ zum unerklärten Anteil der Varianz der AV ist. Daneben gibt es auch das Eta-Quadrat (ohne „partiell“), das schlichtweg den Anteil der Gesamtvarianz der AV angibt, der durch den betrachteten Faktor erklärt werden kann. Dieses könnten wir uns selbst aus den Quadratsummen in der Spalte „Type III Sum of Squares“ berechnen. Beide Effektstärken sind in der Literatur immer wieder anzutreffen. Für diese Übungen beschränken wir uns allerdings auf das partielle Eta-Quadrat. Auch für dieses existieren wieder Heuristiken nach Cohen (1988). Wie schon im vorhergehenden Kapitel für η^2 im Rahmen einfaktorieller Varianzanalysen werden auch für η_p^2 Werte zwischen 0.01 und 0.06 als klein, Werte zwischen 0.06 und 0.14 als mittel, und Werte ab 0.14 als groß bezeichnet.

In der Zeile „Alter_Gruppe“ finden wir die entsprechenden Ergebnisse für unsere Variable *Alter_Gruppe*. Auch hier ist ν_1 = 1, der F-Wert entspricht 31.70, der p-Wert ist kleiner als 0.001 und damit auch signifikant, die Effektstärke ist mit η_p^2 = .23 vergleichbar groß zum Faktor Geschlecht. Auch für diesen Faktor liegt also ein signifikanter Haupteffekt vor.

In der Zeile „Geschlecht*Alter_Gruppe“ finden wir schließlich unsere Ergebnisse zur Interaktion. Auch für diese ist ν_1 = 1 (da $(m-1)(q-1) = (2-1)(2-1) = 1 \cdot 1 = 1$), der F-Wert ist allerdings mit 0.39 sehr klein, der p-wert mit 0.54 entsprechend nicht signifikant, die Effektstärke vernachlässigbar (Effektstärken unterhalb der Kategorie „klein“ werden in der Literatur häufig als vernachlässigbar bezeichnet).

In der Zeile „Error“ finden wir schließlich noch den Wert für die Freiheitsgrade ν_2 = 105.

Tests of Between-Subjects Effects

Dependent Variable: Kraft von 0-100

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	18827.443[a]	3	6275.814	20.952	<.001	.374
Intercept	204360.707	1	204360.707	682.258	<.001	.867
Geschlecht	10793.207	1	10793.207	36.033	<.001	.255
Alter_Gruppe	9496.288	1	9496.288	31.703	<.001	.232
Geschlecht * Alter_Gruppe	116.245	1	116.245	.388	.535	.004
Error	31451.272	105	299.536			
Total	263420.000	109				
Corrected Total	50278.716	108				

a. R Squared = .374 (Adjusted R Squared = .357)

Abbildung 7.5. Wesentlicher Teil der Ausgabe für unsere zweifaktorielle Varianzanalyse.

In Abbildung 7.6 sehen wir auch sehr schön, was es bedeutet, wenn lediglich zwei Haupteffekte aber keine Interaktion vorliegt. Jüngere Versuchspersonen sind stärker als ältere, männliche sind stärker als weibliche. Beides gilt jeweils unabhängig vom anderen Faktor, weshalb auch die beiden Linien in Abbildung 7.6 nahezu parallel zueinander sind.

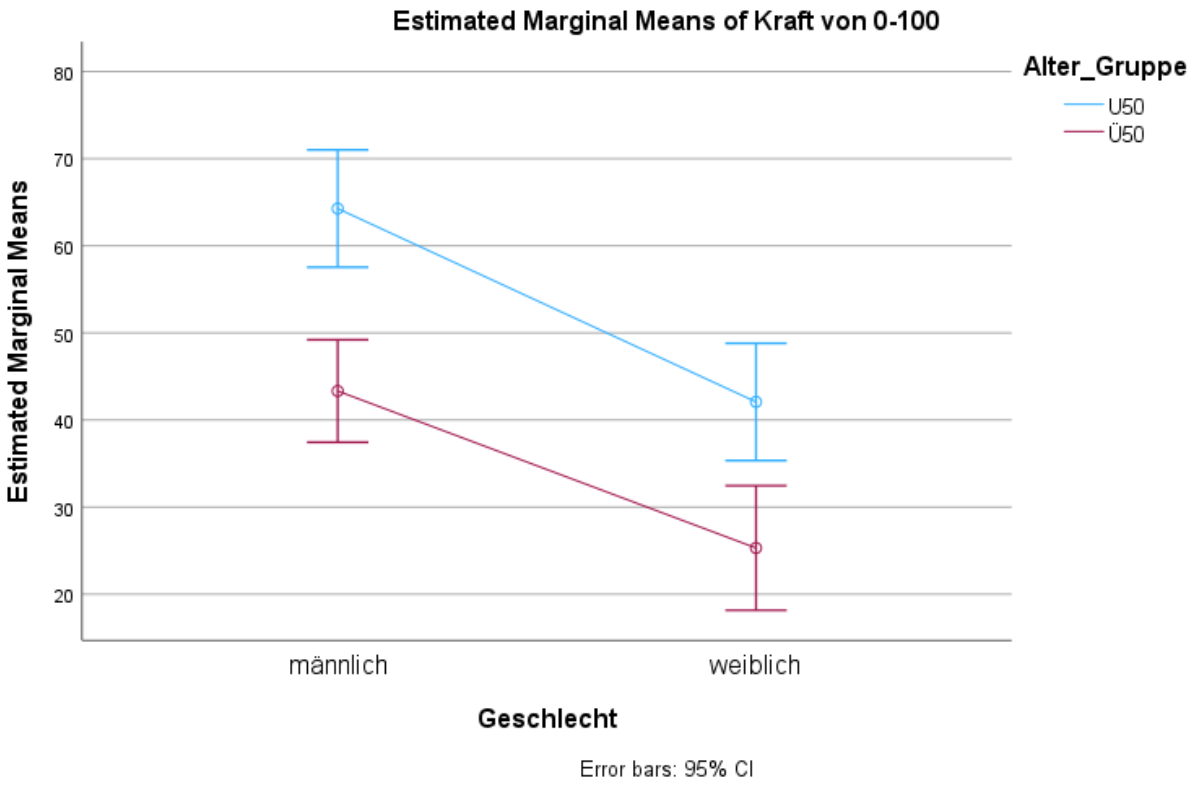


Abbildung 7.6. Grafischer Vergleich der Mittelwerte unserer vier Stichproben. Fehlerbalken entsprechen 95%-KI.

Aus diesem Grund (keine Interaktion, lediglich Haupteffekte) werden häufig in Ergebnisberichten schlichtweg die Verhältnisse der in Abbildung 7.6 gezeigten Randmittel berichtet, wobei aber die Angabe der Stichprobenmittelwerte für alle vier Stichproben zumindest in Form einer entsprechenden Tabelle dennoch empfohlen wird. Schließlich geben die Randmittel (d.h. lediglich der Vergleich der Kraft zwischen Männern und Frauen bzw. zwischen Jüngeren und Älteren) keinen Einblick in die Verhältnisse, die zwischen je zwei der 2x2=4 spezifischen Stichproben bestehen. Spezifische Einzelvergleiche dieser 4 Untergruppen sind in SPSS im Rahmen von post-hoc Vergleichen im Nachgang zu den drei soeben besprochenen Omnibustests möglich. Damit werden wir uns im nächsten Beispiel näher befassen. Davor schauen wir uns allerdings noch ein Beispiel für einen möglichen Ergebnisbericht für die soeben erläuterten Ergebnisse an.

Ergebnisbericht

Ein Ergebnisbericht für dieses Beispiel könnte wie folgt aussehen: „Sowohl das Alter ($F(1,105) = 31.70$, $p < .001$, $\eta_p^2 = .23$) als auch das Geschlecht ($F(1,105) = 36.03$, $p < .001$, $\eta_p^2 = .26$) haben (mit $\alpha = .005$) einen signifikanten Einfluss auf das Ausmaß an Körperkraft. Zwischen Alter und Geschlecht besteht keine signifikante Wechselwirkung ($F(1,105) = 0.39$, $p = .535$, $\eta_p^2 < .01$). Personen jünger als 50 Jahre ($M = 53.17$, $SD = 20.99$) haben im Mittel mehr Kraft als Personen älter als 50 Jahre ($M = 36.05$, $SD = 18.82$), und Männer ($M = 52.40$, $SD = 20.40$) haben im Mittel mehr Kraft als Frauen ($M = 32.40$, $SD = 18.70$). Deskriptive Statistiken für alle untersuchten Stichproben sind in Tabelle 7.1 angegeben. Abbildung 7.6 zeigt einen grafischen Vergleich der Mittelwerte."

Tabelle 7.1

Deskriptive Statistiken

Altersgruppe	Geschlecht	*M*	*SD*	*n*
Unter 50	Männlich	64.27	19.81	26
	Weiblich	42.08	15.81	26
Über 50	Männlich	43.32	15.84	34
	Weiblich	25.30	17.98	23

Beispiel mit Wechselwirkung, 2x3 Design

Zur Illustration eines 2x3 Designs verwenden wir einen weiteren fiktiven Datensatz, der ursprünglich auf Andy Field (2024) zurückgeht, wenn er auch in der aktuellsten Version seines Buchs nicht mehr in dieser Form vorkommt. Sie finden den Datensatz in der Datendatei „musikgeschmack.sav“, die Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

In dieser Datei finden wir Angaben von insgesamt 90 Personen dazu, wie gut ihnen Musik der Bands bzw. Interpreten Nirvana, AC/DC oder Bon Jovi gefällt. Positive Zahlen drücken Gefallen aus, negative Zahlen Missfallen und je größer der Betrag der Zahl, desto höher Gefallen oder Missfallen. Die Personen wurden dabei zufällig aus zwei Altersgruppen ausgewählt: 45 Personen sind älter als 40 Jahre, 45 Personen sind maximal 40 Jahre alt. Die Forschungsfrage lautet: Unterscheidet sich der Musikgeschmack dieser beiden Altersgruppen?

Eine Möglichkeit diese Frage auf Basis dieses Datensatzes zu erhellen, besteht in der Durchführung einer zweifaktoriellen Varianzanalyse ohne Messwiederholung. Dazu wählen wir erst einmal wieder *Analyze >> General Linear Model >> Univariate…* aus und ziehen die Variable *Gefallen* in das Feld „Dependent Variable“ und die Variablen *Musik* und *Altersgruppe* in das Feld „Fixed Factor(s)“. Daraufhin fordern wir unter „Plots…“ wieder zwei verschiedene Grafiken an, einmal mit der Variable *Musik* auf der horizontalen Achse und der Variable *Altersgruppe* durch unterschiedliche Linien dargestellt, und einmal umgekehrt. Dieses Mal werden wir auch gut erkennen können, wie die beiden Darstellungen denselben Sachverhalt unterschiedlich darstellen. Wir verlangen auch wieder, dass Fehlerbalken dargestellt werden, die 95%-KI entsprechen sollen. Danach wählen wir unter „Options…“ auch wieder „Descriptive statistics“, „Homogeneity tests“ sowie „Estimates of effect size“ aus.

Post-hoc Vergleiche

Schließlich fordern wir post-hoc Vergleiche an, indem wir auf „EM Means…“ (und *nicht* auf „Post Hoc…“) klicken. „EM Means“ steht hier für „Estimated Marginal Means“ und bezeichnet den Vergleich der Randmittel, d.h. jener Mittelwerte, die für die einzelnen Faktoren bzw. Faktorstufen ohne Berücksichtigung der jeweils anderen Faktoren bzw. Faktorstufen gebildet werden. In dem sich

öffnenden Fenster ziehen wir die Variablen Kombination *Musik*Altersgruppe* von links in das Feld „Display Means For…“ und wählen anschließend noch „Compare simple main effects“ aus. Für die Korrektur der p-Werte haben wir die Wahl zwischen Fishers LSD (keine Korrektur), Bonferroni und Sidak. Da für die Sidak-Korrektur die zu testenden Hypothesen unabhängig sein müssen (für eine exakte Korrektur der FWER), was im Allgemeinen nicht der Fall ist, wählen wir Bonferroni, da für diese Korrektur die FWER jedenfalls nicht unterschätzt wird (d.h. sie ist bei einem gewünschten α von 0.5% sicher nicht größer als dieser Wert). Die in diesem Untermenü vorgenommenen Einstellungen sind noch einmal in Abbildung 7.7 zusammengefasst.

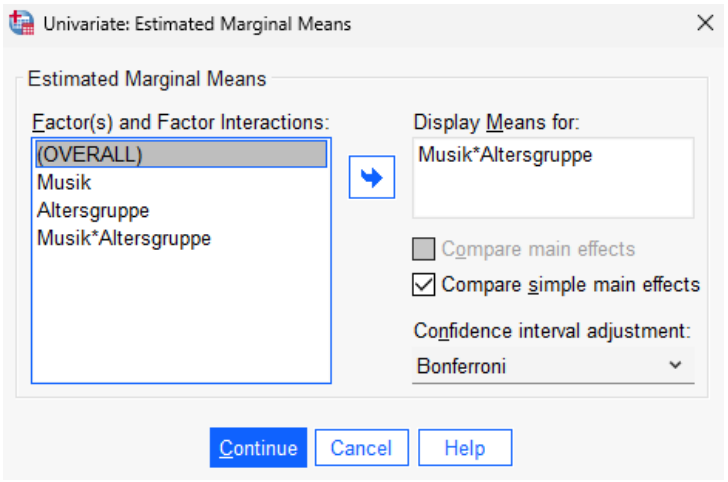


Abbildung 7.7. Anforderung von post-hoc Vergleichen im Nachgang einer zweifaktoriellen Varianzanalyse.

Ergebnisse

Nach dem Einfügen in die Syntax und Ausführen der entsprechenden Kommandozeilen bekommen wir wieder eine umfangreiche Ausgabe. An der Tabelle „Levene’s Test of Equality of Error Variances“ sehen wir, dass Levenes Test auch in diesem Beispiel nicht signifikant ist und freuen uns, da daher nicht noch mehr als ohnehin schon zu tun bleibt. In der Tabelle „Tests of Between-Subjects Effects“ erkennen wir einen signifikanten Haupteffekt für die Variable Musik sowie eine signifikante Interaktion. Wir finden dort auch wieder alle Zahlen, die wir für einen ordnungsgemäßen Ergebnisbericht brauchen, siehe nächster Abschnitt.

Ab der Überschrift „Estimated Marginal Means“ finden wir schließlich alle Ergebnisse, die für unsere angeforderten post-hoc Vergleiche relevant sind und noch einiges mehr. Bei letzterem handelt es sich z.B. schon um die verschiedenen Randmittel und deren Konfidenzintervalle, die in der Tabelle „Estimates“ angegeben sind, siehe Abbildung 7.8. An diesen ließen sich einige Unterschiede zwischen den Altersgruppen sofort ablesen. Gleichzeitig wären die hier gegebenen Punktschätzungen und Intervallschätzungen auch sehr gut brauchbar, wenn man schlichtweg an einer Schätzung der jeweiligen Populationsmittelwerte interessiert ist und gar nicht unbedingt an paarweisen Tests der Mittelwertsunterschiede. Am Standardfehler in der Tabelle erkennt man auch, dass es sich bei diesem um eine Schätzung mittels der gepoolten Varianz aus allen Stichproben handelt, da der Wert für alle Stichproben derselbe ist. Zu beachten ist bei dieser Tabelle schließlich noch, dass es sich bei dem Konfidenzniveau der Konfidenzintervalle nicht um ein korrigiertes Konfidenzniveau (zur Kontrolle der FWER) handelt.

Estimates

Dependent Variable: Angabe wie gut die Musik gefällt

Interpret	Altersgruppe (1=älter, 2=jünger)	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Nirvana	40+	-75.867	5.083	-85.975	-65.759
	0-40	66.200	5.083	56.092	76.308
AC/DC	40+	59.933	5.083	49.825	70.041
	0-40	64.133	5.083	54.025	74.241
Bon Jovi	40+	74.267	5.083	64.159	84.375
	0-40	-71.467	5.083	-81.575	-61.359

Abbildung 7.8. Punkt- und Intervallschätzungen der einzelnen Populationsmittelwerte für alle sechs Stichproben.

Zu beachten bei den Ergebnissen der angeforderten post-hoc Vergleiche ist auch, dass diese für eine zweifaktorielle Varianzanalyse in zweifacher Ausführung kommen. In der ersten Ausführung, die bei der Überschrift „1. Interpret * Altersgruppe (1=älter,2=jünger)“ beginnt, werden zuerst einfaktorielle Varianzanalysen für jeweils jede Stufe der Variable *Altersgruppe* für den Faktor *Musik* durchgeführt (zu finden allerdings in der Tabelle ganz am Schluss dieses Abschnitts). Diese Ergebnisse finden sich in der Tabelle „Univariate Tests“ und wir sehen, dass sich der Gefallen an der jeweiligen Musik sowohl in der jüngeren als auch der älteren Gruppe zwischen den drei Interpreten unterscheidet, siehe auch Abbildung

7.9. In der Tabelle „Pairwise Comparisons“, siehe auch Abbildung 7.10, finden wir schließlich alle paarweisen Tests auf Mittelwertsunterschiede für beide Altersgruppen. Hier sehen wir u.a., dass der älteren Gruppe Nirvana im Mittel signifikant weniger gefällt als AC/DC und Bon Jovi, welche der älteren Gruppe ähnlich gut gefallen (die Mittelwerte unterscheiden sich auch nicht signifikant), während der jüngeren Gruppe Bon Jovi im Mittel signifikant weniger gefällt als die beiden anderen Interpreten, welche wiederum der jüngeren Gruppe ähnlich gut gefallen (auch hier unterscheiden sich die beiden Gruppenmittelwerte nicht signifikant).

Univariate Tests

Dependent Variable: Angabe wie gut die Musik gefällt

Altersgruppe (1=älter,2=jünger)		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
40+	Contrast	205935.511	2	102967.756	265.695	<.001	.864
	Error	32553.467	84	387.541			
0-40	Contrast	186718.711	2	93359.356	240.902	<.001	.852
	Error	32553.467	84	387.541			

Each F tests the simple effects of Interpret within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

Abbildung 7.9. Einfaktorielle Varianzanalysen für jede Stufe des Faktors *Altersgruppe*.

Pairwise Comparisons

Dependent Variable: Angabe wie gut die Musik gefällt

Altersgruppe (1=älter, 2=jünger)	(I) Interpret	(J) Interpret	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
40+	Nirvana	AC/DC	-135.800*	7.188	<.001	-153.360	-118.240
		Bon Jovi	-150.133*	7.188	<.001	-167.693	-132.573
	AC/DC	Nirvana	135.800*	7.188	<.001	118.240	153.360
		Bon Jovi	-14.333	7.188	.148	-31.893	3.227
	Bon Jovi	Nirvana	150.133*	7.188	<.001	132.573	167.693
		AC/DC	14.333	7.188	.148	-3.227	31.893
0-40	Nirvana	AC/DC	2.067	7.188	1.000	-15.493	19.627
		Bon Jovi	137.667*	7.188	<.001	120.107	155.227
	AC/DC	Nirvana	-2.067	7.188	1.000	-19.627	15.493
		Bon Jovi	135.600*	7.188	<.001	118.040	153.160
	Bon Jovi	Nirvana	-137.667*	7.188	<.001	-155.227	-120.107
		AC/DC	-135.600*	7.188	<.001	-153.160	-118.040

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Abbildung 7.10. Alle paarweisen Tests für Unterschiede zwischen Gruppenmittelwerten für jede Stufe des Faktors *Altersgruppe*.

In der zweiten Ausführung, deren Ergebnisse ab „2. Interpret * Altersgruppe (1=älter,2=jünger)“ aufgeführt sind, werden zuerst (auch wenn die entsprechende Tabelle wieder erst ganz am Ende dieses Abschnitts zu finden ist) einfaktorielle Varianzanalysen für jede Stufe der Variablen *Musik* für den Faktor *Altersgruppe* durchgeführt, siehe Abbildung 7.11. Da es nur zwei Altersgruppen gibt, handelt es sich hierbei schlichtweg um Vergleiche der Mittelwerte für die beiden Altersgruppen auf jeder Stufe der Variable *Musik*. Wir sehen, dass sich die Altersgruppen für die Interpreten Nirvana und Bon Jovi signifikant unterscheiden und für AC/DC nicht. Allerdings erkennen wir in dieser Tabelle nicht die Richtung des Unterschieds und haben auch keine Konfidenzintervalle für die Mittelwertsunterschiede. Diese finden wir nur in der Tabelle mit allen paarweisen Unterschieden (diese Konfidenzintervalle sind nun für multiple Vergleiche korrigiert), siehe Abbildung 7.12. Aber wir erkennen an der Gleichheit der p-Werte in diesen beiden Tabellen durchaus, dass es sich bei den F-Tests und den t-Tests für den Vergleich von zwei Mittelwerten um zwei äquivalente Testverfahren handelt.

Univariate Tests

Dependent Variable: Angabe wie gut die Musik gefällt

Interpret		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Nirvana	Contrast	151372.033	1	151372.033	390.596	<.001	.823
	Error	32553.467	84	387.541			
AC/DC	Contrast	132.300	1	132.300	.341	.561	.004
	Error	32553.467	84	387.541			
Bon Jovi	Contrast	159286.533	1	159286.533	411.018	<.001	.830
	Error	32553.467	84	387.541			

Each F tests the simple effects of Altersgruppe (1=älter,2=jünger) within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

Abbildung 7.11. Einfaktorielle Varianzanalysen für jede Stufe des Faktors *Musik*.

Pairwise Comparisons

Dependent Variable: Angabe wie gut die Musik gefällt

Interpret	(I) Altersgruppe (1=älter, 2=jünger)	(J) Altersgruppe (1=älter, 2=jünger)	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
Nirvana	40+	0-40	-142.067*	7.188	<.001	-156.361	-127.772
	0-40	40+	142.067*	7.188	<.001	127.772	156.361
AC/DC	40+	0-40	-4.200	7.188	.561	-18.495	10.095
	0-40	40+	4.200	7.188	.561	-10.095	18.495
Bon Jovi	40+	0-40	145.733*	7.188	<.001	131.439	160.028
	0-40	40+	-145.733*	7.188	<.001	-160.028	-131.439

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Abbildung 7.12. Alle paarweisen Tests für Unterschiede zwischen Gruppenmittelwerten für jede Stufe des Faktors *Musik*.

Ergebnisbericht

Ein Ergebnisbericht für eine zweifaktorielle Varianzanalyse mit einem 2x3 Design inkl. post-hoc Vergleichen könnte beispielsweise folgendermaßen aussehen: „Zur Erhellung der Fragestellung wurde eine zweifaktorielle Varianzanalyse ohne Messwiederholung durchgeführt. Levenes Test war (mit α = .05) nicht signifikant (p = .322), weshalb von Varianzhomogenität ausgegangen wurde.

Jüngere (bis 40 Jahre) und ältere (ab 41 Jahre) Personen unterscheiden sich im Mittel nicht signifikant darin, wie sehr ihnen die untersuchte Musik insgesamt gefällt (F(1,84) < 0.01, p = .966, η_p^2 < .01). Die drei Bands Nirvana, AC/DC und Bon Jovi werden aber im Mittel signifikant unterschiedlich bewertet (F(2,84) = 105.62, p < .001, η_p^2 = .72). Zudem besteht für die im Mittel resultierende Bewertung eine signifikante Interaktion zwischen dem Alter der bewertenden Person und der gehörten Band (F(2,84) = 400.98, p < .001, η_p^2 = .91).

Zur weiteren Analyse paarweiser Mittelwertsunterschiede wurden paarweise post-hoc Vergleiche mit einer Korrektur der p-Werte für multiple Vergleiche gemäß Bonferroni durchgeführt. Im Folgenden werden lediglich korrigierte p-Werte berichtet. Aus diesen ergab sich, dass Nirvana von jüngeren Personen signifikant lieber gehört wird als von älteren Personen (p < .001). Bon Jovi hingegen wird von älteren signifikant lieber gehört als von jüngeren Personen (p < .001). AC/DC wird nicht signifikant unterschiedlich gerne gehört (p = .561).

Innerhalb der Altersgruppen ergaben sich folgende Unterschiede. Jüngere Personen hören sowohl Nirvana als auch AC/DC signifikant lieber als Bon Jovi (jeweils p < .001). Nirvana und AC/DC werden von jüngeren Personen nicht signifikant unterschiedlich gerne gehört (p > .999). Ältere Personen bevorzugen Bon Jovi und AC/DC signifikant gegenüber Nirvana (jeweils p < .001). Bon Jovi und AC/DC werden von älteren Personen nicht signifikant unterschiedlich gerne gehört (p = .148).

Die resultierenden deskriptiven Statistiken sind in Tabelle 7.2 zusammengestellt. Abbildung 7.13 zeigt einen Vergleich der geschätzten Randmittelwerte und deren 95%-KI für die unterschiedlichen Bands und Altersgruppen.“

Da dieser Ergebnisbericht bereits sehr umfangreich ausfällt, bietet es sich an, kurz darüber zu reflektieren, was die grundsätzlichen Bestandteile eines entsprechenden Ergebnisberichts sind und in

welcher Reihenfolge sie erläutert werden sollten. Im ersten Teil wird die Methode der Auswertung und die Überprüfung etwaiger Voraussetzungen kurz erläutert. Gerade bei komplexeren Analysen kann dies wichtig sein, weil dann nicht unbedingt mehr klar aus den angegebenen Teststatistiken, Freiheitsgraden etc. ersichtlich ist, um welches Verfahren es sich an welcher Stelle handelt. In einem zweiten Teil werden die Ergebnisse der drei Omnibustests für die beiden Haupteffekte und die Interaktion berichtet. In einem dritten Teil werden die durchgeführten post-hoc Vergleiche erläutert. Dabei genügt es meist, die verschiedenen paarweisen Vergleiche und die entsprechenden p-Werte anzugeben. Eine Angabe über die Korrektur der letzteren ist aber für eine angemessene Interpretation derselben notwendig. Schließlich sollten auch die deskriptiven Statistiken dargestellt werden. Bei komplexeren Analysen ist dafür eine Tabelle meist zweckdienlich. Eine graphische Darstellung kann zudem die Bedeutung der Ergebnisse oft klarer erhellen als das bloße Auflisten von Zahlen. Dabei erlauben Konfidenzintervalle auch einen guten Überblick über plausible Werte für einzelne Populationsmittelwerte.

Tabelle 7.2

Deskriptive Statistiken

Altersgruppe	Interpret	*M*	*SD*	*n*
> 40 Jahre	Nirvana	-75.87	14.37	15
	AC/DC	59.93	19.98	15
	Bon Jovi	74.27	22.29	15
0-40 Jahre	Nirvana	66.20	19.90	15
	AC/DC	64.13	17.00	15
	Bon Jovi	-71.47	23.18	15

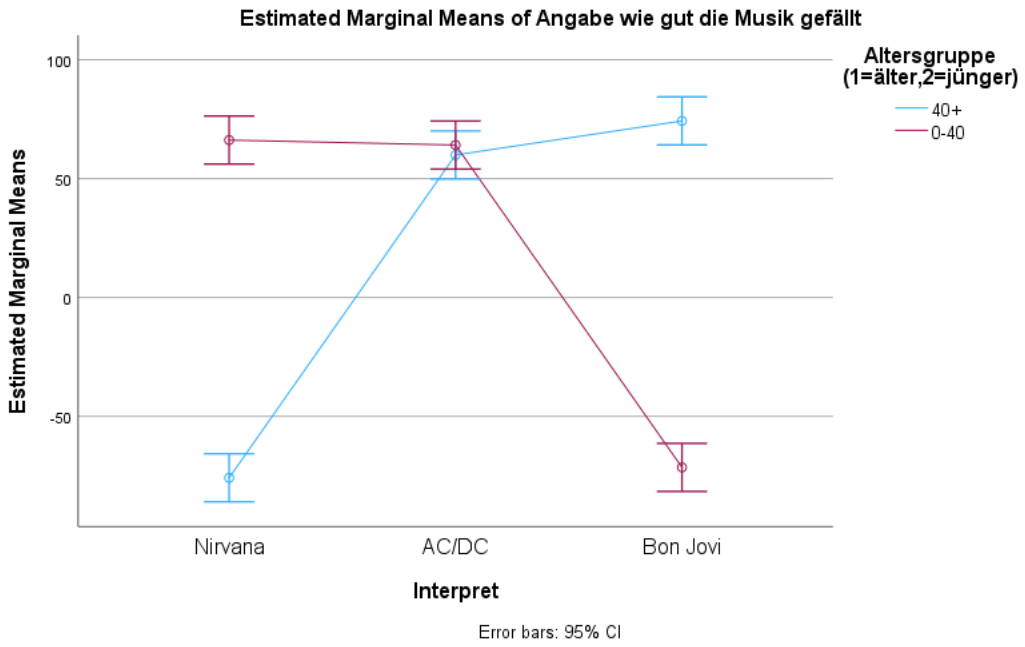


Abbildung 7.13. Visuelle Darstellung der unterschiedlichen Bewertung der drei Interpreten durch die beiden Altersgruppen.

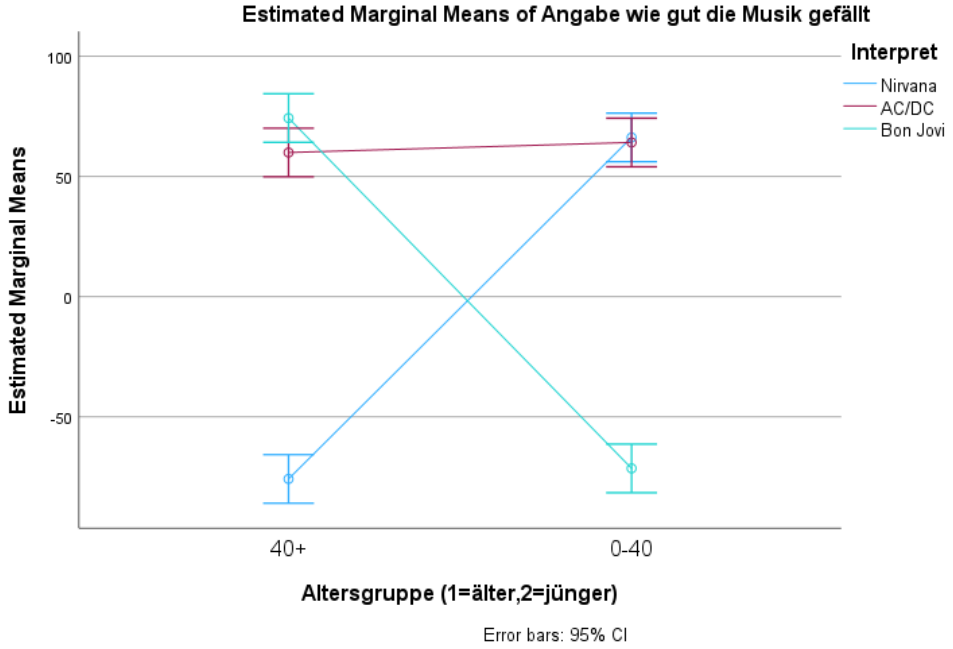


Abbildung 7.14. Alternative Darstellung mit den beiden Altersgruppen links und rechts und den drei Interpreten durch verschiedene Linien.

Stichprobenplanung

Zwar kann auch für mehrfaktorielle Varianzanalysen eine Stichprobenplanung in G*Power durchgeführt werden, diese wird aber nicht im Detail besprochen, da die statistischen Hypothesen der Omnibustests nur selten relevante inhaltliche Forschungsfragen beantworten können (Bühner et al., 2025).

Soll dennoch eine Stichprobenplanung durchgeführt werden, so ist unter „Test family“ wiederum „F tests“ auszuwählen. Unter „Statistical test“ ist „ANOVA: Fixed effects, special, main effects and interactions“ auszuwählen. Unter “Type of power analysis” ist wieder “A priori: Compute required sample size – given α, power, and effect size” auszuwählen. Die Effektstärke f für den interessierenden Effekt (also einen der Haupteffekte oder die Interaktion) kann am besten mittels der Schaltfläche „Determine“ aus η_p^2 direkt in G*Power berechnet werden. Bei „Number of groups“ ist neben der obligatorischen Angabe des gewünschten Signifikanzniveaus und der Teststärke schließlich noch die Anzahl der Stichproben anzugeben, die untersucht werden sollen. Im Falle eines 2 x 2 Designs wären das also 4 Stichproben, im Falle eines 2 x 3 Designs 6 Stichproben usw. Schließlich müssen noch die Zählerfreiheitsgrade („numerator df“) angegeben werden. Dabei handelt es sich um das Produkt der Anzahl der Stufen minus 1 aller Faktoren. D.h., bei einem 2 x 2 Design wäre hier $(2 – 1)(2 – 1) = 1$ einzutragen, bei einem 2 x 3 Design $(2 – 1)(3 – 1) = 2$ usw. Der Grund dafür, dass diese Angabe zusätzlich zur Anzahl der Stichproben gemacht werden muss, liegt darin, dass es Designs mit derselben Anzahl an Gruppen, aber unterschiedlichen Zählerfreiheitsgraden geben kann, z.B. 2 x 2 x 3 (2 Zählerfreiheitsgrade) und 2 x 6 (5 Zählerfreiheitsgrade). Beide Informationen werden aber zur Berechnung des F-Werts und daher auch für die Stichprobenplanung benötigt, daher reicht nur die Angabe der Anzahl der Stichproben nicht aus.

Übungsaufgaben

Beispiel 7.1

Worin unterscheiden sich die Voraussetzungen für eine mehrfaktorielle Varianzanalyse von jenen für eine einfaktorielle Varianzanalyse?

Beispiel 7.2

Welche Voraussetzungen müssen für eine mehrfaktorielle Varianzanalyse erfüllt sein?

Beispiel 7.3

Welche Möglichkeiten gibt es im Rahmen von paarweisen post-hoc Vergleichen nach einer zweifaktoriellen Varianzanalyse in SPSS, um p-Werte und Konfidenzniveaus für den Interaktionseffekt für multiple Vergleiche zu korrigieren?

(a) Bonferroni-Korrektur.
(b) Fishers Least-Significant-Difference Test.
(c) Sidak-Korrektur.
(d) Tukeys Honestly-Significant-Difference Test.

Beispiel 7.4

Welche Aussage(n) trifft(treffen) zu?

(a) Unter einem Haupteffekt versteht man die Auswirkung des einen Faktors auf die Wirkung des anderen Faktors auf die AV.
(b) Gemäß Cohens Heuristik (1988) gelten Effektstärken η_p^2 zwischen 0.01 und 0.06 als klein, zwischen 0.06 und 0.14 als mittel, und ab 0.14 als groß.
(c) Es kann sein, dass es keinen Haupteffekt für einen Faktor gibt, dieser aber trotzdem eine Auswirkung auf die AV auf einzelnen Stufen eines anderen Faktors hat.
(d) Es kann sein, dass es einen Haupteffekt für einen Faktor gibt, dieser aber auf einzelnen Stufen eines anderen Faktors keine Auswirkung auf die AV hat.

Beispiel 7.5

In Kapitel 14 (vielleicht in neueren Auflagen nicht in Kapitel 14, aber jedenfalls im Kapitel zu mehrfaktoriellen Varianzanalysen ohne Messwiederholung) des Buchs „Discovering Statistics Using IBM SPSS Statistics“ von Andy Field (2024) findet man das folgende Beispiel (Übersetzung d. Verf.):

„Eine Anthropologin interessierte sich für die Auswirkungen von Alkohol auf die Partner:innenwahl in Nachtclubs. Ihre Überlegung war, dass nach dem Genuss von Alkohol die subjektive Wahrnehmung der körperlichen Attraktivität ungenauer wird (der bekannte „Bier-Brillen-Effekt“). Außerdem wollte sie wissen, ob dieser Effekt bei Männern und Frauen unterschiedlich ist. Daraufhin nahm sie die Studienteilnehmer:innen in einen Nachtclub mit und gab ihnen keinen Alkohol (die Teilnehmer:innen erhielten stattdessen Plazebogetränke aus alkoholfreiem Lagerbier), 2 Pints starkes Lagerbier oder 4 Pints starkes Lagerbier zu trinken. Am Ende des Abends machte sie ein Foto von der Person, mit der der:die jeweilige Teilnehmer:in geplaudert hatte. Anschließend ließ sie eine Gruppe unabhängiger Beurteiler:innen die Attraktivität der Person auf jedem Foto auf einer Skala von 100 bewerten.“

Die Fragestellung der (fiktiven) Studie lautete also: Unterscheidet sich die Attraktivität des:der ausgewählten Gesprächspartners:in in Abhängigkeit vom (eigenen) Geschlecht und der Menge getrunkenen Alkohols? Sie finden den zugehörigen Datensatz in der Datendatei „Goggles.sav“, die Sie von der frei zugänglichen Webseite mit ergänzenden Ressourcen für Fields Buch „Discovering Statistics Using IBM SPSS Statistics“ unter https://edge.sagepub.com/field5e/student-resources/datasets herunterladen können. Beantworten Sie die Fragestellung der (fiktiven) Studie und wählen Sie dafür ein Signifikanzniveau von $\alpha = .05$, und wählen Sie eine geeignete Methode um p-Werte im Rahmen von post-hoc Vergleichen zu korrigieren.

Beispiel 7.6

Eine (fiktive) Forschungsgruppe möchte untersuchen, wie sich verschiedene Unterrichtsmethoden in unterschiedlichen Unterrichtsfächern auf das Wissen von Schüler:innen am Semesterende auswirken. Dazu soll der Einsatz von Tafel bzw. Powerpoint-Folien als Unterrichtsmethoden im Geschichts- und Mathematikunterricht untersucht werden. Am Ende des Semesters wird das Wissen aller teilnehmenden

Schüler:innen mit standardisierten Wissenstests erhoben, die einen Vergleich des Mathematik- und Geschichtswissens auf einer Skala von 0-100 erlauben. Eine der zentralen Fragestellungen der Studie lautet: Hängt es von der Art des Unterrichtsfachs ab, wie effektiv (bezogen auf die Wissensvermittlung) die eingesetzten Präsentationsmethoden sind?

Den Datensatz für dieses Beispiel finden Sie in der Datei „Kap7UE6.sav“, die Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können. Beantworten Sie die folgenden Fragen:

(a) Was sind die UV und die AV in diesem Beispiel?

(b) Wie viele Stufen haben die Faktoren in diesem Beispiel jeweils?

(c) Wie lässt sich die genannte zentrale Forschungsfrage prinzipiell beantworten?

(d) Führen Sie für dieses Beispiel eine entsprechende Varianzanalyse mit SPSS durch und verfassen Sie einen entsprechenden Ergebnisbericht.

Wählen Sie für alle statistischen Analysen ein Signifikanzniveau von $\alpha = .005$, wobei p-Werte im Rahmen von post-hoc Vergleichen entsprechend zu korrigieren sind.

Beispiel 7.7

Eine (fiktive) Forschungsgruppe untersuchte, ob sich unterschiedliche Fitnessprogramme unterschiedlich auf die allgemeine Fitness von Personen aus verschiedenen Altersgruppen auswirkt. Unterschieden wurden junge Erwachsene (bis inkl. 30 Jahre), Erwachsene mittleren Alters (31-50 Jahre), und ältere Erwachsene (> 50 Jahre). Verglichen wurden konventionelles Krafttraining mit Geräten und HIIT-Programme mit dem eigenen Körpergewicht. Die allgemeine Fitness wurde mit einem Fitnessindex auf einer Skala von 0-100 Punkten erfasst.

Die (fiktive) Forschungsgruppe hat bereits einen (fiktiven) Ergebnisbericht erstellt und Sie darum gebeten, diesen zu kontrollieren. Aufgrund des wachsenden Misstrauens unter Forscher:innen wegen des hohen Publikationsdrucks und den üblen Machenschaften von Dr. Publish-Perish (Gigerenzer, 2004) hat Ihnen die Forschungsgruppe den Ergebnisbericht nur in Papierform zukommen lassen (wobei Ihnen selbst nicht ganz klar ist, wie das verhindern soll, dass Sie die Ergebnisse einfach kurzerhand selbst publizieren) und leider haben Sie Ihren am Morgen bitter benötigten Kaffee darüber

verschüttet. Trotz großer Bemühungen sind daher einige Stellen unleserlich geworden. Um diese Stellen vervollständigen zu können, hat Ihnen die Forschungsgruppe zähneknirschend den Originaldatensatz überlassen. Diesen finden Sie in der Datendatei „Kap7UE7.sav“, die Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können. Verwenden Sie den Datensatz und geeignete statistische Verfahren, um den folgenden Ergebnisbericht an den gekennzeichneten Stellen zu vervollständigen.

Ergebnisbericht: Es wurde eine zweifaktorielle Varianzanalyse ohne Messwiederholung mit den Faktoren Altersgruppe (drei Stufen: jung, d.h. 18-30 Jahre, mittel, d.h. 31-50 Jahre, alt, d.h. > 50 Jahre) und Trainingsmethode (zwei Stufen: konventionelles Krafttraining mit Gewichten vs. HIIT mit eigenem Körpergewicht) durchgeführt. Das Signifikanzniveau wurde zu $\alpha = .005$ gewählt.

Insgesamt wurden Daten von _____ Personen in einem balancierten Design erhoben. Levenes Test war (mit $\alpha = .05$) nicht signifikant ($p = .831$), daher wurde von _____ in den einzelnen Populationen ausgegangen.

Im Mittel war die allgemeine Fitness zwischen den unterschiedlichen Altersgruppen signifikant verschieden (F(_____) = 54.15, $p < .001$, η_p^2 = _____, d.h. ein _____ Effekt gemäß Cohen (1988)). Im Mittel war die erzielte allgemeine Fitness auch zwischen den beiden Fitnessprogrammen signifikant unterschiedlich (F(_____) = _____, _____, η_p^2 = .08, d.h. ein _____ Effekt gemäß Cohen (1988)). Die Interaktion zwischen den beiden Faktoren war _____ (F(_____) = _____, _____, η_p^2 = _____, d.h. ein _____ Effekt gemäß Cohen (1988)). Zur weiteren Analyse paarweiser Mittelwertsunterschiede wurden post-hoc Tests mit einer Korrektur der p-Werte für multiple Vergleiche gemäß Bonferroni durchgeführt. Im Folgenden werden lediglich korrigierte p-Werte berichtet.

Sowohl bei konventionellem Krafttraining mit Gewichten als auch bei HIIT-Programmen mit dem eigenen Körpergewicht nahm die erzielte, allgemeine Fitness mit fortschreitendem Alter ab. Bei konventionellem Krafttraining waren alle paarweisen Mittelwertsunterschiede zwischen den unterschiedlichen Altersgruppen signifikant ($p \leq .001$). Bei HIIT-Programmen war der Unterschied

zwischen jungen und mittleren Erwachsenen nicht signifikant ($p > .999$), während die Unterschiede zwischen jungen und alten sowie mittleren und alten Erwachsenen jeweils signifikant waren (_____). Zudem unterschieden sich konventionelles Krafttraining und HIIT-Programme sowohl bei mittleren (_____) als auch älteren (_____) Erwachsenen signifikant, jedoch nicht bei jungen Erwachsenen (_____). Bei allen Altersgruppen war die erzielte allgemeine Fitness jedoch bei HIIT-Programmen _____ als bei konventionellem Krafttraining.

Punkt- und Intervallschätzungen für die erzielte allgemeine Fitness in Abhängigkeit von Altersgruppe und verwendeter Trainingsmethode sind in Abbildung 7.15 dargestellt. Mittelwerte, Standardabweichungen und Gruppengrößen sind in Tabelle 7.3 zusammengefasst.

Tabelle 7.3

Deskriptive Statistiken

Altersgruppe	Training	*M*	*SD*	*n*
Jung: 18-30 Jahre	Konv. Kraft			
	HIIT	77.64		
Mittel: 31-50 Jahre	Konv. Kraft			
	HIIT			45
Alt: > 50 Jahre	Konv. Kraft			
	HIIT		15.36	

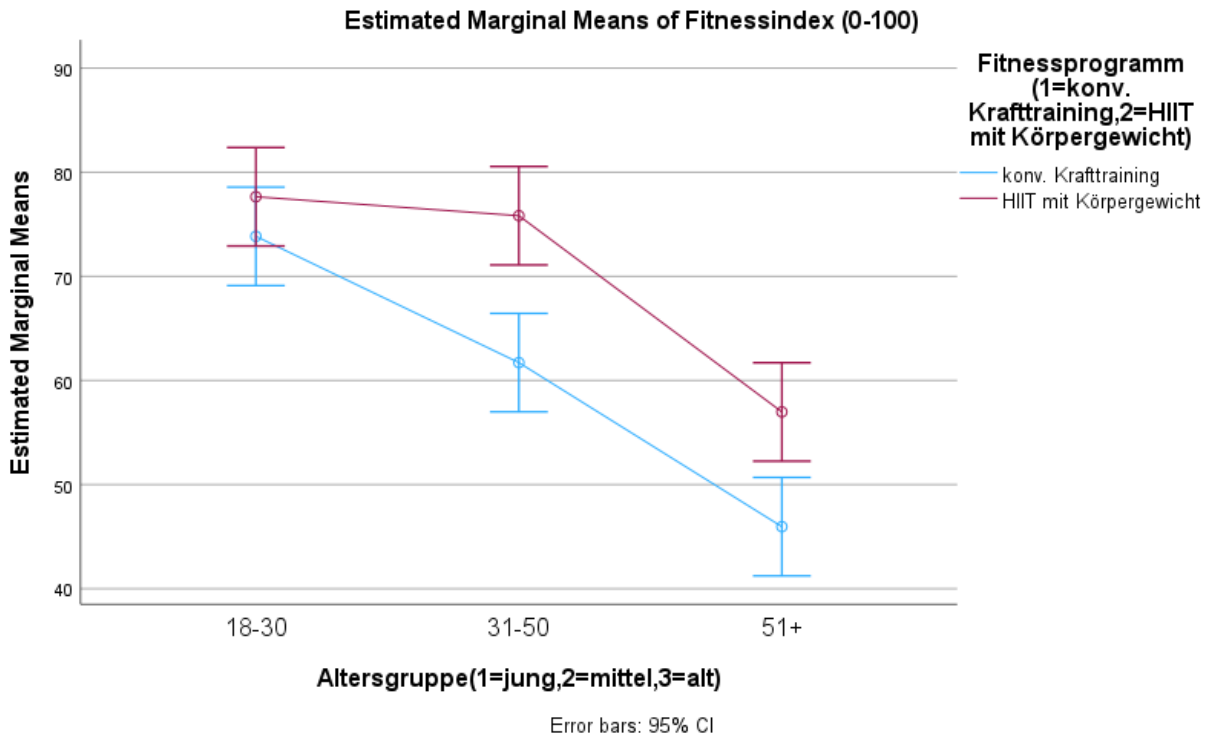


Abbildung 7.15. Punkt- und Intervallschätzungen für die erzielte allgemeine Fitness in Abhängigkeit von Altersgruppe und verwendeter Trainingsmethode.

Beispiel 7.8

Eine therapeutische Intervention soll mit einer entsprechenden Kontrollbedingung (tau = treatment as usual) verglichen werden. Zudem soll untersucht werden, ob sich die Wirksamkeit der Intervention für Frauen und Männer unterscheidet. Dazu wird für 120 Personen die Veränderung der Symptomstärke durch die Intervention bzw. die Kontrollbedingung erhoben. Die entsprechenden Daten sind im Datensatz „Kap7UE8.sav“ zu finden.

Wählen Sie ein geeignetes statistisches Verfahren um zu untersuchen, ob sich die Wirksamkeit (gemessen an der Änderung der Symptomstärke) der Intervention von der Kontrollbedingung unterscheidet und ob dieser Unterschied davon abhängt, ob Frauen oder Männer untersucht werden. Erstellen Sie anschließend einen entsprechenden Ergebnisbericht.

Kapitel 8

Varianzanalysen mit Messwiederholung

Stefan E. Huber

Auch in diesem Kapitel werden wir uns wieder hauptsächlich auf die Durchführungsaspekte, diesmal aber von Varianzanalysen *mit* Messwiederholung in SPSS beschränken. Da wir diesbezüglich sowohl ein- und zweifaktorielle Varianzanalysen mit Messwiederholung als auch ein gemischtes Design, d.h. eine Varianzanalyse mit einem Zwischensubjektfaktor (auch: nicht-messwiederholter Faktor, Zwischen-Personen-Faktor oder engl.: Between-subjects-factor) und einem Innersubjektfaktor (auch Messwiederholungsfaktor, Innerhalb-Personen-Faktor oder engl.: Within-subjects-factor), besprechen werden, ist dafür ohnehin genug zu tun. Gleichzeitig bleiben viele Aspekte ganz analog zum Fall von Varianzanalysen ohne Messwiederholung, auch wenn sich konzeptuell, sozusagen im Hintergrund durchaus einiges ändert (Quadratsummenzerlegung, Variation der AV zwischen Personen etc.). Im Vordergrund gibt es nach wie vor für jeden Faktor einen F-Wert, zwei Freiheitsgrade, einen p-Wert, eine Testentscheidung. Eine Sache, die aber jedenfalls von großer Relevanz bleibt, sind die Voraussetzungen, die für Varianzanalysen mit Messwiederholung erfüllt sein müssen.

Voraussetzungen für Varianzanalysen mit Messwiederholung

Die Voraussetzungen für Varianzanalysen mit Messwiederholung sind (Bühner et al., 2025):

- Intervallskalenniveau der AV.
- Normalverteilung der AV auf jeder Faktorstufe.
    - Präziser: sowohl Personeneffekte als auch Fehler sind jeweils unabhängig voneinander unabhängig und identisch normalverteilt mit Mittelwert Null und jeweils bestimmter (unbekannter) Varianz.
- Kovarianz der Personeneffekte und Fehler ist Null.
- Compound Symmetry (CS): Kovarianzen der Messwerte zwischen den Messzeitpunkten sind identisch und Varianzen der Messwerte sind zwischen den Messzeitpunkten homogen.

- Bei gemischten Designs: Varianzhomogenität sowie Gleichheit der Kovarianzmatrizen über die Faktorstufen des Zwischensubjektfaktors.

Während die ersten drei dieser Voraussetzungen üblicherweise nicht überprüft werden (auch die Varianzanalyse mit Messwiederholung ist relativ robust gegenüber Verletzung der Normalverteilungsannahme, solange die übrigen Annahmen gut erfüllt sind; die anderen beiden Voraussetzungen sollten durch ein passendes Studiendesign gewährleistet werden), ist die Annahme der CS vor der Durchführung einer Varianzanalyse jedenfalls zu prüfen.

Wie auch schon bei der Prüfung der Varianzhomogenität im Falle der Varianzanalyse ohne Messwiederholung (Levenes Test), wird aber auch für diesen Fall eine entsprechende Prüfung standardmäßig in SPSS durchgeführt und korrigierte Freiheitsgrade für die Berechnung des p-Werts mittels der F-Statistik für die gegebene Stichprobe berechnet. Strenggenommen handelt es sich bei dem dafür verwendeten Mauchly-Test nicht um einen Test der CS, sondern der abgeschwächten Annahme der Sphärizität, was für die meisten Fälle aber einen hinreichenden Test darstellen sollte. Ist der Mauchly-Test signifikant (üblicherweise mit $\alpha = .05$), so wird von Verletzung der Sphärizität (und, in Extension, von Verletzung der CS) ausgegangen und korrigierte Werte für die Freiheitsgrade der F-Verteilung zur Berechnung des p-Wertes verwendet. Dafür stehen mehrere Korrekturverfahren zur Verfügung. Häufig wird dabei auf das Verfahren nach Greenhouse-Geisser zurückgegriffen. Dieses ist aber sehr konservativ, weshalb stattdessen die Verwendung des etwas liberaleren Huynh-Feldt Verfahrens empfohlen wird.

Das klingt alles recht kompliziert, die Praxis ist aber in diesem Fall deutlich einfacher als die Theorie. Davon werden wir uns im Folgenden an je einem Beispiel für eine einfaktorielle und eine zweifaktorielle Varianzanalyse mit Messwiederholung sowie für ein gemischtes Design überzeugen.

Durchführung einer einfaktoriellen Varianzanalyse mit Messwiederholung in SPSS

Gegeben sind (fiktive) Depressionswerte (gemessen mit Becks Depressionsinventar) für eine Stichprobe von $n = 100$ zufällig ausgewählten Patient:innen zu vier Messzeitpunkten während einer Psychotherapie. Die Fragestellung lautet: Unterscheiden sich die Erwartungswerte der Depressionswerte zwischen den Messzeitpunkten in der Population der Patient:innen? Bei den

Messzeitpunkten handelt es sich um einen Innersubjektfaktor: für jede Person gibt es vier Messwerte. Dieser Innersubjektfaktor ist auch der einzige Faktor. Ist der Omnibustest für diesen Faktor signifikant, so gehen wir von einem Unterschied der Erwartungswerte für mindestens zwei der Messzeitpunkte aus und können die Frage demnach affirmativ beantworten. Das geeignete Verfahren für diese Fragestellung ist also eine einfaktorielle Varianzanalyse mit Messwiederholung. Die Daten für dieses Beispiel finden sich in der Datei „Kap8daten1.sav“, die Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Zur Durchführung einer einfaktoriellen Varianzanalyse mit Messwiederholung wählen wir in SPSS nun *Analyze >> General Linear Model >> Repeated Measures…*. Im sich öffnenden Menü müssen wir nun erst einmal unseren Messwiederholungsfaktor definieren. Dazu geben wir diesem einmal einen Namen, z.B. „Messzeitpunkt“, und geben anschließend an, dass er über 4 Stufen verfügt, indem wir bei „Number of Levels“ die Zahl 4 eintragen, siehe Abbildung 8.1 links. Dann klicken wir auf „Add“, woraufhin die Angabe „Messzeitpunkt(4)“ in dem Feld rechts erscheint, siehe Abbildung 8.1 rechts. Nun klicken wir auf „Define“.

Im sich öffnenden Fenster weisen wir die vier Variablen „BDI1“, „BDI2“, „BDI3“ und „BDI4“ den vier Stufen des soeben definierten Messwiederholungsfaktors zu, indem wir alle vier Variablen markieren und in das Fenster „Within-Subjects Variables (Messzeitpunkt)“ ziehen, siehe Abbildung 8.2.

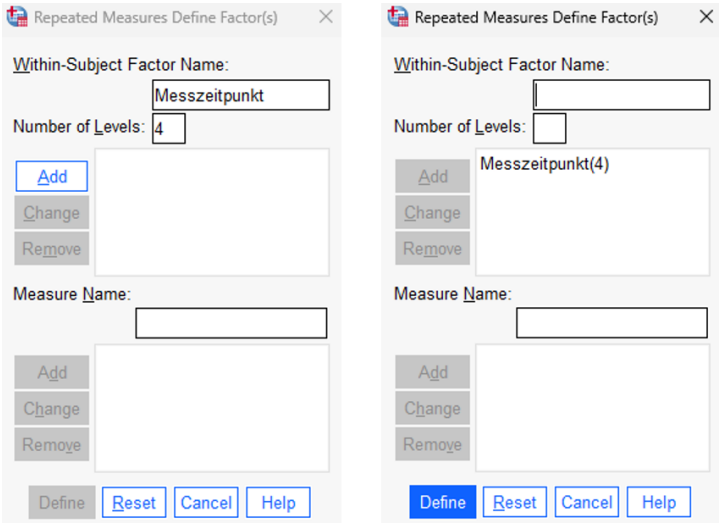


Abbildung 8.1. Definition unseres Messwiederholungsfaktors.

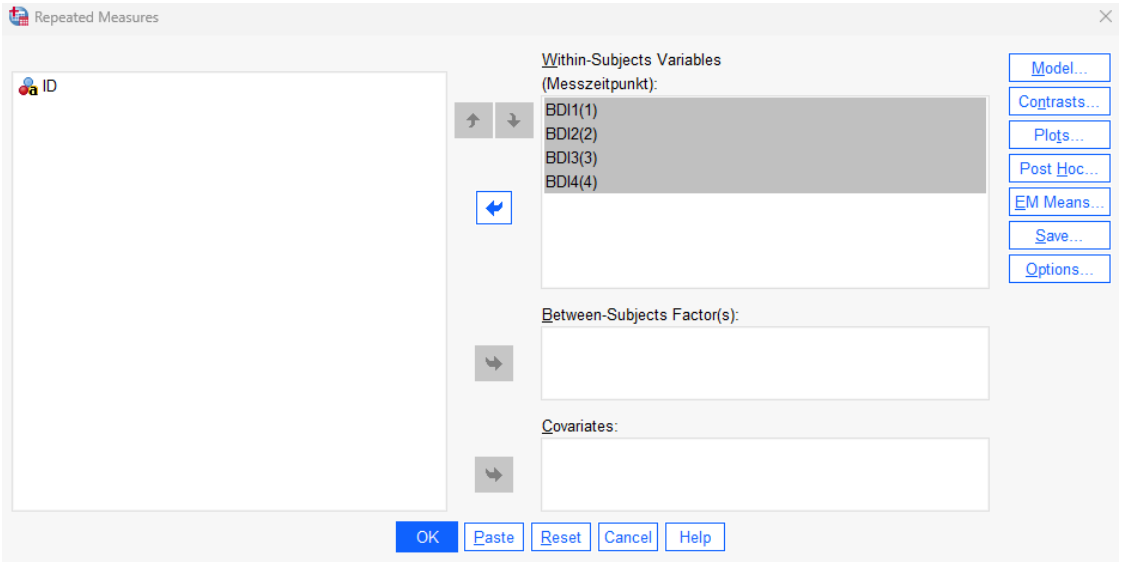


Abbildung 8.2. Zuweisung unserer vier Variablen zu den vier Stufen unseres soeben definierten Messwiederholungsfaktors.

Unter „Plots…“ wählen wir dann noch aus, dass wir den Messwiederholungsfaktor auf der horizontalen Achse darstellen wollen und auch gerne wieder Fehlerbalken hätten, die 95%-KI entsprechen, siehe Abbildung 8.3. Unter „Options…“ wählen wir „Descriptive statistics“ und „Estimates of effect size“. Unter „EM Means…“ ziehen wir schließlich noch den Faktor *Messzeitpunkt* in das Feld „Display Means for“, wählen „Compare main effects“ und fordern eine Bonferroni-Korrektur an, um p-Werte und Konfidenzniveaus für alle paarweisen Vergleiche für die Mittelwerte der vier Messzeitpunkte entsprechend einer FWER von maximal 5% zu adjustieren. Dann fügen wir wieder alles in eine Syntaxdatei ein, führen die entsprechenden Kommandozeilen aus und wenden uns der dadurch erzeugten, recht umfangreichen Ausgabe zu (die hier nicht dargestellt, sondern nur beschrieben wird).

In der Tabelle „Within-Subjects Factors“ finden wir noch einmal die Definition unseres Messwiederholungsfaktors. In der Tabelle „Descriptive Statistics“ finden wir deskriptive Statistiken in Form von Mittelwerten und Standardabweichungen unserer AV für die vier Messzeitpunkte. Die Tabelle „Multivariate Tests“ können wir ignorieren (Interessierte finden mehr Informationen zu dieser Tabelle z.B. bei Field, 2024). In der Tabelle „Mauchly’s Test of Sphericity“ finden wir das Ergebnis des Mauchly-Tests. Wir sehen, dass dieser signifikant ist und werden daher unten die Ergebnisse für gemäß Huynh-Feldt korrigierte Freiheitsgrade und p-Werte berichten.

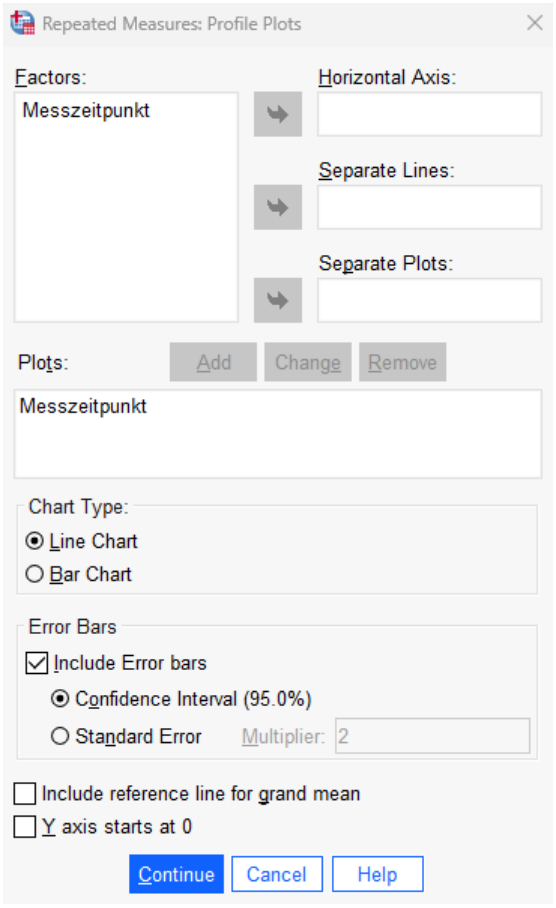


Abbildung 8.3. Anforderung einer grafischen Darstellung von Punkt- und Intervallschätzungen unserer Populationsmittelwerte für die vier Messzeitpunkte.

In der Tabelle „Tests of Within-Subjects Effects“ finden wir die eigentlichen Ergebnisse unserer Varianzanalyse. Hier schauen wir uns aufgrund der Signifikanz des Mauchly-Tests lediglich die Zeilen an, die mit „Huynh-Feldt“ bezeichnet werden, um Freiheitsgrade, F- und p-Wert sowie Effektstärke für unsere Varianzanalyse abzulesen. In unserem Fall also: $F(2.86,283.40) = 146.84$, $p < .001$, $\eta_p^2 = .60$. Die Effektstärke sagt uns in diesem Fall, dass 60% der Variabilität im Depressionswert durch den Messzeitpunkt erklärt werden können. Das ist ein sehr großer Wert. Auch hier gelten gemäß Cohens Heuristik (1988) wieder Effektstärken zwischen 0.01 und 0.06 als klein, zwischen 0.06 und 0.14 als mittel, und ab 0.14 als groß.

Die Tabelle „Tests of Within-Subjects Contrasts“ können wir wieder ignorieren. Die Tabelle „Tests of Between-Subjects Effects“ ebenfalls, da sie uns ohne Zwischensubjektfaktoren für dieses Beispiel keine inhaltlich relevanten Ergebnisse liefert.

Im Abschnitt „Estimated Marginal Means“ finden wir die Ergebnisse für unsere angeforderten paarweisen Vergleiche. In der Tabelle „Estimates“ finden wir Punkt- und Intervallschätzungen für die auf Basis der vier Stichproben ermittelten Populationsmittelwerte. In der Tabelle „Pairwise Comparisons“ finden wir paarweise Tests für Gleichheit der jeweiligen Populationsmittelwerte. Die p-Werte und Konfidenzniveaus der Konfidenzintervalle für die Mittelwertdifferenzen sind hier jeweils nach der Methode korrigiert, die wir angefordert haben; in unserem Fall also nach Bonferroni. Wir sehen, dass sich die Mittelwerte der AV für den ersten und den letzten Messzeitpunkt von den Mittelwerten zu den anderen Messzeitpunkten signifikant unterscheiden ($p < .001$). Die Mittelwerte der AV zu den Messzeitpunkten 2 und 3 unterscheiden sich allerdings nicht signifikant voneinander ($p = .766$). Die Tabelle „Multivariate Tests“ können wir auch hier wieder ignorieren.

Schließlich bekommen wir auch unsere angeforderte graphische Darstellung inklusive der gewünschten 95%-KI (die numerischen Werte für diese haben wir in der Tabelle „Estimates“ etwas weiter oben gegeben), siehe Abbildung 8.4. Auch an dieser Darstellung erkennen wir, dass die Annahme einer Verringerung der Depressionswerte über den Therapieverlauf ganz plausibel erscheint.

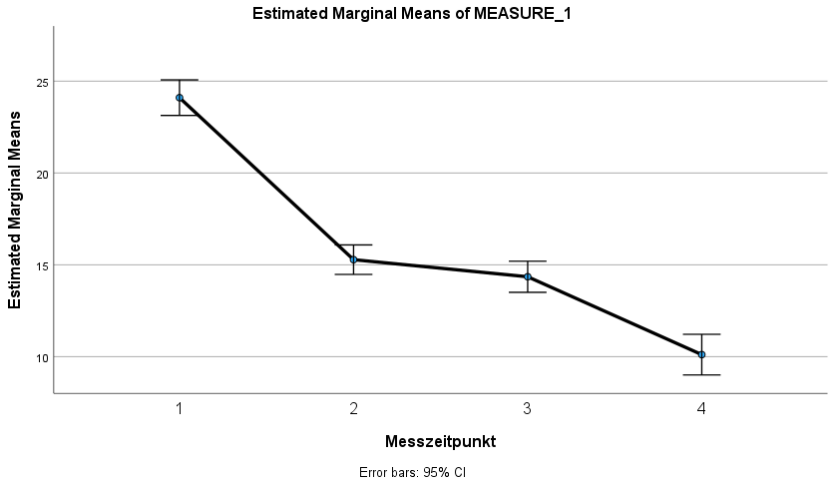


Abbildung 8.4. Graphische Darstellung der mittleren Depressionsniveaus und derer 95%-KI über die vier Messzeitpunkte.

Ergebnisbericht

Ein Ergebnisbericht für diese Ergebnisse könnte wie folgt aussehen: „Da die Voraussetzung der Sphärizität verletzt war (p = .043), werden im Folgenden Huynh-Feldt-korrigierte Werte berichtet. Der Messzeitpunkt hat einen signifikanten Einfluss auf den Depressionswert, F(2.86,283.40) = 146.84, p < .001, η_p^2 = .60, d.h. 60% der Variabilität im Depressionswert können durch den Messzeitpunkt erklärt werden. Post-hoc Tests mit p-Wert-Korrektur für multiple Vergleiche gemäß Bonferroni ergaben zudem, dass sich die Depressionswerte zu Messzeitpunkt 1 und 4 von den Depressionswerten zu allen anderen Messzeitpunkten signifikant unterscheiden (p < .001), während sich die Depressionswerte zu den Messzeitpunkten 2 und 3 nicht signifikant voneinander unterscheiden (p = .766). Deskriptive Statistiken sind in Tabelle 8.1 gegeben. Der Verlauf der mittleren Depressionswerte sowie derer 95%-KI über die Messzeitpunkte hinweg ist in Abbildung 8.4 dargestellt."

Tabelle 8.1

Deskriptive Statistiken

Messzeitpunkt	*M*	*SD*	*n*
1	24.10	4.87	100
2	15.28	4.05	100
3	14.35	4.27	100
4	10.11	5.58	100

Durchführung einer zweifaktoriellen Varianzanalyse mit vollständiger Messwiederholung in SPSS

Gegeben sind Leistungsindizes von 26 (fiktiven) Personen, die an einem Aerobic-Kurs teilgenommen haben. Bei diesem Kurs wurde an unterschiedlichen Tagen unter unterschiedlichen Bedingungen trainiert. Zu einem Zeitpunkt wurde ohne Musik und mit 2kg-Hanteln, zu einem anderen Zeitpunkt mit Musik und mit 2kg-Hanteln, wieder zu einem anderen Zeitpunkt ohne Musik und mit 5kg-Hanteln, und zu einem vierten Zeitpunkt mit Musik und mit 5kg-Hanteln trainiert. Die Forschungsfrage lautete: Wie wirkt sich Musik und das Gewicht der verwendeten Hanteln auf die Leistung von Teilnehmer:innen in einem Aerobic-Kurs aus?

Da sowohl für den Faktor Musik als auch für den Faktor Gewicht jeweils 2 Stufen vorliegen, handelt es sich in diesem Fall um eine zweifaktorielle Varianzanalyse mit Messwiederholung mit einem 2x2 Design (d.h. 2 Faktoren mit jeweils 2 Stufen). Es handelt sich in beiden Fällen um Messwiederholungsfaktoren, da die Leistung für jede Person für jede Stufe jeden Faktors erhoben wurde; daher wird die entsprechende Varianzanalyse auch als zweifaktorielle Varianzanalyse mit *vollständiger* Messwiederholung bezeichnet. Die Daten für dieses Beispiel finden sich in der Datei „Kap8daten2.sav“, die Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Um eine entsprechende Varianzanalyse in SPSS durchzuführen, wählen wir wieder *Analyze* >> *General Linear Model* >> *Repeated Measures…* aus. Dort müssen wir jetzt allerdings zwei Messwiederholungsfaktoren definieren. Dazu geben wir dem ersten erst einmal einen Namen, z.B. „Musik“, und geben bei „Number of Levels“ die Zahl 2 an, da der Faktor 2 Stufen hat. Dann klicken wir auf „Add“. Anschließend definieren wir einen zweiten Faktor. D.h. wir geben ihm einen Namen, z.B. „Gewicht“, und geben wiederum an, dass er 2 Stufen hat, und klicken wieder auf „Add“. Die ganze Prozedur ist graphisch in Abbildung 8.5 dargestellt (von links nach rechts). Anschließend klicken wir auf „Define“.

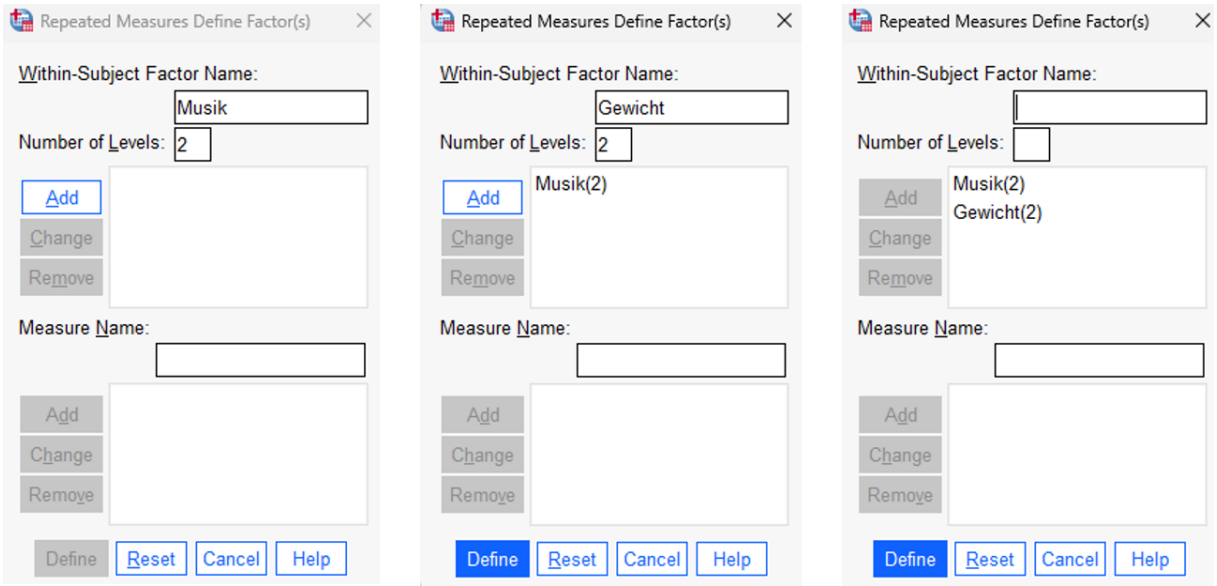


Abbildung 8.5. Definition von zwei Messwiederholungsfaktoren, jeweils mit 2 Stufen.

Im sich öffnenden Menü müssen wir nun die vier Variablen passend zu den vier Kombinationen aus den jeweils 2 Stufen der beiden soeben definierten Faktoren zuordnen. Der erste Faktor ist *Musik*, der zweite *Gewicht*, d.h. die Kombinationen (1,1) und (1,2) stehen für (leise,2kg) und (leise,5kg) mit den Bezeichnungen aus der Datendatei. D.h. wir ziehen zuerst die Variable leise2kg in das Feld „Within-Subjects Variables (Musik,Gewicht)“ und dort in die erste Zeile „_?_(1,1)“. Dann ziehen wir die Variable leise5kg in das Feld „Within-Subjects Variables (Musik,Gewicht)“ und dort in die zweite Zeile „_?_(1,2)“. Dann kümmern wir uns um die verbleibenden beiden Variablen, die die Leistungsindizes für die beiden Bedingungen enthalten, in denen mit Musik trainiert wurde. D.h. wir ziehen die Variable musik2kg in das Feld „Within-Subjects Variables (Musik,Gewicht)“ und dort in die dritte Zeile „_?_(2,1)“. Schließlich ziehen wir die Variable musik5kg in das Feld „Within-Subjects Variables (Musik,Gewicht)“ und dort in die vierte Zeile „_?_(2,2)“.

Unter „Plots…“ fordern wir analog zum vorhergehenden Kapitel zwei Grafiken an. Einmal mit dem Faktor *Musik* auf der horizontalen Achse („Horizontal Axis“) und verschiedenen Linien („Separate Lines“) für den Faktor *Gewicht*, und einmal umgekehrt. In beiden Fällen wollen wir aber auch wieder Fehlerbalken, die 95%-KI entsprechen.

Unter „Options…“ fordern wir wieder „Descriptive statistics“ sowie „Estimates of effect size“ an. Unter „EM Means…“ wollen wir diesmal alle paarweisen Vergleiche für beide Faktoren. Wir ziehen daher „Musik*Gewicht“ in das Feld „Display Means for“, wählen „Compare simple main effects“ und wählen dann für die Korrektur von p-Werten und Konfidenzniveaus wieder „Bonferroni“ aus. Dann fügen wir wieder alles in eine Syntaxdatei ein und führen die Kommandozeilen aus, woraufhin wieder eine umfangreiche Ausgabe erzeugt wird (die hier wiederum nicht abgebildet, sondern nur beschrieben wird). Bei dieser beschränken wir uns im Folgenden nur mehr auf die wesentlichen Aspekte.

Deskriptive Statistiken für alle möglichen Kombinationen aus Faktorstufen finden wir wieder in der Tabelle „Descriptive Statistics“. Mauchly Tests bekommen wir im Fall einer zweifaktoriellen Varianzanalyse mit Messwiederholung für jeden Faktor sowie für deren Interaktion. Hier sieht das Ergebnis allerdings seltsam aus, was aber in diesem Fall tatsächlich so sein sollte. Der Mauchly-Test prüft die Gleichheit der Varianzen der Differenzen zwischen mehreren Faktorstufen von abhängigen

Variablen (das ist – vereinfacht gesagt – was der Test auf Sphärizität macht). Für jede der beiden abhängigen Variablen gibt es allerdings hier nur zwei Faktorstufen, d.h. nur eine Differenzvariable. Bei lediglich einer Differenzvariable mit einer dazugehörigen Varianz gibt es aber keine andere Differenzvariable bzw. Varianz, mit der diese verglichen werden kann, d.h. der Mauchly-Test kann bei zwei Faktorstufen nicht durchgeführt werden. Man kann auch sagen, dass bei zwei Faktorstufen Sphärizität immer erfüllt ist (bei nur einer Differenzvariable haben selbstverständlich alle Differenzvariablen dieselbe Varianz, da es ja nur eine gibt). Aus diesem Grund ist die Tabelle „Mauchly's Test of Sphericity" leer bzw. sind alle Korrekturfaktoren gleich Eins.

In der Tabelle „Tests of Within-Subjects Effects" finden wir die Ergebnisse unserer Varianzanalysen (da wir ja zwei Faktoren haben, führen wir wiederum drei Omnibustests durch: einen für jeden Faktor und einen für die Interaktion). Im Abschnitt Musik finden wir die Ergebnisse für den Test unseres Faktors *Musik*, d.h. $F(1,25) = 7.43$, $p = .012$, $\eta_p^2 = .23$. Was bedeutet die Effektstärke η_p^2 im Fall mehrerer Faktoren? In diesem Fall entspricht die Effektstärke dem Anteil der Varianz, den dieser Faktor aufklären kann, der nicht bereits durch andere Faktoren oder die Interaktion aufgeklärt werden kann. Die Heuristik nach Cohen (1988) bleibt wie bisher bestehen, hier liegt also ein großer Effekt für den Faktor Musik vor.

Im Abschnitt Gewicht finden wir die Ergebnisse für den Test unseres Faktors *Gewicht*, d.h. $F(1,25) = 56.08$, $p < .001$, $\eta_p^2 = .69$. Auch hier liegt also ein großer Effekt nach Cohen (1988) vor. Im Abschnitt „Musik * Gewicht" finden wir die Ergebnisse für den Test der Interaktion zwischen beiden Faktoren, d.h. $F(1,25) = 18.57$, $p < .001$, $\eta_p^2 = .43$. Auch hier liegt also ein großer Effekt nach Cohen (1988) vor.

Im Abschnitt „Estimated Marginal Means" finden wir schließlich wieder alle Informationen zu unseren angeforderten paarweisen Vergleichen. Da wir zwei Faktoren haben, gibt es wieder zwei Unterabschnitte. Zuerst finden wir die paarweisen Vergleiche für die zwei Stufen des Faktors *Musik* für jede Stufe des Faktors *Gewicht*. Wir sehen, dass sich die mittlere Leistung nur bei hohem Gewicht zwischen den beiden Musikbedingungen signifikant unterscheidet ($p < .001$). Bei niedrigem Gewicht gibt es keinen signifikanten Unterschied ($p = .234$). Im zweiten Unterabschnitt sehen wir, dass sich die

mittlere Leistung bei niedrigerem und höherem Gewicht in beiden Musikbedingungen signifikant unterscheidet, sowohl ohne Musik ($p < .001$) als auch mit Musik ($p = .001$).

Im Abschnitt „Profile Plots“ finden wir wiederum unsere angeforderten graphischen Darstellungen. Diese illustrieren sehr schön, dass beide Faktoren eine Rolle spielen, dass zudem aber der Faktor Musik eine erhebliche Rolle nur dann spielt, wenn mit größerem Gewicht trainiert wird, siehe Abbildung 8.6. Bei niedrigerem Gewicht unterscheiden sich die mittleren Leistungen kaum (und die plausiblen Bereiche überlappen stark). Das entspricht gerade unserem Interaktionseffekt von oben und illustriert ein weiteres Mal, weshalb eine einfache Interpretation der Haupteffekte bei bestehender Interaktion nicht ohne weiteres möglich ist.

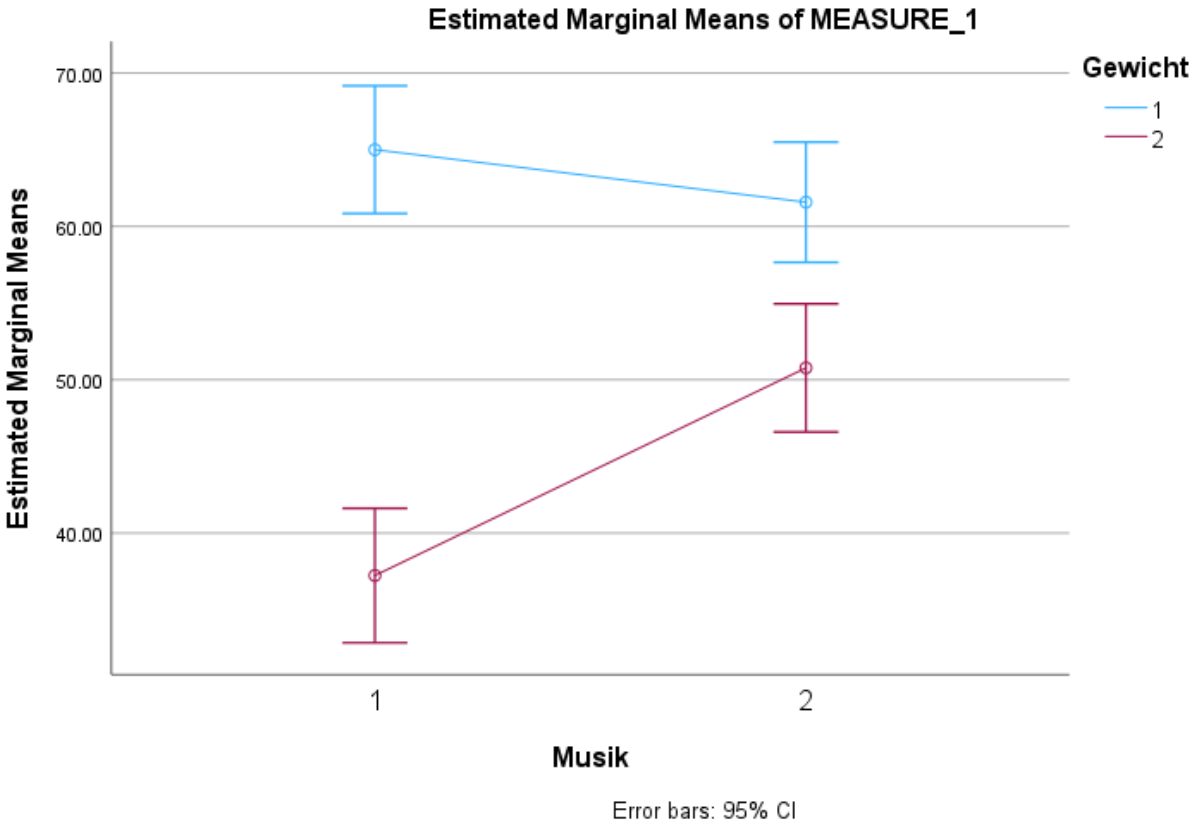


Abbildung 8.6. Mittlere Leistungsindizes für einen Aerobic-Kurs in Abhängigkeit vom Gewicht der Hanteln, mit denen trainiert wurde, sowie von der Tatsache, ob mit oder ohne Musik trainiert wurde. Fehlerbalken entsprechen 95%-KI.

Ergebnisbericht

Ein Ergebnisbericht für dieses Beispiel könnte wie folgt aussehen: „Im Mittel gibt es einen signifikanten Unterschied in der Aerobic-Leistung abhängig davon, ob Musik gespielt wird oder nicht, $F(1,25) = 7.43$, $p = .012$, $\eta_p^2 = .23$. Einen signifikanten Unterschied gibt es auch abhängig vom Gewicht der verwendeten Hanteln, $F(1,25) = 56.08$, $p < .001$, $\eta_p^2 = .69$. Ferner besteht eine signifikante Wechselwirkung zwischen den Faktoren Musik und Gewicht der Hanteln, $F(1,25) = 18.57$, $p < .001$, $\eta_p^2 = .43$. Um paarweise Unterschiede zu untersuchen wurden post-hoc Tests mit p-Wert-Korrektur für multiple Vergleiche gemäß Bonferroni berechnet. Bei 2kg-Hanteln macht es für die Leistung keinen signifikanten Unterschied, ob Musik gespielt wird ($M = 61.58$, $SD = 9.70$, $n = 26$ in jeder Bedingung) oder nicht ($M = 65.00$, $SD = 10.31$), $p = .234$. Bei 5kg-Hanteln zeigen die Teilnehmer:innen mit Musik ($M = 50.77$, $SD = 10.34$) eine signifikant höhere mittlere Leistung als ohne Musik ($M = 37.23$, $SD = 10.86$), $p < .001$. Die Leistung ist mit 2kg-Hanteln immer signifikant höher als mit 5kg-Hanteln, egal ob Musik gespielt wird ($p = .001$) oder nicht ($p < .001$). Eine graphische Darstellung dieser Ergebnisse inklusive 95%-KI für die mittleren Leistungsindizes ist in Abbildung 8.6 gegeben."

Durchführung einer zweifaktoriellen Varianzanalyse mit gemischtem Design in SPSS

Zur Illustration eines gemischten Designs greifen wir auf einen ähnlichen Datensatz wie oben für die einfaktorielle Varianzanalyse mit Messwiederholung zurück. Gegeben sind wiederum Depressionswerte, dieses Mal allerdings zu drei Messzeitpunkten und von 200 fiktiven Patient:innen, von welchen jeweils 100 entweder eine Verhaltenstherapie oder eine Mischung aus verschiedenen Therapien (= Therapiemix) in Anspruch genommen haben. Bei den Messzeitpunkten handelt es sich um den Beginn der Therapie, sowie sechs und zwölf Wochen nach Beginn der Therapie. Die Fragestellung lautet diesmal: Unterscheidet sich die Wirkung der Therapiemethoden abhängig vom zeitlichen Verlauf? Die Daten für dieses Beispiel finden sich in der Datei „Kap8daten3.sav", die Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

In diesem Fall liegt ein sog. gemischtes Modell vor, da es sich nur bei einem der beiden Faktoren um einen Messwiederholungsfaktor handelt, nämlich beim Messzeitpunkt, der über drei Stufen verfügt. D.h. für jede:n Patient:in liegen drei Depressionswerte zu jeweils einem der drei Messzeitpunkte vor.

Bei dem anderen Faktor, der Therapieform, handelt es sich hingegen um einen Zwischensubjektfaktor, da jede:r Patient:in entweder in der einen Stufe (Verhaltenstherapie) oder in der anderen Stufe (Therapiemix) vorliegt. Insgesamt wird also eine zweifaktorielle Varianzanalyse in einem gemischten 2x3 Design durchgeführt.

Zur Durchführung wählen wir erst wieder *Analyze >> General Linear Model >> Repeated Measures…* und definieren anschließend unseren 3-stufigen Messwiederholungsfaktor, dem wir z.B. wieder den Namen „Messzeitpunkt“ geben. Nach Klick auf „Define“ weisen wir die drei Variablen „BDI1“, „BDI2“ und „BDI3“ wieder (in der richtigen Reihenfolge) unseren drei Faktorstufen zu. Zudem haben wir in diesem Beispiel noch einen Zwischensubjektfaktor, die Variable *Therapie*, die wir daher in das Feld „Between-Subjects Factor(s)“ ziehen, siehe Abbildung 8.7.

Unter „Plots…“ fordern wir wieder eine graphische Darstellung unserer Resultate an, mit dem Messzeitpunkt auf der horizontalen Achse und den Therapieformen in unterschiedlichen Linien (dies sollte für diesen Fall in dieser einen Form genügen, um unsere Fragestellung zu erhellen). Selbstverständlich wollen wir neben der Punktschätzung (Mittelwert) auch eine Darstellung plausibler Bereiche durch entsprechende Konfidenzintervalle.

Unter „Options…“ fordern wir dieses Mal neben „Descriptive statistics“ und „Estimates of effect size“ auch noch „Homogeneity tests“ an, da wir auch einen Zwischensubjektfaktor vorliegen haben. Unter „EM Means…“ verlangen wir wieder paarweise Vergleiche für beide Therapieformen bzw. zu allen Messzeitpunkten, indem wir „Therapie * Messzeitpunkt“ in das Feld „Display Means for“ ziehen, „Compare simple main effects“ anwählen und die Bonferroni-Methode zur Korrektur von p-Werten und Konfidenzniveaus auswählen.

Nach Ausführen der entsprechenden Kommandozeilen in der Syntaxdatei bekommen wir wieder eine sehr umfangreiche Ausgabe, von der wir hier wieder nur die wesentlichen Bestandteile erläutern. In der Tabelle „Descriptive Statistics“ haben wir wieder deskriptive Statistiken in Form von Mittelwerten und Standardabweichungen für alle Kombinationen an Faktorstufen gegeben.

In der Tabelle „Box’s Test of Equality of Covariance Matrices“ haben wir einen Test auf Gleichheit der Kovarianzmatrizen für alle Faktorstufen unseres Zwischensubjektfaktors. Diese

Voraussetzung ist im Fall eines gemischten Designs zusätzlich zu allen bisher besprochenen Voraussetzungen bei Varianzanalysen zu prüfen. Wie bei Levenes Test (siehe unten) gilt: ist dieser Test signifikant (üblicherweise mit $\alpha = .01$) kann davon ausgegangen werden, dass die Voraussetzung nicht erfüllt ist. In diesem Fall müsste dann auf ein sog. robustes Verfahren zurückgegriffen werden (siehe z.B. Mair & Wilcox, 2020), die im Rahmen dieser Übungen aber nicht behandelt werden. Im vorliegenden Fall ist der Box-Test nicht signifikant, $p = .996$, und wir können ohne weitere Umschweife mit der Überprüfung der Sphärizität weitermachen und sehen, dass auch der Mauchly-Test nicht signifikant ist, $p = .556$. Bleibt noch die Überprüfung der Varianzhomogenität in der Tabelle „Levene's Test of Equality of Error Variances". Auch hier können wir erleichtert aufatmen, da Levenes Test zu keinem der drei Messzeitpunkte signifikant ist, es ergibt sich jeweils $p > .05$.

Damit können wir schließlich zu den eigentlichen Ergebnissen unserer Varianzanalyse kommen. Einen Teil davon finden wir wieder in der Tabelle „Tests of Within-Subjects Effects". Dort sehen wir, dass wir einen signifikanten Haupteffekt für den Faktor Messzeitpunkt haben, $F(2, 396) = 2897.67$, $p < .001$, $\eta_p^2 = .94$, also wiederum einen (sehr) großen Effekt des Messzeitpunkts. Wir sehen auch, dass eine signifikante Interaktion zwischen dem Messzeitpunkt und der Therapieform besteht, $F(2, 396) = 26.25$, $p < .001$, $\eta_p^2 = .12$, d.h. mit einem mittleren Effekt gemäß Cohens Heuristik (1988). Dies beantwortet im Prinzip schon unsere Forschungsfrage: Die Wirkung der Therapieformen scheint sich in der Tat abhängig vom Zeitverlauf zu unterscheiden!

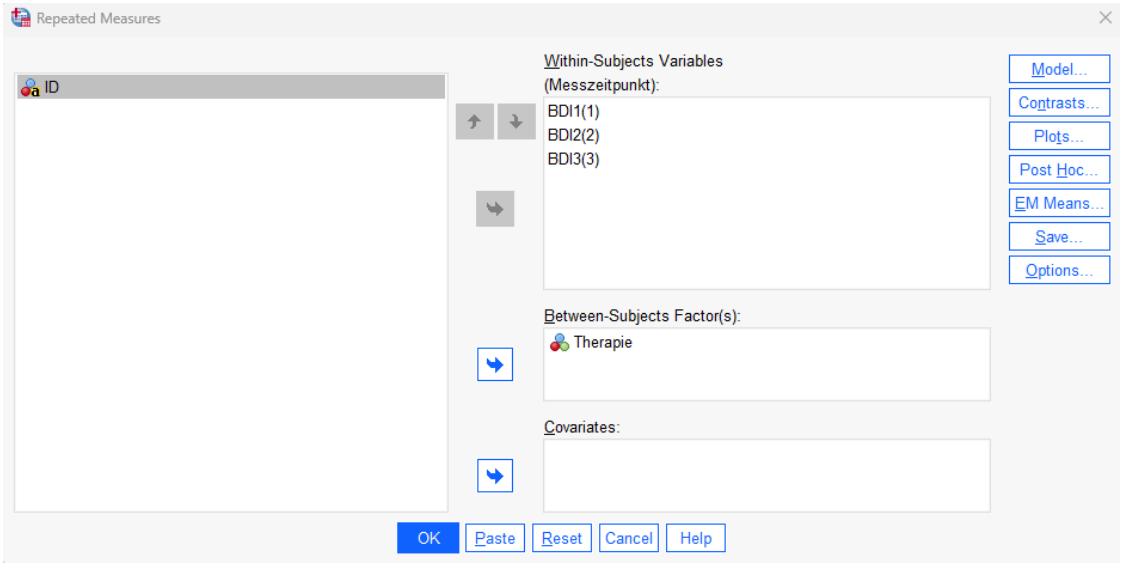


Abbildung 8.7. Definition einer Varianzanalyse mit gemischtem Design in SPSS.

Allerdings wollen wir auch noch wissen, ob sich die mittleren Depressionswerte für die beiden Therapieformen ohne Rücksicht auf den Messzeitpunkt voneinander unterscheiden. D.h. wir sind auch daran interessiert, ob es einen Haupteffekt für die Therapieform gibt. Diese Frage kann uns nun die Tabelle „Tests of Between-Subjects Effects“ erhellen, die für uns nun auch endlich interessant geworden ist, da wir mit der Therapieform einen Zwischensubjektfaktor vorliegen haben. In dieser Tabelle können wir ablesen, dass es auch einen signifikanten Haupteffekt für die Therapieform gibt, $F(1, 198) = 168.31$, $p < .001$, $\eta_p^2 = .46$, d.h. ein großer Effekt gemäß Cohens Heuristik (1988).

Im Abschnitt „Estimated Marginal Means“ finden wir wieder sämtliche Informationen zu den angeforderten paarweisen Mittelwertvergleichen. Hier erkennen wir z.B. in der (ersten) Tabelle „Pairwise Comparisons“ (diese Tabelle gibt es ja wieder zweimal), dass sich die mittleren Depressionswerte für Verhaltenstherapie und Therapiemix zum ersten Messzeitpunkt nicht signifikant unterscheiden ($p = .058$), während sie es für den Zeitpunkt 2 und 3 jeweils tun ($p < .001$). Dies macht auch durchaus Sinn, da ja der erste Zeitpunkt den Therapiebeginn bezeichnet und sich da zufällig gezogene Stichproben depressiver Patient:innen in ihren mittleren Depressionswerten nur selten stark unterscheiden sollten, da der Populationsmittelwert ja ein- und derselbe sein sollte. Zu den anderen beiden Zeitpunkten sehen wir aber, dass sich zwischen den beiden Therapieformen eine Lücke auftut, was es ganz so aussehen lässt, als würde sich der Therapiemix besser auf die Depressionssymptomatik auswirken als die Verhaltenstherapie alleine. Mit zunehmender Zeit scheint diese Lücke auch größer zu werden, die Bereiche für plausible Werte überlappen nicht, d.h. es scheint auch unwahrscheinlich, dass es sich bei der Zunahme des Unterschieds nur um eine Zufallsschwankung handelt.

In der zweiten Tabelle mit der Überschrift „Pairwise Comparisons“ sehen wir schließlich noch, dass sich die mittleren Depressionswerte in beiden Therapieformen zwischen allen drei Zeitpunkten signifikant unterscheiden ($p < .001$); in beiden Therapieformen nimmt die Depressionssymptomatik also über die Zeit hinweg ab. Diesen Verlauf sowie die oben konstatierte Tatsache der sich öffnenden Lücke zwischen den beiden Therapieformen sehen wir auch in der graphischen Darstellung der Resultate in Abbildung 8.8 noch einmal schön illustriert. Der Interaktionseffekt zeitigt sich in diesem Beispiel also insofern, dass die Abnahme der Depressionssymptomatik über die Zeit hinweg je nach Therapieform unterschiedlich stark ausgeprägt ist.

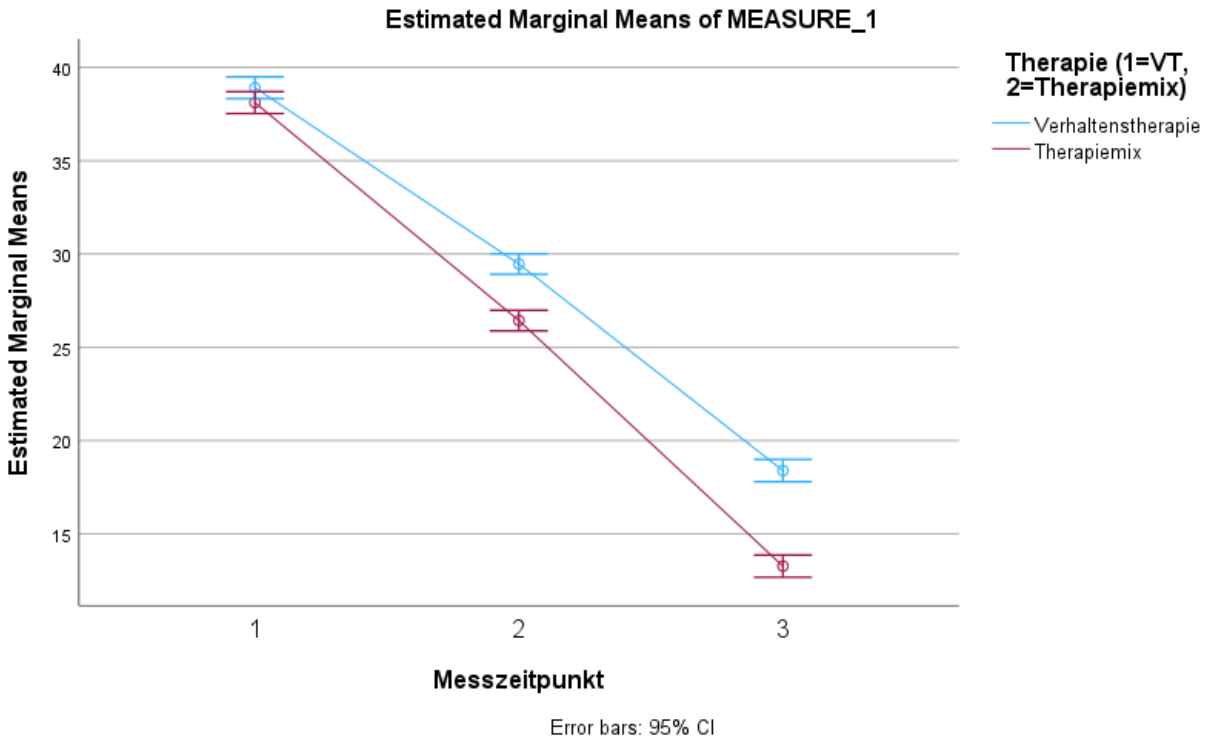


Abbildung 8.8. Mittlere Depressionswerte über die drei Messzeitpunkte für beide Therapieformen. Fehlerbalken entsprechen 95%-KI.

Ergebnisbericht

Ein Ergebnisbericht für dieses Beispiel könnte wie folgt aussehen: „Weder der Box Test (p = .996) noch Mauchlys Test (p = .556) noch Levenes Tests auf jeder Faktorstufe des Messwiederholungsfaktors (jeweils p > .05) waren signifikant, weshalb davon ausgegangen wird, dass die Voraussetzungen für eine Varianzanalyse mit einem gemischten Design erfüllt sind. Die mittleren Depressionswerte unterscheiden sich signifikant für die drei Messzeitpunkte, $F(2, 396) = 2897.67$, $p < .001$, $\eta_p^2 = .94$, was einem großen Effekt gemäß Cohen (1988) entspricht. Die mittleren Depressionswerte unterscheiden sich zudem signifikant zwischen den beiden Therapieformen, $F(1, 198) = 168.31$, $p < .001$, $\eta_p^2 = .46$, was ebenfalls einem großen Effekt gemäß Cohens Heuristik (1988) entspricht. Schließlich gibt es eine signifikante Interaktion zwischen Therapieform und Messzeitpunkt, $F(2, 396) = 26.25$, $p < .001$, $\eta_p^2 = .12$, was gemäß Cohens Heuristik (1988) einem mittleren Effekt entspricht. Für paarweise post-hoc Vergleiche werden gemäß Bonferroni korrigierte p-Werte berichtet. Für beide Therapieformen unterscheiden sich die mittleren Depressionswerte signifikant voneinander zwischen allen Messzeitpunkten ($p < .001$). Insbesondere nehmen die Depressionswerte für beide Therapieformen über

die Zeit hinweg ab. Zu Messzeitpunkt 1, d.h. zu Beginn der jeweiligen Therapie, unterscheiden sich die mittleren Depressionswerte für die beiden Therapieformen nicht signifikant voneinander ($p = .058$), während sie sich für die anderen beiden Messzeitpunkte signifikant voneinander unterscheiden ($p < .001$). Insbesondere sind die mittleren Depressionswerte jeweils niedriger für den Therapiemix als für die Verhaltenstherapie und der Unterschied nimmt in der Zeit zu. Deskriptive Statistiken für alle Kombinationen aus Therapieform und Messzeitpunkt sind in Tabelle 8.2 zu finden. Eine graphische Darstellung von Punkt- und Intervallschätzungen für die mittleren Depressionswerte zu den einzelnen Messzeitpunkten für beide Therapieformen ist in Abbildung 8.8 gegeben."

Tabelle 8.2

Deskriptive Statistiken

Messzeitpunkt	Therapieform	*M*	*SD*	*n*
1	Verhaltenstherapie	38.91	2.89	100
	Therapiemix	38.11	3.04	100
2	Verhaltenstherapie	29.46	2.77	100
	Therapiemix	26.43	2.81	100
3	Verhaltenstherapie	18.39	3.00	100
	Therapiemix	13.27	3.06	100

Stichprobenplanung

Eine Stichprobenplanung für die Omnibustests der varianzanalytischen Modelle mit Messwiederholung ist noch etwas komplizierter als im Fall ohne Messwiederholung. Da die statistischen Hypothesen der Omnibustests ohnehin nur selten inhaltliche Forschungsfragen beantworten können (Bühner et al., 2025), verzichten wir hier auf die Beschreibung.

Für einfache Parameterdifferenzen kann auf die besprochenen Verfahren für Hypothesentests und Konfidenzintervalle im Zwei-Stichprobenfall zurückgegriffen werden (Bühner et al., 2025).

Übungsaufgaben

Alle im Folgenden benötigten Datendateien können Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Beispiel 8.1

Was gehört zu den Voraussetzungen der Varianzanalyse mit vollständiger Messwiederholung (d.h. kein gemischtes Design)?

(a) Die UV muss intervallskaliert sein.

(b) Die Voraussetzung der Compound Symmetry muss erfüllt sein.

(c) Die Kovarianz der Personeneffekte und der Fehler muss Null sein.

(d) Die AV muss auf jeder Stufe aller Messwiederholungsfaktoren gleichverteilt sein.

Beispiel 8.2

Was gehört zu den Voraussetzungen der Varianzanalyse mit gemischtem Design?

(a) Varianzhomogenität der AV bezüglich des Zwischensubjektfaktors.

(b) Gleichheit der Kovarianzmatrizen über die Faktorstufen des Innersubjektfaktors.

(c) Gleichheit der Kovarianzmatrizen über die Faktorstufen des Zwischensubjektfaktors.

(d) Sphärizität.

Beispiel 8.3

Geben Sie für jede der folgenden Aussagen an, ob sie richtig oder falsch ist.

Nr.	**Aussage**	**R/F**
1)	Bei der Effektstärke η_p^2 werden Werte ab 0.01/0.06/0.14 gemäß Cohen (1988) als klein/mittel/groß bezeichnet.	
2)	Beim Box Test handelt es sich um einen Test der Sphärizität.	
3)	Die Greenhouse-Geisser-Korrektur ist zu konservativ, weshalb besser die Huynh-Feldt-Korrektur verwendet werden sollte.	
4)	Die Gleichheit der Kovarianzmatrizen kann mit Mauchlys Test überprüft werden.	

Beispiel 8.4

Ein (fiktives) Nationalparkprogramm möchte sein Veranstaltungsprogramm evaluieren. Insbesondere soll herausgefunden werden, ob sich durch die Teilnahme am Programm, das Umweltverhalten von Teilnehmer:innen kurz- bzw. langfristig ändert. Dazu wird allen Teilnehmer:innen über einen bestimmten Zeitraum u.a. ein Fragebogen zum Umweltverhalten vor Beginn sowie ein Monat und ein Jahr nach einer entsprechenden Veranstaltung ausgehändigt. Bei der Skala zum Umweltverhalten kann jede Person 0 bis maximal 24 Punkte erhalten. Die Fragestellung lautet: Unterscheidet sich das Umweltverhalten von Teilnehmer:innen zu diesen drei Zeitpunkten und falls ja, wie?

Sie finden die Daten für dieses Beispiel in der Datei „Kap8UE4.sav“. Wählen Sie ein angemessenes statistisches Verfahren, um die Fragestellung auf Basis der gegebenen Daten zu beantworten, und fertigen Sie einen entsprechenden Ergebnisbericht an.

Beispiel 8.5

Ein Psychologe erforscht wie sich die Bedeutung des Lernmaterials auf die Gedächtnisleistung auswirkt. Dazu lässt er 50 Studierende sowohl 17 Paare sinnloser Silben sowie 17 Paare aus deutschen und japanischen Begriffen erlernen. Die Behaltensleistung fragt der Forscher danach zu drei Zeitpunkten ab: (i) eine halbe Stunde nach dem Erlernen der Begriffe, (ii) einen Tag später, (iii) eine Woche später. Die Forschungsfrage lautet: Hängt die Zeitabhängigkeit der Behaltensleistung von der Bedeutung des Lernmaterials ab?

Sie finden die Daten für dieses Beispiel in der Datei „Kap8UE5.sav“. Wählen Sie ein angemessenes statistisches Verfahren, um die Fragestellung auf Basis der gegebenen Daten zu beantworten, und fertigen Sie einen entsprechenden Ergebnisbericht an.

Beispiel 8.6

Ein Patient leidet neuerdings an erhöhtem Blutdruck und seine Medikation muss so eingestellt werden, dass der Ruheblutdruck sich stabil in einem Normalbereich befindet, d.h. der systolische (= der obere/höhere) Blutdruckwert sollte sich zwischen 115 und 125 mmHg und der diastolische (= der untere/niedrigere) Wert zwischen 70 und 80 mmHg befinden.

In der Datei „Kap8UE6.sav“ finden Sie Blutdruckwerte für diesen Patienten, die über längere Zeiträume bei drei verschiedenen Medikationen sowohl am linken als auch am rechten Arm gemessen wurden. Bei den drei verschiedenen Medikationen handelt es sich um Gabe von (i) 8 mg Candesartan abends und 8 mg Candesartan morgens, (ii) 8 mg Candesartan abends und 16 mg Candesartan morgens, (iii) 8 mg Candesartan sowie 5 mg Amlodipin abends und 16 mg Candesartan morgens.

Sie können für dieses Beispiel davon ausgehen, dass die Voraussetzungen für Varianzanalysen mit Messwiederholung erfüllt sind (Sie können sich selbst davon überzeugen, dass sie es eigentlich nicht sind, aber da wir für diesen Fall keine angemessenen robusten Verfahren im Rahmen dieser Übungen besprechen, können Sie für dieses Beispiel so tun, als wäre alles in Ordnung). Verwenden Sie ferner ein Signifikanzniveau von $\alpha = .05$ sowie ein Konfidenzniveau von 95% für dieses Beispiel.

Bilden Sie zuerst jeweils einen mittleren systolischen und diastolischen Blutdruckwert aus den beiden Werten für die beiden Arme. Untersuchen Sie anschließend, ob die beiden Blutdruckwerte sich für die drei Medikationen unterscheiden, und falls ja wie. Erstellen Sie einen angemessenen Ergebnisbericht und berichten Sie auch plausible Werte für die beiden Blutdruckwerte für die drei Medikationen. Welche Medikation erscheint Ihnen aufgrund Ihrer Resultate für diesen Patienten am passendsten?

Beispiel 8.7

Eine (fiktive) Forschungsgruppe möchte wissen wie sich unterschiedliche Lehrmethoden auf die Entwicklung von Statistik-Expertise auswirken. Dazu werden jeweils 120 Studierende mit zwei unterschiedlichen Lehrmethoden ein Semester lang in Statistik unterrichtet. In einer Gruppe wird traditionelle Lehre eingesetzt (Powerpoint-Vortrag und schriftliche Klausur), in der anderen Gruppe wird das flipped-classroom Konzept verwendet (eigenständige Aneignung der Inhalte zu Hause und gemeinsame Diskussion und praktische Übungen in den jeweiligen Unterrichtseinheiten).

Die Statistikkenntnisse werden zu drei Zeitpunkten erhoben. Eine Überprüfung zu Beginn der Lehrveranstaltung soll das Vorwissen der Studierenden erfassen. Eine Überprüfung nach Ende der Lehrveranstaltung soll den Lernerfolg mit der jeweiligen Methode erfassen. Eine dritte Überprüfung ein halbes Jahr nach der zweiten Überprüfung soll Rückschlüsse darauf erlauben wie gut die Inhalte mit den

jeweiligen Methoden über einen längeren Zeitraum behalten werden. Für die drei Überprüfungen wird jeweils derselbe Leistungstest verwendet. Dieser ergibt als Maß für die Leistung eine Zahl zwischen 0 und 100.

Die Daten für das Experiment sind in der Datei „Kap8UE7.sav“ enthalten. Die Variable *ID* enthält den Proband:innencode. Die Variable *Lernmethode* gibt an, in welcher der beiden Gruppen sich die jeweiligen Studierenden befanden (0 = traditionelle Lehrmethode, 1 = flipped classroom). Die Variablen *t1*, *t2* und *t3* beinhalten die Ergebnisse der Statistik-Leistungstests jeweils zu den Zeitpunkten am Beginn der Lehrveranstaltung (*t1*), an deren Ende (*t2*) und ein halbes Jahr nach Ende der Lehrveranstaltung (*t3*).

Führen Sie eine Varianzanalyse inklusive paarweiser post-hoc Vergleiche durch um die unten angegebenen Fragen zu beantworten. Begründen Sie Ihre Antworten dabei jeweils durch Angabe der entsprechenden statistischen Kennwerte. Berichten Sie bei statistischen Ergebnissen immer alle relevanten Kennwerte (Teststatistiken, Freiheitsgrade, *p*-Werte, Effektstärken). Mittelwerte und Standardabweichungen können Sie entweder im Text anführen oder in APA-konformen Tabellen angeben und auf diese verweisen. Berichten Sie Ihre Resultate gemäß APA-Richtlinien. Das Signifikanzniveau soll für alle statistischen Tests zu 0.05 gewählt werden und für post-hoc Tests unter Angabe des Verfahrens zur Korrektur angemessen korrigiert werden.

(a) Welche spezifische Form der Varianzanalyse ist zu wählen, um zu prüfen, ob sich die resultierenden mittleren Leistungen der Studierenden für die beiden Lehrmethoden zu irgendeinem der drei Zeitpunkte unterscheiden? Wie viele Faktoren liegen vor und welcher Art sind die Faktoren? Sind die Voraussetzungen (abgesehen von der Normalverteilung der abhängigen Variablen, von der Sie für diese Aufgabe ausgehen können) für das statistische Verfahren erfüllt? Begründen Sie Ihre Antwort.

(b) Unterscheiden sich die Ergebnisse im Leistungstest signifikant zwischen den Testzeitpunkten?

(c) Unterscheiden sich die Ergebnisse im Leistungstest signifikant zwischen den beiden Lehrmethoden?

(d) Gibt es eine Interaktion zwischen Lehrmethode und Zeitpunkt? Wenn ja, welche Art der Interaktion liegt vor (begründen Sie Ihre Wahl)?

(e) Untersuchen Sie mittels paarweisen Mittelwertvergleichen, ob und wie sich die beiden Lehrmethoden zu den einzelnen Zeitpunkten voneinander unterscheiden bzw. wie sich die Ergebnisse zu den unterschiedlichen Zeitpunkten in der jeweiligen Lehrmethode voneinander unterscheiden. Berichten Sie für jeden Mittelwertvergleich auch dessen statistische Signifikanz.

Beispiel 8.8

Eine Gruppe von (fiktiven) Forscher:innen untersuchte die folgende Fragestellung: Wie wirken sich unterschiedliche Therapieformen (kognitive Verhaltenstherapie und achtsamkeitsbasierte Therapie) und der Zeitpunkt (vor und nach der Therapie) auf die Depressionsschwere bei Erwachsenen aus? Die Depressionsschwere wurde für jeweils 50 Personen für jede der beiden Therapieformen mit Becks Depressionsinventar erhoben und ergibt eine Zahl von 0 bis 63 für jede Person und jeden Zeitpunkt. Die entsprechenden Daten befinden sich in der Datei „Kap8UE8.sav“.

(a) Welches Verfahren ist zur inferenzstatistischen Untersuchung der Daten für die oben angegebene Fragestellung geeignet? Erläutern/beschreiben Sie kurz die Bestandteile des Verfahrens (Was ist/sind UV/AV bzw. Faktoren und Art der Faktoren?).

(b) Analysieren Sie die Daten in SPSS und verfassen Sie einen entsprechenden Ergebnisbericht.

(c) Wie würden Sie das Ergebnis inhaltlich interpretieren (in 1-2 Sätzen)?

<u>*Hinweis*</u>: Sie können für dieses Beispiel davon ausgehen, dass die Voraussetzungen für das benötigte Verfahren erfüllt sind (d.h., Sie müssen keine Voraussetzungsprüfungen durchführen).

Beispiel 8.9

Eine (fiktive) Forscherin fragt sich, wie das Spielen von bestimmten Computerspielen mit moralischem Verhalten und dem Bedürfnis nach kognitiven Anforderungen zusammenhängt. Um dieser Frage nachzugehen, rekrutiert sie insgesamt 300 Versuchspersonen. Je 100 von diesen Versuchspersonen spielen entweder besonders gerne (i) Egoshooter, (ii) fantasiebasierte Computerrollenspiele oder (iii) Aufbau-Strategiespiele. Alle Versuchspersonen füllen einen Moralfragebogen aus, bei dem sie zwischen

0 und 100 Punkten erreichen können. Eine höhere Punktezahl bedeutet dabei moralischeres Antwortverhalten. Zudem füllen alle Versuchspersonen einen Fragebogen zu ihrem Bedürfnis nach kognitiven Anforderungen aus, bei dem sie wiederum zwischen 0 und 100 Punkten erreichen können. Eine höhere Punktezahl bedeutet ein größeres Bedürfnis nach kognitiven Anforderungen. Die erhobenen Daten und Gruppenzugehörigkeiten finden sich in der Datei „Kap8UE9.sav“.

Da sich die Forscherin hinsichtlich statistischer Auswertungen nicht mehr ganz sicher ist, zieht sie für die Auswertung ChatGPT zurate. Aus dieser Zusammenarbeit ergibt sich der folgende Ergebnisbericht. Dieser ist leider in mehreren Punkten fehlerhaft. Identifizieren und korrigieren Sie die Fehler. Streichen Sie dazu die fehlerhaften Stellen durch und ersetzen Sie sie durch die korrekten Angaben.

Hinweis: Es müssen keine Formulierungen geändert werden. Alle Fehler lassen sich durch Änderung/Ersetzung einzelner Wörter oder Zahlen beheben.

Ergebnisbericht: Um Unterschiede im moralischen Verhalten und dem Bedürfnis nach kognitiven Anforderungen je nach Spielegenre zu untersuchen, wurde eine zweifaktorielle Varianzanalyse ohne Messwiederholung durchgeführt. Dabei wies der Zwischensubjektfaktor „Spielegenre“ zwei Faktorstufen auf. Der Innersubjektfaktor berücksichtigte, um welchen der beiden Fragebögen es sich handelte. Die Interaktion zwischen den beiden Faktoren lässt darauf schließen, ob es zwischen den Spielegenres Unterschiede im Antwortverhalten auf die beiden Fragebögen gibt.

Die Varianzanalyse ergab einen signifikanten Haupteffekt für die Art des Fragebogens, $F(1,297) = 9.64$, $p = .031$, $\eta_p^2 = .03$. Es ergab sich auch ein signifikanter Haupteffekt für das Spielegenre, $F(2,297) = 27.21$, $p < .001$, $\eta_p^2 = .16$. Die Interaktion zwischen Fragebogenart und Spielegenre war allerdings nicht signifikant, $F(2,297) = 19.42$, $p = .116$, $\eta_p^2 = .12$, weshalb die Effekte der beiden Faktoren nicht unabhängig voneinander interpretiert werden können.

Bei Spieler:innen, die besonders gerne Egoshooter spielen, wurden sowohl beim Moralfragebogen ($M = 49.90$, $SD = 9.09$) als auch beim Kognitionsfragebogen ($M = 51.34$, $SD = 9.52$) vergleichsweise geringe Werte erreicht, die sich auch nicht signifikant voneinander unterschieden, $p = .188$. Auch die Punktwerte der Spieler:innen, die besonders gerne Strategiespiele spielen, waren sehr

ähnlich bei Moralfragebogen ($M = 51.34$, $SD = 9.52$) und bei Kognitionsfragebogen ($M = 58.38$, $SD = 9.45$), fielen aber vergleichsweise deutlich höher aus und unterschieden sich in beiden Fällen signifikant von den jeweiligen Werten der Egoshooter-Spieler:innen ($p < .001$). Rollenspieler:innen erzielten hingegen ganz andere Werte im Moralfragebogen ($M = 59.87$, $SD = 9.01$) als im Kognitionsfragebogen ($M = 52.41$, $SD = 9.43$), der Unterschied war nicht signifikant ($p < .001$). Rollenspieler:innen erzielten im Kognitionsfragebogen ähnliche Werte wie Egoshooter-Spieler:innen ($p = .631$), während sie im Moralfragebogen ähnliche Werte wie Strategie-Spieler:innen erzielten ($p > .999$).

Beispiel 8.10

Eine (fiktive) Forscherin fragt sich, wie die allgemeine Selbstwirksamkeit mit dem Lernergebnis in spielerischen und nichtspielerischen Lernumgebungen zusammenhängt. In einer Vorerhebung wird die allgemeine Selbstwirksamkeit von 232 Versuchspersonen erfasst. Von allen Versuchspersonen werden daraufhin diejenigen 160 ausgewählt, deren Selbstwirksamkeit entweder besonders hoch oder besonders niedrig war. Die 80 Versuchspersonen mit besonders niedriger Selbstwirksamkeit werden der Gruppe „niedrig“ zugeteilt, die 80 Versuchspersonen mit besonders hoher Selbstwirksamkeit der Gruppe „hoch“ (die Versuchspersonen selbst wissen zu keinem Zeitpunkt über die Gruppenzugehörigkeit Bescheid). Beide Gruppen beschäftigen sich zu zwei verschiedenen Zeitpunkten mit einer Lernaufgabe (eine Fremdsprache erlernen). Zu einem Zeitpunkt beschäftigen sie sich mit einer spielerischen Version der Lernaufgabe (mittels des Online-Tools Duolingo), zum anderen Zeitpunkt mit einer nichtspielerischen Version (üblicher Sprachunterricht). Die Zuweisung der spielerischen und nichtspielerischen Varianten zu den beiden Zeitpunkten erfolgt randomisiert. Direkt nach den Lerneinheiten wird jeweils ein Test zum jeweiligen Lernfortschritt durchgeführt, der das Lernergebnis auf einer Skala von 0 bis 100 angibt. Die erhobenen Daten und Gruppenzugehörigkeiten finden sich in der Datei „Kap8UE10.sav“.

Da sich die Forscherin hinsichtlich statistischer Auswertungen nicht mehr ganz sicher ist, zieht sie für die Auswertung ChatGPT zurate. Aus dieser Zusammenarbeit ergibt sich der folgende Ergebnisbericht. Dieser ist leider in mehreren Punkten fehlerhaft. Identifizieren und korrigieren Sie die

Fehler. Streichen Sie dazu die fehlerhaften Stellen durch und ersetzen Sie sie durch die korrekten Angaben.

Hinweis: Es müssen keine Formulierungen geändert werden. Alle Fehler lassen sich durch Änderung einzelner Wörter oder Zahlen beheben.

Ergebnisbericht: Zur statistischen Analyse wurde eine zweifaktorielle Varianzanalyse ohne Messwiederholung (gemischtes Design) durchgeführt. Beim zweistufigen Innersubjektfaktor handelt es sich um die Variable, die angibt, ob es sich um eine Person mit besonders niedriger oder hoher Selbstwirksamkeit handelt. Beim ebenfalls zweistufigen Zwischensubjektfaktor handelt es sich um die Variable, die angibt, ob es sich um das Lernergebnis zur Spiel- oder zur Nichtspielversion der Lernaufgabe handelt.

Es ergibt sich ein (mit $\alpha = .05$) signifikanter Haupteffekt für die Selbstwirksamkeit (niedrig oder hoch), $F(1,158) = 0.28$, $p = .002$, $\eta_p^2 = .60$. Es ergibt sich ein signifikanter Haupteffekt für die Version der Lernaufgabe (Spiel oder Nichtspiel), $F(1,158) = 0.10$, $p = .001$, $\eta_p^2 = .75$. Es ergibt sich eine signifikante Interaktion zwischen den beiden Faktoren, $F(1,158) = 3769.10$, $p < .001$, $\eta_p^2 = .96$.

Paarweise post-hoc Vergleiche mit gemäß Bonferroni korrigierten p-Werten ergeben, dass Personen mit niedriger Selbstwirksamkeit in der Spielversion ($M = 45.56$, $SD = 9.81$) signifikant höhere Ergebnisse als in der Nichtspielversion ($M = 40.71$, $SD = 9.74$) erzielen, $p < .001$. Bei Personen mit hoher Selbstwirksamkeit ist es gerade umgekehrt: Diese erzielen in der Spielversion ($M = 41.44$, $SD = 8.01$) signifikant niedrigere Ergebnisse als in der Nichtspielversion ($M = 46.34$, $SD = 8.18$), $p < .001$. Zudem erzielen in der Spielversion Personen mit niedriger Selbstwirksamkeit signifikant höhere Ergebnisse als Personen mit hoher Selbstwirksamkeit, $p < .001$. In der Nichtspielversion hingegen erzielen Personen mit niedriger Selbstwirksamkeit signifikant niedrigere Ergebnisse als Personen mit hoher Selbstwirksamkeit, $p < .001$.

Beispiel 8.11

Es wird ein Experiment durchgeführt, mit dem die Wirkung von Silikonen in Haarpflegeprodukten auf die Haarqualität geprüft werden soll. Für das Experiment werden 160 Personen rekrutiert, die bislang keine Haarpflegeprodukte mit Silikonen verwendet haben. Die Personen werden mit Haarpflegeprodukten mit Silikonen ausgestattet und gebeten, diese streng nach Gebrauchsanweisung für zwei Monate zu verwenden. Danach sollen weitere 10 Monate lang keine Haarpflegeprodukte mit Silikonen verwendet werden. Die Haarqualität wird für alle Personen von einer Gruppe von Expert:innen auf einer Skala von 0 bis 100 zu drei Zeitpunkten beurteilt: (i) zu Beginn des Experiments, (ii) zwei Monate nach Beginn des Experiments, (iii) ein Jahr nach Beginn des Experiments.

Die Daten sind in der Datei „Kap8UE11.sav“ gegeben. Wählen Sie ein geeignetes statistisches Verfahren, um die Frage zu erhellen, ob sich die Haarqualität im Mittel zu den drei verschiedenen Zeitpunkten unterscheidet und falls ja, wie. Erstellen Sie anschließend einen entsprechenden Ergebnisbericht. Wie würden Sie das Ergebnis inhaltlich interpretieren?

Beispiel 8.12

Eine Forschungsgruppe untersucht Vor- und Nachteile verschiedener Lernmethoden. In einem Experiment werden die beiden Lernmethoden „Massiertes Lernen“ und „Verteiltes Lernen“ miteinander verglichen. Dazu werden 100 Schüler:innen auf zwei gleich große Gruppen aufgeteilt. Beide Gruppen werden für ein Semester lang mit denselben Inhalten unterrichtet. Bei beiden Gruppen wird am jeweils am Anfang und am Ende des Semesters ein Test durchgeführt. Eine Gruppe wird instruiert sich auf den Abschlusstest mit der Methode des massierten Lernens vorzubereiten. Die andere Gruppe soll sich auf den Abschlusstest mit der Methode des verteilten Lernens vorbereiten. Bei beiden Tests werden sowohl lexikalisches und prozedurales Wissen (Variablen *LexWis* und *ProWis*) als auch die Durchführungseffizienz (Variable *DurEff*) erfasst. Die erhobenen Daten sind in der Datei „Kap8UE12.sav“ gegeben.

Erstellen Sie sowohl für den Test zu Semesterbeginn einen Index für die Prüfungsleistung, indem Sie den Mittelwert aus den drei Variablen *LexWis_vorher*, *ProWis_vorher* und *DurEff_vorher* bilden. Verfahren Sie anschließend ganz analog für die drei Variablen *LexWis_nachher*,

ProWis_nachher und *DurEff_nachher*, um einen entsprechenden Index für die Prüfungsleistung für den Test am Semesterende zu erstellen.

Wählen Sie anschließend ein geeignetes statistisches Verfahren, um die Abhängigkeit der Prüfungsleistung vom Zeitpunkt und der Lernmethode inferenzstatistisch zu untersuchen. Erstellen Sie schließlich einen entsprechenden Ergebnisbericht.

Beispiel 8.13

Es wird eine neue VR-Methode (VR steht für „Virtuelle Realität“) für den Mathematikunterricht untersucht. Die Wirksamkeit der VR-Methode wird dafür mit einer klassischen Lehrmethode (Tafelunterricht) verglichen. Dazu werden jeweils 50 Schüler:innen ein Semester lang entweder mit der VR-Methode oder der klassischen Lehrmethode unterrichtet. Zudem wird die Leistung der Schüler:innen zu Beginn und am Ende eines Semesters mit einem standardisierten Mathematiktest erhoben.

Nach Erhebung der Daten (siehe „Kap8UE13.sav“) wurde ein geeignetes statistisches Verfahren verwendet, um den folgenden lückenhaften Ergebnisbericht zu erstellen. Ihre Aufgabe besteht nun darin, diesen Ergebnisbericht zu vervollständigen. Die Lücken sind jeweils durch grau hinterlegte Bereiche markiert.

Lückenhafter Ergebnisbericht:

Alle folgenden inferenzstatistischen Ergebnisse beziehen sich auf ein Signifikanzniveau von $\alpha = .005$.

Die mittleren Leistungen beim Mathematiktest unterscheiden sich signifikant für die beiden Messzeitpunkte, $F(1, 98) =$ [ ], $p < .001$, $\eta_p^2 =$ [ ], was einem großen Effekt gemäß Cohen (1988) entspricht. Die mittleren Leistungen unterscheiden sich hingegen nicht signifikant zwischen den beiden Lehrmethoden, $F(1,$ [ ]$) = 2.31$, $p =$ [ ], $\eta_p^2 = .02$, was einem [ ] Effekt gemäß Cohens Heuristik (1988) entspricht. Zwischen Messzeitpunkt und Lehrmethode besteht eine [ ] Interaktion, $F($[ ]$, 98) = 8.86$, $p =$ [ ], $\eta_p^2 =$ [ ], was gemäß Cohens Heuristik (1988) einem [ ] Effekt entspricht.

Für paarweise post-hoc Vergleiche werden gemäß Bonferroni korrigierte p-Werte berichtet. Für beide Lehrmethoden unterscheiden sich die mittleren Leistungen signifikant zwischen beiden Messzeitpunkten ($p < .001$). Insbesondere nimmt die Leistung im Mittel für beide Lehrmethoden von Beginn zu Ende des Semesters . Am Beginn des Semesters unterscheiden sich die mittleren Leistungen für die beiden Lehrmethoden nicht signifikant voneinander ($p =$). Auch am Ende des Semesters unterscheiden sich die mittleren Leistungen für die beiden Lehrmethoden voneinander ($p =$).

Deskriptive Statistiken sind in der Tabelle unten zusammengefasst. Wir sehen, dass die Mathematikleistung für beide Lehrmethoden (Haupteffekt Messzeitpunkt). Die Leistungszunahme ist allerdings stärker ausgeprägt für die Methode (Interaktion).

Deskriptive Statistiken

Deskriptive Statistiken für beide Messzeitpunkte und Interventionsformen

Messzeitpunkt	Interventionsform	*M*	*SD*	*n*
1	VR		4.76	
	Klassisch	45.10		
2	VR			
	klassisch			

Kapitel 9

Einführung in die Regressionsanalyse: Einfache und multiple lineare Regression

Stefan E. Huber

Bislang haben wir uns ausschließlich mit inferenzstatistischen Fragestellungen befasst, bei der die unabhängige(n) Variable(n) in Form einer oder mehrerer kategorialer Variablen vorlagen. Dabei fragten wir uns, ob und wie typische Ausprägungen der abhängigen Variablen von der Zuordnung zu den einzelnen Kategorien der unabhängigen Variablen abhingen. Als Maß für typische Ausprägungen haben wir jeweils den Mittelwert in der jeweiligen Kategorie herangezogen.

In den verbleibenden Kapiteln werden wir uns nun mit inferenzstatistischen Verfahren befassen, die solche Fragestellungen für den Fall metrischer Variablen für die unabhängigen Variablen verallgemeinern. An allem Übrigen wird sich nichts ändern. Uns wird weiterhin interessieren, ob und wie typische Ausprägungen der abhängigen Variablen (in Form von Mittelwerten) von einer oder mehreren unabhängigen Variablen abhängen. Beispielsweise haben wir uns in Kapitel 6 mit der Frage befasst, ob und wie das mittlere Depressionsniveau von der Zugehörigkeit zu einer von drei Altersgruppen abhängt. In diesem und den folgenden Kapiteln werden wir diese Frage auf die Frage verallgemeinern, ob und wie das mittlere Depressionsniveau vom Alter der Versuchspersonen abhängt. Das Alter ist bekanntlich eine metrische Variable, für die Messwerte auf einer kontinuierlichen Skala vorliegen können. Mit dem Verfahren, das wir in diesem und den folgenden Kapiteln kennenlernen werden, werden wir also z.B. Fragen beantworten können wie: Welches mittlere Depressionsniveau erwarten wir uns für Personen, die 35 Jahre alt sind? Oder für jemanden wird mit Becks Depressionsinventar ein Depressionsniveau von 32 Punkten ermittelt: Entspricht dies einem für das Alter der Person typischen Wert? Oder: Haben wir Anlass zur Vermutung, dass sich das mittlere Depressionsniveau überhaupt mit dem Alter verändert? Wenn ja, wie? Steigt es mit zunehmendem Alter an oder nimmt es mit zunehmendem Alter ab? Falls ja, wie stark? Und falls ja, wie sicher können wir uns dieses Befunds für den Zusammenhang auf Populationsebene auf der Grundlage unserer (einfachen Zufalls-)Stichprobe sein?

Bei dem Verfahren, das uns erlauben wird, solcherlei Fragen (inferenzstatistisch) zu erhellen, handelt es sich um die sog. Regressionsanalyse. Im Gegensatz zu varianzanalytischen Modellen, bei denen die unabhängigen Variablen diskret sind, können in regressionsanalytischen Modellen die unabhängigen Variablen sowohl stetig oder diskret sein. Zudem können eine oder mehrere unabhängige Variablen vorliegen. In diesem Kapitel werden wir uns zunächst mit dem einfachsten Fall beschäftigen: einer stetigen unabhängigen Variablen. Dieser Fall wird auch als einfache lineare Regression bezeichnet. Anschließend werden wir uns mit der sogenannten multiplen linearen Regression befassen, in der mehrere stetige unabhängige Variablen vorliegen werden. In späteren Kapiteln werden wir uns schließlich den Fall ansehen, dass eine oder mehrere dieser unabhängigen Variablen in Form von diskreten Variablen vorliegt.

Im Fall der linearen Regression werden unabhängige Variablen (UV) manchmal auch als Prädiktoren und die abhängige Variable (AV) als Kriterium bezeichnet. Dabei handelt es sich also um keine neuen Konzepte, nur um zusätzliche Bezeichnungen für bereits bekannte Konzepte.

Einfache lineare Regression: Das regressionsanalytische Modell und seine Voraussetzungen

Wird zwischen einer UV und typischen Ausprägungen einer AV bis auf einen identisch und unabhängig normalverteilten Fehler ein linearer Zusammenhang vermutet, so kann dies in folgendem mathematischen Modell zum Ausdruck gebracht werden:

$$Y_i \sim N(\mu_i, \sigma^2) \text{ mit } \mu_i = E(Y_i|X_i = x_i) = \alpha + \beta x_i.$$

Der Index i bezeichnet hier wiederum den i-ten von insgesamt n Fällen. Wir gehen im Folgenden wiederum davon aus, dass es sich bei diesen Fällen jeweils um unterschiedliche Personen aus einer einfachen Zufallsstichprobe handelt. Mit Y_i wird die zufällige Ausprägung der AV der zufällig gezogenen Person i bezeichnet. Die konkrete Realisation dieser Zufallsvariable wird wiederum mit y_i bezeichnet. Für den Erwartungswert der AV wird angenommen, dass es sich dabei um eine lineare Funktion der UV handelt, mit dem Achsenabschnitt α und der Steigung β. Für diesen Erwartungswert wird kurz μ_i geschrieben. Dabei handelt es sich gemäß dem Modell also um den Mittelwert einer Normalverteilung mit Varianz σ^2 an der Stelle $X_i = x_i$ der unabhängigen Variablen. Das heißt, wir erwarten uns, dass für einen bestimmten Wert der UV die AV durch eine normalverteilte Zufallsvariable

approximiert werden kann, deren Mittelwert linear von der Ausprägung der UV abhängt und deren Varianz unabhängig von der Ausprägung der UV ist.

Gemäß diesem Modell kann die AV also in zwei Anteile zerlegt werden: dem systematischen Zusammenhang zwischen UV (Prädiktor) und AV (Kriterium), $\alpha + \beta x_i$, und einem unsystematischen Fehler $\varepsilon_i = Y_i - (\alpha + \beta x_i)$, der der Abweichung der Zufallsvariable Y_i von ihrem Erwartungswert an der Stelle $X_i = x_i$ entspricht. Da der Erwartungswert der Zufallsvariable Y_i gerade dem zweiten Term im Ausdruck für den Fehler ε_i entspricht, handelt es sich bei ε_i um eine normalverteilte Zufallsvariable mit Mittelwert Null und Varianz σ^2.

Im regressionsanalytischen Modell werden demnach die folgenden Annahmen getroffen:

1. Zwischen UV (Prädiktor) und dem Erwartungswert der AV für eine bestimmte Ausprägung der UV besteht ein linearer Zusammenhang.
2. Die Abweichung konkreter Ausprägungen der AV von ihrem Erwartungswert kann durch eine identisch und unabhängig normalverteilte Zufallsvariable mit Mittelwert Null und einer von der UV unabhängigen, konstanten Varianz modelliert werden.

Mit der Prüfung der Plausibilität dieser Annahmen werden wir uns im nächsten Kapitel befassen. Für die noch folgenden Beispiele im vorliegenden Kapitel nehmen wir schlichtweg an, dass diese Voraussetzungen erfüllt sind.

Das regressionsanalytische Modell für die einfache lineare Regression enthält insgesamt also drei Parameter: den Achsenabschnitt α, die Steigung β, und die Varianz σ^2. Wie bei allen bisher behandelten Fragestellungen sind auch diese Parameter im Regelfall unbekannt und müssen mittels einer endlichen Stichprobe und geeigneten Schätzfunktionen geschätzt werden. Das heißt, wir kennen den linearen Zusammenhang zwischen UV und typischen Ausprägungen der AV in Abhängigkeit der UV nicht, sondern wollen ihn ermitteln, indem wir eine einfache Zufallsstichprobe erheben und dann die unbekannten Parameter α und β schätzen. Die Schätzung der Varianz σ^2 erlaubt uns schließlich abzuschätzen, wie weit die AV um die Regressionsgerade, die den Zusammenhang zwischen UV und

mittleren Ausprägungen der AV beschreibt, in einzelnen, konkreten Fällen streut. Mit Schätzungen für alle drei Parameter können wir dann Daten simulieren, die mit den Stichprobendaten kompatibel sind.

Da es sich bei allen resultierenden Schätzwerten aber um Schätzungen auf der Basis einer endlichen Zufallsstichprobe handelt, bleiben damit wie immer rein statistisch Unsicherheiten zurück, die im Rahmen eines frequentistischen Zugangs in entsprechenden Konfidenzintervallen und p-Werten zum Ausdruck kommen. Mit dem Vorgehen, wie wir diese nebst Schätzungen für die Parameter selbst im konkreten Fall mit SPSS ermitteln können, befasst sich der nächste Abschnitt.

Schätzung und Testung der Modellparameter der einfachen linearen Regression mit SPSS

Die Schätzung und Testung der Modellparameter der einfachen linearen Regression mit SPSS wird an folgendem Beispiel illustriert. Für eine Stichprobe von 50 Personen wurde sowohl die negative Selbstbewertung als auch die Depressionsschwere mit geeigneten psychometrischen Instrumenten erhoben. Die beiden Variablen (die Variable *nsb* entspricht der negativen Selbstbewertung, die Variable *bdi* der Depressionsschwere) für die 50 Personen sind im Datensatz „Kap9daten.sav“ zu finden, den Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können. Zwischen den beiden Variablen wird ein linearer Zusammenhang vermutet. Bzw. präziser: es wird vermutet, dass höhere Werte der negativen Selbstbewertung typischerweise mit höheren Werten der Depressionsschwere einhergehen. Das heißt insbesondere, es liegt eine gerichtete Hypothese vor (bei einer ungerichteten Hypothese würde lediglich ein positiver oder negativer linearer Zusammenhang zwischen den beiden Variablen vermutet werden, aber über das Vorzeichen würde keine entsprechende Vermutung bestehen). Durch Schätzung und Testung der Parameter eines einfachen linearen Regressionsmodells für die gegebenen Daten soll diese Vermutung inferenzstatistisch überprüft werden.

Die entsprechende Regressionsanalyse können wir nach dem Öffnen des Datensatzes in SPSS unter *Analyze >> Regression >> Linear…* anfordern. Im sich öffnenden Menü übertragen wir anschließend die Variable *bdi* in das Feld „Dependent“. Die Variable *nsb* übertragen wir in das Feld „Block 1 of 1“, siehe Abbildung 9.1. Danach klicken wir auf „Paste“, dokumentieren unser Vorgehen in der sich öffnenden Syntax-Datei, siehe Abbildung 9.2, und führen anschließend die dort eingefügten

Kommandozeilen aus. Daraufhin wird die in Abbildung 9.3 dargestellte Ausgabe erzeugt. In dieser Ausgabe finden wir sämtliche benötigten Informationen zur Schätzung und Testung unserer Modellparameter für die soeben durchgeführte Regressionsanalyse. In Abbildung 9.3 sind ferner alle Bereiche bzw. Werte farblich hervorgehoben, die wir für die Formulierung eines entsprechenden Ergebnisberichts benötigen (siehe nächster Abschnitt).

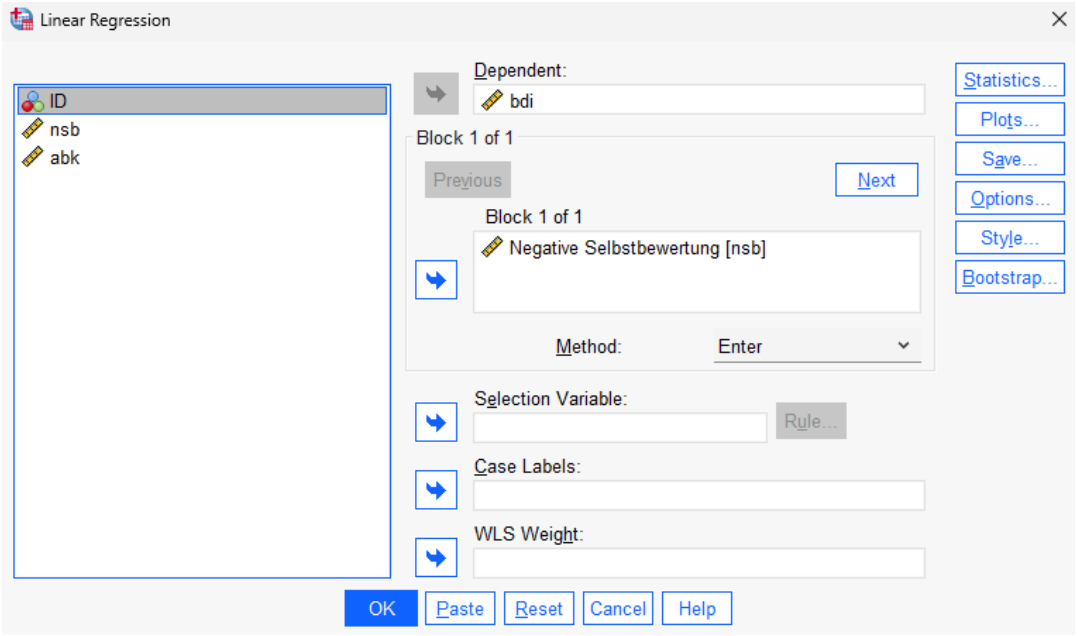


Abbildung 9.1. Anforderung einer einfachen linearen Regression in SPSS im Menü *Analyze >> Regression >> Linear…*.

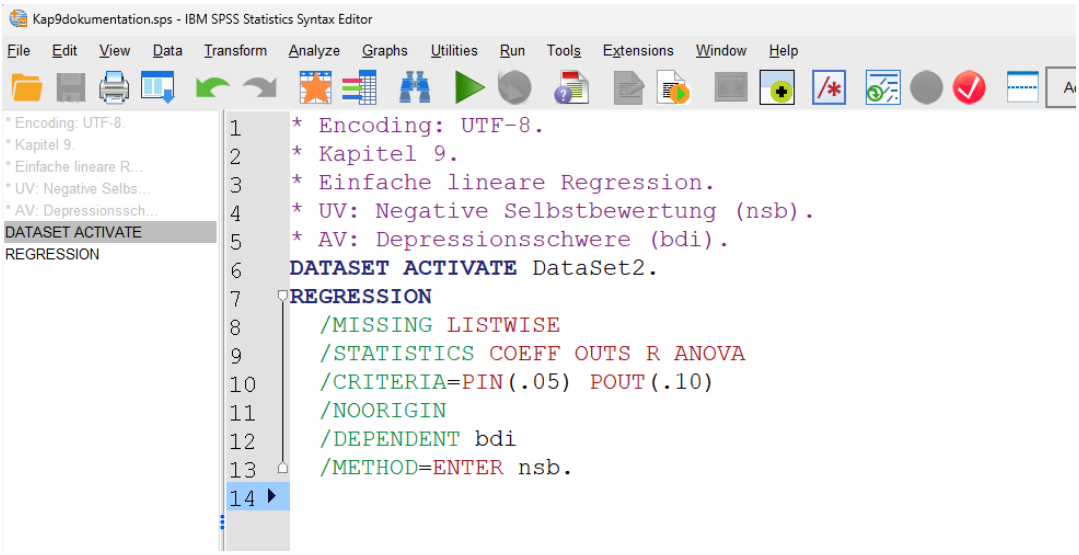


Abbildung 9.2. Syntax-Datei für die angeforderte einfache lineare Regressionsanalyse.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Negative Selbstbewertung[b]	.	Enter

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.841[a]	.708	.702	6.756

a. Predictors: (Constant), Negative Selbstbewertung

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5304.324	1	5304.324	116.203	<.001[b]
	Residual	2191.056	48	45.647		
	Total	7495.380	49			

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. Predictors: (Constant), Negative Selbstbewertung

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-.004	1.939		-.002	.998
	Negative Selbstbewertung	.843	.078	.841	10.780	<.001

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung 9.3. Ausgabe für die durchgeführte einfache lineare Regressionsanalyse mit dem Prädiktor *Negative Selbstbewertung* und dem Kriterium *Depressionsschwere*.

In der Tabelle „Variables Entered/Removed“ finden wir noch einmal sämtliche Prädiktoren aufgelistet, die im Regressionsmodell vorkommen. Da wir im vorliegenden Beispiel nur eine UV haben, finden wir dort nur die *Negative Selbstbewertung*.

In der Tabelle „Model Summary“ finden wir den multiplen Korrelationskoeffizienten („R“), den sog. Determinationskoeffizienten, d.h. den Anteil der Varianz in der AV, der durch den geschätzten linearen Zusammenhang mit der UV in der Stichprobe aufgeklärt werden kann („R Square“), sowie den Standardschätzfehler („Std. Error of the Estimate“). Die ersten beiden dieser Größen werden in den folgenden Kapiteln noch genauer erläutert werden. Den Determinationskoeffizienten, d.h. R^2, werden wir aber für den Ergebnisbericht brauchen. Daher soll jetzt schon erwähnt werden, dass es sich dabei um eine Effektstärke handelt, die analog zum η^2 in der Varianzanalyse den Anteil der Varianz in der AV angibt, der durch die UV (für die gegebene Stichprobe) erklärt werden kann. Auch für diese Effektstärke gibt es wieder Heuristiken nach Cohen (1988), um die relative Größe eines Effekts schnell einschätzen zu können: ab einem Wert von .02 spricht man von einem kleinen, ab einem Wert von .13 von einem mittleren, und ab einem Wert von .26 von einem großen Effekt. Da der Anteil der erklärten Varianz nur zwischen 0 und 1 variieren kann, wird auch bei dieser Größe wieder entsprechend APA-Richtlinien die führende Null weggelassen. Auch diese Kategorisierung werden wir wieder in einem entsprechenden Ergebnisbericht vornehmen. Beim Standardschätzfehler handelt es sich schließlich um die Wurzel aus dem Schätzwert für die Varianz σ^2, d.h. einem unserer Modellparameter. Da dieser inhaltlich jedoch kaum interpretierbar ist (Bühner et al., 2025), wird er kaum je berichtet und auch wir werden diesen Schätzwert hier nicht weiter verwenden. Er könnte allerdings verwendet werden, um Daten zu simulieren, die mit den Stichprobendaten kompatibel sind (Bühner et al., 2025).

In der Tabelle „ANOVA“ finden wir die Ergebnisse des Omnibustests für das gesamte regressionsanalytische Modell. Wie im nächsten Kapitel noch erläutert werden wird, wird mit diesem Omnibustest geprüft, ob sich irgendeiner der Steigungsparameter von Null unterscheidet. Hier haben wir nur einen Steigungsparameter, da nur ein Prädiktor vorliegt, weshalb das Ergebnis des Omnibustests auch dem Ergebnis der Testung dieses einen Steigungsparameters entsprechen wird, wie wir unten noch sehen werden. Das heißt, wirklich relevant werden die Ergebnisse dieses Omnibustests erst im Falle mehrerer Prädiktoren werden. Der Omnibustest entspricht jedenfalls formal einer Varianzanalyse, die prüft, ob mit dem Regressionsmodell ein signifikanter Anteil der Varianz in der AV erklärt wird (d.h., ob sich R^2 von Null unterscheidet). Daher handelt es sich bei der Teststatistik (Spalte „F“) um einen F-Wert aufgrund der F-Verteilung dieser Größe unter Geltung der Nullhypothese. Für eine F-Verteilung

sind wiederum zwei Freiheitsgrade anzugeben, die wir in der Spalte „df“ in den Zeilen „Regression“ und „Residual“ finden. Der p-Wert für die Teststatistik findet sich ganz rechts in der Spalte „Sig.“. Ist dieser p-Wert kleiner als das vorab gewählte Signifikanzniveau, so wird ein signifikanter Anteil an Varianz in der AV aufgeklärt. Bzw. kann man auch sagen, dass sich ein signifikanter Anteil an Varianz in der AV auf den Prädiktor zurückführen (= „regredieren“) lässt (daher auch der Name „Regression“). Unter der Annahme eines Signifikanzniveaus von α = .005 ist das hier auch in der Tat der Fall (da der p-Wert kleiner als dieser Wert ist).

In der Tabelle „Coefficients“ finden wir schließlich die Schätzwerte für unsere Modellparameter sowie entsprechende Ergebnisse von Signifikanztests. Den Schätzwert für unseren Achsenabschnitt finden wir in der Zeile „(Constant)“. Er beträgt a = -0.004. In der Zeile „Negative Selbstbewertung“ finden wir den Schätzwert b = 0.843. Aus der Theorie wissen wir (Bühner et al., 2025), dass wir für beide Schätzwerte mit einer auf der t-Verteilung beruhenden Teststatistik die Kompatibilität mit dem Vergleichswert Null prüfen können. Die Ergebnisse dieser Tests finden wir in den Spalten „t“ und „Sig.“. Insbesondere sehen wir, dass sich für den Schätzwert des Steigungsparameters ein t-Wert von 10.78 ergibt (die Anzahl der Freiheitsgrade entsprechen den Nennerfreiheitsgraden aus der Tabelle „ANOVA“, d.h. der Anzahl an Freiheitsgraden, die dort in der Zeile „Residual“ zu finden ist) sowie ein p-Wert kleiner als 0.001. Wird der t-Wert quadriert, ergibt sich der Wert 116.2, was dem F-Wert aus der Tabelle „ANOVA“ von oben entspricht. In der Tat handelt es sich bei der F-Statistik des Omnibustests im Falle eines einzigen Prädiktors um das Quadrat der t-Statistik für den Vergleich des einzigen Steigungsparameters mit dem Vergleichswert Null. Genauso gilt, dass sich in diesem Fall der Steigungsparameter genau dann von Null unterscheidet, wenn sich R^2 signifikant von Null unterscheidet, d.h., wenn ein signifikanter Anteil der Varianz in der AV auf die UV zurückgeführt werden kann. Das zeigt sich schließlich auch am standardisierten Regressionskoeffizienten für den Steigungsparameter, den wir in der Tabelle „Coefficients“ in der Spalte „Standardized Coefficients Beta“ finden. Dieser entspricht exakt dem multiplen Korrelationskoeffizienten (d.h. der Wurzel aus R^2) und damit im Falle einer einfachen linearen Regression exakt dem Pearson Korrelationskoeffizienten zwischen UV und AV (für die einfache Regression ist der multiple Korrelationskoeffizient eben einfach nur ein ganz normaler einfacher Korrelationskoeffizient). Der standardisierte Regressionskoeffizient ist auch der

Steigungsparameter, den man erhält, wenn UV und AV beide z-transformiert („standardisiert“) werden. Inhaltlich kann der standardisierte Korrelationskoeffizient β_{stand} wie folgt interpretiert werden: Eine Erhöhung des Prädiktors um eine Standardabweichung geht mit einer Erhöhung des Kriteriums um β_{stand} Standardabweichungen einher.

Exkurs: Korrelation

Der standardisierte Regressionskoeffizient entspricht in der einfachen linearen Regression dem Pearson-Korrelationskoeffizienten zwischen UV und AV? Auch davon können wir uns leicht in SPSS überzeugen. Die Berechnung des Pearson-Korrelationskoeffizienten können wir unter *Analyze >> Correlate >> Bivariate…* anfordern. Dazu schieben wir einfach jene Variablen, zwischen denen wir die Korrelationskoeffizienten ermitteln wollen, ins Feld „Variables“. Im Falle des vorliegenden Datensatzes „Kap9daten.sav“ können wir das zu Illustrationszwecken einfach einmal für alle drei metrischen Variablen tun, siehe Abbildung 9.4. Zur Berechnung von Korrelationskoeffizienten stehen drei Möglichkeiten zur Auswahl. Hier belassen wir es bei der Voreinstellung „Pearson“, da das genau der Korrelationskoeffizient ist, den wir mit dem standardisierten Regressionsgewicht von oben vergleichen möchten.

Nach Einfügen und Ausführen der entsprechenden Kommandozeilen in der Syntaxdatei erhalten wir die in Abbildung 9.5 gezeigte Ausgabe. In der Tat sehen wir, dass der Pearson-Korrelationskoeffizient zwischen Negativer Selbstbewertung und Depressionsschwere sich zu 0.841 ergibt. Zudem bekommen wir einen p-Wert, der sich auf die Nullhypothese eines Korrelationskoeffizienten von Null bezieht. Mittels Doppelklicks auf die Tabelle in der Ausgabe sowie Doppelklick auf den p-Wert selbst können wir uns davon überzeugen, dass es sich exakt um denselben p-Wert wie für den Omnibustest und den Regressionskoeffizienten aus dem vorherigen Abschnitt handelt. Im Falle einer einfachen linearen Regression unterscheidet sich also der Anteil erklärter Varianz, der durch R^2 zum Ausdruck kommt, genau dann signifikant von Null, wenn sich der (multiple) Korrelationskoeffizient bzw. das Regressionsgewicht des (einzigen) Prädiktors signifikant von Null unterscheidet.

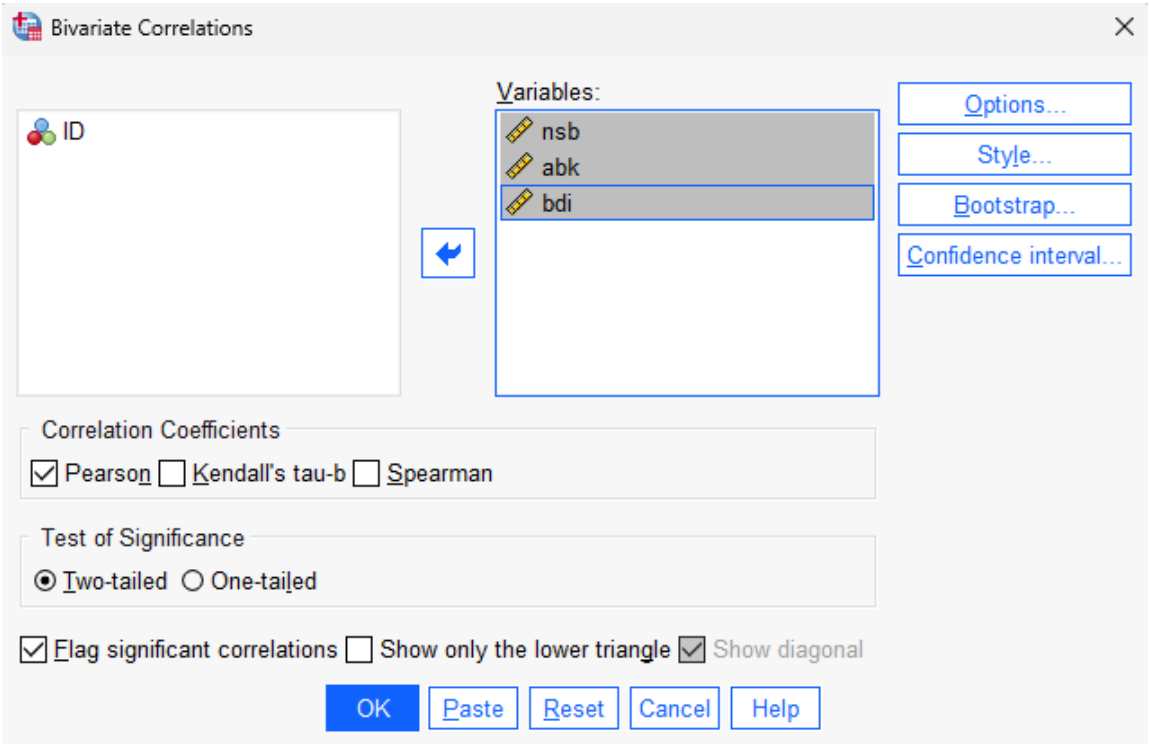


Abbildung 9.4. Anforderung von Pearson-Korrelationskoeffizienten für die drei Variablen *Negative Selbstbewertung*, *Abhängigkeitskognitionen*, und *Depressionsschwere*, die im Datensatz „Kap9daten.sav“ enthalten sind.

Correlations

		Negative Selbstbewertung	Abhängigkeitskognitionen	Depressionsschwere (Gesamtwert für Becks Depressionsinventar)
Negative Selbstbewertung	Pearson Correlation	1	.163	.841**
	Sig. (2-tailed)		.257	<.001
	N	50	50	50
Abhängigkeitskognitionen	Pearson Correlation	.163	1	.287*
	Sig. (2-tailed)	.257		.043
	N	50	50	50
Depressionsschwere (Gesamtwert für Becks Depressionsinventar)	Pearson Correlation	.841**	.287*	1
	Sig. (2-tailed)	<.001	.043	
	N	50	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Abbildung 9.5. Pearson-Korrelationskoeffizienten und deren p-Werte für alle möglichen Kombinationen aus den drei untersuchten Variablen.

In der Ausgabe für die Korrelationskoeffizienten sehen wir zudem noch, dass auch zwischen den Abhängigkeitskognitionen und der Depressionsschwere sowie den Abhängigkeitskognitionen und der negativen Selbstbewertung positive Korrelationen in der erhobenen Stichprobe bestehen. Es könnte also durchaus auch interessant sein, zu untersuchen, welcher Anteil der Varianz in der Depressionsschwere auf die Intensität von Abhängigkeitskognitionen zurückgeführt werden kann. Dieser Fragestellung widmet sich das Übungsbeispiel 9.3. Zudem könnte es interessant sein, der Frage nachzugehen, wie viel Varianz in der Depressionsschwere durch beide Variablen (der negativen Selbstbewertung und den Abhängigkeitskognitionen) erklärt werden kann, wenn man zwischen beiden Variablen und dem Kriterium jeweils lineare Zusammenhänge annimmt. Bei letzterer Fragestellung handelt es sich um eine Fragestellung für eine sog. multiple lineare Regression (mit zwei Prädiktoren), der wir uns im nächsten Kapitel zuwenden werden.

Exkurs: Zentrierung, Skalierung, Standardisierung von Variablen

Wir haben oben gesehen, dass sich für den Schätzwert des Achsenabschnitts ein kleiner negativer Wert ergab (a = -0.004). Die inhaltliche Interpretation dieses Achsenabschnitts würde lauten: Auf der Basis des einfachen Regressionsmodells würde für einen Wert von Null auf der Skala der negativen Selbstbewertung eine mittlere Depressionsschwere von -0.004 erwartet werden. Rein rechnerisch ist an dieser Aussage nichts problematisch. Die Regressionsanalyse ergibt eine Regressionsgerade und selbstverständlich kann man ermitteln, wo diese Regressionsgerade die y-Achse schneidet, d.h., welchen Wert man für die mittlere Depressionsschwere erhalten würde, wenn die negative Selbstbewertung gleich Null (d.h. x = 0) wäre. Wirklich von Interesse sind diese konkreten Werte aber kaum, da Becks Depressionsinventar keine negativen Werte zulässt, und auch die negative Selbstbewertung auf der entsprechenden psychometrischen Skala gar nicht Null sein kann. Manche Autor:innen sagen deshalb auch, dass der Schätzwert für den Achsenabschnitt α daher inhaltlich nicht sinnvoll interpretiert werden kann (Bühner et al., 2025).

Grundsätzlich ist das kein Problem, weil Kenntnis beider Schätzwerte a und b ja die Schätzung mittlerer Ausprägungen im Kriterium im gesamten inhaltlich sinnvoll interpretierbaren Bereich der UV zulässt. Möchte man aber auch für die Schätzung des Achsenabschnitts einen Wert, der bereits im inhaltlich sinnvollen Bereich für die UV liegt, so kann man dafür den Nullpunkt der Prädiktorvariable

entsprechend verschieben. Die Regressionsgerade schneidet dann die y-Achse bei diesem neuen Nullpunkt $x' = 0$.

Eine Möglichkeit dafür ist die sogenannte Zentrierung am Stichprobenmittelwert. Dazu wird eine neue Prädiktorvariable gebildet, indem von der ursprünglichen Variablen deren Mittelwert subtrahiert wird: $X_i' = X_i - \bar{X}$ bzw. für die beobachteten Werte $x_i' = x_i - \bar{x}$. Für die negative Selbstbewertung ergibt sich in unserem vorliegenden Beispiel ein Stichprobenmittelwert von 21.58. Unter *Transform >> Compute Variable…* können wir damit nun eine zentrierte Prädiktorvariable erzeugen, siehe Abbildung 9.6. Führen wir nun neuerlich eine einfache lineare Regressionsanalyse mit diesem zentrierten Prädiktor durch, erhalten wir die in Abbildung 9.7 gezeigte Ausgabe. Wir sehen, dass unser geschätzter Achsenabschnitt nun $a = 18.18$ beträgt. Die inhaltliche Bedeutung dieses Achsenabschnitts ist nun: bei einer mittleren Ausprägung der negativen Selbstbewertung wird auf der Basis der durchgeführten Regressionsanalyse ein mittleres Depressionsniveau von 18.18 Punkten auf der Skala von Becks Depressionsinventar erwartet. An der inhaltlichen Interpretation der geschätzten Steigung b ändert sich nichts.

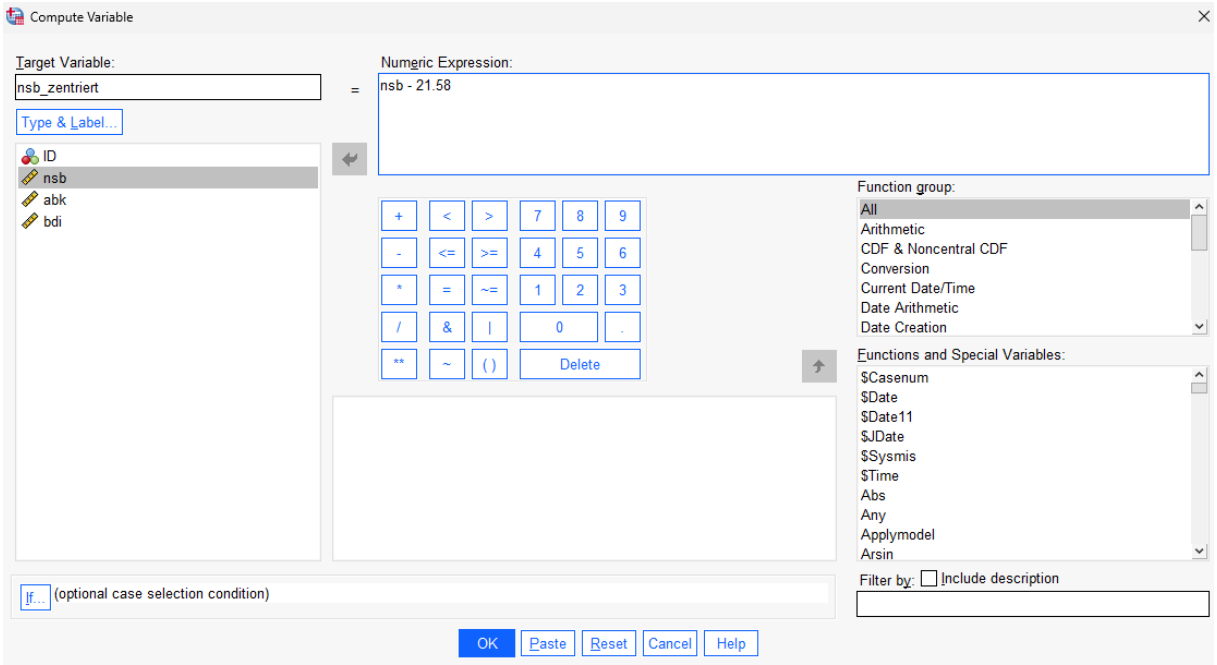


Abbildung 9.6. Erzeugung eines zentrierten Prädiktors am Beispiel der negativen Selbstbewertung.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	18.180	.955		19.027	<.001
	nsb_zentriert	.843	.078	.841	10.780	<.001

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung 9.7. Teil der Ausgabe für die einfache lineare Regressionsanalyse mit zentriertem Prädiktor.

Neben der Zentrierung gibt es noch andere Variablentransformationen, die sich manchmal im Zusammenhang mit Regressionsanalysen anbieten, um die Resultate inhaltlich einfacher interpretieren zu können. Durch Skalierung einer Variablen kann etwa die Einheit, mit der diese Variable gemessen wird, verändert werden. Zur Skalierung kann etwa der gesamte in der Stichprobe vorliegende Variablenbereich verwendet werden: $x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$. Die auf diese Weise skalierte Variable x_i' variiert dann im Intervall 0 bis 1, wobei der Wert 0 dem kleinsten gemessenen Wert und der Wert 1 dem größten gemessenen Wert entspricht. Dadurch ändern sich die inhaltlichen Interpretationen für beide Schätzwerte a und b. Der Schätzwert a entspricht der erwarteten mittleren Ausprägung des Kriteriums für den kleinsten Wert der UV. Der Schätzwert b entspricht der Änderung in der erwarteten mittleren Ausprägung des Kriteriums, wenn sich die UV vom kleinsten zum größten Wert in der Stichprobe ändert.

Eine weitere häufige Form der Variablentransformation ist die z-Transformation oder auch „Standardisierung“ (bei der es sich – jedenfalls in der in SPSS implementierten Form – eigentlich um eine Studentisierung handelt, da für die Skalierung nicht die empirische Standardabweichung, sondern der Schätzwert der Populationsstandardabweichung auf Basis der Stichprobe mittels der (erwartungstreuen) Schätzfunktion für die Populationsvarianz σ^2 verwendet wird, siehe Kapitel 3). Die z-Transformation einer Variablen lässt sich in SPSS sehr einfach über *Analyze* >> *Descriptive Statistics* >> *Descriptives…* durchführen, siehe Abbildung 9.8.

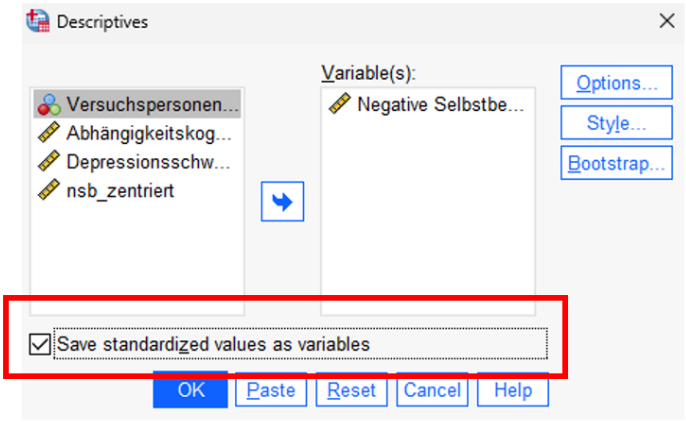


Abbildung 9.8. Z-Transformation einer Variablen in SPSS.

Ergebnisbericht für die einfache lineare Regression

Ein Ergebnisbericht für die einfache lineare Regression für das oben erläuterte Beispiel könnte wie folgt aussehen: „Ein signifikanter Anteil der Varianz in der Depressionsschwere der untersuchten 50 Personen kann (mit $\alpha = .005$) auf die negative Selbstbewertung zurückgeführt werden, $F(1,48) = 116.20$, $p < .001$, $R^2 = .71$. Gemäß Cohens Heuristiken (1988) entspricht dies einem großen Effekt. Der Regressionskoeffizient für den Zusammenhang zwischen negativer Selbstbewertung und Depressionsschwere ist signifikant positiv, $b = 0.84$ (stand. $\beta = 0.84$), $t(48) = 10.78$, $p < .001$ (einseitig); d.h., je höher die negative Selbstbewertung, desto höher die Depressionsschwere. Eine Erhöhung der negativen Selbstbewertung um einen Punkt geht gemäß dem geschätzten einfachen Regressionsmodell im Mittel mit einer Erhöhung der Depressionsschwere um 0.84 Punkte auf der Skala für die Depressionsschwere einher."

Aufgrund der Kleinheit der sich ergebenden p-Werte geht leider in diesem Beispiel etwas unter, dass es sich hierbei um die Prüfung einer gerichteten Hypothese gehandelt hat und wie dabei grundsätzlich für die in SPSS gegebene Ausgabe vorzugehen ist. In der Ausgabe wird ja nur ein p-Wert für den Regressionskoeffizienten (Steigungsparameter) angegeben. Dieser bezieht sich auf eine Unterschiedshypothese („der Regressionskoeffizient unterscheidet sich von Null"). Bei Vorliegen einer einseitigen Hypothese kann dieser p-Wert halbiert werden (wenn allerdings der zweiseitige p-Wert bereits kleiner als .001 ist, gilt dies auch für den einseitigen). Zusätzlich ist selbstverständlich noch zu prüfen, ob der Steigungsparameter auch tatsächlich das vermutete Vorzeichen hat; ansonsten kann die Nullhypothese im Falle einseitiger Testung natürlich nicht „verworfen" werden.

Regressionsanalyse für zwei Prädiktoren

Enthält das Regressionsmodell mehr als einen Prädiktor, spricht man von einer multiplen Regression. Ein typisches Ziel der multiplen Regression ist die Verbesserung von Vorhersagen über die Kriteriumsvariable durch Berücksichtigung mehrerer Prädiktoren. Außerdem sind häufig die folgenden Fragen interessant (Bühner et al., 2025):

- Wie viel Varianz im Kriterium kann durch die Prädiktoren gemeinsam erklärt werden?
- Welcher der Prädiktoren weist dabei den größten Vorhersagebeitrag auf?
- Wie groß ist der eigenständige Vorhersagebeitrag eines Prädiktors, wenn der Prädiktor mit anderen Prädiktoren korreliert?
- Verändert sich die Stärke, Richtung und damit die Interpretation des Effekts eines Prädiktors durch die Berücksichtigung eines anderen Prädiktors?

In diesem Abschnitt und im folgenden Kapitel werden wir uns mit der inferenzstatistischen Erhellung solcher Fragen im Rahmen multipler linearer Regressionsanalysen mit zwei Prädiktoren befassen. Die Erweiterung auf mehr als zwei Prädiktoren ist vergleichsweise einfach und wird im Rahmen einiger entsprechender Übungsbeispiele abgedeckt.

Zur Illustration verwenden wir weiterhin den Datensatz „Kap9daten.sav“. In diesem Datensatz sind neben der Depressionsschwere und der negativen Selbstbewertung auch die Intensität von Abhängigkeitskognitionen (erfasst jeweils mit entsprechend geeigneten psychometrischen Instrumenten) von 50 (fiktiven) Personen gegeben. Mithilfe dieser Daten möchten wir der Frage nachgehen, ob und inwieweit die negative Selbstbewertung und Abhängigkeitskognitionen die Depressionsschwere von Personen erklären (bzw. vorhersagen) können. Wir nehmen dabei an, dass zwischen den beiden Prädiktorvariablen (negative Selbstbewertung und Abhängigkeitskognitionen) und dem Erwartungswert des Kriteriums (Depressionsschwere) bis auf einen normalverteilten Fehler jeweils lineare Zusammenhänge bestehen.

Regressionsanalytisches Modell für zwei Prädiktoren

Das regressionsanalytische Modell im Falle zweier Prädiktoren lautet:

$$Y_i \sim N(\mu_i, \sigma^2) \text{ mit } \mu_i = E(Y_i|X_{i1} = x_{i1}, X_{i2} = x_{i2}) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

Das Modell bringt zum Ausdruck, dass für jede bestimmte Realisation x_{i1} und x_{i2} der beiden Prädiktoren X_{i1} und X_{i2} die Ausprägung des Kriteriums als identisch und unabhängig normalverteilte Zufallsvariable Y_i beschrieben werden kann, wobei der Erwartungswert (Mittelwert) der Normalverteilung linear von den beiden Prädiktoren abhängt, mit (unbekanntem) Achsenabschnitt α und den beiden (unbekannten) Steigungsparametern β_1 und β_2, und die Varianz unabhängig von den beiden Prädiktoren konstant den (unbekannten) Wert σ^2 hat. Die beiden Annahmen der multiplen linearen Regression sind damit wie schon im Fall der einfachen linearen Regression gegeben durch (i) den linearen Zusammenhang zwischen Prädiktoren und Erwartungswert des Kriteriums und (ii) die identische und unabhängige Normalverteilung der Abweichungen von diesem Erwartungswert mit konstanter Varianz für jede beliebige Realisation der Prädiktoren. Im nächsten Kapitel werden wir uns mit Möglichkeiten der Überprüfung der Plausibilität dieser Annahmen im Falle einer konkreten Stichprobe befassen.

Die inhaltliche Bedeutung der Modellparameter α, β_1 und β_2 ist wie folgt. Der Achsenabschnitt α gibt den erwarteten Wert des Kriteriums für $x_{i1} = x_{i2} = 0$ an: $E(Y_i|X_{i1} = 0, X_{i2} = 0) = \alpha + \beta_1 \cdot 0 + \beta_2 \cdot 0 = \alpha$. Im Falle zweier Prädiktoren kann der lineare Zusammenhang $E(Y_i|X_{i1} = x_{i1}, X_{i2} = x_{i2}) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}$ als Ebene im dreidimensionalen, kartesischen Koordinatensystem mit den Achsen x_{i1}, x_{i2} und Y_i dargestellt werden. Bei dieser Ebene handelt es sich um die Menge aller Erwartungswerte des Kriteriums für alle möglichen Kombinationen der beiden stetigen Prädiktorvariablen. In diesem Koordinatensystem entspricht der Achsenabschnitt α dann dem Abstand des Punkts, in dem die Regressionsebene die Y_i-Achse schneidet, zum Ursprung des Koordinatensystems. Der Steigungsparameter β_1 gibt die Steigung der Regressionsebene in Richtung der x_{i1}-Achse an, d.h. die Änderung des erwarteten Werts des Kriteriums in Abhängigkeit der Änderung des Prädiktors x_{i1}: $\frac{\partial E(Y_i|X_{i1}=x_{i1}, X_{i2}=x_{i2})}{\partial x_{i1}} = \frac{\partial}{\partial x_{i1}}(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2}) = \beta_1$. D.h., β_1 gibt an, wie stark sich der erwartete Wert des Kriteriums ändert, wenn ausschließlich der Prädiktor x_{i1} um eine Einheit vergrößert wird

(„ausschließlich“ bedeutet, dass dabei nicht gleichzeitig auch x_{i2} verändert werden darf, d.h., die Veränderung ausschließlich in Richtung der x_{i1}-Achse vonstattengeht). Analog gibt der Steigungsparameter β_2 gibt die Steigung der Regressionsebene in Richtung der x_{i2}-Achse an. Alternativ kann man auch sagen: Eine Erhöhung des Werts für Prädiktor 1 um eine Einheit unter Konstanthaltung des Werts für Prädiktor 2 geht im Mittel mit einer Erhöhung des Kriteriums um β_1 einher; eine Erhöhung des Werts für Prädiktor 2 um eine Einheit unter Konstanthaltung des Werts für Prädiktor 1 geht im Mittel mit einer Erhöhung des Kriteriums um β_2 einher (Bühner et al., 2025).

Dabei ist zu betonen, dass mit einer „Erhöhung“ des Kriteriums „durch Erhöhung“ eines Prädiktors kein Kausalzusammenhang zum Ausdruck gebracht werden soll, sondern lediglich eine *bedingte Assoziation*. Mit der Veränderung eines Prädiktors (unter der Bedingung, dass andere Prädiktoren sich nicht ändern) geht lediglich (im Mittel) eine Änderung des Kriteriums einher, wird aber nicht (zwingendermaßen) durch die Veränderung des Prädiktors verursacht. Eine präzisere Formulierung des Sachverhalts würde lauten (Bühner et al., 2025): „Vergleicht man Personen aus der Population, die sich in ihren Werten auf UV1 (oder UV2) um genau eine Einheit unterscheiden, aber alle den gleichen Wert auf dem anderen Prädiktor aufweisen, dann haben die Personen mit dem höheren UV-Wert im Mittel einen um β_1 (oder β_2) Einheiten höheren Wert auf der AV.“

Schätzung und Testung der Modellparameter für die multiple Regressionsanalyse mit zwei Prädiktoren in SPSS

Zur Schätzung und Testung der Modellparameter können wir ganz analog zum Vorgehen bei der einfachen linearen Regressionsanalyse im vorhergehenden Kapitel verfahren. Das heißt, wir können die Regressionsanalyse wiederum unter *Analyze >> Regression >> Linear…* anfordern. Im sich öffnenden Menü können wir dann die Variable *bdi* wieder in das Feld „Dependent“ und dieses Mal die beiden Variablen *nsb* und *abk* in das Feld „Block 1 of 1“ schieben, siehe Abbildung 9.9. Einfügen und Ausführen der entsprechenden Kommandozeilen in der Syntax-Datei ergibt dann die in Abbildung 9.10 dargestellte Ausgabe. In Abbildung 9.10 sind auch wieder alle Bestandteile farblich hervorgehoben, die wir für die Erstellung eines entsprechenden Ergebnisberichts bzw. für die Erläuterung der Ausgabe im Folgenden brauchen werden.

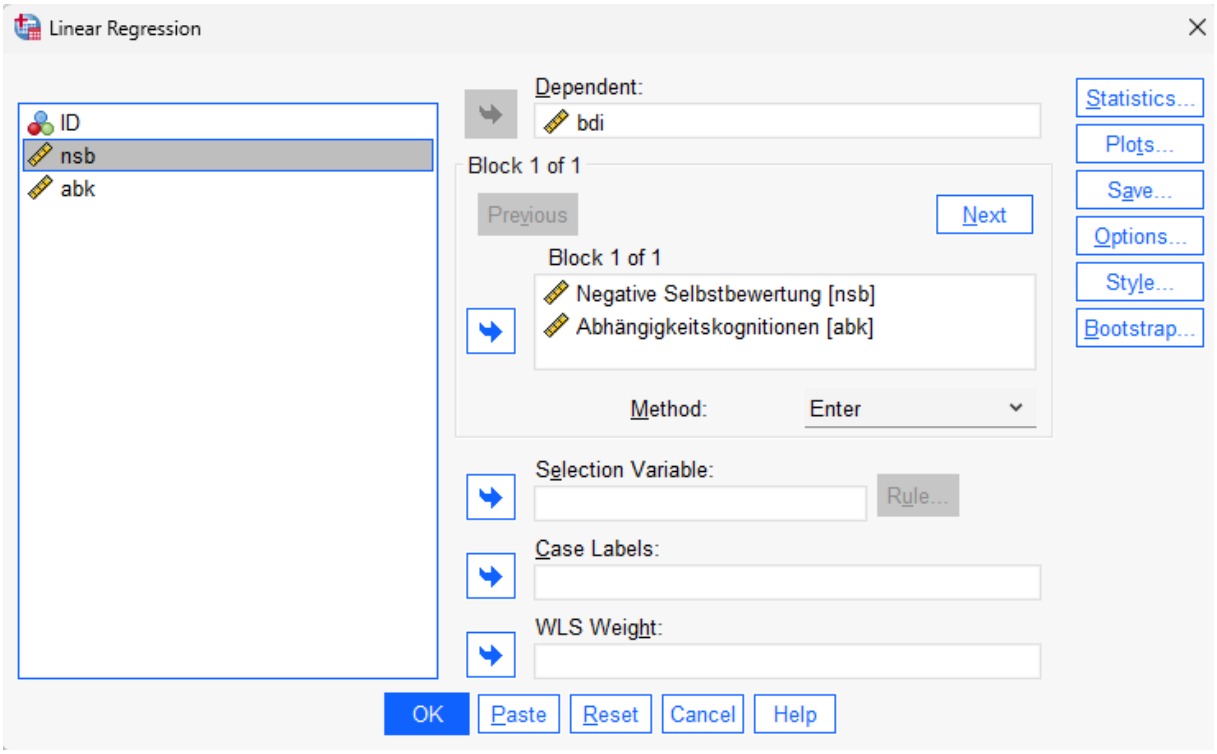


Abbildung 9.9. Ausführung einer multiplen linearen Regression mit zwei Prädiktoren in SPSS.

In der Tabelle „Model Summary“ sehen wir sogleich, dass unsere beiden Prädiktoren zusammen 73.1% der Varianz der Depressionsschwere in der Stichprobe erklären können. Wir finden hier zudem wieder den Schätzwert für die Fehlervarianz σ^2 aus dem Regressionsmodell, der hier den Wert $6.552^2 \approx 42.93$ hat. Der Wert unter „Std. Error of the Estimate“ muss hierbei quadriert werden, da es sich dabei um einen Schätzwert für die Standardabweichung (Achtung: nicht erwartungstreu!) und nicht die Varianz handelt.

In der Tabelle „ANOVA“ finden wir das Ergebnis des Omnibustests, mit dem überprüft wird, ob für mindestens einen der Prädiktoren der Regressionskoeffizient ungleich Null ist (Bühner et al., 2025). Dies ist gleichbedeutend mit der Überprüfung, ob der Anteil insgesamt erklärter Varianz, d.h. R^2, ungleich Null ist. Wir sehen, dass der Unterschied zu Null signifikant ist mit $F(2, 47) = 63.79$, $p < .001$.

In der Tabelle „Coefficients“ finden wir schließlich Schätzungen und Testungen unserer Modellparameter α, β_1 und β_2. Wir sehen, dass der Achsenabschnitt mit $a = -3.05$ geschätzt wird und nicht signifikant von Null abweicht, $t(47) = -1.26$, $p = .213$. Für die Steigungsparameter ergibt sich: $b_1 = 0.82$, $t(47) = 10.64$, $p < .001$, sowie $b_2 = 0.18$, $t(47) = 2.01$, $p = .050$.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Abhängigkeitskognitionen, Negative Selbstbewertung[b]	.	Enter

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.855[a]	.731	.719	6.552

a. Predictors: (Constant), Abhängigkeitskognitionen, Negative Selbstbewertung

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5477.548	2	2738.774	63.792	<.001[b]
	Residual	2017.832	47	42.933		
	Total	7495.380	49			

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. Predictors: (Constant), Abhängigkeitskognitionen, Negative Selbstbewertung

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-3.048	2.415		-1.262	.213
	Negative Selbstbewertung	.817	.077	.816	10.638	<.001
	Abhängigkeitskognitionen	.183	.091	.154	2.009	.050

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung 9.10. Ausgabe für eine multiple lineare Regressionsanalyse mit zwei Prädiktoren in SPSS.

Inhaltlich lassen sich die Schätzwerte wie folgt interpretieren: Für eine Population, in der das durch diese Schätzwerte spezifizierte Regressionsmodell den Zusammenhang zwischen den beiden Prädiktoren und dem Kriterium beschreibt, beträgt die mittlere Depressionsschwere für Personen mit einer negativen Selbstbewertung und einer Abhängigkeitskognition von jeweils 0 Punkten -3.05 Punkte. Eine Erhöhung der negativen Selbstbewertung um einen Punkt bei konstanter Abhängigkeitskognition geht im Mittel mit einer Erhöhung der Depressionsschwere um 0.82 Punkte einher. Eine Erhöhung der

Abhängigkeitskognitionen um einen Punkt bei konstanter negativer Selbstbewertung geht mit einer Erhöhung der Depressionsschwere um 0.18 Punkte einher.

Inhaltlich ergeben Werte von Null für die beiden Prädiktoren und ein negativer Wert für die mittlere Depressionsschwere natürlich wenig Sinn. Hier könnte wieder die Zentrierung der beiden Prädiktoren Abhilfe schaffen. Häufig ist man aber ohnehin hauptsächlich an der bedingten Assoziation des Kriteriums mit den Prädiktoren und weniger am Wert für den Achsenabschnitt interessiert.

Was ist hier mit „bedingter Assoziation“ gemeint? Bedingte Assoziation bezieht sich auf den Zusammenhang zwischen einem Prädiktor und einem Kriterium unter der Bedingung, dass zwischen dem Kriterium und einem anderen Prädiktor ein bestimmter Zusammenhang besteht und im statistischen Modell auch als solcher berücksichtigt wird. Das heißt, für die Interpretation unserer Schätzwerte und insbesondere derer p-Werte ist es ganz wesentlich, dass jeder Wert sich unter der Bedingung ergibt, dass alle anderen Modellparameter ihren resultierenden Schätzwerten entsprechen. D.h. unter der Voraussetzung, dass $\alpha = a =$ -3.048, $\beta_2 = b_2 = 0.183$, und $\sigma^2 = s^2 = 42.93$ ergibt sich b_1 = 0.817. Im dazugehörigen Nullhypothesentest, $t(47) = 10.64$, $p < .001$, ist dabei der p-Wert ein Maß dafür, wie plausibel uns auf Basis der gegebenen Stichprobe die Entscheidung erscheint, dass sich der „wahre“ Modellparameter β_1 von Null unterscheidet, wenn im betrachteten Regressionsmodell neben dem Prädiktor x_{i1} auch noch der Prädiktor x_{i2} berücksichtigt wird. Die Argumentationslinie folgt dabei wieder der typischen Logik des Nullhypothesensignifikanztestens: Unter der Voraussetzung, dass β_1 = 0 würden sich bei zufälliger Ziehung einfacher Zufallsstichproben nur selten Regressionskoeffizienten mit Betrag $|\hat{\beta}_1| \geq b_1 = 0.817$ ergeben, und d.h., die Annahme $\beta_1 = 0$ erscheint unter der *Bedingung der Gültigkeit aller anderen Annahmen* unplausibel. Auf Grundlage dieser Argumentation kann die Nullhypothese $\beta_1 = 0$ „abgelehnt“ werden. Typischerweise wird dafür wieder im Vorhinein ein Signifikanzniveau festgelegt um festzulegen, wie selten etwas unter Gültigkeit der Nullhypothese der Fall sein müsste, um die Gültigkeit der Nullhypothese zu bezweifeln, falls ein so seltener Fall für eine konkrete Zufallsstichprobe tatsächlich eintritt. Ist der p-Wert für den entsprechenden Regressionskoeffizienten kleiner als dieses vorab festgelegte Signifikanzniveau, wird die entsprechende Nullhypothese abgelehnt.

Für unseren zweiten Steigungsparameter haben wir den Schätzwert b_2 = 0.18 und den zugehörigen p-Wert p = 0.050 erhalten. Wir können uns in diesem Fall also deutlich weniger sicher sein, dass – wiederum unter der Bedingung der Berücksichtigung genau der im Modell enthaltenen Prädiktoren und der mit der Regressionsanalyse verbundenen Annahmen – ein entsprechend positiver Zusammenhang in der Population auch tatsächlich besteht. Mit dem üblichen Signifikanzniveau von .005 würden wir in diesem Fall gemäß der oben erläuterten Argumentationslinie die Nullhypothese β_2 = 0 nicht „verwerfen".

Aber heißt das, dass die Abhängigkeitskognitionen keine Rolle für die Vorhersage der Depressionsschwere spielen? Nein, es heißt lediglich, dass, wenn wir die negative Selbstbewertung bereits kennen, uns die Abhängigkeitskognitionen im Mittel nicht mehr viel zusätzliche Information für die Vorhersage der Depressionsschwere liefern und wir uns deshalb auch bezüglich des Vorzeichens der zusätzlichen Auswirkung auf die Vorhersage nicht sehr sicher sind. Kennen wir umgekehrt die Abhängigkeitskognitionen und erfahren nun zusätzlich von der negativen Selbstbewertung, erlaubt uns das eine deutliche bessere Vorhersage der Depressionsschwere und wir sind uns sehr sicher, dass eine höhere negative Selbstbewertung im Mittel mit einem höheren Depressionsniveau für eine beliebige, gegebene Ausprägung der Abhängigkeitskognitionen einhergeht. Auf diese Bedeutung des zusätzlichen Informationsmehrwerts, den ein Prädiktor relativ zu anderen Prädiktoren für die Vorhersage der AV bietet und wie dieser in den Ergebnissen für die einzelnen Schätzwerte und deren p-Werte bereits abgebildet ist, werden wir im nächsten Kapitel wieder zurückkommen.

Da wir alle Modellparameter geschätzt haben, können wir zur Vorhersage von Werten der AV für beliebige Werte der Prädiktoren diese Schätzwerte nun auch in unsere Modellgleichung einsetzen:

$$Y_i \sim N(\mu_i, 6.55^2),\ \ \mu_i = E(Y_i|X_{i1} = x_{i1}, X_{i2} = x_{i2}) = -3.05 + 0.82 \cdot x_{i1} + 0.18 \cdot x_{i2}.$$

Damit können wir nun typische Fragestellungen, die sich auf die Vorhersage des Kriteriums beziehen, beantworten. Z.B.: Welche Depressionsschwere erwarten wir im Mittel für Personen mit einer negativen Selbstbewertung von 20 Punkten und einer Intensität von Abhängigkeitskognitionen von 25 Punkten? Einsetzen dieser Zahlenwerte für X_{i1} und X_{i2} ergibt:

$$\mu_i = E(Y_i|X_{i1} = 20, X_{i2} = 25) = -3.05 + 0.82 \cdot 20 + 0.18 \cdot 25 = 17.85.$$

Wir können dies nun auch mit dem Vorhersagewert für die mittlere Depressionsschwere vergleichen, den wir erhalten würden, wenn wir die negative Selbstbewertung um einen Punkt erhöhen, die Abhängigkeitskognitionen aber unverändert lassen, d.h. $X_{i1} = 21$ und $X_{i2} = 25$:

$$\mu_i = E(Y_i|X_{i1} = 21, X_{i2} = 25) = -3.05 + 0.82 \cdot 21 + 0.18 \cdot 25 = 18.67.$$

Wir sehen, dass in diesem Fall die mittlere Depressionsschwere um $18.67 - 17.85 = 0.82$ Punkte zugenommen hat. D.h. die Differenz entspricht genau dem Schätzwert $b_1 = 0.82$. Das illustriert, weshalb die inhaltliche Interpretation des Schätzwerts genau so lautet wie oben angegeben: Eine Erhöhung der negativen Selbstbewertung um einen Punkt bei konstanter Abhängigkeitskognition geht mit einer Erhöhung der Depressionsschwere um 0.82 Punkte einher. Die Interpretation ist eben nur korrekt, wenn der jeweils andere Prädiktor konstant gehalten wird. Das war auch bereits ganz zu Anfang des Kapitels mit der Auswirkung der Änderung ausschließlich in Richtung eines der beiden Prädiktoren auf den erwarteten Wert des Kriteriums gemeint. Werden beide Prädiktoren geändert, findet die Änderung nicht mehr ausschließlich entlang einer der beiden Koordinatenachsen statt.

Ergebnisbericht für die multiple lineare Regression mit zwei Prädiktoren

Ein Ergebnisbericht für eine multiple lineare Regression mit zwei Prädiktoren könnte wie folgt formuliert werden: „Eine multiple lineare Regressionsanalyse ergab, dass ein (mit α = .005) signifikanter Anteil der Varianz in der Depressionsschwere der untersuchten n = 50 Personen durch die negative Selbstbewertung und die Abhängigkeitskognitionen der Personen erklärt werden kann, F(2, 47) = 63.79, p < .001, R^2 = .73; ein großer Effekt gemäß Cohen (1988). Gemäß des resultierenden Regressionsmodells geht eine Erhöhung der negativen Selbstbewertung um einen Punkt (bei Konstanthaltung der Abhängigkeitskognitionen) mit einer (mit α = .005) signifikanten Erhöhung der Depressionsschwere um 0.82 Punkte auf der Skala von Becks Depressionsinventar einher, b_{nsb} = 0.82 (stand. β = 0.82), $t(47) = 10.64$, $p < .001$. Eine Erhöhung der Abhängigkeitskognitionen um einen Punkt geht (bei Konstanthaltung der Depressionsschwere) hingegen mit einer (mit α = .005) nicht-signifikanten Erhöhung der Depressionsschwere um 0.18 Punkte einher, b_{abk} = 0.18 (stand. β = 0.15), $t(47) = 2.01$, $p = .050$. Schätzwerte und Teststatistiken für das Gesamtmodell sind in Tabelle 10.1 gegeben."

Im folgenden Kapitel werden wir sehen, wie wir diesen Ergebnisbericht noch um den Bericht der Varianzanteile, die jeder Prädiktor für sich genommen aufklären kann, ergänzen können.

Tabelle 10.1

Schätzwerte und Teststatistiken für alle Modellparameter des regressionsanalytischen Modells

Parameter	Schätzwert	Standardfehler	Stand. Koeff.	$t(47)$	p
Achsenabschnitt (a)	-3.05	2.42		-1.26	.213
Neg. Selbstbewertung (b_1)	0.82	0.08	0.82	10.64	< .001
Abhängigkeitskogn. (b_2)	0.18	0.09	0.15	2.01	.050

Übungsaufgaben

Die Datensätze für Beispiele 9.4 und 9.5 gehen ein weiteres Mal auf die bewundernswerte Fantasie von Andy Field (2024) zurück. Die für diese Beispiele benötigten Datensätze „Album Sales.sav“ und „Supermodel.sav“ können auf der frei zugänglichen Webseite mit ergänzenden Ressourcen für sein Buch „Discovering Statistics Using IBM SPSS Statistics“ unter https://edge.sagepub.com/field5e/student-resources/datasets heruntergeladen werden.

Beispiel 9.1

Was gehört zu den Voraussetzungen der linearen Regressionsanalyse?

(a) Die UV muss diskret sein.

(b) Die UV muss stetig sein.

(c) Zwischen UV und (Erwartungswerten der) AV (bei gegebenen Ausprägungen der UV) muss ein linearer Zusammenhang bestehen.

(d) Die Fehler (= Abweichungen der Ausprägungen der AV in Ordinatenrichtung von der wahren Regressionsgeraden) müssen unabhängig von der Ausprägung der UV normalverteilt mit Mittelwert Null und konstanter Standardabweichung σ^2 sein.

Beispiel 9.2

Welche Aussage/n trifft/treffen zu?

(a) Im Rahmen der Regressionsanalyse wird die UV auch häufig als Kriterium bezeichnet.

(b) Eine Regressionsanalyse kann nur mit einer stetigen UV durchgeführt werden.

(c) Die AV wird bei Regressionsanalysen manchmal auch als Prädiktor bezeichnet.

(d) Gemäß Cohen (1988) handelt es sich bei R^2 im Bereich von .02 bis .13 um kleine, im Bereich von .13 bis .26 um mittlere, und ab .26 um große Effekte.

Beispiel 9.3

Im Datensatz „Kap9daten.sav“, den Sie in dem elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können, ist neben der negativen Selbstbewertung und der Depressionsschwere noch eine weitere Variable gegeben. Bei dieser Variablen handelt es sich um die Intensität von Abhängigkeitskognitionen. Es liegt nahe, dass die Depressionsschwere auch von der Intensität von Abhängigkeitskognitionen abhängt. Untersuchen Sie diese Fragestellung mit einer einfachen linearen Regressionsanalyse und beantworten Sie insbesondere die folgenden Fragen:

(a) Wie lauten die statistischen Hypothesen bezogen auf β?

(b) Für welche statistische Hypothese entscheiden Sie sich mit einem Signifikanzniveau von $\alpha = .05$?

(c) In welchem Wert haben sich die Schätzfunktionen für den Achsenabschnitt und den Steigungsparameter in der gegebenen Stichprobe realisiert?

(d) Wie werden die Schätzwerte für den Achsenabschnitt und den Steigungsparameter inhaltlich interpretiert?

(e) Wie lautet die geschätzte Regressionsgerade?

Formulieren Sie schließlich einen Ergebnisbericht für die obige Fragestellung gemäß APA-Richtlinien.

Beispiel 9.4

Im Datensatz „Album Sales.sav“ sind die Verkaufszahlen (in Tausenden von Stück) von Musikalben diverser Bands (Variable *Sales*) sowie das Werbebudget für diese Alben (Variable *Adverts*; in Tausenden von Englischen Pfund) gegeben. Untersuchen Sie mit einem geeigneten statistischen Verfahren, ob und zu welchem Anteil sich die Verkaufszahlen auf das verwendete Werbebudget zurückführen lassen. Formulieren Sie einen entsprechenden Ergebnisbericht nach APA-Richtlinien. Beantworten Sie zusätzlich folgende Fragen: wenn das Werbebudget für ein Album um eine Million Pfund gesteigert wird, mit welcher Veränderung der Verkaufszahlen kann man im Mittel rechnen? Mit welcher Streuung um diese mittlere Veränderung kann man auf der Basis der gegebenen Daten rechnen?

Beispiel 9.5

Im Datensatz „Supermodel.sav“ sind die Gehälter (Variable *salary*) von 231 Models sowie die Anzahl an Jahren, für die sie bereits als Model tätig sind (Variable *years*), gegeben. Untersuchen Sie mit einem geeigneten statistischen Verfahren, ob und zu welchem Anteil sich die Gehälter auf die Berufserfahrung (gemessen mit der Variable *years*) zurückführen lassen. Formulieren Sie einen entsprechenden Ergebnisbericht nach APA-Richtlinien.

Beispiel 9.6

Transformieren Sie die Intensität der Abhängigkeitskognitionen im Datensatz „Kap9daten.sav“ indem Sie (a) eine zentrierte Variable erzeugen, (b) die Variable so skalieren, dass Werte von Null dem kleinsten erhobenen Wert und Werte von 1 dem größten erhobenen Wert entsprechen, und (c) eine z-transformierte Variable erzeugen. Lassen Sie sich dann jeweils für ein einfaches lineares Regressionsmodell mit dem Kriterium Depressionsschwere und der jeweiligen transformierten Variablen als Prädiktor Schätzwerte für Achsenabschnitt und Steigung ausgeben. Machen Sie sich mittels der jeweiligen Ausgaben bewusst, wie sich die jeweiligen Transformationen auf die inhaltliche Interpretation der beiden Modellparameter auswirken.

Beispiel 9.7

Ein Statistikprofessor vermutet, dass zwischen der aufgewendeten Lernzeit (Variable *AnzahlStunden*) und dem Logarithmus der Bestehensquote für eine seiner Vorlesungsprüfungen (= das Verhältnis des Anteils an Studierenden, die die Prüfung bestehen, zum Anteil derjenigen, die nicht bestehen) ein positiver linearer Zusammenhang besteht. Daraus folgert er, dass auch zwischen der aufgewendeten Lernzeit und einer regularisierten Logit-Transformierten der bei der Prüfung erreichten Punktzahl (Variable *LogitPunkte*) ein positiver linearer Zusammenhang bestehen sollte. Daher erhebt er seit einiger Zeit bei der entsprechenden Vorlesungsprüfung auch die Anzahl an aufgewendeten Lernstunden aller Studierenden, die die Prüfung absolvieren. Die entsprechenden Daten sind in der Datei „Kap9UE7.sav“ gegeben.

Veranschaulichen Sie zuerst mittels eines Streudiagramms, dass die Annahme eines linearen Zusammenhangs zwischen den Variablen *AnzahlStunden* (UV) und *LogitPunkte* (AV) gerechtfertigt

erscheint. Überprüfen Sie anschließend mit einem geeigneten statistischen Verfahren die Vermutung des Professors bezüglich des positiven linearen Zusammenhangs und verfassen Sie einen entsprechenden Ergebnisbericht. Beantworten Sie zudem die folgenden Fragen: Welcher Anteil an der Varianz der abhängigen Variablen kann durch die Anzahl der Lernstunden erklärt werden? Um welchen Betrag und in welche Richtung ändert sich die abhängige Variable für eine Zunahme der aufgewendeten Lernzeit um 20 Stunden?

Beispiel 9.8

Ein Ernährungswissenschaftler vermutet einen Zusammenhang zwischen dem Anteil an Gemüse (in %), den Studierende im Mittel pro Tag zu sich nehmen und deren Leistung bei Prüfungen. Um dieser Vermutung nachzugehen, erhebt er u.a. entsprechende Daten von 200 Studierenden, die im Datensatz „Kap9UE8.sav“ zu finden sind.

Überprüfen Sie unter der Annahme eines linearen Zusammenhangs mit einem geeigneten statistischen Verfahren die Vermutung des Ernährungswissenschaftlers und verfassen Sie einen entsprechenden Ergebnisbericht. Beantworten Sie zudem die folgenden Fragen: Welcher Anteil an der Varianz der Prüfungsleistung kann in der Stichprobe im Mittel durch den Anteil an Gemüse erklärt werden? Um welchen Betrag und in welche Richtung ändert sich die Prüfungsleistung für eine Zunahme des Gemüseanteils um 10%? Erstellen Sie schließlich ein Streudiagramm für die beiden Variablen und fügen Sie dieses zu Ihrem Ergebnisbericht hinzu.

Beispiel 9.9

Für das folgende (fiktive) Beispiel können Sie davon ausgehen, dass die für die lineare Regression notwendigen Annahmen allesamt erfüllt sind.

Eine Forscherin vermutet, dass zwischen der Menge an Brokkoli, die Personen wöchentlich zu sich nehmen, und der Intelligenz der Personen ein Zusammenhang besteht. Daher vermutet die Forscherin, dass sich die Leistung bei einem Intelligenztest zum Teil auf die Menge an wöchentlich verzehrtem Brokkoli zurückführen lässt. Um diese Hypothese zu testen, rekrutiert die Forscherin 464 Personen, um deren IQ und die wöchentlich verzehrte Menge an Brokkoli (in g) zu ermitteln. Die

erhobenen Daten befinden sich in der Datei „brokkoli.sav“. Verwenden Sie für alle inferenzstatistischen Tests ein Signifikanzniveau von .005.

(a) Führen Sie eine einfache lineare Regression durch, um die Hypothese der Forscherin inferenzstatistisch zu prüfen. Fassen Sie Ihr Ergebnis in einem entsprechenden Ergebnisbericht zusammen. Wie viel Varianz bezogen auf die Gesamtvarianz der Intelligenzleistung kann die Menge wöchentlich verzehrten Brokkolis erklären? Wie sehr verändert sich die Intelligenzleistung pro g wöchentlich verzehrten Brokkolis?

(b) Ein anderer Forscher vermutet, dass neben der Menge wöchentlich verzehrten Brokkolis auch die Menge wöchentlich verzehrter Karotten ein bedeutsamer Prädiktor für die Intelligenz einer Person ist. Der Forscher befragt daher dieselben 464 Personen nach der Menge wöchentlich verzehrter Karotten (ebenfalls in g). Zeigen Sie, dass sowohl die Menge wöchentlich verzehrten Brokkolis als auch die Menge wöchentlich verzehrter Karotten signifikante Prädiktoren für die Intelligenzleistung sind. Wie viel Varianz bezogen auf die Gesamtvarianz der Intelligenzleistung können die beiden Prädiktoren gemeinsam erklären?

Beispiel 9.10

Ein Club organisiert regelmäßig Konzerte. Um den Umsatz zu optimieren möchten die Konzertveranstalter:innen herausfinden, welche Faktoren zum Erfolg (= Anzahl Besucher; Variable *Besucher*) eines Konzertes beitragen. Aus ihrer langjährigen Erfahrung wissen sie, dass der Erfolg unter anderem vom Ticketpreis (in Schweizer Franken; Variable *Preis*), dem Werbeaufwand (in Schweizer Franken; Variable *Werbung*), sowie dem Erfolg der Band (Anzahl verkaufter CDs; Variable *CD_Verkauf*) abhängt. Dies möchten die Veranstalter nun statistisch überprüfen, um künftig den Erfolg eines Konzertes im Voraus besser abschätzen zu können. Führen Sie eine lineare Regressionsanalyse für diese Fragestellung durch und verfassen Sie einen Ergebnisbericht gemäß APA-Richtlinien. Die Daten für dieses Beispiel finden Sie in der Datendatei „konzertbesuche.sav“. Hinweis: Sie können für dieses Beispiel davon ausgehen, dass die für die lineare Regression notwendigen Annahmen allesamt erfüllt sind.

Beispiel 9.11

In der Datei „sterne.sav“ sind die Logarithmen der Oberflächentemperatur und der Leuchtkraft von 47 Sternen gegeben. Zwischen dem Logarithmus der Oberflächentemperatur und dem Logarithmus der Leuchtkraft eines Sterns im Hauptreihenstadium besteht laut Theorie näherungsweise ein linearer Zusammenhang: mit steigender Oberflächentemperatur nimmt die Leuchtkraft zu. Für die folgenden Berechnungen können Sie davon ausgehen, dass die für die lineare Regression notwendigen Annahmen allesamt erfüllt sind.

(a) Führen Sie eine einfach lineare Regressionsanalyse durch und erstellen Sie einen entsprechenden Ergebnisbericht. Wie würden Sie das Resultat in Hinsicht auf die theoretische Vorhersage interpretieren?

(b) Bei der Inspektion eines Streudiagramms für die 47 Sterne stellt ein Astrophysiker fest, dass das Diagramm vier Sterne enthält, die sehr hohe Leuchtkraft (> 5.5) bei sehr geringer Oberflächentemperatur (< 3.6) aufweisen. Da es sich bei diesen Sternen vermutlich nicht um Hauptreihensterne, sondern um sogenannte Rote Riesen handelt, empfiehlt der Astrophysiker die Quantifizierung des linearen Zusammenhangs unter Ausschluss dieser vier Sterne zu wiederholen. Zu welchem Ergebnis kommen Sie in diesem Fall und was schließen Sie daraus für den theoretisch postulierten Zusammenhang zwischen den Logarithmen von Oberflächentemperatur und Leuchtkraft?

Dieses Beispiel wurde inspiriert von der Erläuterung desselben Sachverhalts im Rahmen der Korrelationsanalyse bei Wilcox (2022; S. 543). Die Daten entsprechen ebenfalls in etwa den dort in Abb. 9.2 abgebildeten Daten, die ursprünglich auf Rousseeuw und Leroy (1987) zurückgehen.

Kapitel 10

Regressionsdiagnostik, Effektstärken, Stichprobenplanung, Kollinearität

Stefan E. Huber

Regressionsdiagnostik

Mit Regressionsdiagnostik ist eine Überprüfung der Annahmen des Regressionsmodells gemeint. Typischerweise wird im Rahmen einer Regressionsdiagnostik auch eine Ausreißeranalyse durchgeführt. Grundsätzlich ist die Regressionsdiagnostik vor dem inferenzstatistischen Verfahren durchzuführen (Bühner et al., 2025), auch wenn wir aus didaktischen Gründen hier die Reihenfolge umgekehrt haben.

Der Einfachheit halber seien hier die Annahmen der linearen Regression noch einmal wiederholt: (i) es besteht ein linearer Zusammenhang zwischen Prädiktoren und Erwartungswert des Kriteriums und (ii) die Abweichungen von der Regressionsgerade (oder Regressionsebene bei zwei Prädiktoren oder Regressionshyperebene bei mehr als zwei Prädiktoren) können durch identische und unabhängig normalverteilte Zufallsvariablen mit Erwartungswert Null und konstanter Varianz für jede beliebige Realisation der Prädiktoren beschrieben werden. Für die Annahme der konstanten Varianz der Fehler werden wie schon für das varianzanalytische Modell häufig auch die Begriffe Varianzhomogenität oder Homoskedastizität verwendet. Liegt keine konstante Varianz vor, spricht man von Varianzheterogenität bzw. Heteroskedastizität.

Während die Unabhängigkeit der Fehler durch die Stichprobenziehung gewährleistet werden muss, können die Annahmen der Linearität, der Normalverteilung und der konstanten Varianz der Fehler allesamt verletzt sein, und sollten daher überprüft werden. Neben der meist zusätzlich durchgeführten Ausreißeranalyse umfasst die Regressionsdiagnostik demnach (Bühner et al., 2025):

- die Überprüfung der Linearitätsannahme,
- die Überprüfung der Normalverteilungsannahme,
- sowie die Überprüfung der Homoskedastizitätsannahme.

Da sich die Fehler ε_i auf die unbekannte wahre Regressionsgerade bzw. -(hyper)ebene beziehen, können ihre Eigenschaften (außer in Simulationsstudien) im konkreten Fall nicht untersucht werden.

Stattdessen werden die sog. Residuen verwendet, d.h. die Differenzen zwischen den durch die Regressionsgleichung mit den geschätzten Modellparametern vorhergesagten Kriteriumswerten und den tatsächlich gemessenen Kriteriumswerten. Aus mathematischen Gründen empfiehlt es sich zudem hierbei sog. studentisierte Residuen zu verwenden.

Zur Illustration der Durchführung einer Regressionsdiagnostik verwenden wir im Folgenden wieder den schon aus dem vorherigen Kapitel bekannten Datensatz „Kap9daten.sav".

Überprüfung der Linearitätsannahme

Im Falle der einfachen linearen Regression kann die Linearitätsannahme mithilfe eines Streudiagramms für AV und UV überprüft werden. Ein solches kann in SPSS unter *Graphs >> Chart Builder…* angefordert werden. Dort kann dann im Reiter „Gallery" die Grafikrubrik „Scatter/Dot" und für diese wiederum die erste Auswahlmöglichkeit ganz links ausgewählt werden. Zu Illustrationszwecken tragen wir die *Depressionsschwere* nach oben (auf die sog. „y-Achse") und die *negative Selbstbewertung* nach rechts („x-Achse") auf, indem wir die beiden Variablen in die entsprechenden Felder im Fenster „Chart preview…" ziehen. Zusätzlich fügen wir dem Streudiagramm noch eine lineare Fitgerade hinzu. All die getroffenen Auswahlen sind auch in Abbildung 10.1 illustriert.

Das Ergebnis ist in Abbildung 10.2 gezeigt. Wir sehen, dass sich die einzelnen Datenpunkte relativ dicht und gleichmäßig um die lineare Fitgerade drängen. Die Annahme eines linearen Zusammenhangs zwischen diesen beiden Variablen scheint also durchaus ihre Berechtigung zu haben.

Zum Vergleich ist in Abbildung 10.3 eine Datensituation dargestellt, die wenig Grund zur Annahme eines linearen Zusammenhangs zwischen den beiden Variablen im Streudiagramm gibt. Die Punktewolke weicht einerseits stark von der linearen Fitgerade ab, aber insbesondere scheinen die Abweichungen dabei auch einem bestimmten Muster zu folgen. Für kleine und große x-Werte weichen die y-Werte eher nach oben ab, während sie für mittlere x-Werte stark nach unten hin abweichen. In der Tat handelt es sich bei den gegen x aufgetragenen y-Werten um eine Überlagerung einer Sinusfunktion mit weißem Rauschen, bzw. präziser: $y = y(x) = \sin(x) + N(0,1)$. Es liegt also tatsächlich kein linearer Zusammenhang vor. Ob aber in einer konkreten Situation die Annahme eines linearen Zusammenhangs eher gerechtfertigt ist oder nicht, ist selten so klar wie im gezeigten Beispiel.

Im Rahmen der Überprüfung der Linearitätsannahme ist es wichtig zu bemerken, dass für eine sinnvolle Interpretation der geschätzten Modellparameter eines linearen Regressionsmodells zumindest ein monoton steigender oder fallender Zusammenhang zwischen Prädiktor und Kriterium bestehen sollte. Ist dieser Zusammenhang in Wahrheit nicht linear (sondern z.B. exponentiell oder quadratisch) wird der wahre Zusammenhang durch die Annahme der Linearität zwar unter- oder überschätzt, aber es wird zumindest immer noch ein wichtiger Aspekt des Zusammenhangs, nämlich das Steigen oder Fallen des Kriteriums mit steigendem Prädiktor, zum Teil erfasst.

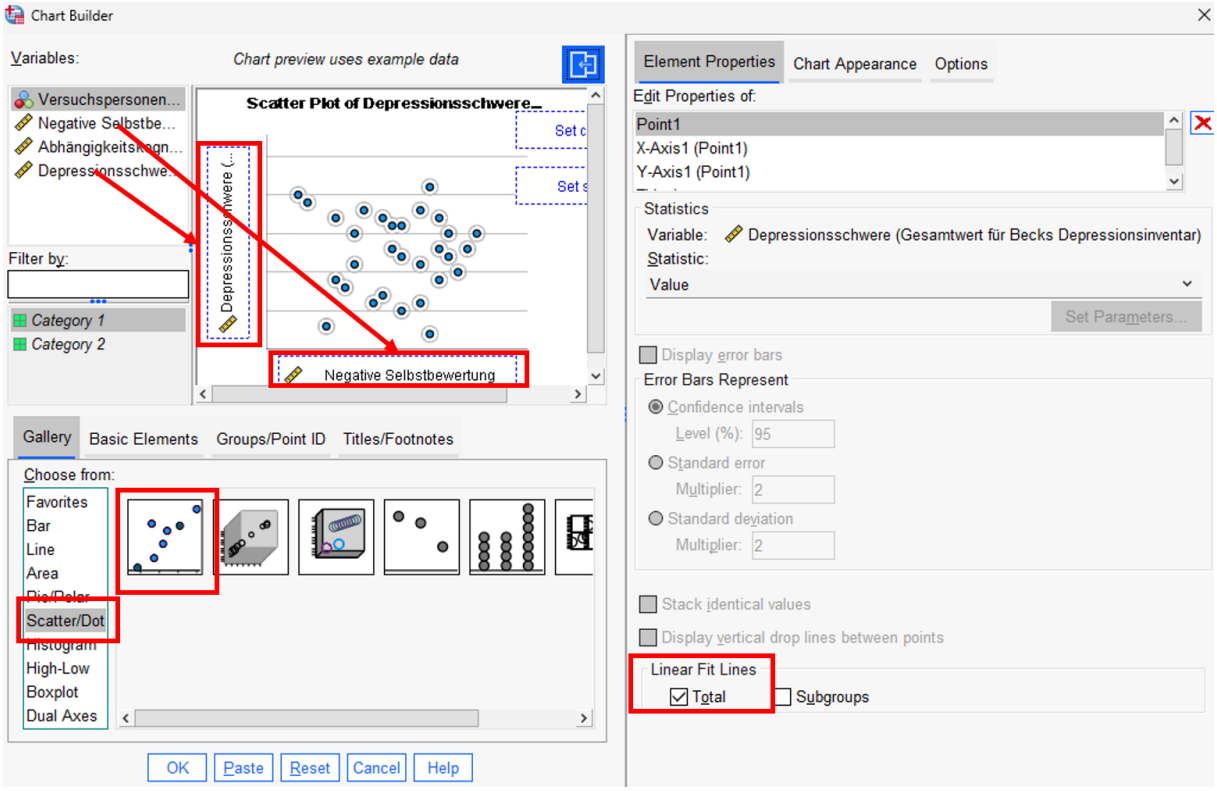


Abbildung 10.1. Anforderung eines Streudiagramms für die *Depressionsschwere* und die *negative Selbstbewertung* inklusive einer linearen Fitgeraden.

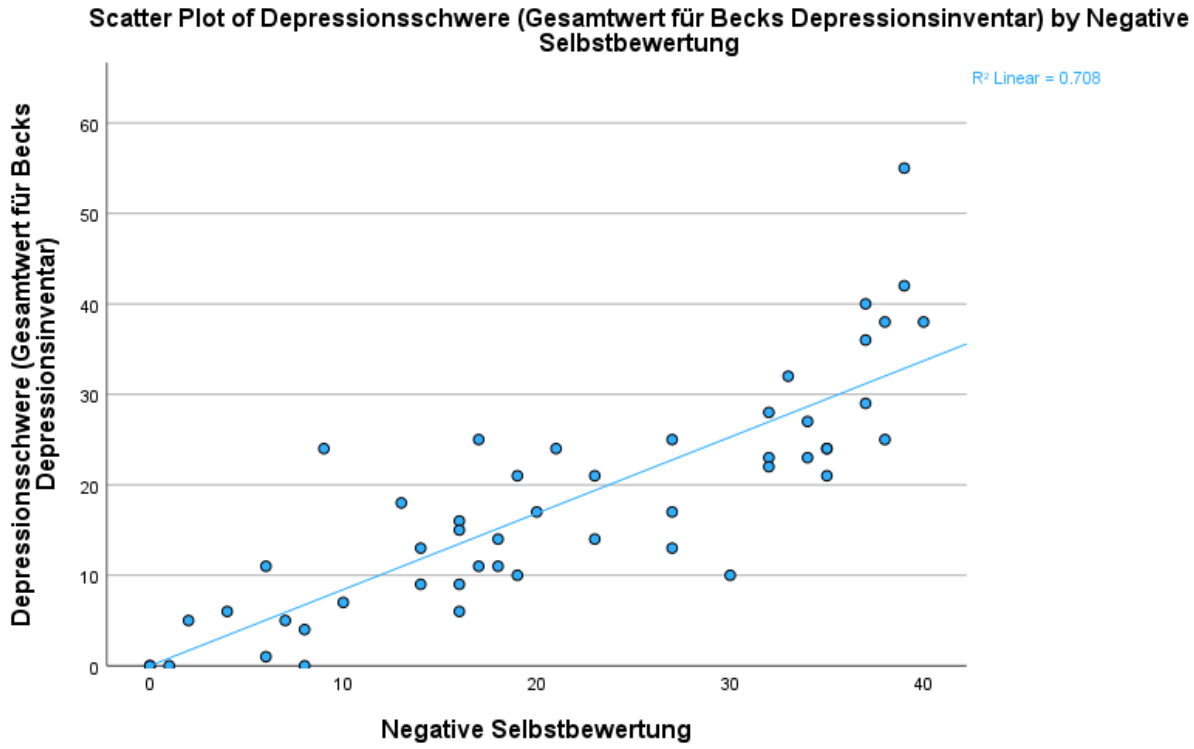


Abbildung 10.2. Streudiagramm für die *Depressionsschwere* und die *negative Selbstbewertung*.

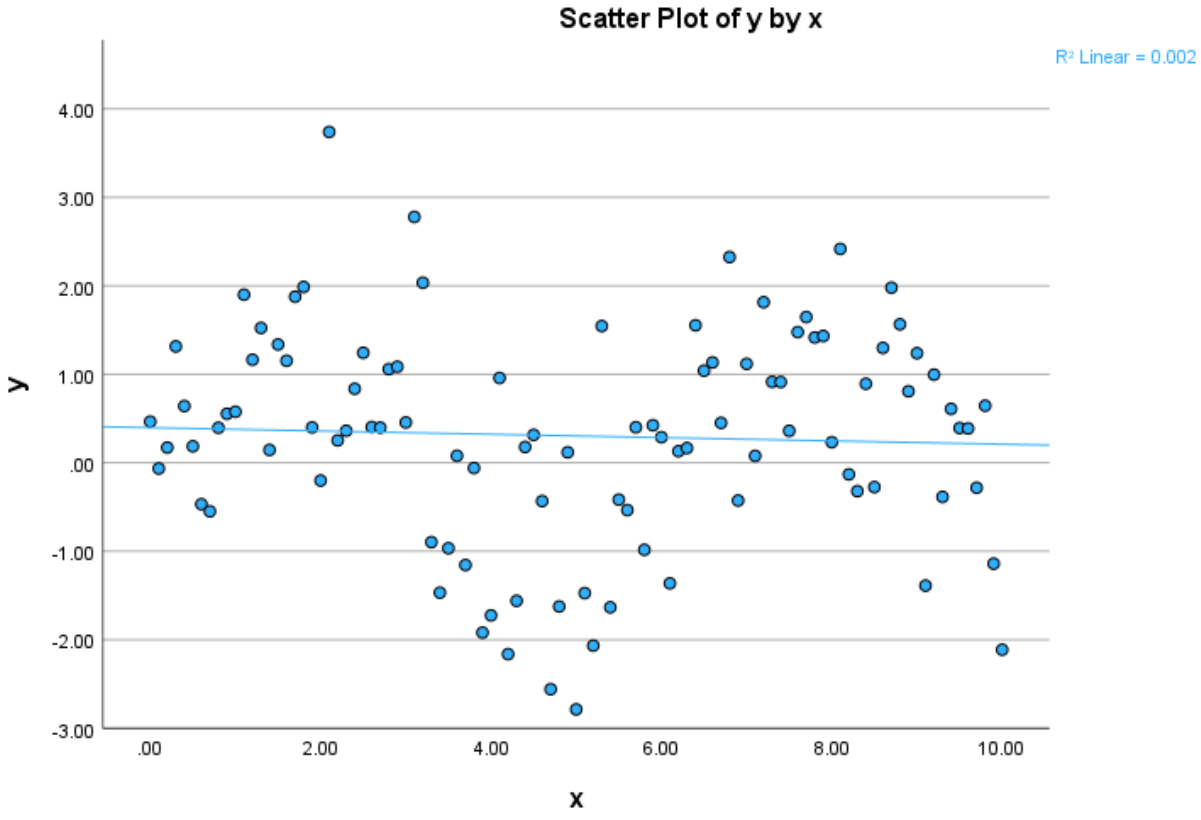


Abbildung 10.3. Streudiagramm für zwei Variablen, zwischen denen eher kein linearer Zusammenhang bestehen dürfte.

Besteht nicht einmal ein monotoner Zusammenhang, lässt sich dieser mit einem linearen Regressionsmodell gar nicht abbilden. D.h., wenn eine lineare Regressionsanalyse kein Indiz für einen linearen Zusammenhang liefert, heißt dies nicht, dass zwischen den beiden Variablen gar kein Zusammenhang besteht. So besteht etwa im in Abbildung 10.3 illustrierten Fall ganz klar ein Zusammenhang zwischen x und y; dieser ist eben gerade durch die Funktionsvorschrift $y(x)$ oben definiert. Allerdings besteht zwischen den beiden Variablen eben kein *linearer* Zusammenhang. Die Durchführung einer linearen Regressionsanalyse ergibt auch eine verschwindend kleine Steigung. Das ist kein Fehler. In der Tat hängt y nicht linear von x ab. Daraus aber abzuleiten, dass y generell nicht von x abhängt wäre ein Fehler. Die bestehende (sinusförmige) Abhängigkeit kann schlichtweg nicht mit der Steigung β einer Geraden $y(x) = \alpha + \beta x$ erfasst werden. Um diesen Zusammenhang zu quantifizieren müssten wir andere Modelle verwenden.

Wie Bühner et al. (2025) richtig folgern, beeinträchtigt die Verletzung der Linearitätsannahme also die Interpretation der Modellparameter selbst und nicht nur die inferenzstatistischen Verfahren zur Testung dieser Parameter. Daran kann auch ein großer Stichprobenumfang nichts ändern. Daher sollte die Linearitätsannahme in jedem Fall (d.h. insbesondere auch für sehr große Stichproben) überprüft werden. Zumindest eine graphische Überprüfung ist in jedem Fall zu empfehlen (Anscombe, 1973).

Wie kann diese Annahme im Fall einer multiplen linearen Regression überprüft werden? Dafür muss einerseits die Linearitätsannahme für jeden Prädiktor einzeln veranschaulicht, andererseits zusätzlich die linearen Zusammenhänge der jeweils anderen Prädiktoren mit dem Kriterium herausgerechnet (oder auch: „herauspartialisiert“) werden. Eine entsprechende Überprüfung der Linearität kann in SPSS mit sog. partiellen Regressions-Plots geleistet werden.

Dazu wird zuerst unter *Analyze >> Regression >> Linear…* zuerst wieder die multiple Regressionsanalyse, für die die partiellen Regressions-Plots inspiziert werden sollen, angefordert. Unter „Plots“ wird dann zusätzlich „Produce all partial plots“ ausgewählt, siehe Abbildung 10.4. Die Ergebnisse sind in Abbildung 10.5 und Abbildung 10.6 gezeigt. Beiden Streudiagrammen wurde nachträglich noch eine Fitgerade durch Doppelklick auf die entsprechende Grafik in der Ausgabe und dann Auswahl der entsprechenden Schaltfläche im Grafikeditor, siehe Abbildung 10.7, hinzugefügt.

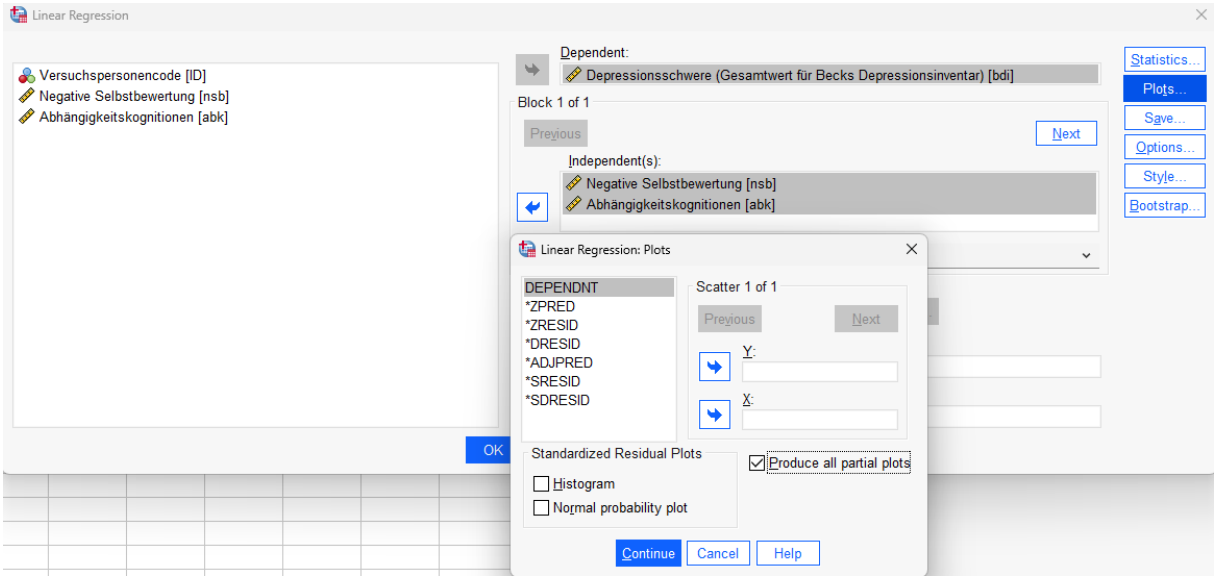


Abbildung 10.4. Auswahl der partiellen Regressions-Plots in SPSS.

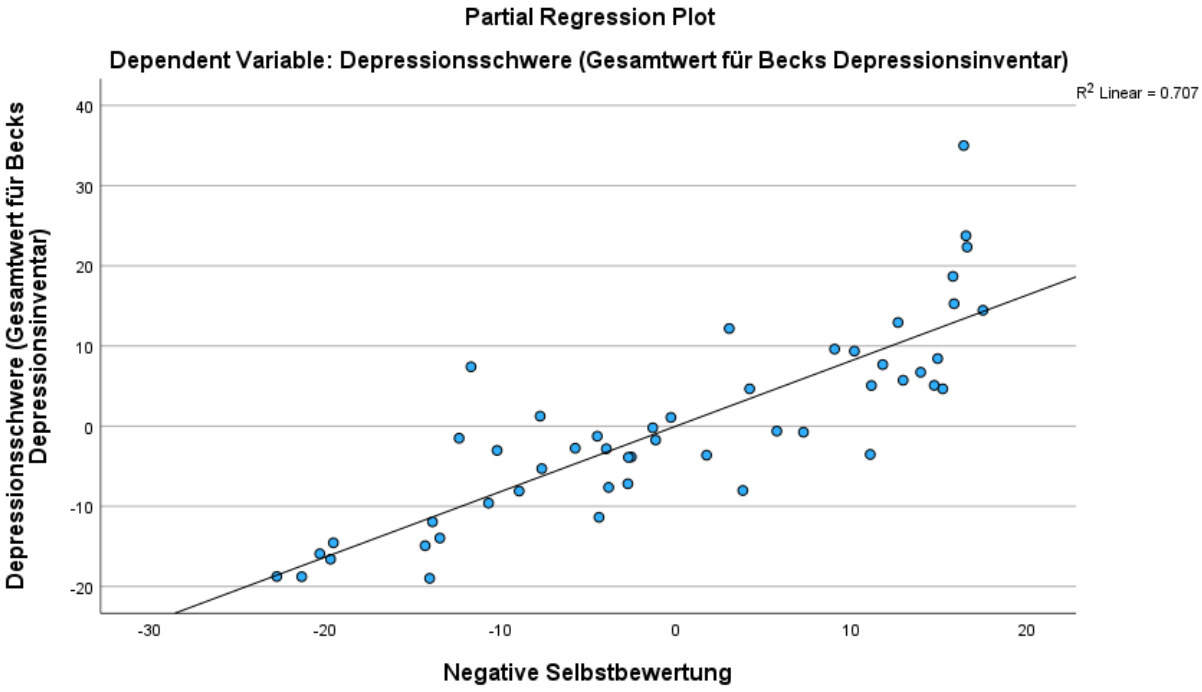


Abbildung 10.5. Partieller Regressions-Plot für die negative Selbstbewertung.

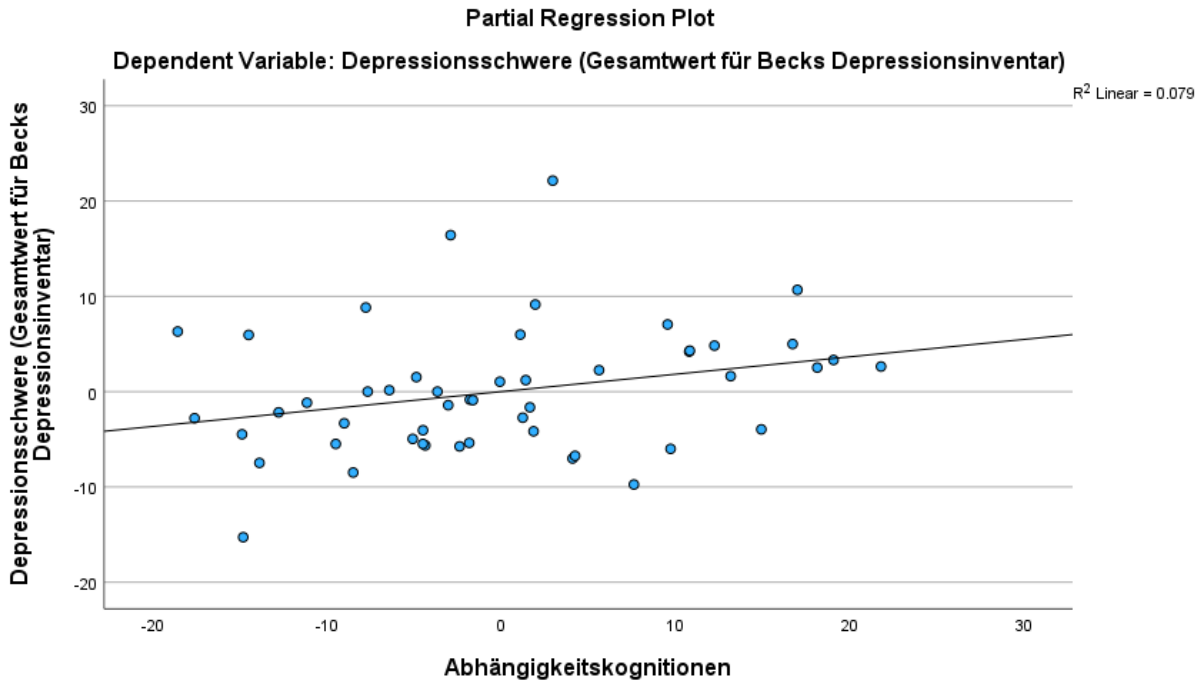


Abbildung 10.6. Partieller Regressions-Plot für die Abhängigkeitskognitionen.

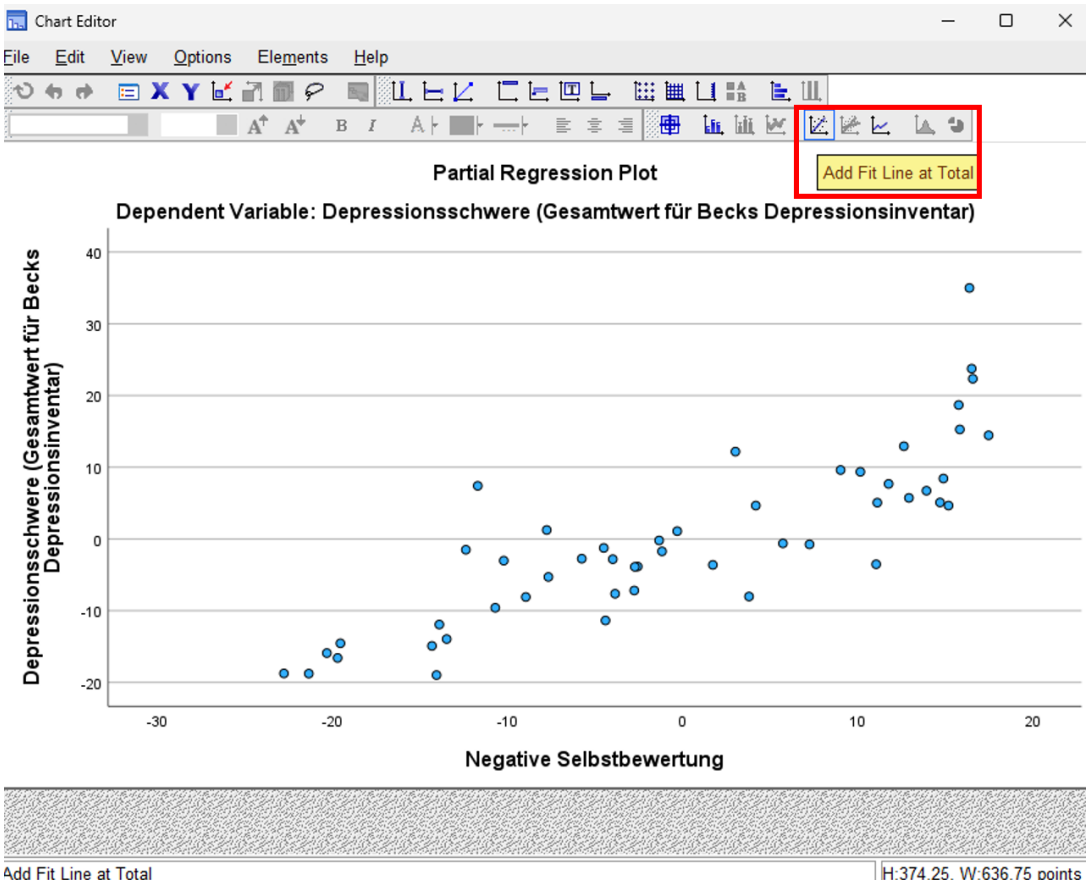


Abbildung 10.7. Hinzufügen einer Fitgeraden zu einem Streudiagramm im Grafikeditor.

Überprüfung der Normalverteilungsannahme

Die Überprüfung der Normalverteilungsannahme kann durch Inspektion eines Histogramms der standardisierten Residuen erfolgen. Dieses kann ebenfalls unter „Plots“ im Menü zur Anforderung der multiplen Regressionsanalyse ausgewählt werden, siehe Abbildung 10.9. Das sich ergebende Histogramm ist in Abbildung 10.8 dargestellt.

Die Überprüfung der Annahme anhand dieser Abbildung ist in der Tat sehr subjektiv. Im vorliegenden Fall scheint die Annahme auch nur mehr schlecht als recht erfüllt zu sein (sie ist es allerdings, weil die fiktiven Daten entsprechend der Annahme erzeugt wurden). Allerdings lässt sich zeigen, dass die inferenzstatistischen Verfahren im Rahmen der linearen Regression relativ robust gegenüber der Verletzung der Normalverteilungsannahme sind (Rajh-Weber et al., 2025). Eine Verletzung der Normalverteilungsannahme ist daher insbesondere in großen Stichproben nicht so schlimm (Bühner et al., 2025).

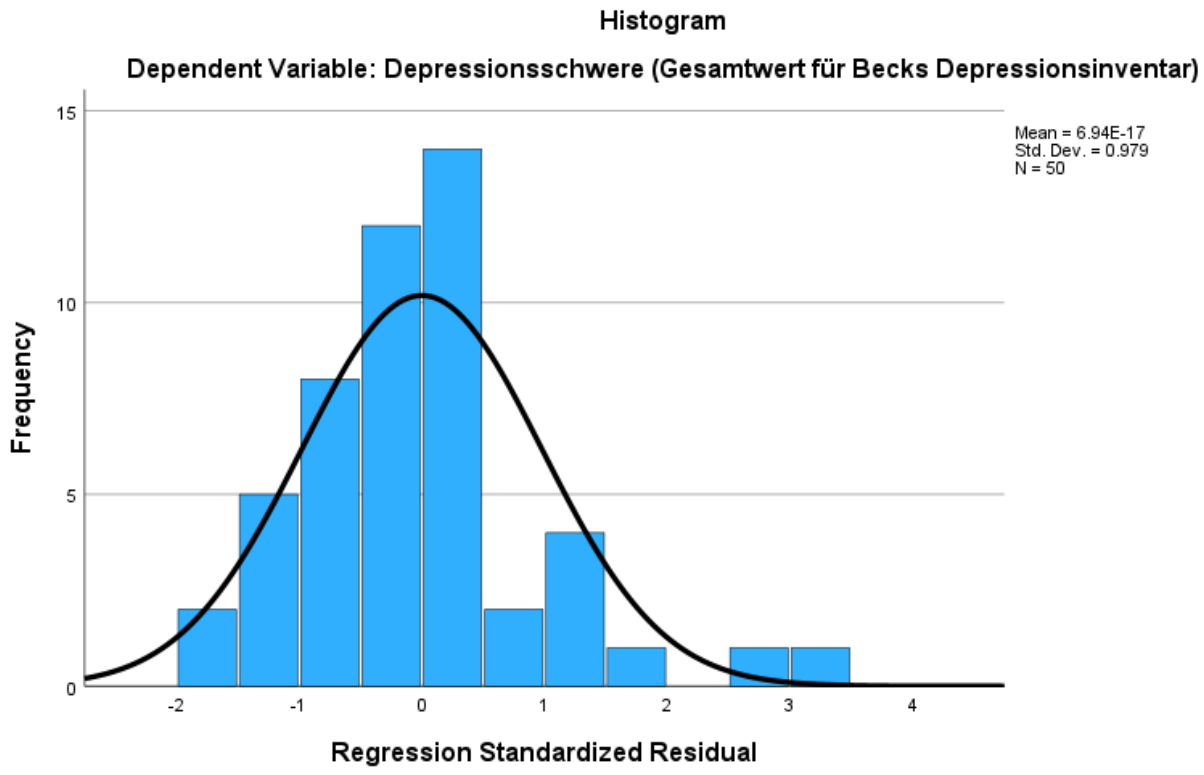


Abbildung 10.8. Histogramm der standardisierten Residuen inklusive einer gefitteten Normalverteilungskurve.

Überprüfung der Homoskedastizität

Auch die Annahme der Homoskedastizität kann mittels Inspektion einer geeigneten Grafik überprüft werden. Dazu wird unter „Plots“ im Menü zur Anforderung der multiplen Regressionsanalyse ein Streudiagramm für die studentisierten Residuen sowie die z-transformierten Vorhersagewerte des Regressionsmodells angefordert, siehe Abbildung 10.9.

Grund zur Annahme von Homoskedastizität liegt dann vor, wenn die studentisierten Residuen (als Schätzwerte der unbekannten Fehler im linearen Regressionsmodell, auf die sich die Annahme der Varianzhomogenität bezieht) sich gleichmäßig über den gesamten Bereich oberhalb und unterhalb der horizontalen Nulllinie verteilen (die rote Nulllinie in Abbildung 10.10 wurde nachträglich zur einfacheren Beurteilung eingefügt). Im vorliegenden Fall scheint die Annahme eher verletzt als erfüllt zu sein (als Erzeuger der Daten weiß der Verfasser aber, dass sie durchaus erfüllt war). Im Regelfall ist die grafische Beurteilung wiederum sehr subjektiv und selten eindeutig.

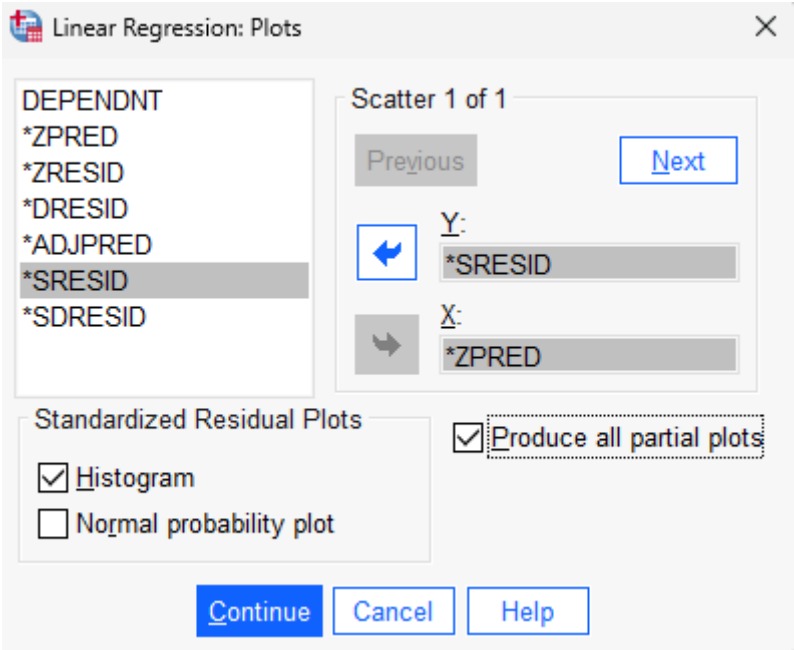


Abbildung 10.9. Anforderung (i) eines Streudiagramms für die studentisierten Residuen und die z-transformierten Vorhersagewerte des Regressionsmodells (rechts oben im Feld „Scatter 1 of 1“), (ii) eines Histogramms für die standardisierten Residuen, und (iii) der partiellen Regressions-Plots.

Im Zweifelsfall ist es aber im Fall der Homoskedastizitätsannahme durchaus ratsam auf alternative inferenzstatistische Verfahren zurückzugreifen, da Heteroskedastizität die inferenzstatistischen Verfahren im Rahmen multipler Regressionsanalysen selbst bei großen Stichproben maßgeblich beeinträchtigen kann (Rajh-Weber et al., 2025). Mögliche Alternativen sind Bootstrap-Verfahren (verfügbar im Menü für die lineare Regression in SPSS, siehe z.B. Bühner & Ziegler, 2017, für Beschreibung und Anleitung), Korrektur der Standardfehler (z.B. mit der HC3 Methode, siehe z.B. Rajh-Weber et al., 2025), oder voraussetzungsrobustere Verfahren (siehe z.B. Wilcox, 2022).

Ausreißeranalyse

Ausreißer, d.h. Datenpunkte mit ungewöhnlich großen oder kleinen Werten für AV oder UV, können die Ergebnisse einer Regressionsanalyse stark verzerren. Dabei sind vor allem Einflusswerte von besonderer Bedeutung, die sowohl ungewöhnlich große oder kleine Werte für die UV aufweisen (= sog. Hebelwerte) als auch ungewöhnlich weit von der (ohne Ausreißer) geschätzten Regressionsgerade abweichen (= sog. Diskrepanzwerte).

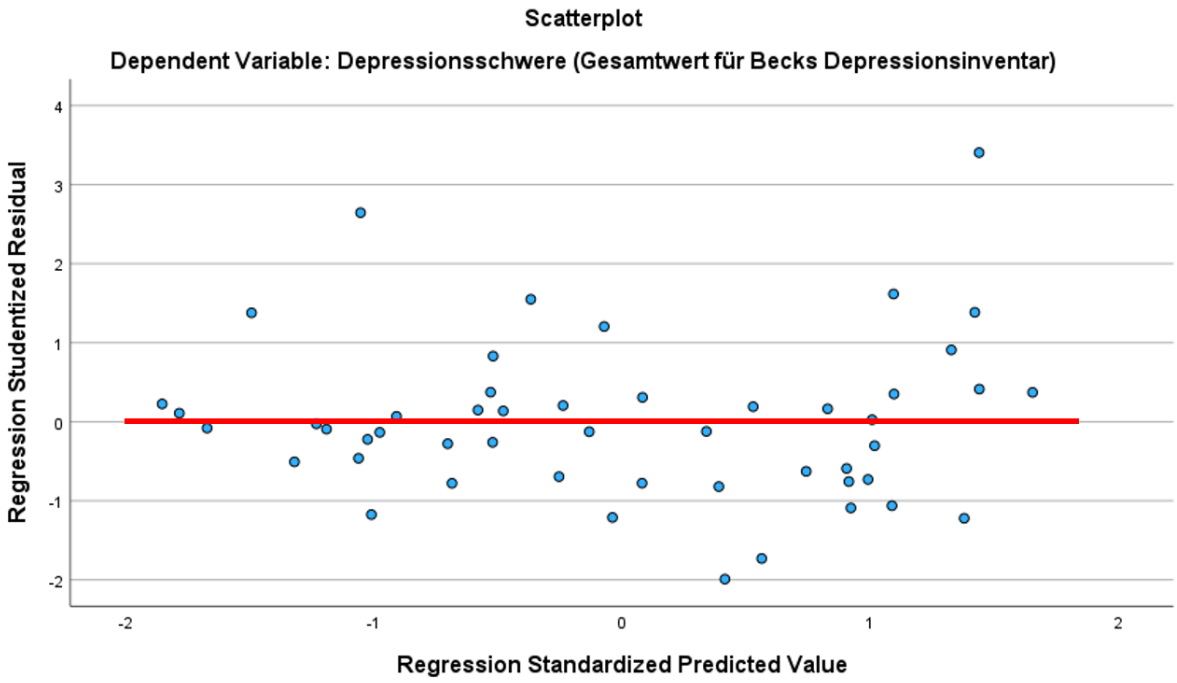


Abbildung 10.10. Streudiagramm für die studentisierten Residuen und die z-transformierten Vorhersagewerte des Regressionsmodells.

Einflusswerte lassen sich mit der sog. Cook'schen Distanz identifizieren. Diese Metrik gibt für jeden Fall (Person) an, wie stark sich die Vorhersagewerte für alle anderen Fälle (Personen) ändern würden, wenn dieser eine Fall (Person) von der Regressionsanalyse ausgeschlossen werden würde. Je größer die Cook'sche Distanz, desto höher der Einflusswert. Manche Autor:innen (Bühner et al., 2025) empfehlen daher, alle Fälle als kritisch zu betrachten, die eine Cook'sche Distanz von über $4/n$ aufweisen, wobei n den Stichprobenumfang bezeichnet. Allerdings ist dieser Zugang etwas problematisch, da solche Werte in 5% aller Falle rein statistisch zu erwarten sind, wenn alle Annahmen der linearen Regressionsanalyse erfüllt sind (siehe Übungsaufgabe 10.3). Als Alternative wird daher empfohlen, Fälle mit den extremsten Cook'schen Distanzen genauer anzusehen (Bühner et al., 2025). Sich eingehender mit den Daten auseinanderzusetzen ist grundsätzlich immer eine gute Idee.

Allerdings sollten Fälle niemals ausschließlich deshalb aus den Daten entfernt werden, weil sie extremere Werte als die Mehrzahl der Fälle aufweisen. Liegen keine offensichtlichen Fehler (etwa offensichtliche Fehler beim Digitalisieren von Papier-und-Bleistift-Fragebogendaten; z.B. Eintragen von Werten, die auf Item-Skalen gar nicht möglich sind) vor, sollte stattdessen besser auf Methoden zurückgegriffen werden, die robust gegenüber den Effekten von einzelnen Ausreißern sind (siehe z.B. Mair & Wilcox, 2020; Wilcox, 2022). Unter Umständen können Analysen auch einmal mit und einmal ohne extreme Ausreißerwerte durchgeführt werden, um immerhin deren Einfluss auf die Schlussfolgerungen quantifizieren zu können.

Die Ermittlung der Cook'schen Distanz für alle Fälle (Personen) erfolgt in SPSS ebenfalls im Menü zur Anforderung der Regressionsanalyse und dort im Untermenü „Save…", siehe Abbildung 10.11. Durch Auswahl dieser Option wird dann eine neue Variable erzeugt (mit dem klingenden Namen „COO_1"). Für den vorliegenden Datensatz beläuft sich der Stichprobenumfang auf n = 50, d.h. $4/n$ = 0.08. In der Datenansicht können alle Fälle durch Rechtsklick auf den Spaltennamen „COO_1" und Auswahl von „Sort Descending" absteigend nach der Cook'schen Distanz sortiert werden. Wir sehen, dass nur drei Fälle eine Cook'sche Distanz größer als 0.08 aufweisen. Da 3/50 = 0.06 entspricht dies sehr genau den erwarteten 5% unter Gültigkeit aller Voraussetzungen für die lineare Regression, woraus der Verfasser dieses Dokuments schlussfolgern würde, dass hier vermutlich keine untypische Datensituation bestehen dürfte.

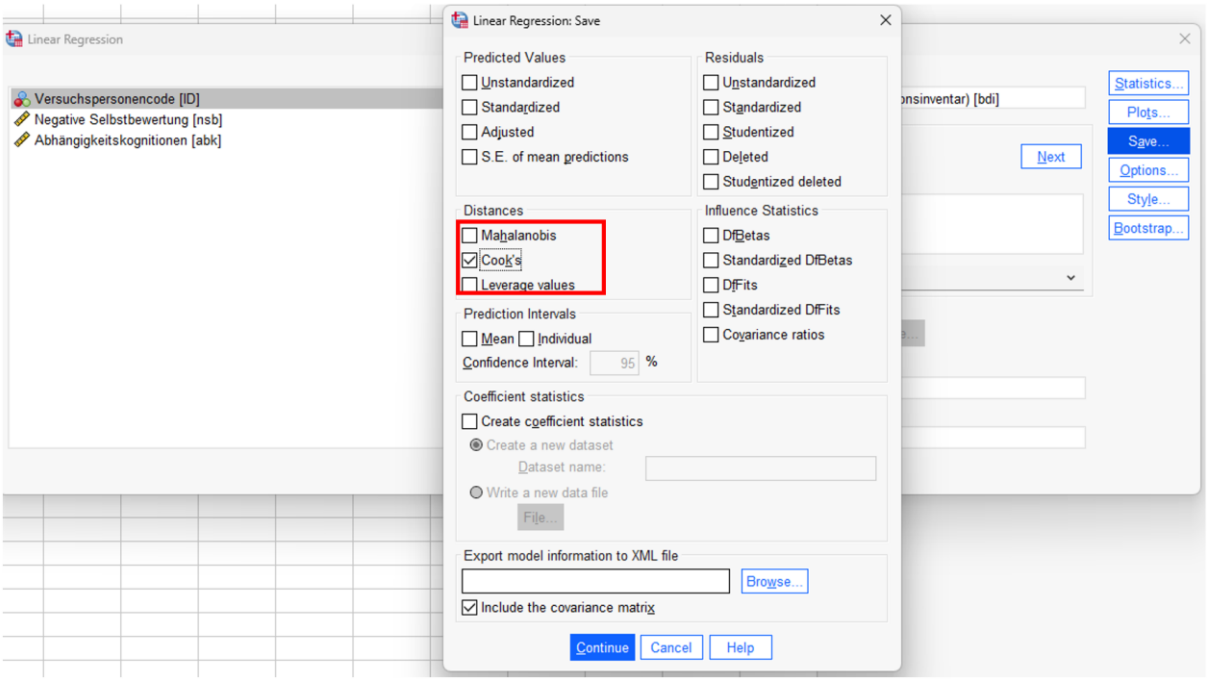


Abbildung 10.11. Ermittlung der Cook'schen Distanz in SPSS.

	ID	nsb	abk	bdi	COO_1
1	14	39	25	55	.25671
2	18	9	15	24	.10424
3	25	30	6	10	.10211
4	17	21	1	24	.07565
5	47	6	3	11	.06481
6	42	37	14	40	.05897
7	16	39	24	42	.04172
8	12	17	36	25	.04115
9	24	27	28	13	.03631

Abbildung 10.12. Fälle mit der größten Cook'schen Distanz in unserem Beispieldatensatz.

Effektstärken

Im Rahmen der multiplen linearen Regression können zwei Typen von Effektstärken unterschieden werden (Bühner et al., 2025): Effektstärken, die die Stärke des Zusammenhangs aller Prädiktoren gemeinsam mit dem Kriterium quantifizieren, sowie Effektstärken, die die Stärke des Zusammenhangs eines einzelnen Prädiktors mit dem Kriterium quantifizieren.

Eine Effektstärke des ersten Typs haben wir mit dem Determinationskoeffizienten R^2 bereits kennengelernt. Auch eine Effektstärke des zweiten Typs haben wir mit dem standardisierten Regressionskoeffizienten bereits kennengelernt, aber bisher noch nicht im Detail erläutert. Dieser gibt uns neben der Stärke des Zusammenhangs auch eine Information über die Richtung des Zusammenhangs (unter Konstanthaltung aller anderen Prädiktoren). Zusätzlich zum standardisierten Regressionskoeffizienten werden wir mit der quadrierten Semipartialkorrelation noch eine weitere Effektstärke des zweiten Typs kennenlernen, die uns wiederum ein Maß für die Stärke des Zusammenhangs, aber nicht für seine Richtung angibt. Sie hat allerdings eine sehr anschauliche Bedeutung in Relation zum Determinationskoeffizienten, weshalb sie für eine Interpretation der Ergebnisse einer multiplen linearen Regressionsanalyse einen Mehrwert darstellt, den der standardisierte Regressionskoeffizient nicht liefern kann.

Der Determinationskoeffizient R^2

In der einfachen linearen Regression ist der Determinationskoeffizient R^2 schlichtweg das Quadrat des Pearson Korrelationskoeffizienten und gibt an, welchen Varianzanteil sich Prädiktor und Kriterium teilen. Für die multiple lineare Regression ist der Determinationskoeffizient R^2 eine direkte Verallgemeinerung der quadrierten Pearson Korrelation aus der einfachen linearen Regression. Zu seiner Berechnung können einfach die durch alle Prädiktoren vorhergesagten Erwartungswerte des Kriteriums ermittelt werden, und deren Varianz dann mit der Varianz des Kriteriums ins Verhältnis gesetzt werden (Bühner et al., 2025). Diese Größe kann dann als jener Anteil der Varianz des Kriteriums interpretiert werden, der durch alle Prädiktoren zusammen „erklärt“ werden kann. Aufgrund der inhaltlichen Nähe zur Pearson Korrelation aus der einfachen linearen Regression wird die Wurzel aus dem Determinationskoeffizienten auch als multiple Korrelation bezeichnet.

Sowohl die multiple Korrelation als auch den Determinationskoeffizienten haben wir bei der Besprechung der Ausgaben für die lineare Regression in SPSS schon wiederholt gesehen und auch schon in entsprechenden Ergebnisberichten verwendet. Zur Wiederholung bemühen wir hier noch einmal den Datensatz aus Kapitel 9, d.h. die Datei „Kap9daten.sav“, und führen eine multiple lineare Regressionsanalyse mit den beiden Prädiktoren negative Selbstbewertung und Abhängigkeitskognitionen und dem Kriterium Depressionsschwere durch. Der für uns hier relevante Teil der Ausgabe ist in Abbildung 10.13 dargestellt. Wir sehen, dass R^2 = .73 und $R = \sqrt{R^2}$ = .86. Bei beiden Größen können wir entsprechend APA-Konventionen wieder die führende Null weglassen, da sie nur zwischen Null und Eins variieren können.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.855[a]	.731	.719	6.552

a. Predictors: (Constant), Abhängigkeitskognitionen, Negative Selbstbewertung

Abbildung 10.13. Ausgabe der multiplen Korrelation und des Determinationskoeffizienten für eine multiple lineare Regressionsanalyse in SPSS.

Es sei hier noch angemerkt, dass manche Autor:innen (Bühner et al., 2025) völlig zu Recht sehr streng zwischen Parameter, Schätzfunktion und Schätzwert unterscheiden und dafür auch ihre je eigene Notation einführen, z.B. griechische Buchstaben ausschließlich für (unbekannte) Modellparameter, große lateinische Buchstaben für Schätzfunktionen, kleine lateinische Buchstaben für Schätzwerte. Auch wenn hier in Anlehnung an die SPSS-Notation (und die übliche Notation nach APA-Richtlinien; American Psychological Association, 2019) für die multiple Korrelation ein großer lateinischer Buchstabe verwendet wird, handelt es sich hier wie auch in der SPSS-Ausgabe natürlich zweifelsfrei jeweils um konkrete Schätzwerte. Grundsätzlich darf man sich von unterschiedlichen Notationen nicht zu sehr verwirren lassen. Viel wichtiger als die Frage, welche Symbole verwendet werden, ist die Frage, was diese Symbole jeweils denotieren. Letzteres ist von inhaltlichem Belang, ersteres hauptsächlich Gewohnheit. Das zweite Newton’sche Axiom bleibt dasselbe, egal, ob es als $F = ma$, $K = mb$ oder $\dot{p} = F$ denotiert wird. Am Naturgesetz ändert seine Schreibweise nichts.

Standardisierte Regressionskoeffizienten

Auch die standardisierten Regressionskoeffizienten sind uns bereits wiederholt begegnet und wir haben sie auch in allen Ergebnisberichten von Regressionsanalysen brav angegeben. Abgesehen von der einfachen linearen Regression haben wir sie aber nicht weiter erläutert.

Bei den standardisierten Regressionsgewichten handelt es sich schlichtweg um die Steigungsparameter, die man erhält, wenn Prädiktoren und Kriterium allesamt z-transformiert werden. Wenn mit β_{zj} der j-te standardisierte Regressionskoeffizient (des wahren Regressionsmodells) bezeichnet wird, kann der standardisierte Regressionskoeffizient wie folgt interpretiert werden (Bühner et al., 2025): Eine Erhöhung der j-ten UV um eine Standardabweichung geht (bei Konstanthaltung aller anderen UV) mit einer Erhöhung der AV um β_{zj} Standardabweichungen einher. Der standardisierte Regressionskoeffizient gibt also für jede UV an, wie stark diese bei Konstanthaltung aller anderen UV mit der AV zusammenhängt. Ferner gibt er die Richtung des Zusammenhangs an: bei positivem Vorzeichen, wächst die AV mit steigender UV, bei negativem Vorzeichen, fällt die AV mit steigender UV.

Während der standardisierte Regressionskoeffizient (bzw. präziser: sein Schätzwert) in der einfachen linearen Regression dem Pearson Korrelationskoeffizienten zwischen UV und AV entspricht, tut er das in der multiplen linearen Regression nicht. Hier ist wichtig, sich daran zu erinnern, dass die einzelnen Regressionsgewichte bedingte Assoziationen zwischen den UV und der AV beschreiben. Das Regressionsgewicht (egal ob standardisiert oder nicht) für den j-ten Prädiktor ist ein Maß für den linearen Zusammenhang zwischen j-ter UV und AV unter Berücksichtigung aller anderen linearen Zusammenhänge zwischen den anderen UV und der AV. Auf die wesentliche Bedeutung dieser bedingten Assoziation werden wir bei der Diskussion der Kollinearität unten auch wieder zurückkommen.

In SPSS erhalten wir die Schätzwerte für die standardisierten Regressionsgewichte standardmäßig in der Ausgabe für Regressionsanalysen. Für die im vorherigen Abschnitt durchgeführte Regressionsanalyse ist der entsprechende Teil der Ausgabe in Abbildung 10.14 dargestellt. Die Darstellung im Ergebnisbericht kann in den beiden vorhergehenden Kapiteln nachgeschlagen werden.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-3.048	2.415		-1.262	.213
	Negative Selbstbewertung	.817	.077	.816	10.638	<.001
	Abhängigkeitskognitionen	.183	.091	.154	2.009	.050

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung 10.14. Ausgabe der standardisierten Regressionsgewichte in der Spalte „Standardized Coefficients Beta“ in der SPSS-Ausgabe für eine multiple lineare Regression.

Die quadrierte Semipartialkorrelation

Die quadrierte Semipartialkorrelation stellt eine Ergänzung zum standardisierten Regressionsgewicht dar, die besonders wegen ihres inhaltlichen Zusammenhangs mit dem Determinationskoeffizienten von Bedeutung ist. In der Tat ist die quadrierte Semipartialkorrelation für den j-ten Prädiktor nichts anderes als die Differenz des Determinationskoeffizienten des gesamten Regressionsmodells und dem Determinationskoeffizienten des Regressionsmodells, das man erhält, wenn man den j-ten Prädiktor aus dem gesamten Regressionsmodell entfernt (Bühner et al., 2025). Sie ist also gerade jener Anteil an der Varianz im Kriterium, der durch Hinzunahme des j-ten Prädiktors über die anderen Prädiktoren hinaus erklärt werden kann. Gleichzeitig ist sie jener Anteil der Varianz des Kriteriums, der auf den einzigartigen Beitrag des j-ten Prädiktors zurückzuführen ist. Dies ist vereinfacht gesagt, deshalb so, weil alle Anteile der Varianz des Kriteriums, die vom j-ten Prädiktor gemeinsam mit den anderen Prädiktoren erklärt werden, bereits im Regressionsmodell mit allen Prädiktoren außer dem j-ten Prädiktor abgedeckt sind (weil sie ja eben *geteilte* erklärte Varianzanteile sind).

D.h., die quadrierten Semipartialkorrelationen geben uns an, welche Anteile an der insgesamt erklärten Varianz im Kriterium jeweils auf die einzelnen Prädiktoren zurückgehen. Die Summe der einzelnen Beiträge muss dabei aber nicht die gesamte erklärte Varianz (= Determinationskoeffizient) ergeben, da auch ein Anteil an erklärter Varianz übrigbleiben kann, der nur durch eine Kombination der Prädiktoren erklärt werden kann. In seltenen Fällen kann die Summe der Semipartialkorrelationen den Anteil insgesamt erklärter Varianz auch übersteigen. Die mathematischen Gründe dafür sind kompliziert und werden hier nicht weiter erläutert.

In SPSS können die Semipartialkorrelationen im Menü „Statistics…“ unter *Analyze >> Regression >> Linear…* durch Auswahl von „Part and partial correlations“ angefordert werden, siehe Abbildung 10.15. Die Schätzwerte der Semipartialkorrelationen sind dann in der Ausgabe in der Tabelle „Coefficients“ in der letzten Spalte unter der Bezeichnung „Part“ zu finden, siehe Abbildung 10.16.

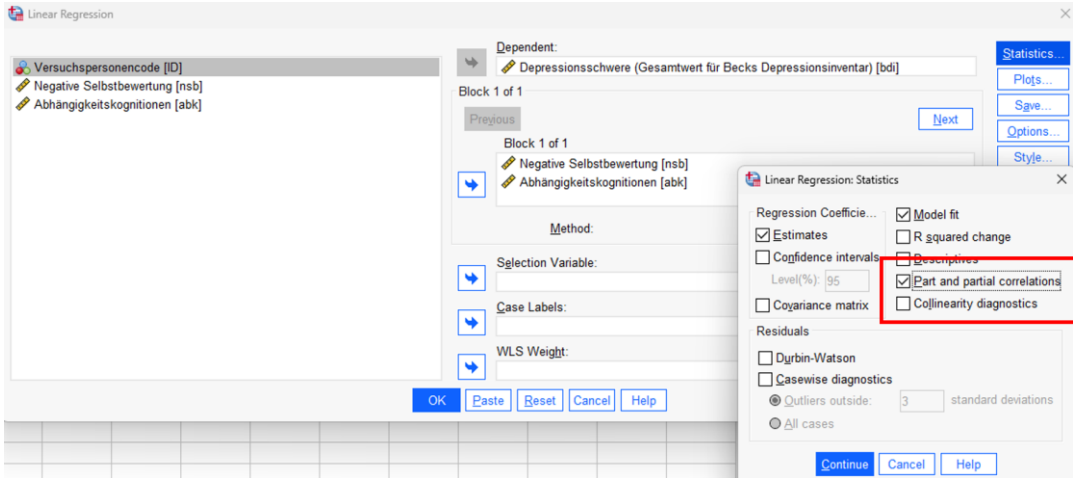


Abbildung 10.15. Anforderung der Semipartialkorrelationen (u.a.) im Rahmen der Regressionsanalyse in SPSS.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Correlations Zero-order	Correlations Partial	Correlations Part
1	(Constant)	-3.048	2.415		-1.262	.213			
	Negative Selbstbewertung	.817	.077	.816	10.638	<.001	.841	.841	.805
	Abhängigkeitskognitionen	.183	.091	.154	2.009	.050	.287	.281	.152

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung 10.16. Schätzwerte für die Semipartialkorrelationen.

Für einen entsprechenden Ergebnisbericht sind diese Werte dann noch zu quadrieren. Damit ergibt sich für die negative Selbstbewertung eine quadrierte Semipartialkorrelation von 64.8% und für die Abhängigkeitskognitionen eine quadrierte Semipartialkorrelation von 2.3%. Unseren Ergebnisbericht aus dem vorherigen Kapitel könnten wir dann noch um die folgenden Zeilen ergänzen: „Während die Stärke der Abhängigkeitskognitionen *eigenständig* lediglich 2.3% der Varianz der Depressionsschwere erklären kann, kann die negative Selbstbewertung *eigenständig* 64.8% der Varianz erklären. Die Erklärung von 73.1% − 2.3% − 64.8% = 6.0% der Varianz kommt daher durch die gemeinsame Wirkung der beiden Prädiktoren zustande.“

Stichprobenumfangsplanung

Einfache lineare Regression

Die Stichprobenumfangsplanung für die einfache lineare Regression wird für das folgende Beispiel illustriert. Der statistische Test soll ein Signifikanzniveau von α = .005 sowie eine Teststärke (power) von 0.8 aufweisen. Als Mindesteffektstärke geben wir den Populationsdeterminationskoeffizienten von ρ^2 = .04 vor. Dieser muss erst noch in eine alternative (hier nicht behandelte) Effektgröße umgerechnet werden:

$$f^2 = \frac{\rho^2}{1-\rho^2} = 0.041\dot{6} \approx 0.0416667.$$

Diese Umrechnung kann allerdings auch direkt in G*Power durchgeführt werden, wie im Folgenden erläutert wird. Dazu wählen wir in G*Power zuerst unter „Test family“ die Option „F tests“ aus. Unter „Statistical test“ wählen wir „Linear multiple regression: Fixed model, R^2 increase“ aus. Unter “Type of power analysis” wählen wir “A priori: Compute required sample size – given α, power, and effect size“ aus. Wenn wir auf die Schaltfläche “Determine =>” links neben dem Feld „Effect size f^2“ klicken, öffnet sich ein weiteres Menü, in dem wir die Umrechnung des Determinationskoeffizienten vornehmen können, indem wir diesen zuerst in das Feld „Partial R^2“ eintragen und dann durch Klick auf „Calculate“ bestätigen. Durch Klick auf „Calculate and transfer to main window“ wird die errechnete Effektstärke in das Hauptfenster transferiert. Im Feld „α err prob“ tragen wir jetzt noch die Zahl 0.005 ein, und anschließend im Feld „Power (1-β err prob)“ die Zahl 0.8. In den beiden verbleibenden Feldern „Number of tested predictors“ und „Total number of predictors“ tragen wir jeweils die Zahl 1 ein. Danach bestätigen wir unsere Eingaben durch Klick auf „Calculate“ und erhalten das Ergebnis, dass wir für diese Wahl an Parametern eine Stichprobe des Umfangs n = 324 benötigen.

Multiple lineare Regression

Stichprobenumfangsplanungen für allgemeine Hypothesen über die Steigungsparameter sind leider kompliziert (Bühner et al., 2025). Wollen wir aber lediglich einzelne Steigungsparameter auf einen Unterschied von Null testen (d.h., $H_0: \beta_j = 0, H_1: \beta_j \neq 0$), dann können wir eine Stichprobenplanung durch Rückgriff auf die entsprechende Semipartialkorrelation ρ_j wie folgt durchführen.

In diesem Fall ermitteln wir die benötigte Effektstärke

$$f_j^2 = \frac{\rho_j^2}{1 - \rho^2}$$

und setzen diesen Wert dann in G*Power für f^2 ein. Falls wir z.B. $\rho_j^2 = 0.02$ und $\rho^2 = 0.1$ vorgeben, erhalten wir $f_j^2 = 0.0\dot{2} \approx 0.0222222$. Nehmen wir zudem an, dass die Anzahl unserer Prädiktoren 2 sei und alle übrigen Parameter dieselben Werte wie im vorhergehenden Fall haben sollen.

In diesem Fall können wir in G*Power dieselben Einstellungen wie vorhin vornehmen, nur dass wir jetzt im Feld „Effect size f²“ direkt den Wert 0.0222222 eintragen und im Feld „Total number of predictors“ den Wert 2. Nach Bestätigung unserer Eingaben durch Klick auf „Calculate“ erhalten wir das Ergebnis, dass wir für diese Wahl an Parametern eine Stichprobe von *n* = 604 Personen benötigen.

Kollinearität

Zu guter Letzt werden wir uns in diesem Kapitel noch mit dem Thema der Kollinearität befassen, die sich stark auf die Standardfehler und damit die Hypothesentests für einzelne Steigungsparameter auswirkt. In diesem Zusammenhang werden wir auf die Frage zurückkommen, wie diese Hypothesentests denn im Einzelnen interpretiert werden können. Als hilfreich wird sich dabei die Frage danach erweisen, was durch die Hinzunahme eines bestimmten Prädiktors über das Kriterium herausgefunden werden kann, wenn die Ausprägung eines anderen Prädiktors bereits bekannt ist. Dies wird uns schließlich auch zur Frage führen, wie entschieden werden kann, ob ein zusätzlicher Prädiktor mit in ein Regressionsmodell aufgenommen werden sollte oder nicht. Wir werden sehen, dass dafür theoretische Überlegungen über kausale Zusammenhänge zwischen Variablen eine wesentliche Rolle spielen und dies grundsätzlich eine inhaltliche, konzeptuelle und keine statistische Frage ist. Bei der Klärung dieser konzeptuellen Fragen können sich allerdings gerichtete, azyklische Graphen (Engl.: Directed Acyclic Graphs, kurz DAGs) als nützlich erweisen, die wir daher auch kennenlernen werden.

Von Kollinearität spricht man, wenn einer oder mehrere der Prädiktoren untereinander stark zusammenhängen. Da Kollinearität einen starken Einfluss auf die Standardfehler der Schätzfunktionen der Regressionskoeffizienten und damit auf die inferenzstatistischen Ergebnisse von Hypothesentests für diese Parameter hat (Bühner et al., 2025) ist die Berücksichtigung von Kollinearität vor allem für

die Schätzung der Regressionsparameter und deren Interpretation von Bedeutung. Punkt- und Intervallschätzungen des Determinationskoeffizienten, der Omnibustest im Rahmen der multiplen linearen Regression, und Vorhersagen auf Basis des gesamten Regressionsmodells sind von der Kollinearität nicht beeinträchtigt (Bühner et al., 2025).

Illustration des Einflusses von Kollinearität auf die Ergebnisse einer multiplen Regressionsanalyse

Um den Einfluss der Kollinearität auf die Ergebnisse einer multiplen linearen Regressionsanalyse zu illustrieren, wird das folgende Beispiel verwendet, das auf McElreath (2020) zurückgeht. Angenommen, wir wollen die (mittlere) Körpergröße mithilfe der Beinlänge von Personen vorhersagen. Da das Verhältnis von Beinlänge zu Körpergröße im Mittel in etwa 0.4-0.5 beträgt (siehe z.B. Bammer, 1998), sollte eine solche Prognose auf Basis eines Regressionsmodells im Mittel ja recht gut funktionieren.

Im Datensatz „Kap10daten.sav“, den Sie im elektronischen Ergänzungsmaterial zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können, sind die linken und rechten Beinlängen sowie die Körpergrößen von 333 weiblichen fiktiven Personen gegeben. Wenn wir nun eine einfache lineare Regressionsanalyse mit der linken Beinlänge als Prädiktor und der Körpergröße als Kriterium durchführen, erhalten wir das in Abbildung 10.17 gezeigte Ergebnis.

Dieses sieht auch ganz plausibel aus. Die linke Beinlänge erklärt 59% der Körpergröße, es bleibt ein Standardschätzfehler von etwa 4 cm, die Körpergröße entspricht ziemlich genau 2 mal der linken Beinlänge, $b = 1.97$, und die Beinlänge ist ein signifikanter (mit $\alpha = .005$) Prädiktor der Körpergröße, d.h. wir würden uns auch in der Population einen positiven Zusammenhang zwischen der linken Beinlänge und der Körpergröße erwarten. Wenn wir anstelle des linken Beins das rechte Bein verwenden, bekommen wir ausgesprochen ähnliche Ergebnisse (hier nicht gezeigt). Das ergibt Sinn, da ja beide Beinlängen sehr ähnlich (wenn auch nicht identisch sind). Ein Zusammenhang zwischen Beinlänge und Körpergröße erscheint also auch nach unserer statistischen Analyse sehr plausibel.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.766[a]	.587	.586	4.022036

a. Predictors: (Constant), Beinlänge links (in cm)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7622.101	1	7622.101	471.176	<.001[b]
	Residual	5354.512	331	16.177		
	Total	12976.612	332			

a. Dependent Variable: Körpergröße (in cm)

b. Predictors: (Constant), Beinlänge links (in cm)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.639	7.706		-.213	.832
	Beinlänge links (in cm)	1.969	.091	.766	21.707	<.001

a. Dependent Variable: Körpergröße (in cm)

Abbildung 10.17. Vorhersage der Körpergröße durch die Länge des linken Beins. Sieht ja ganz plausibel aus.

Aber was geschieht, wenn wir nun beide Beinlängen als Prädiktoren verwenden? Die entsprechende Ausgabe ist in Abbildung 10.18 gezeigt. Wir sehen, es werden immer noch ca. 59% an Varianz aufgeklärt, die Schätzung der Körpergröße auf Basis der Beinlängen ist auch immer noch ca. 4 cm genau. Das Gesamtmodell ist immer noch signifikant, d.h. der Anteil erklärter Varianz unterscheidet sich signifikant von Null (was jetzt keine bahnbrechende Information ist, aber auch nicht nichts). Aber: Keiner der beiden Prädiktoren ist signifikant! Und die Schätzwerte sind für keinen von beiden auch nur in der Nähe von 2. Und die Standardfehler sind so groß wie die Schätzwerte selbst! Hilfe!

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.767[a]	.588	.586	4.023870

a. Predictors: (Constant), Beinlänge links (in cm), Beinlänge rechts (in cm)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7633.407	2	3816.703	235.722	<.001[b]
	Residual	5343.205	330	16.192		
	Total	12976.612	332			

a. Dependent Variable: Körpergröße (in cm)

b. Predictors: (Constant), Beinlänge links (in cm), Beinlänge rechts (in cm)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.261	7.723		-.163	.870
	Beinlänge rechts (in cm)	-3.525	4.218	-1.370	-.836	.404
	Beinlänge links (in cm)	5.489	4.214	2.136	1.303	.194

a. Dependent Variable: Körpergröße (in cm)

Abbildung 10.18. Vorhersage der Körpergröße durch die Länge beider Beine. Was zum…?!

Um sich dieses auf den ersten (und vielleicht auch zweiten, dritten oder vierten) Blick seltsam anmutende Ergebnis zu erklären, ist es gut, sich noch einmal in Erinnerung zu rufen, was uns die Schätzung und Testung einzelner Prädiktoren an Information liefern. Im Kontext dieses Beispiels geben sie darüber Auskunft, was uns die Länge des jeweils anderen Beins zusätzlich für die Schätzung der Körpergröße bringt, wenn wir die Länge eines Beins bereits kennen. D.h., wenn wir bereits wissen, dass das linke Bein 85 cm lang ist, was nützt es uns dann für die Schätzung der Körpergröße, wenn wir zusätzlich erfahren, dass das rechte Bein ebenfalls (nicht exakt, aber nahezu) 85 cm lang ist? Die Antwort ist: nicht sonderlich viel (wenn überhaupt irgendwas), denn die Körpergröße kann ja bereits durch die linke Beinlänge sehr genau geschätzt werden, die rechte Beinlänge bietet darüber hinaus kaum noch Zusatzinformation. Der Unterschied zwischen den Beinlängen hängt sogar kaum (im fiktiven Beispiel gar nicht) systematisch mit der Körpergröße zusammen. Das was den Zusammenhang zwischen Körpergröße und Beinlänge ausmacht, ist gerade das, was die beiden Beinlängen miteinander gemeinsam haben, und nicht die mehr oder weniger zufällige Schwankung im exakten Wert.

Das heißt, das Ergebnis der multiplen Regressionsanalyse ist keineswegs falsch. Es sagt genau das aus: kenne ich die Länge des linken Beins, dann kann mir die Zusatzinformation über die Länge des rechten Beins nicht mehr viel zur Schätzung der Körpergröße beitragen. Selbst darüber, ob sie mir überhaupt etwas beitragen kann – etwa ob die rechte Beinlänge die Körpergröße positiv oder negativ beeinflusst – bin ich mir sehr unsicher. Das ist die Aussage des p-Werts: Selbst wenn zwischen Körpergröße und rechter Beinlänge unter Berücksichtigung des linearen Zusammenhangs zwischen Körpergröße und linker Beinlänge überhaupt kein Zusammenhang besteht, würde in 40.4% aller Fälle (bei gleicher Varianz) ein gleich extremer oder sogar noch extremerer Regressionskoeffizient resultieren. Also in diesem Fall würde ich mich auf das negative Vorzeichen (b_{rechts} = -3.53) nicht verlassen. Das heißt, wir können uns in diesem Fall auf die Schätzungen der einzelnen Parameter (und noch nicht einmal darauf, ob sie überhaupt positiv oder negativ sind) überhaupt nicht verlassen. Das ist aber nach all dem Besprochenen völlig klar, weil sie ja jeweils bedingte Assoziationen sind.

Das erklärt auch, warum beide Parameter zusammen trotz individueller Unverlässlichkeit immer noch eine gute Schätzung der Körpergröße erlauben. Unter der Voraussetzung, dass der Zusammenhang zwischen linker Beinlänge und Körpergröße durch b_{links} = 5.49 beschrieben wird, ist der Zusammenhang zwischen rechter Beinlänge und Körpergröße durch b_{rechts} = -3.53 gegeben. Unter der hier gültigen Voraussetzung nahezu gleicher Beinlängen ergibt das, dass der Zusammenhang zwischen Körpergröße und durchschnittlicher Beinlänge ca. gleich 5.49 – 3.53 = 1.96 ist, was in der Tat sehr genau unserem Wert aus den einfachen linearen Regressionsmodellen entspricht.

Konzeptuell wäre es auch eine sinnvolle Wahl zur Vorhersage der Körpergröße aufgrund der Beinlänge weder die linke noch die rechte, sondern den Mittelwert aus beiden zu verwenden. Die zufälligen Schwankungen, die ja überhaupt erst für einen Unterschied zwischen den beiden sorgen, werden dadurch einerseits zum Teil reduziert und andererseits wird exakt das in das Regressionsmodell aufgenommen, was von Bedeutung ist, nämlich der Anteil an Beinlänge, der dem linken und rechten Bein gemeinsam ist. Wenn wir diese einfache lineare Regressionsanalyse durchführen, bekommen wir erwartungsgemäß auch nahezu wieder dasselbe Ergebnis wie oben, siehe Abbildung 10.19.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.766[a]	.587	.586	4.024740

a. Predictors: (Constant), Mittlere Beinlänge

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7614.899	1	7614.899	470.098	<.001[b]
	Residual	5361.714	331	16.199		
	Total	12976.612	332			

a. Dependent Variable: Körpergröße (in cm)

b. Predictors: (Constant), Mittlere Beinlänge

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.655	7.716		-.215	.830
	Mittlere Beinlänge	1.969	.091	.766	21.682	<.001

a. Dependent Variable: Körpergröße (in cm)

Abbildung 10.19. Vorhersage der Körpergröße durch die mittlere Beinlänge.

Um es ganz deutlich auszusprechen: der bedeutendste Fehler, den es bei der Interpretation des Ergebnisses der multiplen Regressionsanalyse zu vermeiden gilt, ist der völlig falsche Schluss, dass zwischen den beiden Prädiktoren und dem Kriterium kein Zusammenhang besteht. Dieser Fehlschluss ist aber leicht zu vermeiden, wenn erinnert wird, dass es sich bei der Schätzung und Testung von Zusammenhängen zwischen einzelnen Prädiktoren und Kriterium jeweils um Schätzung und Testung von bedingten Assoziationen geht.

Kollinearitätsdiagnostik in SPSS

Im soeben erläuterten Beispiel bestand eine große Korrelation zwischen den beiden Prädiktoren im multiplen Regressionsmodell. Im Allgemeinen muss aber zum Vorliegen von Kollinearität keine große Korrelation für ein Paar von Prädiktoren bestehen. Kollinearität liegt auch vor, wenn eine Linearkombination aus zwei oder mehr Prädiktoren einen anderen Prädiktor sehr gut beschreibt. Das heißt aber nichts anderes als dass ein Prädiktor sehr gut durch die jeweils anderen (oder einige der anderen) beschrieben bzw. vorhergesagt werden kann.

Ob dies der Fall ist, kann also dadurch überprüft werden, dass von den m Prädiktoren eines multiplen Regressionsmodells jeweils ein Prädiktor entfernt wird und die übrigen Prädiktoren dann als Prädiktoren in einem Regressionsmodell verwendet werden, in dem der entfernte Prädiktor als Kriterium fungiert. Ergibt sich für eines dieser insgesamt m Regressionsmodelle ein sehr großer Determinationskoeffizient, so ist dies ein Zeichen von Kollinearität.

Diese Regressionsmodelle müssen in SPSS allerdings nicht einzeln durchgeführt werden. Stattdessen kann die Option „Collinearity diagnostics“ im Menü zur Anforderung des multiplen Regressionsmodells unter „Statistics“ ausgewählt werden, siehe Abbildung 10.20. Im Output werden der Tabelle „Coefficients“ dann zwei Spalten mit sog. „Collinearity Statistics“ hinzugefügt, siehe Abbildung 10.21.

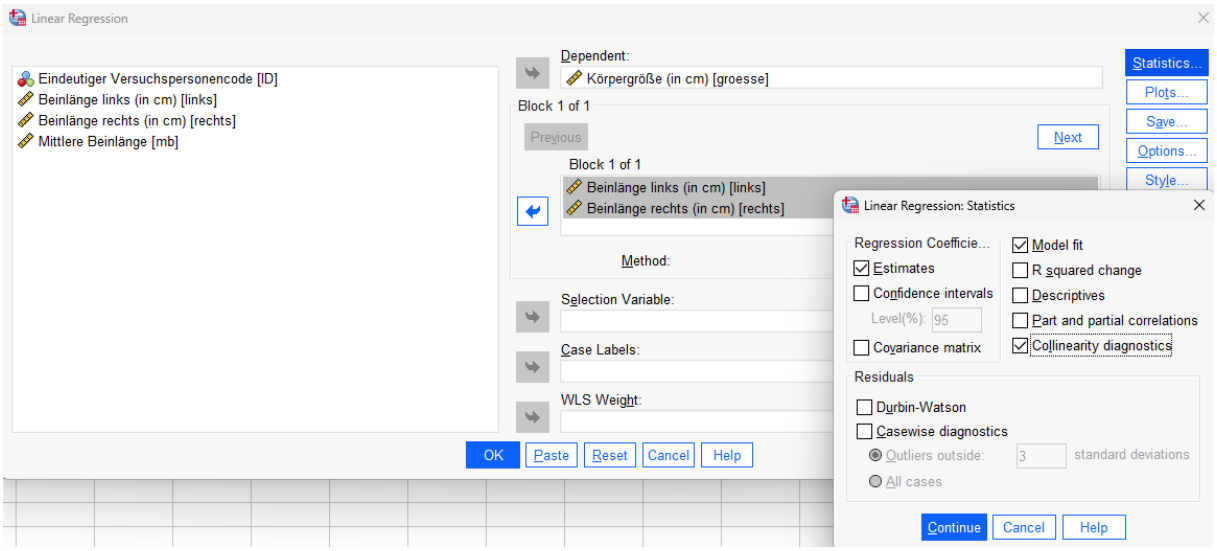


Abbildung 10.20. Anforderung einer multiplen Regressionsanalyse inklusive einer Kollinearitätsdiagnostik in SPSS.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-1.261	7.723		-.163	.870		
	Beinlänge links (in cm)	5.489	4.214	2.136	1.303	.194	.000	2155.419
	Beinlänge rechts (in cm)	-3.525	4.218	-1.370	-.836	.404	.000	2155.419

a. Dependent Variable: Körpergröße (in cm)

Abbildung 10.21. Teil der Ausgabe für ein multiples Regressionsmodell inklusive Kollinearitätsdiagnostik (letzte beide Spalten ganz rechts).

Bei diesen Kennwerten zur Beurteilung darüber, ob Kollinearität vorliegt, handelt es sich um die sog. Toleranz und den Varianzinflationsfaktor. Bei der Toleranz handelt es sich schlichtweg um $1 - r_j^2$, wobei r_j^2 der Determinationskoeffizient eines Regressionsmodells ist, in dem der j-te Prädiktor als Kriterium und alle anderen Prädiktoren als Prädiktoren fungieren. Kann ein Prädiktor also sehr gut durch die übrigen vorhergesagt werden, ist der Determinationskoeffizient nahe 1 und die Toleranz entsprechend nahe Null. Der Varianzinflationsfaktor (VIF) ist schlichtweg der Kehrwert der Toleranz, d.h. VIF = 1/Toleranz. Ist die Toleranz sehr klein, ist der VIF sehr groß und umgekehrt. Üblicherweise betrachtet man Toleranzen < 0.25 bzw. VIF > 4 als Indizien für Kollinearität, und Toleranzen < 0.1 bzw. VIF > 9 als deutliche Anzeichen von Kollinearität (Bühner & Ziegler, 2017).

Gerichtete azyklische Graphen (DAGs)

Das Beispiel im vorhergehenden Abschnitt illustriert, dass eine unüberlegte Hinzunahme von Prädiktoren in Regressionsmodelle die Interpretation erschweren und Fehlschlüsse erleichtern kann. Allgemein gilt, dass ein Hinzunehmen oder Weglassen von Prädiktoren in erster Linie eine konzeptuelle bzw. inhaltliche und keine statistische Frage ist. Sowohl das Hinzunehmen als auch das Weglassen von Prädiktoren kann einerseits wirklich vorliegende Zusammenhänge verschleiern oder verzerren und andererseits Scheinzusammenhänge überhaupt erst erzeugen. Man muss sich also gut überlegen, weshalb und wozu man welche Prädiktoren in einem multiplen Regressionsmodell berücksichtigen möchte.

Ein Werkzeug, dass diese grundsätzlich alles andere als triviale Entscheidungen erleichtern kann, sind sog. gerichtete azyklische Graphen (Engl.: directed acyclic graphs, kurz: DAGs). DAGs dienen der grafischen Veranschaulichung kausaler Zusammenhänge. Kausale Wirkrichtungen zwischen Variablen werden dabei durch Pfeile dargestellt, Variablen als beschriftete „Boxen“ oder Felder.

Aus vier grundlegenden DAGs (mit den Bezeichnungen „Fork“, „Collider“, „Pipe“, „Descendant“) lassen sich alle möglichen, komplexen kausalen Zusammenhänge zwischen beliebigen Variablen konstruieren (McElreath, 2020). Die systematische Analyse dieser komplexen DAGs erlaubt dann abzuleiten, welche Variablen berücksichtigt werden müssen, um eine bestimmte Fragestellung zu erhellen. Ob ein DAG allerdings zutreffend ist, ist wiederum eine konzeptuelle Frage; repräsentiert der

DAG ein reales Netzwerk aus Kausalzusammenhängen einfach nicht oder falsch, so sind auch die daraus abgeleiteten Prädiktoren unter Umständen für die Erhellung der Fragestellung irreführend (Bühner et al., 2025).

Im Folgenden erläutern wir die vier grundlegenden DAGs anhand konkreter Beispiele und entsprechender Regressionsanalysen in SPSS.

Die Gabel (Fork; auch: Confounder)

Zur Illustration einer sog. konfundierenden Variablen (Engl.: confounder) betrachten wir den in Abbildung 10.22 dargestellten DAG für das Beispiel bei Bühner et al. (2025). Einen dazu passenden illustrativen Datensatz finden Sie in der Datendatei „fork.sav“, die Sie wiederum im elektronischen Ergänzungsmaterial zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können. Für dieses Beispiel stellen wir uns vor, ein Forscher interessiere sich für den Zusammenhang zwischen der Intensität von Symptomen eines Sonnenbrands und Eiskonsum. Dazu erhebt er beide Variablen an 365 Tagen an einem grundsätzlichen sehr sonnigen Ort und führt im Anschluss eine einfache lineare Regressionsanalyse durch.

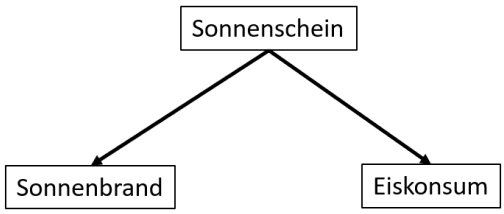


Abbildung 10.22. DAG für eine konfundierende Variable (Engl.: confounding variable); auch als „Gabel“ (Engl.: Fork) bezeichnet.

Die Ergebnisse sind in Abbildung 10.23 dargestellt. Der Forscher stellt in der Tat einen (mit α = .005) signifikanten Zusammenhang zwischen Sonnenbrandsymptomen und Eiskonsum fest, b = 0.32 (stand. β = 0.32), $t(363) = 6.36$, $p < .001$. Der Forscher freut sich, publiziert sein Ergebnis in einer namhaften Zeitschrift unter dem Titel „Sunburn causes ice cream consumption“ und wird viele Mal zitiert.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Sonnenbrand[b]	.	Enter

a. Dependent Variable: Eiskonsum

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.316[a]	.100	.098	11.746

a. Predictors: (Constant), Sonnenbrand

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5573.743	1	5573.743	40.398	<.001[b]
	Residual	50082.931	363	137.970		
	Total	55656.674	364			

a. Dependent Variable: Eiskonsum

b. Predictors: (Constant), Sonnenbrand

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	24.266	2.015		12.044	<.001
	Sonnenbrand	.316	.050	.316	6.356	<.001

a. Dependent Variable: Eiskonsum

Abbildung 10.23. Ergebnis einer einfachen linearen Regression ohne Berücksichtigung einer konfundierenden Variablen.

Leider wird er aber hauptsächlich als Gegenbeispiel für gute wissenschaftliche Forschung zitiert. Denn kurz nach der Veröffentlichung seiner Studie hat eine Kollegin sich die erhobenen Daten und durchgeführten statistischen Analysen (die der Forscher dankenswerter Weise auf einem Open Science Repository zur Verfügung gestellt hat) noch einmal genauer angeschaut und festgestellt, dass der Forscher nicht die ebenfalls erhobene Variable „Sonnenschein“ in seiner Regressionsanalyse berücksichtigt hat. Mithilfe dieser Variablen wurde die Intensität des Sonnenscheins an jedem der 365 Tage erhoben. Im Gegensatz zu unserem berühmt-berüchtigten Forscher argumentiert die Forscherin entsprechend des DAGs in Abbildung 10.22, dass die Intensität des Sonnenscheins sich sowohl auf die Intensität von Sonnenbränden als auch den Eiskonsum auswirkt. Zwischen den letzteren beiden Variablen bestehe gar kein direkter Zusammenhang, ein Zusammenhang komme nur scheinbar

zustande, wenn die konfundierende Variable „Sonnenschein“ nicht berücksichtigt werde. Um die Variable zu berücksichtigen, führt die Forscherin eine multiple lineare Regressionsanalyse mit den beiden Prädiktoren Sonnenschein und Sonnenbrand sowie dem Kriterium Eiskonsum durch. Die entsprechenden Ergebnisse sind in Abbildung 10.24 gezeigt.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Sonnenschein, Sonnenbrand[b]	.	Enter

a. Dependent Variable: Eiskonsum

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.527[a]	.278	.274	10.538

a. Predictors: (Constant), Sonnenschein, Sonnenbrand

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15458.606	2	7729.303	69.606	<.001[b]
	Residual	40198.068	362	111.044		
	Total	55656.674	364			

a. Dependent Variable: Eiskonsum

b. Predictors: (Constant), Sonnenschein, Sonnenbrand

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1.871	2.983		.627	.531
	Sonnenbrand	.006	.055	.006	.110	.912
	Sonnenschein	.456	.048	.523	9.435	<.001

a. Dependent Variable: Eiskonsum

Abbildung 10.24. Ergebnisse der multiplen linearen Regression unter Berücksichtigung der konfundierenden Variablen.

Wir sehen, dass nun in der Tat zwischen Sonnenbrand und Eiskonsum nur noch ein verschwindender Zusammenhang besteht (b = 0.006), der auch nicht mehr signifikant ist (p = .912). Zwischen Sonnenschein und Eiskonsum besteht ein deutlicher (und auch signifikanter) Zusammenhang, b = 0.46 (stand. β = .52), $t(362) = 9.44$, $p < .001$. Ein einfaches lineares Regressionsmodell zeigt schließlich, dass auch zwischen Sonnenschein und Sonnenbrand ein deutlicher, signifikanter Zusammenhang besteht, b = 0.52 (stand. β = .59), $t(363) = 14.03$, $p < .001$, siehe Abbildung 10.25. Die

Forscherin argumentiert, dass es auf Basis dieser Resultate plausibler erscheint, dass der Sonnenschein sowohl die Intensität von Sonnenbränden als auch den Eiskonsum erhöht, und der Zusammenhang zwischen den letzten beiden Variablen gar kein ursächlicher ist. Der höhere Eiskonsum geht nicht auf die Sonnenbrände zurück, sondern bloß auf die sonnigeren Tage. Leider wird der Artikel der Forscherin nicht so häufig zitiert wie der des Forschers, da sie ihre Arbeit nur in einem spezialisierteren Methodenjournal veröffentlichen konnte. Dies bringt uns schon zu unserem nächsten DAG.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Sonnenschein[b]	.	Enter

a. Dependent Variable: Sonnenbrand

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.593[a]	.352	.350	9.989

a. Predictors: (Constant), Sonnenschein

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	19642.278	1	19642.278	196.863	<.001[b]
	Residual	36218.911	363	99.777		
	Total	55861.189	364			

a. Dependent Variable: Sonnenbrand

b. Predictors: (Constant), Sonnenschein

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-.389	2.828		-.138	.891
	Sonnenschein	.518	.037	.593	14.031	<.001

a. Dependent Variable: Sonnenbrand

Abbildung 10.25. Einfache lineare Regression mit Kriterium Sonnenbrand und Prädiktor Sonnenschein.

Die Kollision (Collider)

Zur Illustration der Kollision bzw. eines sog. Colliders wird auf ein Beispiel bei McElreath (2020) zurückgegriffen, das der Frage nachgeht, weshalb besonders bahnbrechende oder innovative Forschung eigentlich so häufig fragwürdig erscheint, was ihre wissenschaftliche Qualität angeht. Und weshalb umgekehrt die langweiligsten Themen offenbar mit den rigorosesten Methoden untersucht werden.

Um diesen scheinbaren Widerspruch zu klären, verwenden wir den Datensatz „collider.sav“, den Sie wiederum im elektronischen Ergänzungsmaterial zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können. In diesem Datensatz finden wir Beurteilungen von 1000 (fiktiven) wissenschaftlichen Arbeiten. Die Beurteiler:innen hatten dabei den Auftrag die Arbeiten nach jeweils zwei Hauptkriterien zu beurteilen: „Innovation“ (Variable *I*) und „Wissenschaftliche Qualität“ (Variable *W*). Neben diesen Hauptkriterien spielten noch eine Reihe fachspezifischer Kriterien eine Rolle, die allerdings nur ein geringes Gewicht im Beurteilungsprozess erhalten sollten. Aus allen Kriterien sollte anschließend ein Gesamtindex für die „Publikationswürdigkeit“ (Variable *P*) der jeweiligen Arbeit gebildet werden, in den die beiden Hauptkriterien additiv mit gleichem Gewicht eingingen.

Was die Beurteiler:innen jedoch nicht wussten, war, dass bei allen beurteilten Arbeiten insgesamt kein systematischer Zusammenhang zwischen Innovation und wissenschaftlicher Qualität bestand. Das zeigt sich auch in guter Übereinstimmung mit einem einfachen linearen Regressionsmodell mit dem Kriterium Wissenschaftlichkeit (kurz für wissenschaftliche Qualität) und dem Prädiktor Innovation, siehe Abbildung 10.26.

Betrachtet man nun aber nur wissenschaftliche Arbeiten mit derselben Publikationswürdigkeit, indem man letztere als weiteren Prädiktor in das Regressionsmodell hinzufügt (da dann die bedingte Assoziation zwischen Innovation und Wissenschaftlichkeit bei *konstanter* Publikationswürdigkeit berechnet wird), ergibt sich ein anderes Bild, siehe Abbildung 10.27. Für wissenschaftliche Arbeiten vergleichbarer Publikationswürdigkeit besteht in der Tat ein negativer Zusammenhang zwischen Innovation und Wissenschaftlichkeit.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Innovation[b]	.	Enter

a. Dependent Variable: Wissenschaftlichkeit

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.014[a]	.000	-.001	15.706

a. Predictors: (Constant), Innovation

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	51.383	1	51.383	.208	.648[b]
	Residual	246195.941	998	246.689		
	Total	246247.324	999			

a. Dependent Variable: Wissenschaftlichkeit

b. Predictors: (Constant), Innovation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	50.786	1.737		29.242	<.001
	Innovation	-.015	.033	-.014	-.456	.648

a. Dependent Variable: Wissenschaftlichkeit

Abbildung 10.26. Grundsätzlich scheint kein Zusammenhang zwischen wissenschaftlicher Qualität und Innovation wissenschaftlicher Arbeiten zu bestehen.

Weshalb ist das so? Der Grund liegt im Beurteilungsverfahren und insbesondere der Addition der Kriterien Innovation und Wissenschaftlichkeit. Um eine gewisse Publikationswürdigkeit zu erreichen, kann eine Arbeit entweder äußerst innovativ und dafür etwas weniger wissenschaftlich sein oder aber auch äußerst wissenschaftlich und dafür etwas weniger innovativ.

Der Zusammenhang zwischen den beiden Variablen kommt in diesem Fall nur scheinbar zustande, wenn ausschließlich Arbeiten einer gewissen Publikationswürdigkeit berücksichtigt werden; also z.B. Arbeiten, die alle in wissenschaftlichen „Top"-Journalen publiziert wurden oder in besonders angesehenen Journalen in einem gewissen Fachbereich. In diese Journale schaffen es nur die besten Artikel des entsprechenden Fachbereichs. Sind sie nicht innovativ genug, kommen sie nicht in Betracht. Sind sie nicht wissenschaftlich genug, kommen sie nicht in Betracht. Sind sie beides über die Maßen, versuchen die Autor:innen sie meist in noch höherrangigen Fächer-übergreifenden Journalen zu

publizieren. Übrig bleibt eine Balance zwischen Innovation und Wissenschaftlichkeit für einen gewissen Grad an Publikationswürdigkeit, die von dem negativen Vorzeichen des Regressionskoeffizienten für den Prädiktor Innovation reflektiert wird.

Ist man also grundsätzlich an dem Zusammenhang zwischen Innovation und Wissenschaftlichkeit von wissenschaftlichen Artikeln interessiert, dann sollte man in diesem Beispiel gerade nicht für die Publikationswürdigkeit „kontrollieren". Dadurch entsteht erst der soeben diskutierte Scheinzusammenhang. Allgemeiner sollte man für keine Variable „kontrollieren", d.h. sie im multiplen Regressionsmodell als Prädiktor hinzunehmen, die sowohl durch einen Prädiktor als auch das Kriterium verursacht bzw. beeinflusst wird (hier die Publikationswürdigkeit, die sich hauptsächlich aus Innovation und Wissenschaftlichkeit ergibt). Bei dieser Variablen handelt es sich um einen sog. Collider.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Publikationswürdigkeit, Innovation[b]	.	Enter

a. Dependent Variable: Wissenschaftlichkeit

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.848[a]	.718	.718	8.341

a. Predictors: (Constant), Publikationswürdigkeit, Innovation

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	176876.717	2	88438.359	1271.043	<.001[b]
	Residual	69370.607	997	69.579		
	Total	246247.324	999			

a. Dependent Variable: Wissenschaftlichkeit

b. Predictors: (Constant), Publikationswürdigkeit, Innovation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13.640	1.181		11.554	<.001
	Innovation	-.715	.022	-.681	-31.851	<.001
	Publikationswürdigkeit	.719	.014	1.078	50.412	<.001

a. Dependent Variable: Wissenschaftlichkeit

Abbildung 10.27. Wird für Publikationswürdigkeit „kontrolliert" stellt sich plötzlich ein negativer Zusammenhang ein.

Der Kausalzusammenhang, der zwischen den Variablen X und Y und dem Collider Z vorliegt, ist durch den DAG in Abbildung 10.28 dargestellt. Beide Variablen X und Y wirken sich kausal auf die Variable Z aus. Wird in diesem Fall die Variable Z als Prädiktor mit in ein Regressionsmodell aufgenommen, so ergibt sich scheinbar ein Zusammenhang zwischen X und Y, selbst wenn zwischen diesen beiden Variablen kein (direkter oder indirekter) ursächlicher Zusammenhang besteht.

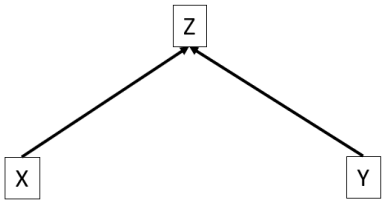


Abbildung 10.28. DAG für den Fall eines Colliders (hier die Variable Z).

Der Nachkomme (Descendant)

Der sog. Nachkomme (Eng.: Descendant) ist eine Variable, die von einer anderen Variablen beeinflusst wird. Wird der Nachkomme als Prädiktor in ein Regressionsmodell mitaufgenommen, so hat dies teilweise dieselben Auswirkungen wie die Aufnahme der Variablen, von welcher der Nachkomme abhängt. Für die Situation, die in Abbildung 10.29 dargestellt ist, hat die Aufnahme der Variablen D als Prädiktor in etwa dieselben Auswirkungen wie die Aufnahme des Colliders Z in ein entsprechendes Regressionsmodell. Der Grund liegt darin, dass die Variable D mit Z zusammenhängt und deshalb deren Wirkung zum Teil (je nach Stärke des Zusammenhangs) vermittelt. Für das Beispiel mit der Publikationswürdigkeit aus dem vorherigen Abschnitt kann man sich etwa vorstellen, dass vom Kriterium der Publikationswürdigkeit ein weiteres Kriterium, z.B. das der Förderungswürdigkeit, abhängt. Würde man nun für letztere „kontrollieren“, würde man denselben Scheineffekt erhalten.

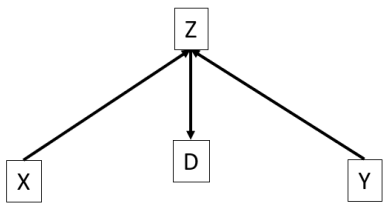


Abbildung 10.29. DAG für den Fall eines Descendants (hier die Variable D) von einem Collider.

Die genaue Wirkung von Descendants hängt allerdings von der Art der Variable ab, von der sie abhängen. Handelt es sich dabei z.B. um einen Confounder, dann vermittelt der Descendant die Wirkung einer konfundierenden Variablen. Descendants sind gerade in den Sozialwissenschaften und der Psychologie sehr häufig, weil hier selten Variablen direkt gemessen werden können, sondern stattdessen Konstrukte erfasst werden, die näherungsweise mit den eigentlich interessierenden latenten Variablen zusammenhängen. Interessiert man sich beispielsweise für den IQ von Personen, wird etwa erfasst, welchen Anteil einer Reihe von Aufgaben Personen lösen können. In allen statistischen Modellen wird dann mit Zahlen gearbeitet, die sich auf diesen Anteil beziehen, der ja nicht ident ist mit dem dahinter liegenden Konstrukt (IQ), sondern (im Idealfall eindeutig) von ihm abhängt. Der Anteil gelöster Aufgaben ist also ein Descendant des eigentlich interessierenden Merkmals IQ.

Die Mediation (Pipe)

Im Falle einer Mediation wird die Wirkung einer Variablen auf eine andere Variable über eine Drittvariable vermittelt. Ein Beispiel ist durch den DAG in Abbildung 10.30 dargestellt. Zum Beispiel führt das Interesse für Aufgaben oder Tätigkeiten einer bestimmten Art zu mehr Übung in diesen Tätigkeiten und damit zu einem höheren Verständnis eines bestimmten Themengebiets (z.B. Statistik). Die Vermittlung der Wirkung von Interesse auf Verständnis muss aber auch nicht komplett über die Variable Übung vermittelt sein, deshalb bleibt in Abbildung 10.30 auch ein direkter Pfeil von Interesse zu Verständnis bestehen.

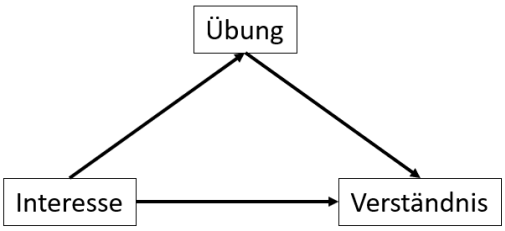


Abbildung 10.30. Beispiel einer Mediation.

Ein Beispieldatensatz für einen Fall einer totalen Mediation ist in der Datendatei „mediation.sav“ gegeben, die Sie wiederum im elektronischen Ergänzungsmaterial zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können. Wird eine einfache lineare Regression ohne die Mediatorvariable (Variable *m*) durchgeführt, resultiert die in Abbildung 10.31

gezeigte Ausgabe. Die einzige Prädiktorvariable (Variable *x*) erklärt einen signifikanten Anteil der Varianz im Kriterium (Variable *y*).

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Prädiktor[b]	.	Enter

a. Dependent Variable: Kriterium

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.566[a]	.320	.319	21.654

a. Predictors: (Constant), Prädiktor

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	220139.904	1	220139.904	469.497	<.001[b]
	Residual	467947.020	998	468.885		
	Total	688086.924	999			

a. Dependent Variable: Kriterium

b. Predictors: (Constant), Prädiktor

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	101.129	2.367		42.723	<.001
	Prädiktor	.990	.046	.566	21.668	<.001

a. Dependent Variable: Kriterium

Abbildung 10.31. Ausgabe für eine einfache lineare Regressionsanalyse ohne die Mediatorvariable als Prädiktor im Falle einer totalen Mediation.

Die Ausgabe für ein multiples Regressionsmodell, in dem auch die Mediatorvariable als Prädiktor hinzugefügt wurde, ist in Abbildung 10.32 gezeigt. Der Zusammenhang mit der Prädiktorvariable x ist (im Vergleich zu Abbildung 10.31 nahezu völlig) verschwunden, und nur die Mediatorvariable ist ein signifikanter Prädiktor des Kriteriums. Da kein direkter Einfluss der Variablen x im multiplen Regressionsmodell auf das Kriterium verbleibt, muss der Einfluss aus dem einfachen linearen Regressionsmodell (nahezu) total über die Mediatorvariable vermittelt sein.

Im Falle der Mediation hängt es jedoch von der Fragestellung ab, ob eine Mediatorvariable als Prädiktor mit in ein Regressionsmodell aufgenommen werden soll oder nicht. Man stelle sich beispielsweise vor, man möchte die Wirksamkeit einer neuen Therapiemethode für Depression

untersuchen. Es stellt sich heraus, dass durch die neue Therapiemethode die negative Selbstbewertung der Klienten sinkt. Nimmt man nun die negative Selbstbewertung als Prädiktor der Therapieeffektivität mit in ein entsprechendes Regressionsmodell auf, verringert sich natürlich der direkte Effekt der Therapiemethode. Dabei handelt es sich aber nicht um eine bessere Schätzung des Effekts der Therapiemethode, da ja die Wirkung über die negative Selbstbewertung gerade ein Wirkungspfad der Therapiemethode ist. Für die Gesamteffektivität der Therapiemethode bleibt dieser also durchaus zu berücksichtigen. Allerdings kann die Hinzunahme des Prädiktors negative Selbstbewertung in diesem Fall gleichzeitig das Verständnis für einen möglichen Wirkprozess der Therapiemethode durchaus erhöhen.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Mediator, Prädiktor[b]	.	Enter

a. Dependent Variable: Kriterium

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.804[a]	.646	.645	15.626

a. Predictors: (Constant), Mediator, Prädiktor

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	444634.011	2	222317.005	910.443	<.001[b]
	Residual	243452.913	997	244.185		
	Total	688086.924	999			

a. Dependent Variable: Kriterium

b. Predictors: (Constant), Mediator, Prädiktor

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	51.759	2.360		21.933	<.001
	Prädiktor	.002	.046	.001	.049	.961
	Mediator	.993	.033	.803	30.321	<.001

a. Dependent Variable: Kriterium

Abbildung 10.32. Ausgabe für eine multiple lineare Regressionsanalyse mit der Mediatorvariablen als Prädiktor im Falle einer totalen Mediation.

Übungsaufgaben

Die im Folgenden benötigten Datensätze finden Sie im elektronischen Ergänzungsmaterial (Engl.: electronic supplementary material) zu diesem Dokument, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Beispiel 10.1

Welche Aussage/n zu Voraussetzungen der linearen Regressionsanalyse trifft/treffen zu?

(a) Bei der multiplen Regression muss neben allen Voraussetzungen für die einfache lineare Regression auch die Sphärizität überprüft werden.
(b) Die Fehlervarianz muss unabhängig von den Prädiktoren konstant sein.
(c) Die Verletzung der Homoskedastizitätsannahme ist im Vergleich zur Verletzung der Annahme der Normalverteilung der Fehler bei der linearen Regression nicht so tragisch.
(d) Eine Verletzung der Linearitätsannahme wirkt sich ausschließlich auf die inferenzstatistischen Verfahren im Rahmen der linearen Regressionsanalyse aus.

Beispiel 10.2

Welche Aussage/n zu Ausreißern trifft/treffen zu?

(a) Einflusswerte können massive Auswirkungen auf die Parameterschätzung in der linearen Regression haben.
(b) Unter Diskrepanzwerten versteht man Datenpunkte mit ungewöhnlich großen oder kleinen UV-Werten im Vergleich zu den übrigen Datenpunkten.
(c) Unter Hebelwerten versteht man Datenpunkte mit ungewöhnlich großen Abweichungen von der (ohne Ausreißer) geschätzten Regressionsgerade.
(d) Diskrepanzwerte können anhand von standardisierten Residuen identifiziert werden.

Beispiel 10.3

Führen Sie eine Regressionsdiagnostik für beide Regressionsanalysen aus dem Beispiel 9.9 durch. Mit dem Wissen, dass die fiktiven Daten alle unter den Voraussetzungen für eine lineare Regressionsanalyse erzeugt wurden: Verwundert Sie die Anzahl der Fälle (Personen) mit einer Cook'schen Distanz größer als $4/n$, wobei n den Stichprobenumfang bezeichnet?

Beispiel 10.4

Führen Sie eine Regressionsdiagnostik für das Beispiel 9.10 durch.

Beispiel 10.5

Florida ist auch als „orange county“ bekannt. Ein ortsansässiger Forscher hat schon lange die Vermutung, dass der regelmäßige Verzehr von Obst die Intelligenz fördert, und dass Orangen dabei eine besonders starke Wirkung haben. Um dieser Vermutung nachzugehen, rekrutiert er 400 Personen, um deren IQ sowie die wöchentlich verzehrte Menge an Orangen und Äpfeln (jeweils in g) zu ermitteln. Die erhobenen Daten befinden sich in der Datei „Kap10UE5.sav“.

Bevor sich der Forscher an das Überprüfen seiner vermuteten Zusammenhänge machen kann, muss er eine Regressionsdiagnostik durchführen. Da er aktuell kaum Zeit für seine Forschung findet, wendet er sich an Sie. Unterstützen Sie den Forscher, indem Sie eine Regressionsdiagnostik inklusive einer Ausreißeranalyse für eine Regressionsanalyse mit den beiden Prädiktoren *Orangenverzehr* und *Äpfelverzehr* und dem Kriterium *IQ* durchführen.

Beispiel 10.6

Eine Forschungsgruppe möchte den Zusammenhang zwischen der Intelligenz und dem Ergebnis beim Aufnahmetest für das Medizinstudium untersuchen. Dazu werden die Daten von 1000 Teilnehmer:innen an dem Aufnahmetest untersucht. Die Daten sind in der Datei „Kap10UE6.sav“ gegeben. Bevor eine Regressionsanalyse durchgeführt werden kann, muss eine Regressionsdiagnostik durchgeführt werden, um die Voraussetzungen für eine Regressionsanalyse zu prüfen. Führen Sie diese Regressionsdiagnostik inklusive einer Ausreißeranalyse für die gegebenen Daten durch und fassen Sie beides in einem kurzen Bericht zusammen.

Beispiel 10.7

Welche Aussage/n trifft/treffen zu?

(a) Beim Determinationskoeffizienten handelt es sich um eine Effektstärke für den durchgeführten Omnibustest im Rahmen einer multiplen linearen Regressionsanalyse.
(b) Bei der quadrierten Semipartialkorrelation handelt es sich um eine Effektstärke für den durchgeführten Omnibustest im Rahmen einer multiplen linearen Regressionsanalyse.
(c) Die quadrierte Semipartialkorrelation gibt Auskunft sowohl über Stärke als auch Richtung des Zusammenhangs einer UV mit der AV.
(d) Der standardisierte Regressionskoeffizient gibt Auskunft über die Richtung des Zusammenhangs einer UV mit der AV, aber nicht über die Stärke des Zusammenhangs.

Beispiel 10.8

Welche Aussage/n trifft/treffen zu?

(a) Der Determinationskoeffizient gibt an, welcher Anteil der Varianz im Kriterium nur gemeinsam durch die Prädiktoren im Regressionsmodell erklärt werden kann.
(b) Die quadrierte Semipartialkorrelation für den Prädiktor *j* gibt den Anteil der Varianz im Kriterium an, der eigenständig durch den Prädiktor *j* erklärt werden kann.
(c) Für eine multiple lineare Regression kann keine Stichprobenumfangsplanung durchgeführt werden.
(d) Für eine Stichprobenumfangsplanung für eine einfache lineare Regression in G*Power wird die Semipartialkorrelation benötigt.

Beispiel 10.9

Ergänzen Sie den Ergebnisbericht zu Teil (b) des Beispiels 9.9 aus dem vorhergehenden Kapitel um Angaben zu den Anteilen an der Varianz des Kriteriums, die eigenständig jeweils durch die beiden Prädiktoren erklärt werden können bzw. die nur durch beide Prädiktoren gemeinsam erklärt werden kann.

Beispiel 10.10

Ergänzen Sie den Ergebnisbericht des Beispiels 9.10 aus dem vorhergehenden Kapitel um Angaben zu den Anteilen an der Varianz des Kriteriums, die eigenständig jeweils durch die beiden Prädiktoren erklärt werden können bzw. die nur durch beide Prädiktoren gemeinsam erklärt werden kann.

Beispiel 10.11

Führen Sie eine Stichprobenumfangsplanung für einfache lineare Regressionsanalyse durch. Der statistische Test soll ein Signifikanzniveau von α = .001 sowie eine Teststärke (power) von 0.8 aufweisen. Als Mindesteffektstärke geben wir den Populationsdeterminationskoeffizienten von ρ^2 = .15 vor.

Beispiel 10.12

Führen Sie eine Stichprobenumfangsplanung für multiple lineare Regressionsanalyse durch, wobei es hier nur um den Effekt eines einzelnen Prädiktors auf das Kriterium gehen soll. Der statistische Test soll ein Signifikanzniveau von α = .005 sowie eine Teststärke (power) von 0.9 aufweisen. Die quadrierte Semipartialkorrelation für den interessierenden Prädiktor betrage mindestens $\rho_j^2 = 0.03$. Das gesamte Regressionsmodell mit drei Prädiktoren soll 28% der Varianz im Kriterium erklären.

Beispiel 10.13

Ein Freund von Ihnen argumentiert, dass man der guten Online-Bewertung eines Restaurants nur trauen dürfe, wenn das Restaurant nicht einfach zu erreichen sei. Umgekehrt sei bei Restaurants in einer guten Lage die Chance recht hoch trotz guter Online-Bewertungen nur mittelmäßiges Essen zu bekommen. Zeichnen Sie ein DAG mit den Variablen *Online-Bewertung*, *Speisenqualität*, und *Lage*. Wie müssen Sie die Wirkrichtung der Pfeile angeben, damit der DAG der Argumentation Ihres Freundes entspricht. Welchem Typ der besprochenen vier grundlegenden Arten von DAGs entspricht dieser Fall?

Beispiel 10.14

Überlegen Sie sich ein Beispiel für eine konfundierende Variable. Zeichnen Sie den entsprechenden DAG. Erfinden Sie anschließend einen geeigneten Datensatz, der die Zusammenhänge zwischen den entsprechenden Variablen abbildet. Veranschaulichen Sie sich schließlich die Auswirkung auf die Resultate entsprechender Regressionsanalysen, indem Sie diese auf Basis Ihrer fiktiven Daten durchführen.

Beispiel 10.15

Wiederholen Sie Beispiel 10.14 für den Fall einer Mediation.

Beispiel 10.16

Im Datensatz „collider.sav“ ist neben der Innovation, der Publikationswürdigkeit und der Wissenschaftlichkeit für 1000 (fiktive) wissenschaftliche Arbeiten auch noch die Variable Förderungswürdigkeit gegeben. Diese Variable hat nur drei Ausprägungen: 2 = sehr förderungswürdig, 1 = unter Umständen förderungswürdig, 0 = nicht förderungswürdig. Veranschaulichen Sie sich, dass auf jeder Stufe der Förderungswürdigkeit ein negativer Zusammenhang zwischen der Innovation und der Wissenschaftlichkeit der beurteilten wissenschaftlichen Arbeiten besteht. Wie erklären Sie sich diesen Befund? Zeigen Sie, dass für die beurteilten wissenschaftlichen Arbeiten im Allgemeinen kein solcher Zusammenhang besteht. Mit welchem DAG würden Sie den Zusammenhang zwischen den drei Variablen Innovation, Wissenschaftlichkeit und Förderungswürdigkeit beschreiben?

Kapitel 11

Regressionsmodelle mit diskreten Prädiktoren und Interaktionen

Hanna Rajh-Weber, Stefan E. Huber

Bisher haben wir uns im Rahmen der Regressionsanalyse ausschließlich mit stetigen Prädiktoren befasst. Allerdings ist es mittels einer sogenannten Dummy-Kodierung keine große Schwierigkeit diskrete Prädiktoren in Regressionsanalysen zu berücksichtigen. Das soll im Folgenden illustriert werden. Dafür werden wir uns zuerst mit einfachen Regressionsanalysen mit nur einem Prädiktor (mit zwei oder mehr Ausprägungen) befassen. Danach werden wir uns mit multiplen Regressionsanalysen beschäftigen, bei welchen entweder alle oder nur manche der Prädiktoren diskret sind. In diesem Zusammenhang werden wir uns wieder mit Interaktionen befassen – d.h. der Auswirkung der Ausprägung eines Prädiktors auf die Wirkung eines anderen Prädiktors auf das Kriterium – die uns schon im Rahmen von Varianzanalysen untergekommen sind. Dabei werden wir zu guter Letzt sehen, dass auch zwei kontinuierliche Prädiktoren miteinander interagieren können.

Um das Vorgehen für all die unterschiedlichen Kombinationsmöglichkeiten von diskreten und stetigen Prädiktoren in diesem Kapitel zu illustrieren, beziehen wir uns auf die fiktiven Datensätze „Kap11daten1.sav“, „Kap11daten2.sav“, „Kap11daten3.sav“ und „Kap11daten4.sav“, die Sie im elektronischen Ergänzungsmaterial zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können. Auch wenn es sich dabei prinzipiell um fiktive Daten handelt, wurden die Daten näherungsweise auf der Grundlage der Zusammenhänge zwischen Geschlecht, Nationalität, Alter, Erhebungsjahr, und dem Jahreseinkommen erstellt, die Sie auf https://www.oecd.org/en/data.html nachschlagen können.

Regressionsmodelle mit einem diskreten Prädiktor

Wir betrachten zuerst den einfachsten Fall eines stetigen Kriteriums und eines diskreten Prädiktors mit zwei kategorialen Ausprägungen. Im Anschluss betrachten wir den Fall eines diskreten Prädiktors mit mehr als zwei kategorialen Ausprägungen.

Ein diskreter Prädiktor mit zwei (kategorialen) Ausprägungen

Wir betrachten die folgende Fragestellung mithilfe des fiktiven Datensatzes „Kap11daten1.sav“: Wie wirkt sich das Geschlecht, wobei hier nur die Kategorien männlich und weiblich berücksichtigt werden (aufgrund der Dürftigkeit an Daten für die Kategorie divers), auf das jährliche Bruttoeinkommen in Österreich für Angestellte mittleren Alters aus?

Aus Kapitel 5 wissen wir, dass wir diese Fragestellung auch mit einem t-Test für unabhängige Stichproben erhellen könnten. Eine entsprechende Berechnung in SPSS ergibt die in Abbildung 11.1 gezeigte Ausgabe. Wir sehen, dass sich die mittleren Jahreseinkommen von Männern und Frauen (mit $\alpha = .005$) signifikant unterscheiden, $t(86.01) = 3.33$, $p = .001$, Cohens $d = 0.666$. Das mittlere Jahreseinkommen von Männern ($M = 77806.86$, $SD = 17608.65$, $n = 50$) ist im Mittel um 10012.66 USD, 95%-KI [4038.51, 15986.81] höher als das von Frauen ($M = 67794.20$, $SD = 11895.41$, $n = 50$).

Group Statistics

	Geschlecht (männlich, weiblich)	N	Mean	Std. Deviation	Std. Error Mean
Jährliches Bruttoeinkommen in USD (inflationsbereinigt)	m	50	77806.86	17608.654	2490.240
	w	50	67794.20	11895.408	1682.265

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Jährliches Bruttoeinkommen in USD (inflationsbereinigt)	Equal variances assumed	6.102	.015	3.332	98	<.001	.001	10012.660	3005.214	4048.912	15976.408
	Equal variances not assumed			3.332	86.014	<.001	.001	10012.660	3005.214	4038.508	15986.812

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval	
				Lower	Upper
Jährliches Bruttoeinkommen in USD (inflationsbereinigt)	Cohen's d	15026.068	.666	.262	1.068
	Hedges' correction	15142.299	.661	.260	1.059
	Glass's delta	11895.408	.842	.412	1.264

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

Abbildung 11.1. Ergebnis eines t-Tests für unabhängige Stichproben.

Wenn wir für die beiden Kategorien der Geschlechtsvariable eine Dummy-Kodierung durchführen (für eine berechtigte und jedenfalls bedenkenswerte Kritik der Dummy-Kodierung, siehe McElreath, 2020), dann können wir die Fragestellung auch mit einer Regressionsanalyse erhellen. Für einen kategorialen Prädiktor mit zwei Ausprägungen sieht eine Dummy-Kodierung wie folgt aus: Eine Referenzkategorie wird mit 0 kodiert, die andere Kategorie mit 1. Im vorliegenden Datensatz wurde die Kategorie männlich mit 0 kodiert, d.h. diese Kategorie fungiert hier als Referenzkategorie. Prinzipiell

ist es egal, welche der beiden Kategorien als Referenzkategorie kodiert wird, wir müssen die Wahl nur entsprechend bei der Interpretation der Ergebnisse berücksichtigen.

Bezeichnen wir die Dummy-Variable als D_i mit $D_i = 0$, falls Person i männlich ist, und $D_i = 1$, falls Person i weiblich ist, dann können wir die regressionsanalytische Modellgleichung wie folgt schreiben:

$$Y_i \sim N(\alpha + \beta D_i, \sigma^2)$$

mit den Modellparametern α, β und σ^2. D.h. insbesondere, der Erwartungswert des Kriteriums hängt von der Dummy-Variablen ab. Für Männer ergibt sich $E(Y_i|D_i = 0) = \alpha + \beta \cdot 0 = \alpha$, während sich für Frauen $E(Y_i|D_i = 0) = \alpha + \beta \cdot 1 = \alpha + \beta$ ergibt. D.h. die Interpretation des Steigungsparameters β ist exakt dieselbe, die wir schon im Fall der einfachen linearen Regression in Kapitel 9 kennengelernt haben: Eine Erhöhung der Dummy-Variablen um den Wert 1 geht im Mittel mit einer Erhöhung β im Kriterium einher. Genauso können wir auch die Ausgabe interpretieren, die wir erhalten, wenn wir eine entsprechende einfache lineare Regression mit dem Prädiktor Geschlecht und dem Kriterium Einkommen in SPSS durchführen (im Datensatz liegt die Variable Geschlecht bereits mit der entsprechenden Dummy-Kodierung vor), siehe Abbildung 11.2.

Auch an dieser Ausgabe erkennen wir, dass das Einkommen von Männern und Frauen sich signifikant unterscheidet, da der Schätzwert für das Regressionsgewicht sich (mit α = .005) signifikant von Null unterscheidet, $t(98) = 3.33$, $p = .001$. Der Schätzwert des Parameters ist negativ, $b = -10012.66$, vom Betrag her jedoch genau gleich wie die Mittelwertdifferenz, die wir im Rahmen des t-Tests oben erhalten haben. Das negative Vorzeichen geht bloß darauf zurück, dass oben die Differenz zwischen Männern und Frauen gebildet wurde, und hier die Änderung des mittleren Einkommens ermittelt wurde, wenn wir die Geschlechtskategorie von „männlich“ (= Referenzkategorie = Wert 0) auf „weiblich“ (= Wert 1) ändern, also genau umgekehrt als im vorhergehenden Fall.

Zum Vergleich haben wir uns nun auch für den Fall der Regressionsanalyse ein Konfidenzintervall für den Regressionsparameter ausgeben lassen. Letzteres kann im Menü „Statistics…“ bei der Anforderung der Regressionsanalyse ausgewählt werden. Wir sehen, dass der

plausible Bereich für das mittlere Bruttoeinkommen der Männer zwischen 73589.85 und 82023.87 USD liegt, und der plausible Bereich für das mittlere Bruttoeinkommen der Frauen um 4048.91 bis 15976.41 USD darunter liegt.

Ein Ergebnisbericht für die lineare Regressionsanalyse mit einem diskreten Prädiktor mit zwei Ausprägungen für das oben erläuterte Beispiel könnte wie folgt aussehen: „Eine einfache lineare Regressionsanalyse ergab, dass ein (mit α = .005) statistisch signifikanter Anteil der Varianz im Einkommen der untersuchten n = 100 Personen dadurch erklärt werden kann, ob die Personen männlich oder weiblich sind, $F(1, 98) = 11.10$, $p = .001$, $R^2 = .10$; ein kleiner Effekt gemäß Cohen (1988). Gemäß des resultierenden Regressionsmodells verdienen Männer jährlich im Mittel etwa 78 000 Euro (b = 77806.86, $t(98) = 36.62$, $p < .001$). Frauen verdienen jährlich im Mittel etwa 10 000 Euro weniger als Männer. Dieser Unterschied ist (mit α = .005) signifikant (b = -10012.66, stand. β = -.32, $t(98) = -3.33$, $p = .001$)."

Am Konfidenzintervall (und prinzipiell auch schon an den Freiheitsgraden für den Signifikanztest des Regressionsparameters) erkennen wir auch, dass es sich um die exakt gleichen Werte wie für das Konfidenzintervalls der Mittelwertdifferenz im Rahmen des Student'schen t-Tests handelt, siehe Abbildung 11.1 oben. Das ist kein Zufall; in der Tat handelt es sich dabei um Ergebnisse einer völlig äquivalenten statistischen Berechnung, da in beiden Fällen die gleichen Annahmen getroffen wurden: intervallskalierte AV, unbekannte Varianz der AV in beiden Gruppen (= für beide Ausprägungen der UV), die Messwerte in beiden Gruppen (= für beide Ausprägungen der UV) sind unabhängig voneinander, die AV kann in beiden Gruppen (= für beide Ausprägungen der UV) durch eine Normalverteilung approximiert werden, und insbesondere ist die Varianz dieser Normalverteilung in beiden Gruppe dieselbe (Varianzgleichheit, -homogenität, Homoskedastizität). Die Regressionsanalyse mit Dummy-Kodierung für einen kategorialen Prädiktor mit zwei Ausprägungen ist für den Steigungsparameter also völlig äquivalent zu einem Student'schen t-Test.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Geschlecht (männlich, weiblich)[b]	.	Enter

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.319[a]	.102	.093	15026.068

a. Predictors: (Constant), Geschlecht (männlich, weiblich)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2506334006.9	1	2506334006.9	11.101	.001[b]
	Residual	22126705188	98	225782706.00		
	Total	24633039195	99			

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

b. Predictors: (Constant), Geschlecht (männlich, weiblich)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	77806.860	2125.007		36.615	<.001	73589.853	82023.867
	Geschlecht (männlich, weiblich)	-10012.660	3005.214	-.319	-3.332	.001	-15976.408	-4048.912

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

Abbildung 11.2. Ergebnis einer einfachen linearen Regression mit dem Prädiktor Geschlecht und dem Kriterium Einkommen.

Das zeigt allerdings einen Nachteil dieses Vorgehens auf: Während ungleiche Varianzen im Falle des t-Tests in der Praxis leicht durch Rückgriff auf den ohnehin durchgeführten t-Test nach Welch einfach berücksichtigt werden können (und beim Bericht der Ergebnisse oben auch berücksichtigt wurden, da der u.a. der signifikante Levene-Test auf einen möglichen Unterschied der Populationsvarianzen hinweist), geht die Regressionsanalyse von Varianzhomogenität aus. Grundsätzlich können auch hier ungleiche Varianzen im Rahmen der durchgeführten inferenzstatistischen Verfahren durch Rückgriff auf heteroskedastizitätskorrigierte Standardfehler berücksichtigt werden (siehe z.B. Rajh-Weber et al., 2025).

Ein diskreter Prädiktor mit mehr als zwei (kategorialen) Ausprägungen

Mittels Dummy-Kodierung kann das oben illustrierte Vorgehen sehr einfach auf mehr als zwei Kategorien verallgemeinert werden. Bei einer kategorialen Variablen mit insgesamt k Katgorien werden dafür $k-1$ Dummy-Variablen für die $j = k-1$ Gruppen definiert, die nicht als Referenzkategorie fungieren sollen. Für die Dummy-Variablen D_{ji} gilt dann jeweils $D_{ji} = 1$, falls Person i zur Gruppe j gehört, und $D_{ji} = 0$, falls Person i nicht zur Gruppe j gehört. Welche Gruppe bzw. Kategorie jeweils als Referenzkategorie gewählt wird, wirkt sich wiederum nur darauf aus, wie die Ergebnisse der entsprechenden Regressionsanalyse zu interpretieren sind.

Dieses Vorgehen und die Interpretation der Ergebnisse wird im Folgenden am Beispiel folgender Fragestellung untersucht: Wie wirkt sich die Nationalität (Österreich, Deutschland, USA) auf das jährliche Bruttoeinkommen in Österreich für Angestellte mittleren Alters aus? Dazu wird der Datensatz „Kap11daten2.sav“ verwendet.

In diesem Datensatz ist die Nationalität der jeweils (fiktiven) befragten Person durch die Variable *Nation* mit den Kategorien 0 = AUT (für Österreich), 1 = GER (für Deutschland) und 2 = USA (für die USA) kodiert. D.h., wir müssen in diesem Fall die oben beschriebene Dummy-Kodierung noch selbst durchführen. Dazu können wir einfach in SPSS unter *Transform >> Compute Variable…* zwei neue Variablen erzeugen.

Für die erste der beiden Variablen wählen wir im Feld „Target Variable:“ z.B. die Bezeichnung AUTvsGER (die Variable soll uns also Unterschiede zwischen Österreich und Deutschland kodieren) und fügen dann im Feld „Numeric Expression:“ den Ausdruck „Nation = 1“ ein. Dabei nutzen wir, dass SPSS intern Bools’sche Variablen, d.h. Variablen, die nur „wahr“ oder „falsch“ sein können, ohnehin mit 1 (für „wahr“) und 0 (für „falsch“) kodiert. D.h., im Fall, dass für Person i die Variable *Nation* den Wert 1 (für Deutschland) hat, ist der Ausdruck „Nation = 1“ wahr und die neue Variable *AUTvsGER* bekommt den Wert 1. Für Personen aus Österreich oder den USA ist der Ausdruck „Nation = 1“ hingegen immer falsch und die Variable bekommt den Wert 0.

Für die zweite der beiden Variablen verfahren wir ganz analog. Wir nennen die Variable AUTvsUSA und fügen im Feld „Numeric Expression:“ nun den Ausdruck „Nation = 2“ ein. D.h. diese

Variable bekommt den Wert 1 genau dann, wenn die jeweilige Person aus den USA kommt (d.h., wenn die Variable *Nation* den Wert 2 hat), und den Wert 0 sonst (d.h., wenn die Person aus Österreich oder Deutschland kommt, d.h. die Variable *Nation* nicht den Wert 2 hat).

Haben wir beide Dummy-Variablen erzeugt (bzw. existieren in der Datendatei „Kap11daten2.sav“ auch bereits zwei entsprechend erzeugte Variablen mit den Bezeichnungen *AUTvsGERvordefiniert* und *AUTvsUSAvordefiniert*), können wir sie unter *Analyze >> Regression >> Linear…* als Prädiktoren in ein multiples Regressionsmodell mit dem Kriterium *Einkommen* einfügen. Im Menü „Statistics…“ fordern wir wiederum 95%-KI für die Regressionsparameter an. Die Ausgabe ist in Abbildung 11.3 gezeigt.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	AUTvsUSA, AUTvsGER[b]	.	Enter

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.327[a]	.107	.101	17093.335

a. Predictors: (Constant), AUTvsUSA, AUTvsGER

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10364185200	2	5182092599.9	17.736	<.001[b]
	Residual	86778087739	297	292182113.60		
	Total	97142272939	299			

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

b. Predictors: (Constant), AUTvsUSA, AUTvsGER

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	72800.530	1709.334		42.590	<.001	69436.590	76164.470
	AUTvsGER	-5911.530	2417.363	-.155	-2.445	.015	-10668.860	-1154.200
	AUTvsUSA	8413.190	2417.363	.220	3.480	<.001	3655.860	13170.520

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

Abbildung 11.3. Ausgabe für eine multiple Regression mit den beiden erzeugten Dummy-Variablen *AUTvsGER* und *AUTvsUSA* als Prädiktoren.

Am Schätzwert für den Achsenabschnitt a = 72800.53 USD können wir das mittlere Einkommen für Angestellte in Österreich ablesen. Der plausible Bereich, der einem 95%-KI entspricht liegt zwischen 69436.59 und 76164.47 USD.

Am Schätzwert für das Regressionsgewicht der Variable *AUTvsGER* erkennen wir, dass das mittlere Einkommen im Mittel um 5911.53, 95%-KI [1154.20, 10668.86], unter demjenigen für Österreich liegt. Wir sehen zudem, dass der Unterschied (mit α = .005) nicht signifikant ist, $t(297) = -2.45$, $p = .015$.

Am Schätzwert für das Regressionsgewicht der Variable *AUTvsUSA* erkennen wir schließlich, dass Angestellte in den USA im Mittel um 8413.19 USD, 95%-KI [3655.86, 13170.52], mehr verdienen als in Österreich. Dieser Unterschied ist (mit α = .005) signifikant ist, $t(297) = 3.48$, $p < .001$.

Auch hier gibt es wieder eine Äquivalenz mit den varianzanalytischen Verfahren, die wir in Kapitel 6 kennengelernt haben. Der Omnibustest für das multiple Regressionsmodell ist äquivalent zum Omnibustest der einfaktoriellen Varianzanalyse. Beide Modelle gehen auch wieder von Varianzhomogenität aus. Scheint diese nicht gegeben, kann für den regressionsanalytischen Ansatz wieder auf für Heteroskedastizität korrigierte Standardfehler zurückgegriffen werden (Rajh-Weber et al., 2025). Für den varianzanalytischen Zugang kann der in Kapitel 6 erläuterte Welch-Test durchgeführt werden.

Interaktionen

Sehr häufig ist gerade die kombinierte Wirkung mehrerer UV auf typische Werte der AV von Interesse. Um bei unserem Beispiel für dieses Kapitel zu bleiben: Unterscheidet sich der Unterschied für das typische Jahreseinkommen zwischen Männern und Frauen (das sog. Gender Wage Gap) etwa je nach Nation? Und falls ja, wie?

Wie wir aus Kapitel 7 bereits wissen, handelt es sich hierbei um die Frage nach einer Interaktion zwischen den beiden Variablen: Ist die Wirkung einer Variablen (hier: Geschlecht) abhängig von der Ausprägung einer anderen (hier: Nationalität)? Für den Fall zweier diskreter Prädiktoren könnten wir diese Fragestellung auch mit den uns bereits bekannten Varianzanalysen untersuchen. Diese bieten häufig sogar den Vorteil der einfacheren Interpretierbarkeit im Vergleich zum regressionsanalytischen

Zugang, den wir im nächsten Abschnitt betrachten werden (Bühner et al., 2025). Die Interpretation des letzteren wird u.a. deshalb erschwert, weil sich die Ergebnisse für die hier illustrierte Dummy-Kodierung der Prädiktoren stets auf eine Referenzkategorie beziehen und daher eine inferenzstatistische Analyse auf den Vergleich mit dieser Kategorie beschränkt bleibt. Falls allerdings hauptsächlich die Vorhersage typischer AV-Werte für gegebene UV-Werte im Vordergrund steht und die Interpretation der Parameter nicht interessiert, ist wiederum das Regressionsmodell einfacher zu handhaben (Bühner et al., 2025).

Das Regressionsmodell bietet zudem den Vorteil größerer Flexibilität, indem es auch die Berücksichtigung von Interaktionen zwischen einer diskreten und einer stetigen oder auch zwischen zwei stetigen Variablen erlaubt. Diese Fälle lassen sich in der Tat mit den varianzanalytischen Methoden, die wir in den Kapiteln 6-8 kennengelernt haben, nicht behandeln. In diesen Fällen spricht man auch von Moderation oder moderierter Regression (Bühner et al., 2025): die Ausprägung einer UV wirkt sich auf den linearen Zusammenhang zwischen der anderen UV und typischen Ausprägungen der AV aus. Damit werden wir uns in den letzten beiden Abschnitten dieses Kapitels befassen.

Mehrere diskrete Prädiktoren

Wir bleiben beim Beispiel mit den beiden Prädiktoren Geschlecht und Nationalität, mithilfe derer wir das typische Jahreseinkommen von Angestellten mittleren Alters vorhersagen möchten. Die Daten dafür finden wir nach wie vor in der Datei „Kap11daten2.sav“, die Sie im elektronischen Ergänzungsmaterial zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Um ein Regressionsmodell mit allen möglichen Interaktionstermen für die beiden Prädiktoren zu illustrieren, werden im Folgenden zuerst alle dafür benötigten Dummy-Variablen erzeugt, und erst im Anschluss erläutert, inwiefern diese alle möglichen Kombinationen aus den beiden kategorialen Variablen berücksichtigen. Die drei Dummy-Variablen für die beiden Prädiktoren wurden im vorhergehenden Abschnitt erzeugt (bzw. lagen für das Geschlecht bereits vor) und liegen in dem Datensatz bereits mit den Bezeichnungen *Geschlecht*, *AUTvsGERvordefiniert*, und *AUTvsUSAvordefiniert* vor. Um die Interaktion zwischen den beiden Variablen für alle möglichen Kombinationen aus ihnen zu berücksichtigen, müssen wir nun noch zwei weitere Dummy-Variablen

erzeugen. Dazu geben wir unter *Transform >> Compute Variable…* erst einmal einen Variablennamen unter „Target Variable:“ an. Dieser kann z.B. „Geschlecht_X_AUTvsGER“ sein. In der Datendatei gibt es bereits eine entsprechende Variable mit der Bezeichnung *Geschlecht_X_AUTvsGERvordefiniert*, die für einen Vergleich mit der eigens erzeugten Variablen verwendet werden kann. Im Feld „Numeric Expression:“ geben wir Folgendes ein: „Geschlecht * AUTvsGERvordefiniert“; eine Erläuterung folgt in Kürze. Ganz analog gehen wir für die andere, zusätzlich noch benötigte Dummy-Variable vor. Diese können wir z.B. mit „Geschlecht_X_AUTvsUSA“ bezeichnen, es existiert aber auch wieder bereits eine entsprechende Variable unter der Bezeichnung „Geschlecht_X_AUTvsUSAvordefiniert“. Im Feld „Numeric Expression:“ geben wir für diese Variable „Geschlecht * AUTvsUSAvordefiniert“ ein.

Schauen wir uns nun alle unsere Dummy-Variablen noch einmal genau an. Die Variable *Geschlecht* hat genau dann den Wert 1, wenn das Geschlecht einer Person weiblich ist, sonst hat sie den Wert 0. Die Variable *AUTvsGERvordefiniert* hat genau dann den Wert 1, wenn die Nationalität einer Person Deutschland ist, sonst hat sie den Wert 0. Die Variable *AUTvsUSAvordefiniert* hat genau dann den Wert 1, wenn die Nationalität einer Person USA ist, sonst hat sie den Wert 0. Die Variable *Geschlecht_X_AUTvsGERvordefiniert* hat genau dann den Wert 1, wenn das Geschlecht einer Person weiblich ist und gleichzeitig die Nationalität der Person Deutschland ist, sonst hat sie den Wert 0. Die Variable *Geschlecht_X_AUTvsUSAvordefiniert* hat genau dann den Wert 1, wenn das Geschlecht einer Person weiblich ist und gleichzeitig die Nationalität der Person USA ist, sonst hat sie den Wert 0.

Inwiefern bildet das alle Kombinationsmöglichkeiten aus den beiden Variablen ab? Dazu betrachten wir das gesamte Regressionsmodell mit allen fünf Dummy-Variablen, die für die Reihenfolge des vorhergehenden Absatzes kurz mit D_{ij} mit $j = 1, \dots 5$ bezeichnet werden:

$$Y_i \sim N(\alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \beta_5 D_{5i}, \sigma^2).$$

Betrachten wir nun eine männliche Person mit Nationalität Österreich. In diesem Fall sind alle Dummy-Variablen gleich Null und es bleibt

$$Y_i \sim N(\alpha, \sigma^2),$$

d.h. der Erwartungswert des Jahreseinkommens für männliche Personen mit Nationalität Österreich ist durch den Achsenabschnitt α gegeben.

Für weibliche Personen mit Nationalität Österreich gilt, dass die Dummy-Variable *Geschlecht* gleich Eins ist, d.h. $D_{1i} = 1$, während alle anderen Dummy-Variablen gleich Null sind. Der Erwartungswert des Jahreseinkommens für weibliche Personen mit Nationalität Österreich ist demnach durch $\alpha + \beta_1$ gegeben.

Für männliche Personen mit Nationalität Deutschland ist die Dummy-Variable *AUTvsGERvordefiniert* gleich Eins, d.h. $D_{2i} = 1$, während alle übrigen Dummy-Variablen gleich Null sind. Daraus ergibt sich der Erwartungswert des Jahreseinkommens für männliche Personen mit Nationalität Deutschland zu $\alpha + \beta_2$.

Für weibliche Personen mit Nationalität Deutschland sind die Dummy-Variablen *Geschlecht, AUTvsGERvordefiniert*, sowie *Geschlecht_X_AUTvsGERvordefiniert* gleich Eins, d.h. $D_{1i} = D_{2i} = D_{4i} = 1$, während die übrigen beiden Dummy-Variablen nach wie vor gleich Null sind. Daraus ergibt sich der Erwartungswert des Jahreseinkommens für weibliche Personen mit Nationalität Deutschland zu $\alpha + \beta_1 + \beta_2 + \beta_4$.

Für männliche Personen mit Nationalität USA ist die Dummy-Variable *AUTvsUSAvordefiniert* gleich Eins, d.h. $D_{3i} = 1$, während alle übrigen Dummy-Variablen gleich Null sind. Daraus ergibt sich der Erwartungswert des Jahreseinkommens für männliche Personen mit Nationalität USA zu $\alpha + \beta_3$.

Für weibliche Personen mit Nationalität USA sind schließlich die Dummy-Variablen *Geschlecht, AUTvsUSAvordefiniert*, sowie *Geschlecht_X_AUTvsUSAvordefiniert* gleich Eins, d.h. $D_{1i} = D_{3i} = D_{5i} = 1$, während die übrigen beiden Dummy-Variablen gleich Null sind. Daraus ergibt sich der Erwartungswert des Jahreseinkommens für weibliche Personen mit Nationalität USA zu $\alpha + \beta_1 + \beta_3 + \beta_5$.

Wir sehen: Durch die fünf Dummy-Variablen sind in der Tat alle Kombinationen der beiden Prädiktoren abgedeckt. Indem wir nun alle Dummy-Variablen als Prädiktoren in ein entsprechendes Regressionsmodell in SPSS hinzufügen, können wir eine Schätzung dieser Parameter vornehmen. Der

entsprechende Teil der Ausgabe ist in Abbildung 11.4 gezeigt. Zum Vergleich sind in Abbildung 11.5 auch die Tabelle mit deskriptiven Statistiken sowie den Resultaten einer zweifaktoriellen Varianzanalyse mit denselben beiden Prädiktoren (bzw. Faktoren) und derselben AV angegeben. Wir sehen, dass es sich bei dem Schätzwert für den Achsenabschnitt a = 77806.86 in der Tat um das mittlere Jahreseinkommen von männlichen Personen mit Nationalität Österreich handelt. Für weibliche Personen mit Nationalität Österreich erhalten wir b_1 = 77806.86 – 10012.66 = 67794.20, was wiederum mit dem entsprechenden Wert der Tabelle für die deskriptiven Statistiken in Abbildung 11.5 übereinstimmt. Genauso können wir uns durch Vergleich mit der Übereinstimmung aller anderen Schätzwerte mit den Mittelwerten entsprechend der obigen Erläuterungen überzeugen.

Zusätzlich sehen wir auch die anfangs angesprochene schwierigere Interpretierbarkeit dieses regressionsanalytischen Zugangs. Wie in Abbildung 11.4 ersichtlich, unterscheiden sich die Regressionskoeffizienten für die beiden Variablen *Geschlecht_X_AUTvsGERvordefiniert* und *Geschlecht_X_AUTvsUSAvordefiniert* nicht signifikant von Null. Aber auf welches mittlere Einkommen beziehen sich diese Regressionsgewichte? Im Falle von weiblichen Personen mit Nationalität Deutschland heißt dieses Ergebnis, dass das mittlere Einkommen dieser Personen im Mittel nicht signifikant größer (positives Vorzeichen des Schätzwerts b_4 = 5346.74) ist als das mittlere Einkommen das der Summe $a + b_1 + b2$ = 59209.30 entspricht. Dieses mittlere Einkommen bezieht sich aber auf keine real existierende Personengruppe wie der Vergleich mit den Erläuterungen zu den einzelnen Populationen und wie diese durch Kombinationen von Regressionsgewichten abgebildet sind oben zeigt. Analog verhält es sich mit den Jahreseinkommen von weiblichen Personen mit Nationalität USA. Aus der Tabelle lässt sich schlichtweg nicht ablesen wie sich die mittleren Einkommen (ob etwa signifikant oder nicht) zwischen vielen einzelnen Personengruppen unterscheiden (z.B. zwischen weiblichen Personen mit Nationalität Österreich und weiblichen Personen mit Nationalität Deutschland). Aufgrund dieser fehlenden Informationen lässt sich also auch nicht einfach darauf schließen, ob über alle Kategorien hinweg signifikante Haupteffekte oder Interaktionen vorliegen. Diese Information lässt sich wiederum leicht an der Ausgabe für eine entsprechende zweifaktorielle Varianzanalyse ablesen, siehe Abbildung 11.5.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	77806.860	2310.699		33.672	<.001	73259.252	82354.468
	Geschlecht (männlich, weiblich)	-10012.660	3267.822	-.278	-3.064	.002	-16443.948	-3581.372
	Dummy-Variable für den Vergleich Österreich-Deutschland	-8584.900	3267.822	-.225	-2.627	.009	-15016.188	-2153.612
	Dummy-Variable für den Vergleich Österreich-USA	10645.180	3267.822	.279	3.258	.001	4213.892	17076.468
	Dummy-Variable für Interaktion zwischen Geschlecht und AUTvsGER	5346.740	4621.398	.111	1.157	.248	-3748.475	14441.955
	Dummy-Variable für Interaktion zwischen Geschlecht und AUTvsUSA	-4463.980	4621.398	-.092	-.966	.335	-13559.195	4631.235

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

Abbildung 11.4. Schätzungen der Regressionsgewichte für ein Regressionsmodell mit Interaktion zweier diskreter Prädiktoren.

Descriptive Statistics

Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

Geschlecht (männlich, weiblich)	Nation (Österreich, Deutschland, USA)	Mean	Std. Deviation	N
m	AUT	77806.86	17608.654	50
	GER	69221.96	15086.375	50
	USA	88452.04	20354.412	50
	Total	78493.62	19373.853	150
w	AUT	67794.20	11895.408	50
	GER	64556.04	16209.353	50
	USA	73975.40	15671.303	50
	Total	68775.21	15135.215	150
Total	AUT	72800.53	15773.984	100
	GER	66889.00	15754.115	100
	USA	81213.72	19481.674	100
	Total	73634.42	18024.720	300

Tests of Between-Subjects Effects

Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	18654117085[a]	5	3730823417.0	13.975	<.001
Intercept	1.627E+12	1	1.627E+12	6092.930	<.001
Geschlecht	7083557110.4	1	7083557110.4	26.534	<.001
Nation	10364185200	2	5182092599.9	19.411	<.001
Geschlecht * Nation	1206374774.9	2	603187387.44	2.259	.106
Error	78488155854	294	266966516.51		
Total	1.724E+12	300			
Corrected Total	97142272939	299			

a. R Squared = .192 (Adjusted R Squared = .178)

Abbildung 11.5. Teil der Ausgabe für eine zweifaktorielle Varianzanalyse mit den Faktoren *Geschlecht* (2 Stufen) und *Nationalität* (3 Stufen) und der AV *Einkommen*.

Ein diskreter Prädiktor und ein stetiger Prädiktor

Wir betrachten nun ein Regressionsmodell mit einem stetigen und einem diskreten Prädiktor. Dazu betrachten wir den Datensatz „Kap11daten3.sav“, der nun neben den Jahreseinkommen für das Jahr 2020 für die bisher betrachteten Populationen auch die entsprechenden Jahreseinkommen für die Jahre 2015, 2010 und 2005 enthält. Die Jahreseinkommen sind jeweils für die unterschiedliche Kaufkraft (für USD) in diesen Jahren bereinigt. D.h. die unterschiedlichen Jahreseinkommen sind nicht auf Veränderungen des Geldwerts selbst zurückzuführen.

Zu Illustrationszwecken möchten wir nun wissen, wie die mittleren Jahreseinkommen sich über die Zeit verändern und ob diese Veränderung unterschiedlich für die beiden betrachteten Geschlechter verläuft (der:die interessierte Leser:in kann sich gerne selbst zusätzlich noch ansehen, ob diese Veränderung unterschiedlich für die beiden Geschlechter und die drei untersuchten Nationen verläuft). Zwar könnten wir direkt mit dem Erhebungsjahr als Prädiktor arbeiten, allerdings würden wir in diesem Fall inhaltlich wenig sinnvolle Schätzwerte für den Achsenabschnitt erhalten, da dieser Wert dann dem mittleren Einkommen (von Männern und Frauen) im Jahre 0 entsprechen würde. Stattdessen möchten wir, dass unsere Achsenabschnitte dem mittleren Einkommen zu Beginn des Untersuchungszeitraums entsprechen sollen. Dafür können wir eine neue Variable erzeugen, indem wir vom Erhebungsjahr den Wert 2005 abziehen (= das am längsten zurückliegende Jahr im Datensatz). Eine solche Variable liegt im Datensatz bereits unter der Bezeichnung „Erhebungsjahr_seit_2005“ vor.

Neben diesem stetigen Prädiktor benötigen wir noch eine weitere Dummy-Variable, die zulässt, dass der Steigungsparameter des linearen Zusammenhangs zwischen Erhebungsjahr und mittlerem Einkommen sich zwischen den beiden untersuchten Geschlechtern unterscheidet. Dafür verwenden wir wieder dieselbe Vorgangsweise wie im vorhergehenden Abschnitt. Wir konstruieren zuerst die Dummy-Variable und machen uns im Anschluss klar, weshalb die so erzeugte Variable genau diese Funktion erfüllt.

Zur Erzeugung der Dummy-Variablen multiplizieren wir wieder unsere beiden Prädiktoren. D.h. wir multiplizieren die Variable *Geschlecht* (die bereits eine Dummy-Variable für die beiden betrachteten Geschlechtskategorien ist) mit der Variable *Erhebungsjahr_seit_2005*. Eine entsprechend

erzeugte Variable liegt im Datensatz bereits mit der nahezu unleserlich sperrigen Bezeichnung „Geschlecht_X_Erhebungsjahr_seit_2005“ vor. An der Bezeichnung lässt sich schon erkennen, dass es sich dabei um die Interaktion zwischen dem diskreten Prädiktor *Geschlecht* und dem stetigen Prädiktor *Erhebungsjahr_seit_2005* handeln soll. Die Bedeutung des Begriffs Interaktion bleibt dabei dieselbe wie schon bei Interaktionen zwischen diskreten Prädiktoren: die Wirkung der einen Variablen hängt von der Ausprägung der anderen Variablen ab. Interaktionen sind symmetrisch. D.h. für das vorliegende Beispiel kann diese Bedeutung auf zwei völlig gleichwertige Arten gelesen werden. Einerseits kann damit gefragt sein, wie sich das Erhebungsjahr auf die Abhängigkeit des Jahreseinkommens vom Geschlecht auswirkt. Andererseits kann damit aber auch gefragt sein, wie sich das Geschlecht auf die Abhängigkeit des Jahreseinkommens vom Erhebungsjahr auswirkt.

Diese Symmetrie ist auch am gesamten Regressionsmodell ersichtlich:

$$Y_i \sim N(\alpha + \beta_1 D_i + \beta_2 X_i + \beta_{12}(D_i \cdot X_i), \sigma^2).$$

Aufgrund der Kommutativität der Multiplikation spielt es dabei keine Rolle, ob im Ausdruck hinter dem Regressionsgewicht β_{12} für die Interaktion $D_i \cdot X_i$ oder $X_i \cdot D_i$ steht. Ferner kann man sich leicht davon überzeugen, dass auch beide sprachlichen Interpretationen von oben in diesem Regressionsmodell ihren Ausdruck finden. Einerseits kann der Ausdruck für den Mittelwert im Regressionsmodell wie folgt geschrieben werden:

$$\mu_i = E(Y_i|D_i = d_i, X_i = x_i) = \alpha + (\beta_1 + \beta_{12}x_i)d_i + \beta_2 x_i,$$

d.h., die Änderung des Erwartungswert für die AV mit der Dummy-Variablen $D_i = d_i$ hängt von der Ausprägung der stetigen Variablen $X_i = x_i$ ab. Dies entspricht der Frage von oben: Wie wirkt sich das Erhebungsjahr auf die Abhängigkeit des Jahreseinkommens vom Geschlecht aus?

Andererseits kann der Ausdruck für den Mittelwert im Regressionsmodell genauso wie folgt geschrieben werden:

$$\mu_i = E(Y_i|D_i = d_i, X_i = x_i) = \alpha + \beta_1 d_i + (\beta_2 + \beta_{12}d_i)x_i,$$

d.h., die Änderung des Erwartungswert für die AV mit der stetigen Variablen $X_i = x_i$ hängt von der Ausprägung der Dummy-Variablen $D_i = d_i$ ab. Dies entspricht der Frage von oben: Wie wirkt sich das Geschlecht auf die Abhängigkeit des Jahreseinkommens vom Erhebungsjahr aus?

Wie sehen die Antworten auf diese Fragen auf Grundlage unseres Datensatzes aus? Dazu fügen wir die Variablen *Geschlecht*, *Erhebungsjahr_seit_2005*, und *Geschlecht_X_Erhebungsjahr_seit_2005* allesamt als Prädiktoren in ein Regressionsmodell in SPSS unter *Analyze >> Regression >> Linear…* ein. Der für uns hier wesentliche Teil der Ausgabe ist in Abbildung 11.6 gezeigt.

Wir sehen, dass das mittlere Einkommen im ersten Erhebungsjahr 2005 für männliche Personen bei 68779.68 USD lag. Das mittlere Einkommen für weibliche Personen lag um 11301.66 USD (mit α = .005) signifikant darunter, $t(1196) = -6.68$, $p < .001$. Mit jedem Jahr seit 2005 nahm das mittlere Einkommen von männlichen Personen um 709.12 USD zu. Dieser Zuwachs unterscheidet sich (mit α = .005) signifikant von Null, $t(1196) = 5.54$, $p < .001$. Für weibliche Personen war der Zuwachs im Mittel um 32.96 USD geringer; dieser Unterschied ist nicht signifikant, $t(1196) = -0.18$, $p = .856$. Kurz und knapp bedeutet das in etwa: 2005 haben Frauen im Mittel pro Jahr in etwa 11000 USD weniger verdient als Männer und daran hat sich bis 2020 nicht viel verändert, auch wenn beide Geschlechter 2020 deutlich mehr pro Jahr verdienen (in etwa $15 \cdot 700 = 10500$ USD mehr als 2005). Relativ (zum Jahreseinkommen der Männer) ist die Differenz demnach kleiner geworden, während die Differenz in absoluter Kaufkraft sich kaum verändert hat.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	68779.678	1196.970		57.461	<.001	66431.283	71128.073
	Geschlecht (männlich, weiblich)	-11301.656	1692.771	-.300	-6.676	<.001	-14622.788	-7980.524
	Vergangene Jahre seit Beginn der im Datensatz enthaltenen Erhebungen	709.116	127.961	.211	5.542	<.001	458.062	960.170
	Interaktion zwischen Geschlecht und Erhebungsjahr (seit 2005)	-32.961	180.965	-.010	-.182	.856	-388.004	322.083

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

Abbildung 11.6. Ausgabe für ein Regressionsmodell mit einem diskreten und einem stetigen Prädiktor und deren Interaktion.

Ein Ergebnisbericht für die lineare Regressionsanalyse mit einem diskreten Prädiktor mit zwei Ausprägungen, einem stetigen Prädiktor und deren Interaktion für das oben erläuterte Beispiel könnte wie folgt aussehen: „Eine multiple lineare Regressionsanalyse ergab, dass ein (mit $\alpha = .005$) statistisch signifikanter Anteil der Varianz im Einkommen der untersuchten $n = 1200$ Personen durch die Prädiktoren Geschlecht, Erhebungsjahr seit 2005 und deren Interaktion aufgeklärt werden kann, $F(3, 1196) = 62.99$, $p < .001$, $R^2 = .14$; ein mittlerer Effekt gemäß Cohen (1988). Betrachtet man die einzelnen Regressionsparameter, verdienten Männer im Jahr 2005 im Mittel etwa 69 000 Euro jährlich ($b = 68779.68$, $t(1196) = 57.46$, $p < .001$). Im selben Jahr verdienten Frauen im Durchschnitt 11 000 Euro weniger ($b = -11301.66$, stand. $\beta = -.30$, $t(1196) = -6.68$, $p < .001$). Diese Differenz ist (mit $\alpha = .005$) statistisch signifikant.

Bei Männern erwartet man bei einem Anstieg um 1 Jahr einen Anstieg im mittleren Jahreseinkommen um etwa 700 Euro ($b = 709.12$, stand. $\beta = .21$, $t(1196) = 5.54$, $p < .001$). Dieser Anstieg ist (mit $\alpha = .005$) ebenfalls statistisch signifikant. Bei Frauen erwartet man Anstieg um 1 Jahr einen Anstieg im mittleren Jahreseinkommen um etwa 670 Euro. Der Unterschied im Anstieg des Einkommens pro Jahr seit Beginn des Erhebungszeitraums zwischen Männern und Frauen ist (mit $\alpha = .005$) statistisch nicht signifikant ($b = -32.96$, stand. $\beta = -.01$, $t(1196) = -0.18$, $p = .856$).“

Zwei stetige Prädiktoren

Als letzten Fall wird die Interaktion zwischen zwei stetigen Prädiktoren betrachtet. Dazu betrachten wir den Datensatz „Kap11daten4.sav“, der sich vom Datensatz im vorhergehenden Abschnitt lediglich dadurch unterscheidet, dass nun auch jährliche Einkommen für Männer und Frauen unterschiedlichen Alters zum Erhebungszeitpunkt vorliegen. Insgesamt werden drei Altersgruppen betrachtet, die hier zu Illustrationszwecken als kontinuierliche Variable aufgefasst werden, die nur mit einer sehr groben Skala (20 Jahre, 40 Jahre, 60 Jahre) gemessen wird.

Die Fragestellung, die wir in diesem Abschnitt erhellen wollen, lautet: Wie verändert sich das mittlere Einkommen über die Zeit hinweg und hängt diese Veränderung vom Alter der untersuchten Angestellten ab? Wir fragen also nach der Interaktion zwischen dem Erhebungsjahr und dem Alter der befragten Personen.

Auch dafür empfiehlt es sich, zuerst wieder das Erhebungsjahr so zu transformieren, dass der Achsenabschnitt das mittlere Einkommen um ersten Erhebungsjahr (d.h. 2005) angibt. Für das Alter bietet es sich an, dieses um das mittlere Alter von 40 Jahren zu zentrieren. Die Interaktion zwischen den beiden stetigen Variablen kann anschließend wieder durch eine neue Variable modelliert werden, die dem Produkt aus dem transformierten Erhebungsjahr und dem zentrierten Alter der befragten Personen entspricht. Alle drei Variablen sind bereits im Datensatz unter den Bezeichnungen „Erhebungsjahr_seit_2005“, „Alter_zentriert“ und „Alter_X_Erhebungsjahr“ enthalten.

Verwendung der drei Variablen *Alter_zentriert*, *Erhebungsjahr_seit_2005*, und *Alter_X_Erhebungsjahr* als Prädiktoren in einer Regressionsanalyse resultiert in der in Abbildung 11.7 gezeigten Ausgabe. Wir sehen, dass Angestellte mittleren Alters (40 Jahre) im ersten Jahr der Erhebungen (d.h. 2005) im Mittel ein Bruttoeinkommen von 56966.07 USD hatten. Für Angestellte dieses Alters bei den jeweiligen Erhebungen erhöhte sich das Einkommen mit jedem Jahr seit 2005 um 713.84 USD, was einer (mit $\alpha = .005$) signifikanten Erhöhung entspricht, $t(3596) = 13.09$, $p < .001$. Für das Erhebungsjahr 2005 erhöhte sich zudem das Einkommen mit jedem zusätzlichem Jahr des Lebensalters der Angestellten um 632.24 USD, was ebenfalls einer (mit $\alpha = .005$) signifikanten Erhöhung entspricht, $t(3596) = 20.24$, $p < .001$. Die Moderation der Veränderung des Einkommens mit dem Erhebungsjahr durch das Alter der befragten Personen zum jeweiligen Erhebungszeitpunkt ist ebenfalls signifikant, $t(3596) = 4.18$, $p < .001$.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	56966.068	510.215		111.651	<.001	55965.728	57966.409
	Vergangene Jahre seit Beginn der im Datensatz enthaltenen Erhebungen	713.839	54.544	.179	13.087	<.001	606.898	820.780
	Um 40 Jahre zentriertes Alter	632.239	31.244	.463	20.235	<.001	570.980	693.497
	Interaktion zwischen (zentriertem) Alter und Erhebungsjahr (seit 2005)	13.974	3.340	.096	4.184	<.001	7.425	20.523

a. Dependent Variable: Jährliches Bruttoeinkommen in USD (inflationsbereinigt)

Abbildung 11.7. Ausgabe für eine Regressionsanalyse mit zwei stetigen Prädiktoren und deren Interaktion.

Die Auswirkung des Lebensalters zum Zeitpunkt der Befragung auf die Veränderung des Einkommens mit dem Erhebungsjahr kann eventuell am einfachsten dadurch illustriert werden, dass einige bestimmte Ausprägungen der Moderatorvariable (hier wegen der Fragestellung das Lebensalter der befragten Personen) herausgegriffen werden und der lineare Zusammenhang zwischen dem (anderen) Prädiktor und typischen Ausprägungen der AV jeweils für diese bestimmten Werte angegeben wird. Im vorliegenden Beispiel wird dafür der lineare Zusammenhang zwischen Erhebungsjahr und mittlerem Einkommen für 20-, 40- und 60-Jährige angegeben.

Für 20-jährige erhöhte sich das mittlere Einkommen mit jedem Jahr seit 2005 im Mittel lediglich um $713.84 - 20 \cdot 13.97 = 434.44$ USD. Für 40-Jährige erhöhte sich das mittlere Einkommen mit jedem Jahr seit 2005 im Mittel hingegen um 713.84 USD, wie oben bereits angegeben. Für 60-Jährige erhöhte sich das mittlere Einkommen mit jedem Jahr seit 2005 im Mittel um $713.84 + 20 \cdot 13.97 = 993{,}24$ USD. D.h. mit jedem zusätzlichen Lebensjahr stieg die Erhöhung des mittleren Einkommens mit jedem Jahr seit 2005 um zusätzliche 13.97 USD an.

Übungsaufgaben

Die im Folgenden eventuell benötigten Datensätze finden Sie im elektronischen Ergänzungsmaterial zu diesem Dokument, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Beispiel 11.1

Welche Aussage/n zu diskreten Prädiktoren in der linearen Regressionsanalyse trifft/treffen zu?

(a) Bei der Dummy-Kodierung fungiert die Kategorie, die mit 0 kodiert wurde, als Referenzkategorie.

(b) Für eine kategoriale Variablen mit insgesamt k Kategorien, müssen k Dummy-Variablen definiert werden.

(c) Der Steigungsparameter einer Dummy-Variablen bildet den Mittelwert der Kategorie ab, die mit 1 kodiert wurde.

(d) Die Teststatistik eines Student'schen t-Tests entspricht der Teststatistik des Steigungsparameters in einer linearen Regressionsanalyse mit Dummy-Kodierung für (exakt) einen dichotomen Prädiktor.

Beispiel 11.2

Welche Aussage/n zu Interaktionseffekten in der linearen Regressionsanalyse trifft/treffen zu?

(a) Mit einem Interaktionseffekt kann beschrieben werden wie sich der Zusammenhang zwischen einem Prädiktor und dem Kriterium je nach Ausprägung eines anderen Prädiktors ändert.

(b) In der linearen Regressionsanalyse kann ein Interaktionseffekt hinzugefügt werden, indem man das Produkt der interessierenden unabhängigen Variablen als weiteren Prädiktor in das Modell aufnimmt.

(c) In einem linearen Regressionsmodell mit einem dichotomen dummy-kodierten Prädiktor X_1, einem stetigen Prädiktor X_2 und deren Interaktion, beschreibt der Steigungsparameter des Interaktionseffekts wie sich der Zusammenhang zwischen X_2 und dem Kriterium Y ändert, wenn X_1 von 0 auf 1 steigt.

(d) In einem linearen Regressionsmodell mit zwei dichotomen dummy-kodierten Prädiktoren X_1, X_2 und deren Interaktion, beschreibt der Achsenabschnitt die mittlere Ausprägung in Y, wenn X_1 den Wert 0 und X_2 den Wert 1 hat.

Beispiel 11.3

Eine Freundin, die bei der Suchtpräventionsstelle arbeitet, bittet Sie, ihr zu zeigen, wie man mittels linearer Regressionsanalyse überprüfen kann, ob sich zwei Gruppen im Mittel unterscheiden. In ihrer Studie möchte sie vergleichen, ob Personen mit und ohne Spielsucht sich hinsichtlich des mittleren Restgeldbetrags unterscheiden, den sie nach einem Besuch ins Casino übrighaben. Ist der Restgeldbetrag positiv, haben die Personen beim Casinobesuch Geld dazugewonnen (relativ zu dem, was sie ausgeben wollten), ist der Restgeldbetrag negativ, haben die Personen Geld verloren.

Berechnen Sie eine lineare Regressionsanalyse mit dem diskreten Prädiktor *addiction* (0 = keine Spielsucht, 1 = Spielsucht) und dem Kriterium *balance_pre* und schreiben Sie einen Ergebnisbericht. Verwenden Sie dafür den Datensatz „Kap11UE3.sav“ und ein Signifikanzniveau von 0.5%. <u>*Hinweis*</u>: Für dieses (fiktive) Beispiel können Sie davon ausgehen, dass die für die lineare Regression notwendigen Annahmen allesamt erfüllt sind.

Beispiel 11.4

Öffnen Sie die Datei „Kap11UE4.sav“. In diesem (fiktiven) Datensatz, wurde Personen ein Frustrationstoleranzfragebogen und ein ADHS-Fragebogen vorgegeben. Zusätzlich wurde ermittelt, ob die Personen eine offizielle ADHS Diagnose vorliegen haben und wenn ja, welches ADHS Medikament (Ritalin oder Adderall) sie nehmen. Ihr Kollege möchte diese Daten verwenden, um einige Analysen in SPSS zu rechnen. Weil er sich mit SPSS allerdings nicht gut auskennt, bittet er Sie einige Variablen für ihn zu transformieren:

(a) Die Variable ADHS Diagnose (diagnosis) soll dummy-kodiert werden, wobei die Referenzkategorie keine ADHS Diagnose sein soll.
(b) Die Variable ADHS Medikation (medication) soll ebenfalls in Dummy-kodierte Variablen umgewandelt werden, wobei keine Medikation die Referenzkategorie sein soll.
(c) Die Variable Frustrationstoleranz (tolerance) soll zentriert werden.

Ferner erzählt Ihr Kollege Ihnen, dass er das folgende lineare Regressionsmodell verwenden möchte: $\widehat{adhd}_i = \hat{\beta}_0 + \hat{\beta}_1 diagnosis_i + \hat{\beta}_2 tolerance_cent_i + \hat{\beta}_3 diagnosis_i tolerance_cent_i$.

Berechnen Sie die dafür notwendige Interaktionsvariable.

Beispiel 11.5

In der Studie der Suchtpräventionsstelle (Beispiel 11.3) wurde außerdem eine Intervention durchgeführt, bei der den Teilnehmer*innen Übungen zur Impulskontrolle gezeigt wurden. Der Datensatz „Kap11UE3.sav" beinhaltet u.a. Informationen über den Restgeldbetrag nach dem letzten Casinobesuch vor der Intervention (*balance_pre*), den Restgeldbetrag nach dem ersten Casinobesuch nach der Intervention (*balance_post*) und über das Vorliegen einer Spielsucht (*addiction*). Die Suchtpräventionsstelle ist daran interessiert, ob der Zusammenhang des Restgeldbetrags vor und nach der Intervention unterschiedlich ist, je nachdem ob die Person spielsüchtig ist oder nicht.

Berechnen Sie, um diese Fragestellung zu beantworten, eine lineare Regressionsanalyse zur Vorhersage von *balance_post*, mit dem diskreten Prädiktor *addiction*, dem stetigen Prädiktor *balance*_pre und deren Interaktion *addict_X_balance_pre* und schreiben Sie dafür einen Ergebnisbericht. Verwenden Sie ein Signifikanzniveau von α = .005. <u>*Hinweis*</u>: Für dieses (fiktive) Beispiel können Sie wiederum davon ausgehen, dass die für die lineare Regression notwendigen Annahmen allesamt erfüllt sind.

Beispiel 11.6

Öffnen Sie die Datei „Kap11UE6.sav". Dieser ist eine Erweiterung zum (fiktiven) Datensatz in Beispiel 11.4, in dem Personen ein Frustrationstoleranzfragebogen und ein ADHS-Fragebogen vorgegeben wurde. Zusätzlich wurde ermittelt, ob die Personen eine offizielle ADHS Diagnose vorliegen haben. Die erweiterte Datei „Kap11UE6.sav" beinhaltet außerdem eine mit 0 und 1 kodierte Dummy-Variable der ADHS Diagnose (NOvsDIAG), die zentrierte Variable Frustrationstoleranz (c_tolerance) und deren Produkt (NOvsDIAG_X_ c_tol).

Verwenden Sie das folgende Regressionsmodell zur Schätzung der Regressionskoeffizienten:

$$\widehat{adhd}_i = \hat{\beta}_0 + \hat{\beta}_1 NOvsDIAG_i + \hat{\beta}_2 c_tolerance_i + \hat{\beta}_3 NOvsDIAG_X_c_tol_i$$

Erstellen Sie im Anschluss einen Ergebnisbericht und verwenden Sie ein Signifikanzniveau von 0.5%. In dem Ergebnisbericht soll, neben APA-Richtlinien konformer Berichterstattung der statistischen Kennwerte, auch explizit beschrieben werden wie der Interaktionseffekt zu interpretieren ist. <u>*Hinweis*</u>: Für dieses (fiktive) Beispiel können Sie wieder davon ausgehen, dass die für die lineare Regression notwendigen Annahmen allesamt erfüllt sind.

Beispiel 11.7

Der Datensatz „Kap11UE6.sav“ enthält neben der Information über eine ADHS-Diagnose auch Information darüber, welches Medikament (Ritalin oder Adderall) eingenommen wird, sofern eine Diagnose vorliegt. Die Variable Medikament hat daher 3 Ausprägungen: kein Medikament, Ritalin, Adderall. Für diese kategoriale Variable wurden die entsprechenden Dummy-Variablen erstellt (*NONEvsRIT*, *NONEvsADD*) und jeweils die Interaktionsvariablen mit der zentrierten Frustrationstoleranz berechnet (*NONEvsRIT_X_c_tol*, *NONEvsADD_X_c_tol*).

Verwenden Sie ein lineares Regressionsmodell, in dem der ADHS-Score (*adhd*) durch die unabhängigen Variablen Frustrationstoleranz (c_tolerance), Medikament (*NONEvsRIT*, *NONEvsADD*) und deren Interaktion vorhergesagt wird, zur Schätzung der Regressionskoeffizienten.

Erstellen Sie im Anschluss einen Ergebnisbericht und verwenden Sie ein Signifikanzniveau von $\alpha = .05$. In dem Ergebnisbericht soll, neben APA-Richtlinien konformer Berichterstattung der statistischen Kennwerte, auch explizit beschrieben werden wie der Interaktionseffekt zu interpretieren ist. <u>*Hinweis*</u>: Für dieses (fiktive) Beispiel können Sie wieder davon ausgehen, dass die für die lineare Regression notwendigen Annahmen allesamt erfüllt sind.

Beispiel 11.8

In Kapitel 5 wurde ein unabhängiger t-Test verwendet, um der Frage nachzugehen, ob sich der IQ von (fiktiven) Psychologiestudierenden im Mittel von den (fiktiven) BWL-Studierenden unterscheidet. Der dafür herangezogene Datensatz ist in der Datei „Kap5daten2.sav“ zu finden. Die Abbildung 5.9 zeigt die Ausgabe für eben jenen t-Test.

Anstelle des Student’schen t-Tests hätte man hier ebenfalls eine lineare Regressionsanalyse mit dem dichotomen Prädiktor Studiengang (*Gruppe* mit 0 = Psychologie & 1 = BWL) und dem Kriterium *IQ* berechnen können:

$$\widehat{IQ}_i = \hat{\beta}_0 + \hat{\beta}_1 Gruppe_i.$$

(a) Vervollständigen Sie basierend auf den Ergebnissen des Student’schen t-Tests in Abbildung 5.9 die folgende Ergebnistabelle für das eben angeführte lineare Regressionsmodell:

	Unstandardized b	*t*	Sig.
(Constant)		53.67	<.001
Gruppe			

(b) Überprüfen Sie Ihre Ergebnistabelle, indem Sie sich die lineare Regressionsanalyse ausgeben lassen und schreiben Sie einen Ergebnisbericht, bei dem Sie ein Signifikanzniveau von 5% verwenden.

Hinweis: Für dieses (fiktive) Beispiel können Sie wiederum davon ausgehen, dass die für die lineare Regression notwendigen Annahmen allesamt erfüllt sind.

Beispiel 11.9

Der Datensatz „tulips.sav“ beinhaltet Informationen über die Größe von Tulpenblüten (Variable: *blooms*), je nach Feuchtigkeit der Erde (wenig (1) bis viel (3) Wasser, Variable: *water*) und Beschattung (niedrige (1) bis hohe (3) Beschattung, Variable: *shade*). Sie möchten herausfinden, ob die Feuchtigkeit und die Belichtung eigenständig, wie auch in Interaktion miteinander, die Größe der Tulpenblüten beeinflussen können.

(a) Bevor Sie die Analyse durchführen, ist es in diesem Fall hilfreich die beiden Variablen *water* und *shade* zu zentrieren, da so der Wert 0 eine mittlere Feuchtigkeit bzw. mittlere Beschattung darstellt.

(b) Erzeugen Sie dann die benötigte Interaktionsvariable aus den beiden zentrierten Prädiktoren.

(c) Führen Sie im Anschluss eine multiple lineare Regressionsanalyse mit Interaktionsterm durch und schreiben Sie einen Ergebnisbericht. Verwenden Sie dafür ein Signifikanzniveau von 0.5%.

Anmerkung: Leider ist nicht bekannt in welcher Einheit die Größe der Tulpenblüten in diesem Datensatz gemessen wurde, daher ist die Interpretation im Ergebnisbericht etwas erschwert.

Beispiel 11.10

Erstellen Sie für das Regressionsmodell aus Beispiel 11.9 eine entsprechende Grafik. Auf der Grafik soll zu sehen sein wie sich der Zusammenhang zwischen Feuchtigkeit und Blütengröße je nach Beschattung (1 = niedrige Beschattung, 2 = mittlere Beschattung, 3 = hohe Beschattung) verändert.

Verwenden Sie dafür ein Streudiagramm bei dem die Datenpunkte je nach Beschattung eingefärbt sein sollen. Weiters soll für die drei Untergruppen der Beschattung eine Regressionsgerade (Linear Fit Line) angezeigt werden.

Hinweis: SPSS kann im Chart-Builder nur andere Farben für Datenpunkte setzen, wenn die entsprechende Variable in der SPSS-Datendatei als nominal-skaliert klassifiziert ist. Verwenden Sie daher eine nominale skalierte Variable für das Ausmaß der Beschattung.

Lösungen zu den Übungsaufgaben

Lösungen der Übungsaufgaben zu Kapitel 1

Beispiel 1.1

Richtig: (a), (b), (c). Falsch: (d).

Beispiel 1.2

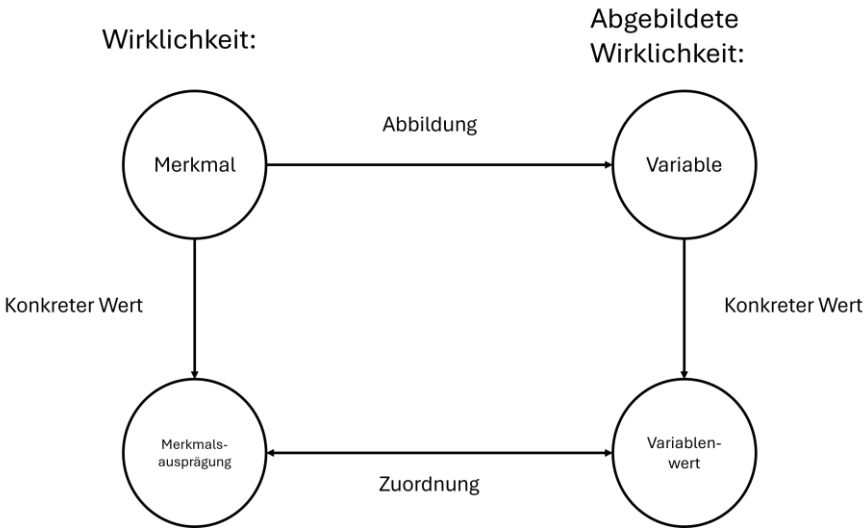


Abbildung L.1. Lösung zu Beispiel 1.2.

Beispiel 1.3

Begriffe	Synonyme
Merkmalsträger:in	Untersuchungseinheit, Untersuchungsobjekt
Merkmalsausprägung	Merkmalswert
Variablenwert	Messwert

Beispiel 1.4

Richtig: (a), (c). Falsch: (b), (d).

Beispiel 1.5

Skalenniveau	Beispiele
Nominalskalenniveau	Erkrankung (Ja/Nein); Haarfarbe; Geschlecht
Ordinalskalenniveau	Wettkampfplatzierung; Schulabschluss
Intervallskalenniveau	Temperatur
Verhältnisskalenniveau	Größe, Gewicht
Absolutskalenniveau	Anzahl (an Fehltagen, Büchern etc.)

Beispiel 1.6

Richtig: (b), (c). Falsch: (a), (d).

Beispiel 1.7

Diskret: Studienfach, Augenfarbe, Anzahl an Fehltagen. Kontinuierlich: Größe, Gewicht, Temperatur.

Beispiel 1.8

Richtig: (c), (d). Falsch: (a), (b).

Beispiel 1.9

Z.B.: Die Anzahl an konkreten Übungsbeispielen, die Schüler:innen im Mathematikunterricht bearbeiten, wirkt sich positiv auf die Mathematikabschlussnote aus.

Beispiel 1.10

Z.B.: Je mehr Bücher Personen zu Hause haben, desto geringer ist das Merkmal Extraversion dieser Personen ausgeprägt.

Beispiel 1.11

Mit Urliste wird die ungeordnete tabellarische Darstellung von Daten bezeichnet. Unter Daten werden hierbei die Ergebnisse von Beobachtungen, Fragebögen, psychologischen Tests oder physikalischen Messinstrumenten bezeichnet, mit deren Hilfe die Ausprägung von Merkmalen untersuchter Merkmalsträger abgebildet werden soll.

Beispiel 1.12

Richtig: (d). Falsch: (a), (b), (c).

Beispiel 1.13

(a) $H_{\text{kum}}(x_4 = 12) = \sum_{j=1}^{4} H(x_j) = H(x_1) + H(x_2) + H(x_3) + H(x_4) = 1 + 2 + 4 + 1 = 8.$

(b) $h_{\text{kum}}(x_4 = 12) = \frac{1}{n}\sum_{j=1}^{4} H(x_j) = \frac{1}{n}[H(x_1) + H(x_2) + H(x_3) + H(x_4)] = \frac{1}{50}(1 + 2 + 4 + 1) = 8/50 = 0.16.$

(c) $h_{\text{kum}}(x_j < 10) = \frac{1}{n}\sum_{j=1}^{2} H(x_j) = \frac{1}{n}[H(x_1) + H(x_2)] = \frac{1}{50}(1 + 2) = \frac{3}{50} = 0.06 = 0.06 \cdot 1 = 0.06 \cdot \frac{100}{100} = (0.06 \cdot 100) \cdot \frac{1}{100} = 6 \cdot \frac{1}{100} = 6\%$. (Die Ausformulierung der letzten Teilschritte dient lediglich der Illustration der Bedeutung des Symbols „%“ für „pro zent“, d.h. wortwörtlich für „in Einheiten von $\frac{1}{100}$“.)

Beispiel 1.14

Anzahl Liegestütz	Absolute Häufigkeit	Relative Häufigkeit	Absolute kumulierte Häufigkeit	Relative kumulierte Häufigkeit
5	1	0.02	1	0.02
9	2	0.04	3	0.06
11	4	0.08	7	0.14
12	1	0.02	8	0.16
…	…		…	…

Beispiel 1.15

Richtig: (a), (b), (c). Falsch: (d).

Lösungen der Übungsaufgaben zu Kapitel 2

Die Lösungen für die Übungsaufgaben dieses Kapitels liegen hauptsächlich in Form elektronischer Dateien vor, die Sie allesamt im elektronischen Ergänzungsmaterial zu diesem Dokument finden, das Sie unter https://osf.io/9tcx3/ herunterladen können.

Beispiel 2.1

Schritt für Schritt in Kapitel 2 erklärt. Syntaxdatei: siehe „Erste_Syntaxdatei. sps“.

Beispiel 2.2

Richtig: (b), (d). Falsch: (a), (c).

Beispiel 2.3

Richtig: (a), (c). Falsch: (b), (d).

Beispiel 2.4

Alle falsch.

Beispiel 2.5

Antworten:

(a) 10.

(b) 5.

(c) 11.

(d) 26 Jahre. 179 cm. Ja.

(e) In cm. In kg.

Beispiel 2.6

Siehe Datei „Kap2UE6.sav“.

Beispiel 2.7

Siehe Datei „Kap2UE7.sav“.

Beispiel 2.8

Siehe Datei „Kap2UE8.sav“.

Beispiel 2.9

Siehe Datei „Kap2UE9.sav“.

Beispiel 2.10

Siehe Datei „Kap2UE10.sps“.

Beispiel 2.11

Siehe Datei „Kap2UE11.sav“.

Beispiel 2.12

Siehe Datei „Kap2UE12.sav“.

Beispiel 2.13

Am kleinen roten Plus-Symbol über dem SPSS-ICON ganz oben links in der Ecke.

Lösungen der Übungsaufgaben zu Kapitel 3

Beispiel 3.1

Die Variablen *mathe_mathe2* und *mathe_mathe3* sind jeweils umzukodieren. Um die Umpolung durchzuführen wählen wir *Transform >> Recode into Different Variables…* und klicken dort mit der rechten Maustaste in das linke Feld, um uns die Variablennamen anzeigen zu lassen. Daraufhin wählen wir beide Variablen aus und schieben sie in das mittlere Feld. Daraufhin markieren wir die erste der beiden Zuweisungen, d.h. „mathe_mathe2 --> ?“ im mittleren Feld und tragen bei „Name“ den Namen der umkodierten Variable ein, z.B. „mathe_mathe2_umk“. Unter Label tragen wir ein: „Umkodierung des Items ‚Ich hasse Statistik‘“. Dann klicken wir auf „Change“ und dann auf „Old and New Values…“.

Im sich öffnenden Fenster tragen wir zuerst die Zahl 1 unter „Value“ bei „Old Value“ ein und die Zahl 5 unter „Value“ bei „New Value“. Der Grund für diese beide Zahlen ist, dass die Skala der ursprünglichen Variablen von 1 bis 5 geht. Würde die Skala von 0 bis 6 gehen, hätten wir die Zahlen 0 und 6 eingetragen. Daraufhin klicken wir auf „Add“. Nun tragen wir die Zahl 2 unter „Value“ bei „Old Value“ ein und die Zahl 4 unter „Value“ bei „New Value“ und klicken wieder auf „Add“. Nun tragen wir die Zahl 3 unter „Value“ bei „Old Value“ ein und die Zahl 3 unter „Value“ bei „New Value“ und klicken wieder auf „Add“ (diesen Schritt könnten wir uns strenggenommen auch sparen). Nun tragen wir die Zahl 4 unter „Value“ bei „Old Value“ ein und die Zahl 2 unter „Value“ bei „New Value“ und klicken wieder auf „Add“. Schließlich tragen wir die Zahl 5 unter „Value“ bei „Old Value“ ein und die Zahl 1 unter „Value“ bei „New Value“ und klicken wieder auf „Add“. Dann klicken wir auf „Continue“.

Nun klicken wir auf die zweite Zuweisung, d.h. auf „mathe_mathe3 --> ?“ im mittleren Feld und gehen ganz analog zur vorhergehenden Variable vor. Im Fenster „Old and New Values“ sehen wir, dass die passenden Zuweisungen bereits eingetragen sind, hier ist also in diesem Fall nichts mehr zu tun (Vorsicht aber falls zwei Items unterschiedliche Skalen haben; dann müssen hier die entsprechenden Anpassungen vorgenommen werden).

Wenn wir mit den Einstellungen für die Umkodierung der zweiten Variable fertig sind, klicken wir auf „Paste“ für unsere Dokumentation. Es öffnet sich eine Syntaxdatei, in der wir die beiden Zeilen „* Kapitel 3, Beispiel 3.1.“ und „* Umkodierung der Items mathe_mathe2 und mathe_mathe3.“

Ergänzen. Danach speichern wir die Syntaxdatei gleich einmal unter der Bezeichnung „Kap3UE1.sps“ ab. Schließlich führen wir die Kommandozeilen in der Syntaxdatei aus, indem wir sie markieren und auf das grüne „Abspielen“-Symbol klicken.

Daraufhin sehen wir, dass zwei neue Variablen in unserem Datensatz hinzugekommen sind. In der Variablenansicht nehmen wir noch sämtliche Einstellungen für die neuen Variablen vor, die noch nicht passen. Dann überprüfen wir, ob die Umkodierung richtig vonstattenging, indem wir in der Datenansicht jeweils die Variablen *mathe_mathe2* und *mathe_mathe3* mit den Variablen *mathe_mathe2_umk* und *mathe_mathe3_umk* vergleichen. Sofern alles richtig aussieht, speichern wir die neue Datendatei unter der neuen Bezeichnung „Kap3daten_bearbeitet_UE1.sav“ ab.

Beispiel 3.2

Zur Bildung einer entsprechenden Summenskala gehen wir wie folgt vor. Unter Transform >> Compute Variable… tragen wir zuerst als Bezeichnung für unsere Summenskala „Affinität_Mathe_Statistik“ unter „Target Variable“ ein. Daraufhin klicken wir rechts im linken Feld mit allen Variablen und lassen uns wieder die Variablennamen anstelle der Labels anzeigen. Nun klicken doppelt mit der linken Maustaste auf die Variable *mathe_mathe1* und ergänzen anschließend hinter der gerade eingefügten Variablen ein „+“ im Feld „Numeric Expression“. Daraufhin klicken wir doppelt auf die Variable *mathe_mathe2_umk*, fügen wiederum ein „+“ nach der hinzugefügten Variable ein und klicken schließlich noch doppelt auf die Variable *mathe_mathe3_umk*. Danach fügen wir mit „Paste“ wieder alles in unsere Syntaxdatei ein, ergänzen diese um eine entsprechende Kommentarzeile und speichern Sie unter der neuen Bezeichnung „Kap3UE2.sps“ ab. Dann führen wir die neuen Kommandozeilen aus und überprüfen anhand einiger Personen in der Datenansicht, ob die Bildung der Summenskala wie gewünscht funktioniert hat. Daraufhin speichern wir die Datendatei unter der neuen Bezeichnung „Kap3daten_bearbeitet_UE2.sav“ ab.

Beispiel 3.3

Bei der in Beispiel 3.2 berechneten Summenskala handelt es sich um eine metrische Variable, die diskrete Werte zwischen 3 und 15 annehmen kann. Für eine metrische Variable sind jedenfalls die Angabe der typischen Ausprägung, der Streuung, sowie Minimum und Maximum informativ. Letztere

dienen insbesondere dazu, zu überprüfen, ob sich alle Werte im erlaubten Bereich für diese Variable befinden. Ist dem nicht so, dann ist das ein Hinweis auf Fehler bei der Dateneingabe oder Fehler bei der Berechnung der Summenskala. Zudem lassen wir uns noch eine Häufigkeitstabelle sowie ein Histogramm und ein Boxplot für die Summenskala ausgeben.

Für die Maßzahlen und die Häufigkeitstabelle wählen wir erst *Analyze* >> *Descriptive Statistics* >> *Frequencies…* und dort unsere Summenskala *Affinität_Mathe_Statistik* aus. Unter „Statistics…" wählen wir den Mittelwert (Mean), den Median, die Standardabweichung (Std. deviation), das Minimum, das Maximum, die Schiefe (Skewness) sowie die Wölbung (Kurtosis) aus. Unter „Charts…" wählen wir Histogramm aus und lassen uns auch eine Normalverteilungskurve für das Histogramm anzeigen. Nach Einfügen der entsprechenden Kommandozeilen in die Syntaxdatei lassen wir uns unter Graphs >> Chart Builder… ein Boxplot für unsere Summenskala ausgeben. Ausführen aller neuen Kommandozeilen generiert eine Ausgabe, die wir unter der Bezeichnung „Kap3UE3.spv" abspeichern.

Beispiel 3.4

Wir sollen die Variable *statistikschmerzen* umkodieren. Dazu können wir prinzipiell ganz analog zu Beispiel 3.1 vorgehen, allerdings ist hier zu beachten, dass die ursprüngliche Variable auf einer Skala von 1 bis 10 zu beantworten war. Dies muss bei der Eingabe alter und neuer Werte unter Transform >> Recode into Different Variables >> Old and New Values entsprechend berücksichtigt werden. Für das Ergebnis, siehe die Dateien „Kap3UE4.sps" sowie „Kap3daten_bearbeitet_UE4.sav".

Beispiel 3.5

Nun sollen wir eine Mittelwertskala aus den drei Items *statistikliebe*, *mathematikliebe*, und dem aus *statistikschmerzen* umkodierten Item generieren. Hierzu kann wiederum analog zu Beispiel 3.2 vorgegangen werden, allerdings ist nun unter „Numeric Expression" im Menü Transform >> Compute Variable… die Mittelwertsfunktion aus der Funktionsgruppe Statistik mit den entsprechenden drei Argumenten zu wählen. Dazu kann entweder die Funktion aus den Feldern rechts ausgewählt und die drei Variablen eingefügt werden oder es kann gleich „mean(statistikliebe, mathematikliebe, statistikschmerzen_umk)" im Feld „Numeric Expression" eingegeben werden. Für das Ergebnis, siehe die Dateien „Kap3UE5.sps" sowie „Kap3daten_bearbeitet_UE5.sav".

Beispiel 3.6

Dieses Beispiel kann völlig analog zu Beispiel 3.3 gelöst werden. Für das Ergebnis, siehe „Kap3UE6.spv“ sowie „Kap3UE6.sps“. Interessant ist bei diesem Beispiel, dass wir im Boxplot einen Ausreißer haben. Die Person mit dem Code 025 scheint so wirklich gar nichts für Statistik übrig zu haben.

Beispiel 3.7

51 Personen wurden zu ihrem Lieblingsfach unter den Hauptfächern beim Schulabschluss befragt. Mit knapp der Hälfte (49%; 25 von 51 Personen) gefällt das Hauptfach Englisch den befragten Personen am häufigsten. Mathematik belegt den zweiten Platz der beliebtesten Hauptfächer mit 29.4% (15 der 51 Personen). Auf Platz landet Deutsch mit 21.6% (11 von 51 Personen). Jede befragte Person hat ein Lieblingsfach angegeben. Abbildung L.2 zeigt die Verteilung der Hauptfächer unter den befragten Personen.

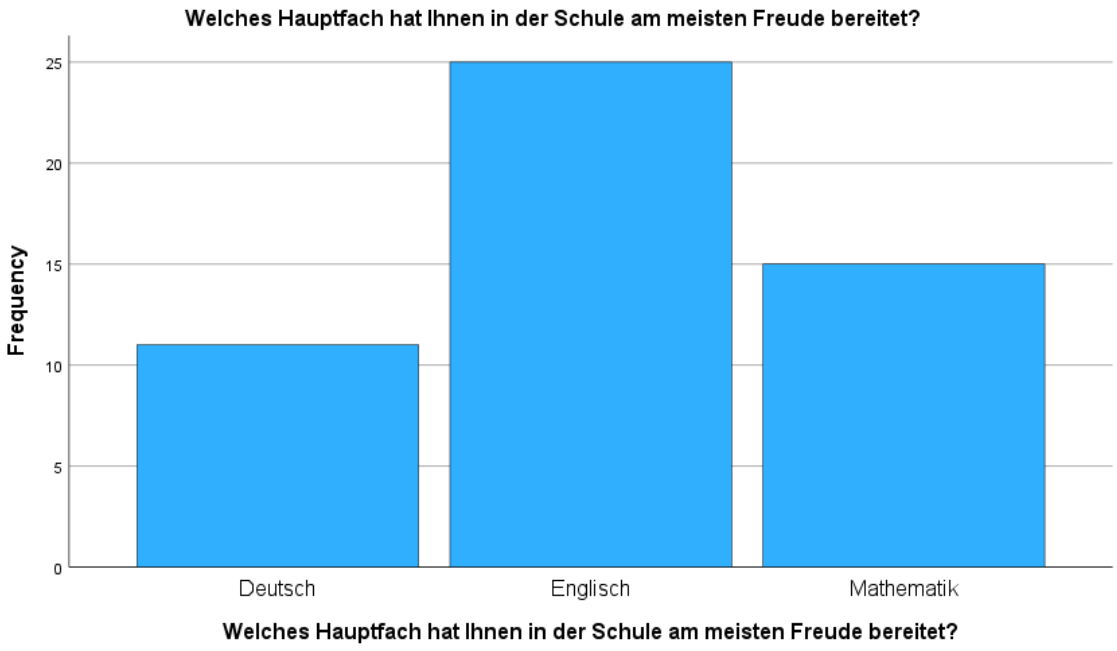


Abbildung L.2. Verteilung der Hauptfächer unter den 51 danach befragten Personen.

Beispiel 3.8

Von insgesamt 17 Leuten aus der Stichprobe, die derzeit in einer Beziehung sind, geben 14 an, aktuell auch verliebt zu sein und 3 wissen es nicht. Von 25 Leuten, die aktuell Single sind, geben hingegen nur 2 an, gerade verliebt zu, während 14 Personen nicht verliebt sind, 8 es nicht wissen und 1 Person meint, dass das die Durchführenden der Untersuchung gar nichts angehe. Es sieht also schon ein bisschen danach aus, als wären Leute in einer Beziehung eher verliebt als Leute, die Single sind. Um den Sachverhalt weiter aufzuklären, ist allerdings noch mehr Forschung nötig. Die gesamte Kreuztabelle ist in Tabelle L.1 gegeben.

Beispiel 3.9

Mittelwerte und Standardabweichungen für die drei Variablen sind in Tabelle L.2 zusammengefasst. Bei der Körpergröße und der Schuhgröße dürfte es das Problem geben, dass es sich mit hoher Wahrscheinlichkeit jeweils um bimodale Verteilungen handelt, da sich weibliche und männliche Befragte in diesen Variablen doch recht deutlich unterscheiden. Zumindest weisen die Histogramme und Häufigkeitstabellen darauf hin. Bei der Körpergröße dürfte es zwei Maxima bei etwa 165 und 175-180 cm geben, bei der Schuhgröße bei 38-39 und 43. Für weitere Details, siehe „Kap3UE9.spv“ sowie „Kap3UE9.sps“.

Tabelle L.1. Resultierende Kreuztabelle in Beispiel 3.8.

		Verliebt?				
		Ja	Nein	Weiß nicht	Privatsache	Total
Beziehung Status	Single	2	14	8	1	25
	In Beziehung	14	0	3	0	17
	Kompliziert	1	0	2	1	4
	Privatsache	0	0	0	5	5
	Total	17	14	13	7	51

Tabelle L.2. Mittelwerte und Standardabweichungen für die drei Variablen Alter, Körpergröße und Schuhgröße aus Beispiel 3.9.

	M	*SD*
Alter (in Jahren)	21.63	3.49
Körpergröße (in cm)	170.80	9.65
Schuhgröße (EU Format)	39.85	2.59

Beispiel 3.10

Siehe „Kap3UE10.spv“ sowie „Kap3UE10.sps“.

Beispiel 3.11

Für das Zusammenfügen der einzelnen Datendateien handelt es sich um den Fall „Hinzufügen neuer Fälle mit denselben Variablen“ aus Kapitel 2. Dazu können wir also wie folgt vorgehen, wir öffnen zuerst die beiden Dateien „Deutschland.sav“ und „Österreich.sav“ und kontrollieren, ob alle Variablen gleich definiert wurden. Daraufhin wählen wir in der Datei „Deutschland.sav“ Data >> Merge Files >> Add Cases… und wählen dort die bereits geöffnete Datei „Österreich.sav“ aus und klicken auf „Continue“. Sofern alle Variablen exakt gleich definiert waren, sollten wir keine Variablen im Feld „Unpaired Variables“ sehen und wir können gleich auf „OK“ klicken. Daraufhin werden die Fälle direkt in der Datei „Deutschland.sav“ hinzugefügt und es gibt dort jetzt 836 Fälle. Wir speichern diese Datei nun neu ab unter der Bezeichnung „Deutschland_Österreich.sav“. Wir können die Datei „Österreich.sav“ jetzt schließen und öffnen die Datei „Schweiz.sav“. Wir kontrollieren wiederum, ob alle Variablen gleich definiert wurden und fügen daraufhin die beiden Dateien ganz analog zu vorhin zusammen. Wir speichern den resultierenden Gesamtdatensatz, nun mit 1194 Fällen, unter der Bezeichnung „dach.sav“ ab.

Wie bereits an der der Datenansicht erkennbar, haben insgesamt 1194 Personen an der Befragung teilgenommen. Unter Analyze >> Descriptive Statistics >> Frequencies… können wir uns eine Häufigkeitstabelle für die drei Nationen ausgeben lassen. An dieser können wir ablesen, dass 423 Personen aus Deutschland teilgenommen haben, was einem Anteil von 35.4% an der Gesamtstichprobe entspricht. Aus Österreich haben 358 Personen teilgenommen, was 30% der Gesamtstichprobe

entspricht. In der Schweiz haben schließlich die verbleibenden 413 Personen teilgenommen, was 34.6% der Gesamtstichprobe entspricht.

Beispiel 3.12

Von den 1194 Befragten gaben 592 (49.6%) an, dass ihr Geschlecht „männlich" sei, 602 (50.4%) gaben an, dass ihr Geschlecht „weiblich" sei. Die Frage wurde von allen Teilnehmenden beantwortet.

Die Befragten waren zwischen 22 und 59 Jahre alt. Das Durchschnittsalter betrug $M = 36.29$ mit einer Standardabweichung von $SD = 4.78$. Der Median betrug $Mdn = 36$ Jahre. Eine Person gab kein Alter an.

Das Bildungsniveau entsprach bei 13 (1.1%) Befragten der Volks- oder Hauptschule, bei 287 (24.0%) der Fachschule oder einer höheren Schule ohne Matura oder einer Lehre, bei 198 (16.6%) der höheren Schule mit Matura oder Meisterprüfung oder Kolleg/Mat., und bei 689 (57.7%) der Universität, Fachhochschule oder Akademie. Von den Befragten machten 7 (0.6%) keine Angabe zum Bildungsniveau.

Bei der Variable *m_dur* handelt es sich um die Ehedauer in Monaten. Sieht man sich einige statistische Maßzahlen für diese Variable an, wird allerdings schnell klar, dass mit dieser etwas nicht stimmen kann. Mittelwert (M = 179.68 Monate) und Median (Mdn = 77.50 Monate) klaffen beispielweise sehr weit auseinander, was auf eine äußerst rechtsschiefe Verteilung hindeutet. Ebenso ist die Standardabweichung im Vergleich zum Mittelwert auffällig groß (SD = 292.37 Monate). An der Häufigkeitsverteilung und am Histogramm kann man schließlich erkennen, dass ein erheblicher Teil der Stichprobe (10.3% der Gesamtstichprobe; 11.1% der Stichprobe, bereinigt für fehlende und ungültige Werte) angeblich eine Ehedauer von 999 Monaten angegeben hat. Dies erscheint aus mehreren Gründen äußerst unplausibel. Erstens entspricht eine Ehedauer von 999 Monaten einer Ehedauer 83.25 Jahren. Das ist zwar prinzipiell nicht unmöglich, aber vermutlich nur äußerst selten der Fall. Zweitens war das maximale Alter der befragten Personen 59 Jahre, also geringer als eine Ehedauer von 999 Monaten. Daraus lässt sich durchaus schlussfolgern, dass es sich bei den angeblichen 999 Monaten um Fehleingaben handeln dürfte.

In der Variablenübersicht sehen wir, dass fehlende Werte bei unterschiedlichen Variablen mit der Zahl 99 definiert wurden (zusätzlich zum ganz gewöhnlichen Fehlen des entsprechenden Eintrags). Wenn wir nun eine weitere Variable erzeugen, die wir z.B. „ehedauer_neu“ nennen und die schlichtweg der alten Variable m_dur entspricht, nur dass für diese Variable nun die Zahl 999 (statt 99; eine Ehedauer von 99 Monaten ist ja durchaus plausibel) fehlende Werte anzeigt, kommen wir zu plausibleren Werten für die Ehedauer in Monaten. Es ergibt sich ein Mittelwert von M = 77.37 Monaten (etwa 6.5 Jahre) mit einer Standardabweichung von SD = 42.11 Monaten. Die Verteilung ist immer noch stark rechtsschief. Auch das ergibt durchaus Sinn. Nach unten hin ist die Ehedauer ja durch 0 begrenzt, nach oben gibt es viel mehr Spielraum und hin und wieder ist die Ehedauer auch sehr lang, siehe Abbildung L.3.

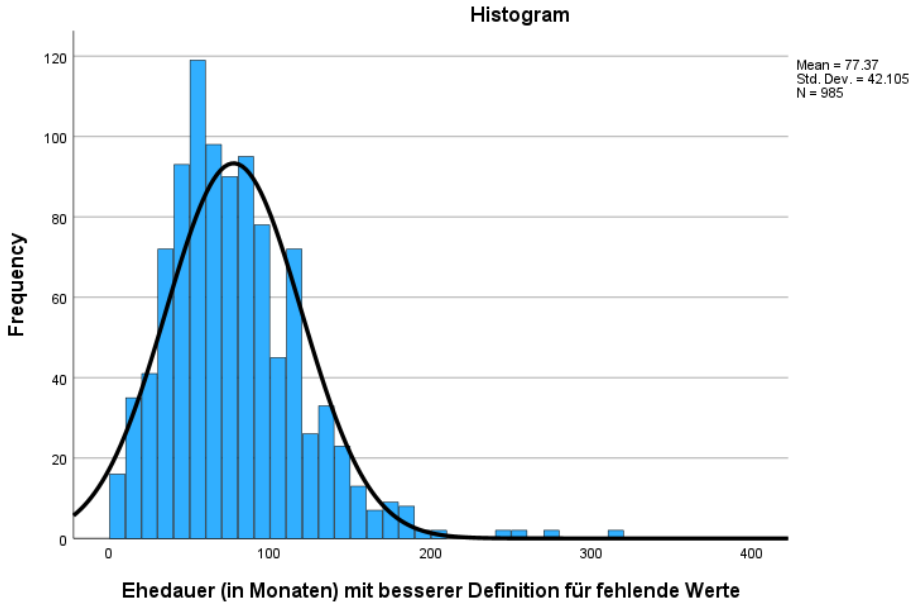


Abbildung L.3. Ein weitaus plausibleres Histogramm für die Ehedauer.

Beispiel 3.13

Dieses Beispiel lässt sich mit einer entsprechenden Kreuztabelle lösen, siehe „Kap3UE13.sps“. Inspektion derselben zeigt, dass in Deutschland 212 Männer und 211 Frauen befragt wurden, während es in der Schweiz 177 Männer und 181 Frauen und in Österreich 203 Männer und 210 Frauen waren.

Beispiel 3.14

Für justice_mean ergibt sich ein Mittelwert von $M = 4.61$ mit einer Standardabweichung von $SD = 1.07$.

Für justice_sum ergibt sich ein Mittelwert von $M = 9.21$ mit einer Standardabweichung von $SD = 2.17$.

Für Details, siehe die Dateien “Kap3UE14.spv” sowie “Kap3UE14.sps”.

Beispiel 3.15

Die Streudiagramme für die vier Variablenpaare sind in den Abbildungen L.4-7 gezeigt.

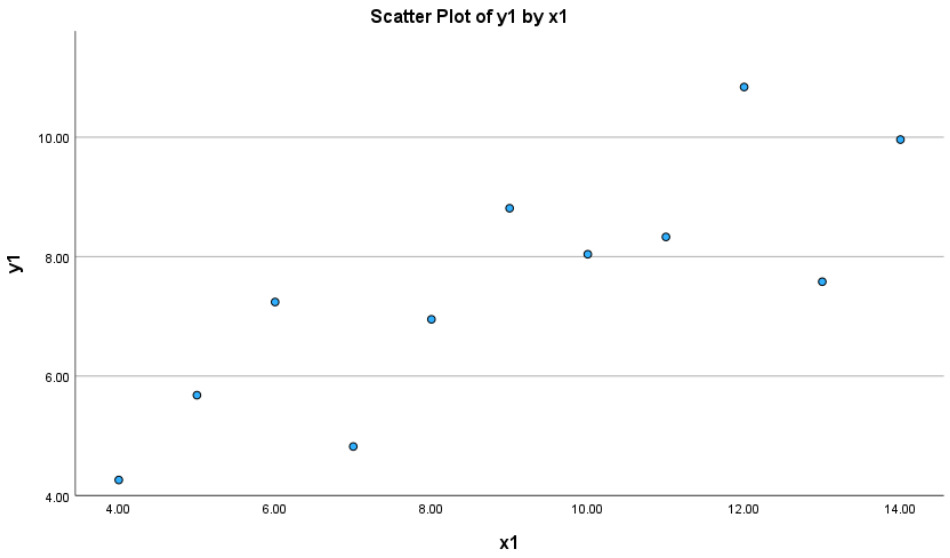


Abbildung L.4. Streudiagramm für das Variablenpaar (x_1, y_1) aus Übungsaufgabe 3.15.

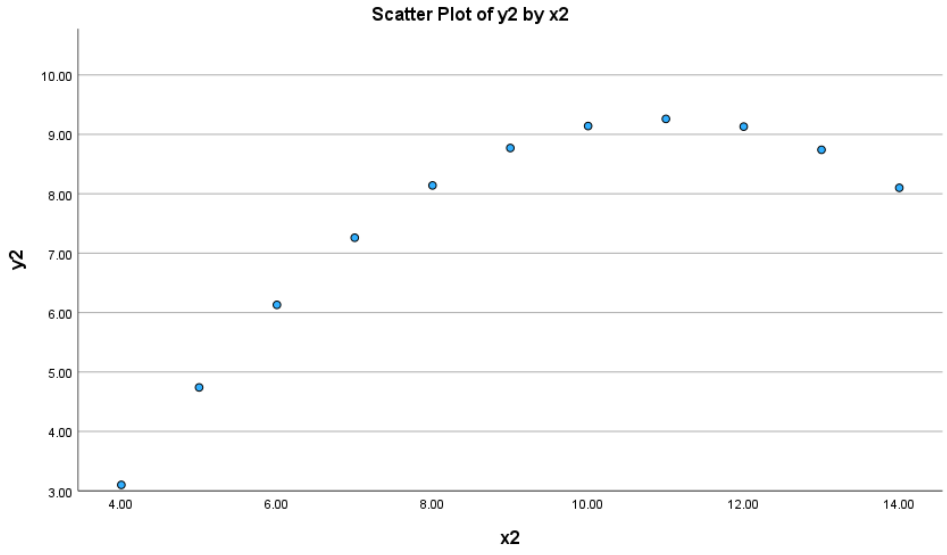


Abbildung L.5. Streudiagramm für das Variablenpaar (x_2, y_2) aus Übungsaufgabe 3.15.

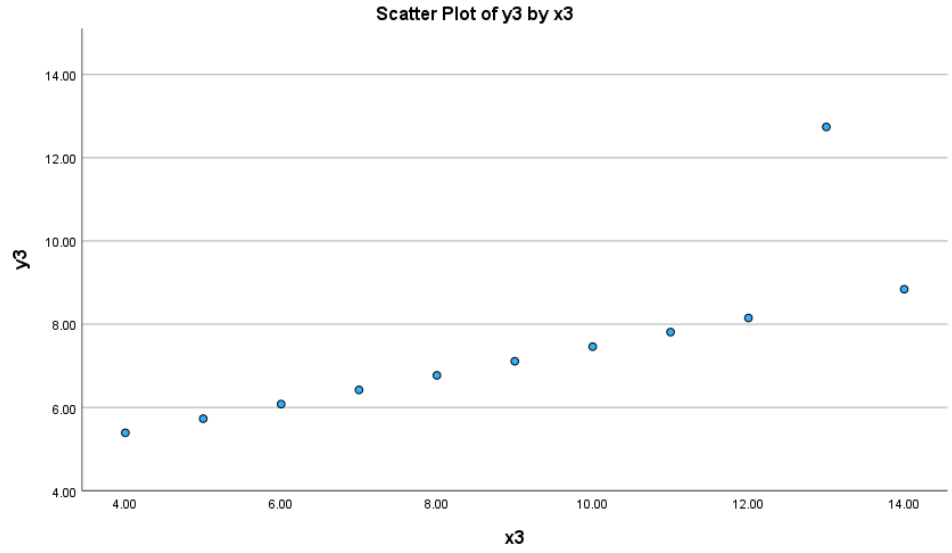


Abbildung L.6. Streudiagramm für das Variablenpaar (x_3, y_3) aus Übungsaufgabe 3.15.

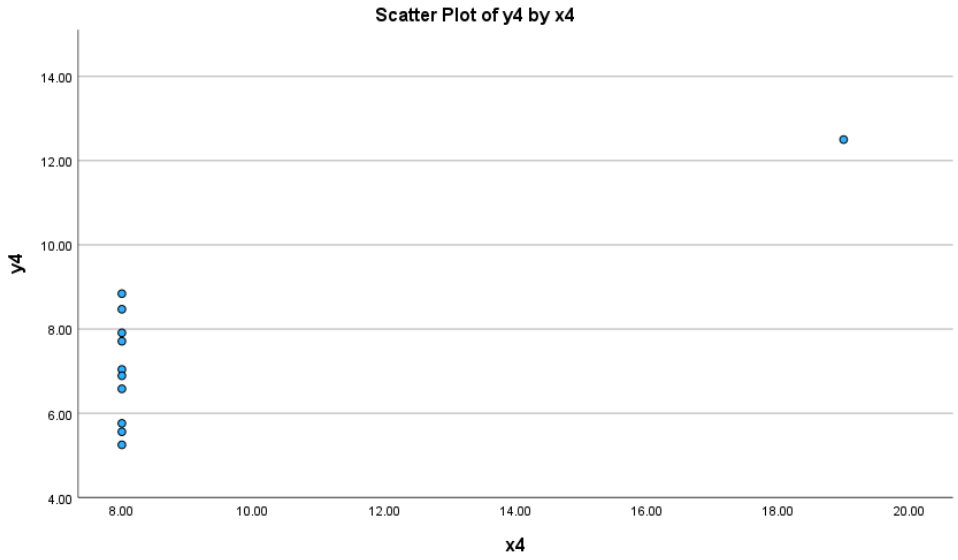


Abbildung L.7. Streudiagramm für das Variablenpaar (x_4, y_4) aus Übungsaufgabe 3.15.

In allen vier Fällen ist der Pearson Korrelationskoeffizient zu r = .82 gegeben, obwohl völlig unterschiedliche Datensituationen vorliegen. Den Datenpunkten könnte im ersten Fall durchaus ein linearer Zusammenhang zugrunde liegen. Der Datensatz im zweiten Fall scheint hingegen exakt (oder immerhin sehr präzise) durch einen quadratischen Zusammenhang beschreibbar. Der Zusammenhang im dritten Datensatz ist vermutlich für den Großteil der Datenpunkte optimal linear und der Koeffizient ungleich 1 kommt nur durch den einzelnen Ausreißer zustande. Im vierten Fall liegt für den Großteil der Datenpunkte gar keine Varianz bezüglich der Variablen x vor. In allen außer dem ersten Fall erscheint eine Charakterisierung der Datenpunkte durch einen linearen Zusammenhang irreführend.

Beispiel 3.16

Die drei Streudiagramme sind in den Abbildungen L.8-10 gezeigt. Mit allen Datenpunkten ergeben sich folgende Korrelationskoeffizienten: $r = .30$, $r_S = .24$, $\tau_b = .17$; allesamt kleine Effekte gemäß Cohen (1988). Ohne die beiden Datenpunkte ganz rechts unten im Streudiagramm ergeben sich die folgenden Werte: $r = .55$, $r_S = .41$, $\tau_b = .28$; ein großer Effekt gemäß Cohen (1988) für Pearsons Korrelationskoeffizienten, ein mittlerer Effekt für Spearmans Rangkorrelationskoeffizienten, ein kleiner Effekt für Kendalls tau-b.

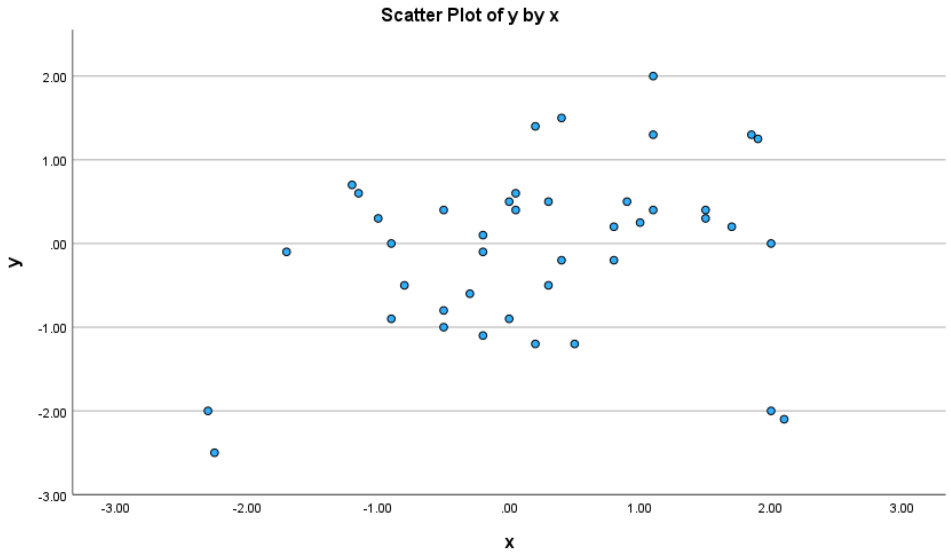


Abbildung L.8. Streudiagramm für den Originaldatensatz aus Übungsaufgabe 3.16.

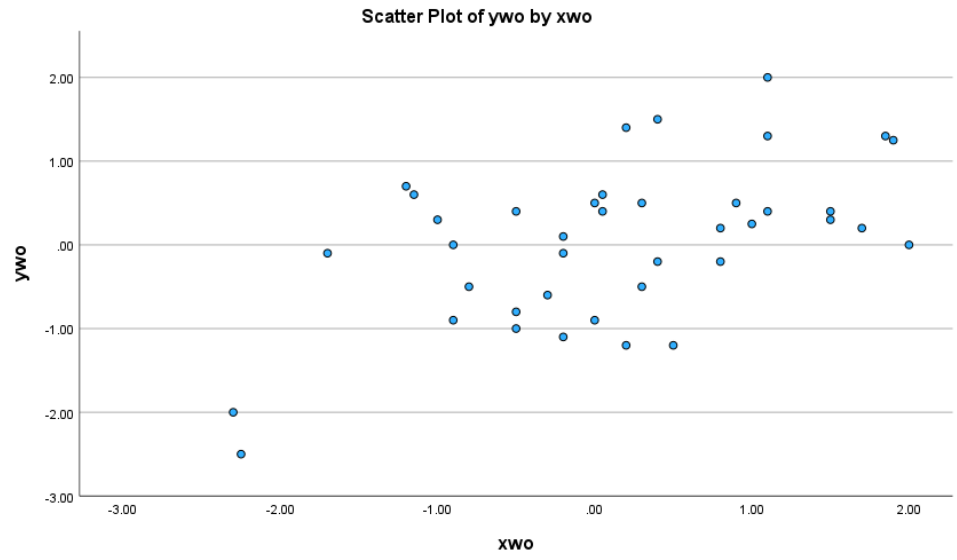


Abbildung L.9. Streudiagramm für Übungsaufgabe 3.16 nach Entfernung zweier Datenpunkte (rechts unten).

Werden auch noch die beiden Datenpunkte links unten entfernt, ergeben sich die folgenden Werte: r = .36, r_S = .31, τ_b = .20; mittlere Effekte gemäß Cohen (1988) für Pearsons Korrelationskoeffizienten und Spearmans Rangkorrelationskoeffizienten, immer noch ein kleiner Effekt für Kendalls tau-b.

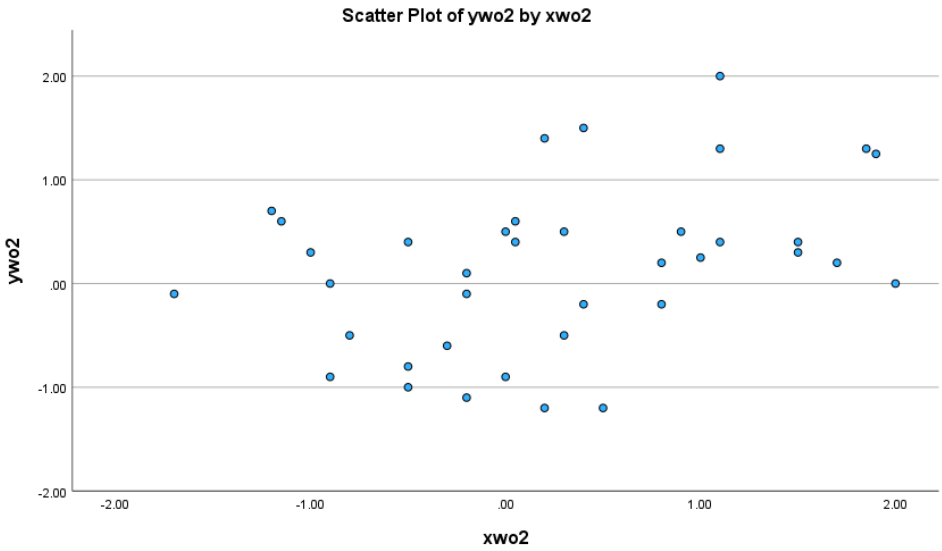


Abbildung L.10. Streudiagramm für Übungsaufgabe 3.16 nach Entfernung zweier weiterer Datenpunkte (links unten).

Man sieht an diesem Beispiel, dass alle drei Korrelationskoeffizienten durch ungewöhnliche Datenpunkte (im Vergleich zu den anderen) beeinflussbar sind, allerdings in deutlich anderem Ausmaß.

Beispiel 3.17

(a) Zwischen den Logarithmen der beiden Variablen besteht kein signifikanter Zusammenhang, $r(45) = -.04$, $p = .782$. Das Ergebnis liefert keine Evidenz für die theoretische Vorhersage.

(b) Unter Ausschluss der entsprechenden vier Sterne besteht ein signifikanter Zusammenhang zwischen den Logarithmen der beiden Variablen, $r(41) = .65$, $p < .001$. In Einklang mit der theoretischen Vorhersage deutet das Ergebnis auf einen positiven Zusammenhang zwischen den Logarithmen der Oberflächentemperatur und der Leuchtkraft hin, d.h. je höher die Oberflächentemperatur, desto höher die Leuchtkraft.

Lösungen der Übungsaufgaben zu Kapitel 4

Beispiel 4.1

Richtig: (a), (b). Falsch: (c), (d).

Beispiel 4.2

Alle Aussagen sind falsch, siehe insbesondere auch Gigerenzer (2004).

Beispiel 4.3

Nr.	**Aussage**	**R/F**
1)	Es kann sein, dass der p-Wert kleiner als das Signifikanzniveau α ist, aber die Teststatistik T nicht im Ablehnungsbereich der Nullhypothese liegt.	F
2)	Für eine ungerichtete Hypothese ist der p-Wert die Wahrscheinlichkeit unter Annahme der Gültigkeit der Nullhypothese dafür, dass sich die Teststatistik in der beobachteten Realisation oder einer extremeren Realisation in Richtung der Alternativhypothese realisiert.	R
3)	Ist der p-Wert klein, dann liegt der wahre Populationsmittelwert weit weg vom Testwert.	F
4)	Ist der p-Wert klein, dann hat man einen Effekt mit großer Effektstärke detektiert.	F

Beispiel 4.4

Nr.	**Aussage**	**R/F**
1)	Ein p-Wert größer als das gewählte Signifikanzniveau bedeutet, dass es keinen Unterschied zwischen dem Populationsmittelwert und dem Testwert gibt.	F
2)	Ein p-Wert größer als das gewählte Signifikanzniveau bedeutet, dass die Nullhypothese stimmt.	F
3)	Ein p-Wert größer als das gewählte Signifikanzniveau bedeutet, dass die Nullhypothese eher stimmt als die Alternativhypothese.	F
4)	Ein p-Wert kleiner als das gewählte Signifikanzniveau bedeutet, dass die Alternativhypothese zutrifft.	F

Für weitere häufige Missverständnisse bezüglich des p-Werts, siehe z.B. Greenland et al. (2016).

Beispiel 4.5

Richtig: (b). Falsch: (a), (c), (d).

Beispiel 4.6

Richtig: (b), (d). Falsch: (a), (c).

Beispiel 4.7

Im Mittel ist das Alter der Kursteilnehmer:innen um 5.47 Jahre geringer als der Vergleichswert von 27.1 Jahren (n = 51, M = 21.63, 95%-KI [20.65, 22.61], SD = 3.49). Das mittlere Alter unterscheidet sich (mit $\alpha = .005$) signifikant vom Vergleichswert, $t(50)$ = -11.21, p < .001, Cohens d = 1.57, 95%-KI [1.15, 1.98]. Gemäß Cohens Heuristik (1988) handelt es sich um einen großen Effekt.

Die Voraussetzungen für einen Einstichproben t-Test sind erfüllt: (i) es handelt sich um eine intervallskalierte Variable, (ii) die Stichprobe ist hinreichend groß (n > 30), (iii) die Varianz des Alters von Studierenden in Österreich (sowie von Kursteilnehmer:innen) ist unbekannt und muss mittels der Stichprobe geschätzt werden.

Beispiel 4.8

Im Mittel beträgt die Reaktionszeitverzögerung bei den 42 Versuchspersonen M = 55.47 ms (SD = 22.03) und überschreitet (mit α = .05) den Wert von 50 ms nicht signifikant, $t(41)$ = 1.61, p = .058 (gerichtete Hypothese). Als Effektstärke ergibt sich Cohens d = 0.25, was gemäß Cohens Heuristik (1988) einem kleinen Effekt entspricht.

Beispiel 4.9

(a) Siehe Abbildung L.11.

(b) Siehe Abbildung L.12.

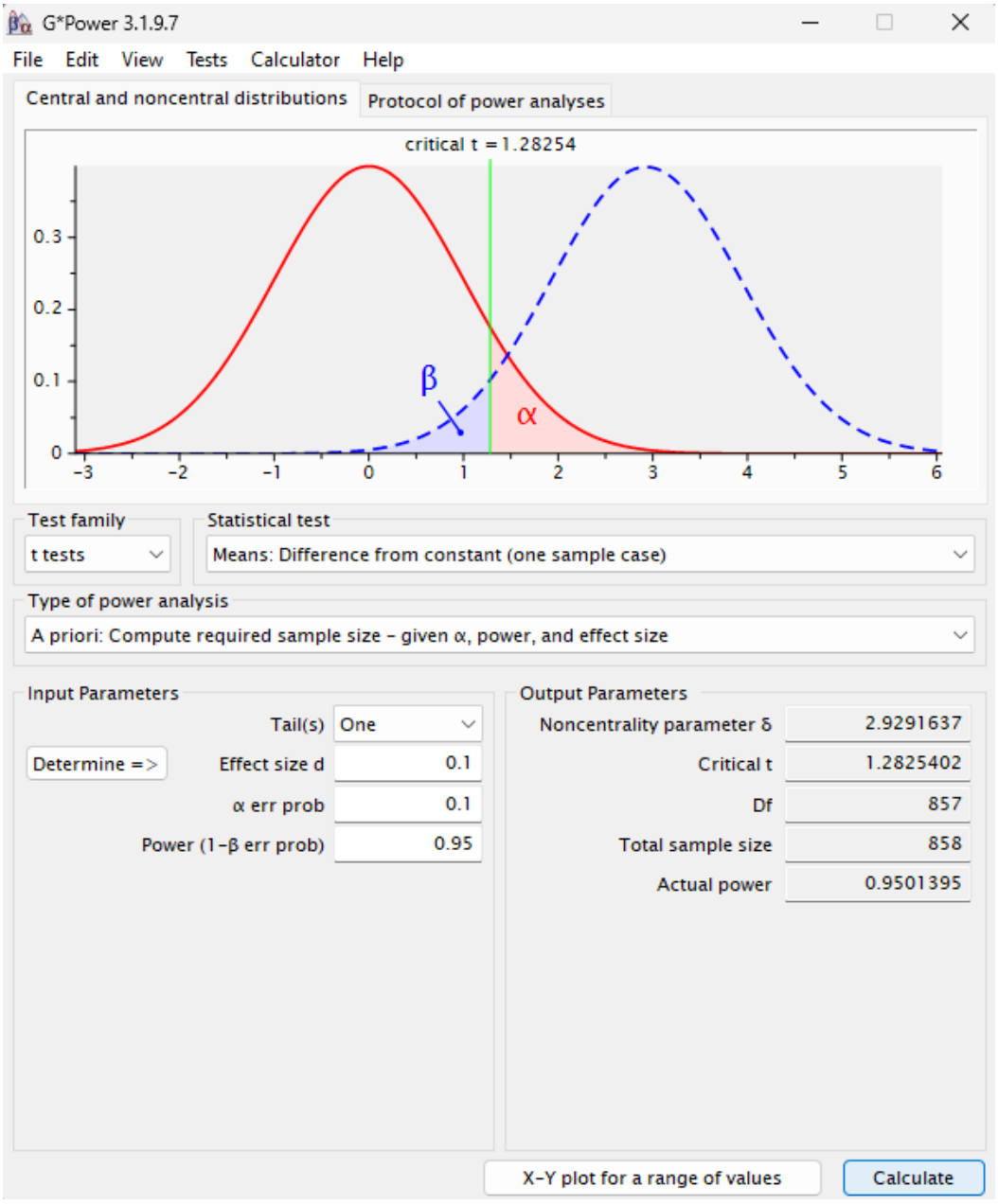


Abbildung L.11. Lösung Beispiel 4.9(a).

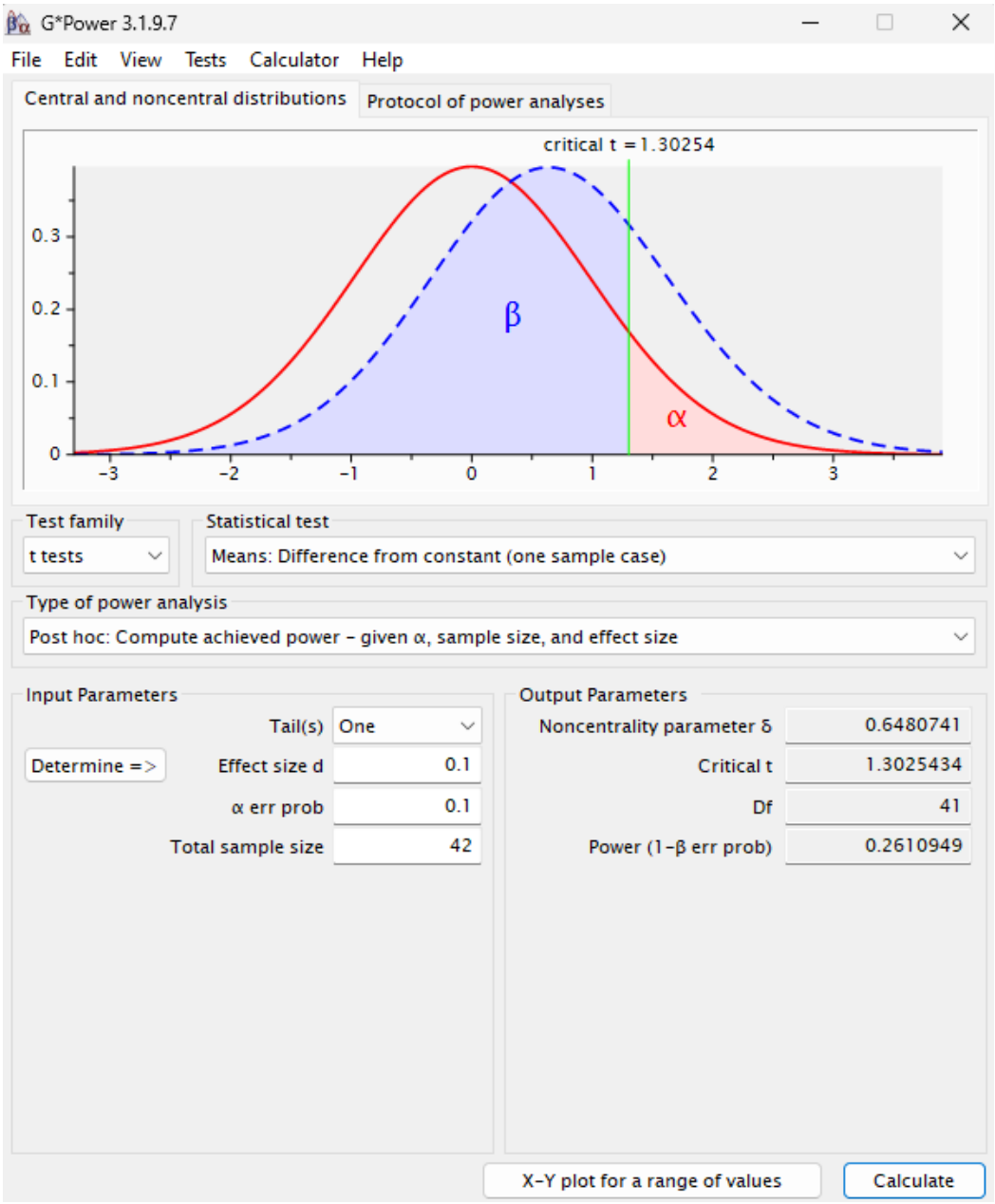


Abbildung L.12. Lösung Beispiel 4.9(b).

Beispiel 4.10

Die mittlere Qualität der $n = 10$ Hopfenproben beträgt $M = 61.70$ ($SD = 11.15$) und liegt (mit $\alpha = .005$) signifikant über dem Testwert von 50, $t(9) = 3.32$, $p = .004$ (gerichtete Hypothese). Gemäß Cohens Heuristik (1988) handelt es sich mit Cohens $d = 1.05$ um einen großen Effekt. Das Signifikanzniveau wurde zu .005 gewählt, da hier ein Fall der Qualitätssicherung vorliegt und wir uns daher vor allem gegen den Fehler 1. Art absichern und die false detection rate (FDR) geringhalten möchten.

Beispiel 4.11

Die Stichprobe muss 217 Personen umfassen.

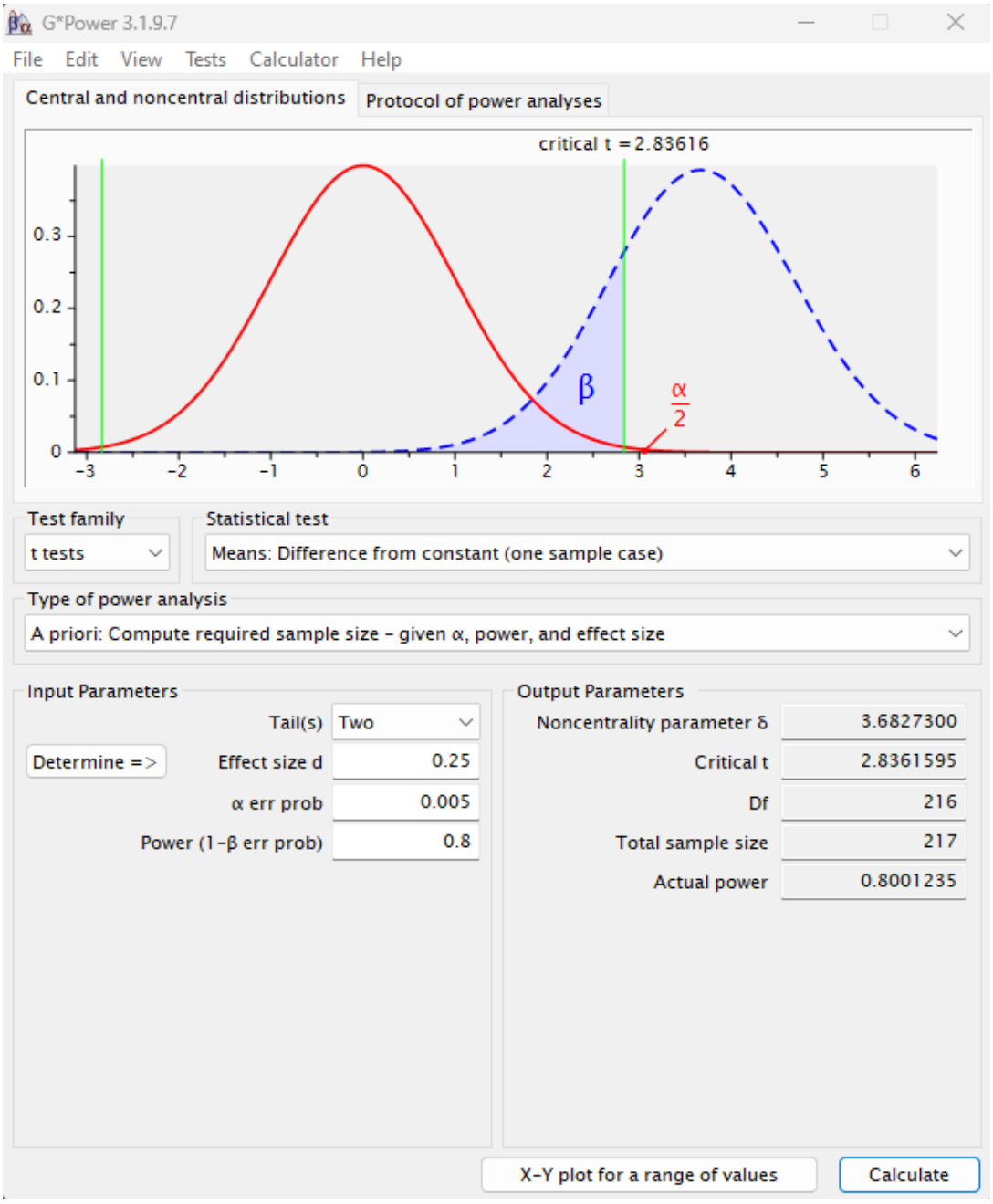


Abbildung L.13. Verlangter Screenshot für Beispiel 4.11.

Beispiel 4.14

Die Voraussetzungen für den Einstichproben-t-Test sind: (i) die Populationsvarianz der untersuchten Variable ist nicht bekannt, sondern muss mittels der Stichprobe geschätzt werden; (ii) die untersuchte Variable ist mindestens intervallskaliert; (iii) die untersuchte Variable kann durch eine normalverteilte

Zufallsvariable approximiert werden oder es handelt sich um eine hinreichend große Stichprobe, dass die Stichprobenkennwerteverteilung des Mittelwerts hinreichend gut durch eine Normalverteilung angenähert werden kann.

Ob Voraussetzung (i) gegeben ist, hängt vom Domänenwissen des:der Forschers:Forscherin ab, der:die die jeweilige Fragestellung untersucht. Geht es beispielsweise um den IQ der Allgemeinbevölkerung, dann ist die Varianz per Definition des IQs bekannt. Bei den meisten anderen psychologischen Konstrukten ist hingegen davon auszugehen, dass die Populationsvarianz kaum bekannt sein dürfte. Die Gültigkeit von Voraussetzung (ii) hängt vom Untersuchungsdesign, insbes. von der Operationalisierung des entsprechenden psychologischen Konstrukts ab: Kann das Konstrukt durch eine metrische Variable beschrieben werden? Die Gültigkeit von Voraussetzung (iii) lässt sich, sofern sie nicht aus fundierter Theorie abgeleitet werden kann (etwa aus der Tatsache, dass sich Abweichungen vom Populationsmittelwert durch Summation sehr vieler kleiner, aber im Einzelfall unbekannter Abweichungen ergeben, siehe z.B. McElreath, 2020, S. 72-74), nur begrenzt mit statistischen Verfahren legitimieren. Dazu wird in der Praxis (jedenfalls zum jetzigen Zeitpunkt, d.h., September, 2025) nach wie vor auf Tests wie den Kolmogorov-Smirnov- oder den Shapiro-Wilk-Test zurückgegriffen. Bei beiden Tests handelt es sich um Signifikanztests, die die Gleichheit einer Teststatistik mit einem für eine normalverteilte Zufallsvariable üblichen Wert testen. D.h., die Argumentation ist prinzipiell dieselbe, die wir für das Testen eines Populationsmittelwerts kennengelernt haben: ergibt sich eine Teststatistik, die so extrem ist (d.h., so groß oder so klein), dass sie sich bei Ziehung einer einfachen Zufallsstichprobe aus einer entsprechenden Referenzverteilung (hier: Normalverteilung) nur selten ergeben würde, dann ist dies ein Indikator dafür, dass eventuell die Annahme der Referenzverteilung (hier: Normalverteilung) keine brauchbare Annahme darstellt und sie in diesem Fall abgelehnt würde. Dieses weit verbreitete Vorgehen ist allerdings nicht unproblematisch. Gerade bei kleinen Stichproben ergibt sich häufig keine signifikante Abweichung der entsprechenden Teststatistik vom Vergleichswert, auch wenn eigentlich keine Normalverteilung gegeben ist. Dies hat den Grund, dass bei kleinen Stichproben die Teststärke der Verfahren nicht groß genug ist, dass sich verlässlich (d.h., in einer angemessenen Anzahl der Fälle) eine signifikante Abweichung ergibt. Bei großen Stichproben ergibt sich hingegen sehr schnell eine signifikante Abweichung, ist dort aber grundsätzlich uninteressant, da bei großen Stichproben ohnehin

die Stichprobenkennwerteverteilung des Mittelwerts aufgrund des zentralen Grenzwerttheorems hinreichend gut durch eine Normalverteilung approximiert werden kann. Andere Verfahren, wie die in Kapitel 3 beschriebene Inspektion von Schiefe und Wölbung haben prinzipiell dasselbe Problem: dort, wo wir die Gültigkeit der Voraussetzung am ehesten brauchen, können wir uns v.a. im Falle eines nicht-signifikanten Ergebnisses nicht auf das Ergebnis der Voraussetzungsprüfung verlassen. Zudem besagt ein nicht-signifikantes Ergebnis bei jeder Stichprobengröße niemals die Gleichheit eines Parameterwerts mit einem bestimmten Referenzwert, sondern lediglich, dass diese Gleichheit nicht mit einer bestimmten Irrtumswahrscheinlichkeit (d.h., α) ausgeschlossen werden kann. Voraussetzung (iii) bezieht sich aber in der Tat auf die Gleichheit der Verteilung der untersuchten Variablen mit einer Normalverteilung (jedenfalls der Form nach). Um sich daher gegen Fehler erster Art abzusichern, wird daher in der Praxis v.a. bei kleinen Stichproben schon bei leichten Indizien für die Ungültigkeit dieser Voraussetzung auf Verfahren zurückgegriffen, die diese Voraussetzung nicht haben.

Welche Konsequenzen hat es, wenn die Voraussetzungen nicht erfüllt sind? Wenn Voraussetzung (i) nicht erfüllt ist, ist im Allgemeinen die Teststärke geringer, da anstelle des t-Tests ein z-Test gerechnet werden könnte (siehe z.B. Bühner & Ziegler, 2017) und der kritische z-Wert kleiner ist als der kritische t-Wert (für alle Stichprobenumfänge und Irrtumswahrscheinlichkeiten). Falls Voraussetzung (ii) nicht erfüllt ist, ist die gesamte Ableitung der Teststatistik T in Kapitel 4 nicht anwendbar, und es lässt sich nicht sagen, wie sich das auf Irrtumswahrscheinlichkeit und Teststärke auswirkt. Ist Voraussetzung (iii) nicht erfüllt, so folgt die Teststatistik T keiner t-Verteilung mit n-1 Freiheitsgraden. D.h., man weiß i.A. nicht wie häufig man eine so extreme oder extremere Teststatistik wie in der Stichprobe unter Gültigkeit der Nullhypothese erhalten würde. Das bedeutet der wahre p-Wert kann ein ganz anderer Wert sein als der Wert, den man unter Annahme der Gültigkeit der Voraussetzung erhält. D.h. auch, dass der Vergleich jenes unter einer falschen Annahme erhaltenen p-Werts mit einer bestimmten vorab festgelegten Irrtumswahrscheinlichkeit nichts bringt, da ja der wahre p-Wert ein völlig anderer Wert sein kann. Das ganze Argument des Nullhypothesensignifikanztestens bricht in sich zusammen. Man kann strenggenommen keine der Aussagen mehr machen, die man überhaupt mit dem Verfahren machen kann. In der Praxis ist es allerdings so, dass sich ab einer hinreichend großen Stichprobengröße die Kennwerteverteilung des Mittelwerts hinreichend gut durch

eine Normalverteilung approximieren lässt, und daher die Teststatistik T in guter Näherung der entsprechenden t-Verteilung folgt. Für Variablen, die sich gut durch symmetrische Verteilungen mit schmalen Rändern beschreiben lassen, ist das in guter Näherung ab einigen wenigen 10 (z.B. 30) Messwerten der Fall. Für Variablen, die hingegen nur durch sehr schiefe Verteilungen mit breiten Rändern gut beschrieben werden können, können hingegen einige hundert Messwerte nötig sein, um von einer hinreichend guten Näherung der Teststatistik T durch eine t-Verteilung ausgehen zu können. Wichtig bleibt dabei immer, dass die Messwerte alle durch dieselbe Verteilung und unabhängig voneinander beschreibbar sein sollten (auch wenn sich beide dieser Voraussetzungen für allgemeinere Formulierungen des zentralen Grenzwerttheorems etwas relaxieren lassen).

Ein alternatives Verfahren, das neben dem in Kapitel 4 bereits angesprochenen Bootstrap-Verfahren zumindest Erwähnung finden sollte, wenn es lediglich darum geht zu untersuchen, ob ein Populationsmittelwert größer oder kleiner als ein bestimmter Referenzwert ist, ist der sog. Vorzeichentest. Besteht zwischen dem Populationsmittelwert und dem Referenzwert in der Tat kein Unterschied, so sollten sich bei zufälliger Ziehung einzelner Personen aus der Population für deren Messwerte im Mittel gleich viele unter dem Referenzwert wie über dem Referenzwert befinden. Weicht das Verhältnis stark in eine Richtung ab, so spricht das gegen die Gleichheit mit dem Referenzwert. Eine detaillierte Beschreibung des Verfahrens findet sich bei Wilcox (2017, S. 364-365).

Lösungen der Übungsaufgaben zu Kapitel 5

Beispiel 5.1

Richtig: (a), (d). Falsch: (b), (c).

Beispiel 5.2

Richtig: (a), (c). Falsch: (b), (d).

Beispiel 5.3

Richtig: (d). Falsch: (a), (b), (c).

Beispiel 5.4

Die Fragestellung wird mit einem t-Test für abhängige Stichproben untersucht, da alle Voraussetzungen für diesen erfüllt sind ($n > 30$; gemessene Variablen sind intervallskaliert; Varianz der Differenzvariable nicht bekannt).

Mittelwert und Standardabweichung für die BDI-Werte der $n = 67$ Patient:innen betragen zu Zeitpunkt 1 $M_1 = 14.36$ und $SD_1 = 9.71$, und zu Zeitpunkt 2 $M_2 = 5.78$ und $SD_2 = 3.70$. Mit einer mittleren Differenz von 8.58 Punkten zwischen den beiden Messzeitpunkten haben die BDI-Werte der Patient:innen signifikant abgenommen, $t(66) = 7.64$, $p < .001$ (gerichtete Hypothese), Cohens $d = 0.93$. Der Unterschied entspricht gemäß Cohens Heuristik (1988) einem großen Effekt.

Beispiel 5.5

Bei dem passenden Test für diese Fragestellung handelt es sich um einen t-Test für abhängige Stichproben, da jedes der beiden Items von jedem Studierenden beantwortet wurde und somit zwischen den Antworten eine Abhängigkeit besteht (2 Messwerte für jede Person). Dies zeigt sich auch in der erheblichen Korrelation zwischen den beiden Items von $r = 0.64$.

Mittelwert und Standardabweichung der $N = 50$ Studierenden für die Zustimmung zur Aussage "Ich hasse Statistik" (auf einer Skala von 1 bis 5) betragen $M_1 = 2.22$ und $SD_1 = 1.03$. Der Aussage "Ich habe Angst vor der nächsten Statistikprüfung" stimmen die Studierenden hingegen im Mittel mehr zu, mit $M_2 = 2.94$ mit Standardabweichung $SD_2 = 1.17$ zu.

Mit einer mittleren Differenz von 0.73 unterscheiden sich die beiden Mittelwerte (mit $\alpha = .005$) signifikant, $t(50) = 5.51$, $p < .001$ (ungerichtete Hypothese), Cohens $d = 0.77$. Gemäß Cohens Heuristik (1988) handelt es sich um einen mittleren Effekt.

Beispiel 5.6

Da in diesem Fall mit nur 14 männlichen Teilnehmern keine hinreichend große Stichprobe vorliegt, wurde die Normalverteilungsvoraussetzung geprüft. Weder der Kolmogorov-Smirnov- noch der Shapiro-Wilk-Test waren signifikant, noch wichen Schiefe und Wölbung mehr als zwei Standardfehler von den Werten für eine Normalverteilung ab. Zudem ergab auch die Inspektion des Q-Q-Plots keine Auffälligkeiten. Daher wird von einer Verträglichkeit mit der Normalverteilungsvoraussetzung ausgegangen.

Mittelwert und Standardabweichung für die Körpergröße der $n = 37$ weiblichen Studierenden betragen $M = 166.84$ cm und $SD = 7.57$ cm. Mittelwert und Standardabweichung für die Körpergröße der $n = 14$ männlichen Studierenden betragen hingegen $M = 181.29$ cm und $SD = 5.99$ cm.

Mit einer Differenz von 14.45 cm zwischen den beiden Mittelwerten sind die männlichen Studierenden der gerichteten Hypothese entsprechend im Mittel (mit $\alpha = 0.005$) signifikant größer als die weiblichen Studierenden, $t(29.56) = 7.12$, $p < .001$, Cohens $d = 2.01$. Gemäß Cohens Heuristik (1988) handelt es sich um einen großen Effekt.

Beispiel 5.7

Als Variable zur Operationalisierung der Größe der Füße wird die Schuhgröße gewählt. D.h. das Merkmal, das uns eigentlich interessiert, ist die Größe der Füße. Die Variable, mit der wir dieses Merkmal quantifizieren (d.h. in Zahlen fassen), ist die Schuhgröße, da bekanntlich größere Füße mit einer größeren Schuhgröße einhergehen.

Mittelwert und Standardabweichung für die Schuhgröße der $n = 37$ weiblichen Studierenden betragen $M = 38.53$ und $SD = 1.33$. Mittelwert und Standardabweichung für die Schuhgröße der $n = 14$ männlichen Studierenden betragen hingegen $M = 43.36$ und $SD = 1.65$.

Mit einer Differenz von 4.83 zwischen den beiden Mittelwerten haben die männlichen Studierenden der gerichteten Hypothese entsprechend im Mittel (mit $\alpha = 0.005$) signifikant größere Füße als die weiblichen Studierenden, $t(19.81) = 9.83$, $p < .001$, Cohens $d = 3.40$. Gemäß Cohens Heuristik (1988) handelt es sich um einen großen Effekt.

Beispiel 5.8

Die Variable Schuhgröße ist intervallskaliert und die Varianzen der Schuhgröße in den beiden Populationen sind nicht bekannt. Diese beiden Voraussetzungen sind also erfüllt. Levenes Test ist nicht signifikant, d.h. es könnte auch die Durchführung eines Studentschen t-Tests gerechtfertigt werden.

Allerdings scheint die Normalverteilungsvoraussetzung für die Schuhgrößen der männlichen Studierenden verletzt zu sein. Sowohl Schiefe (= 1.74, *SE* = 0.60) als auch Wölbung (= 4.67, *SE* = 1.15) weichen mehr als zwei Standardabweichungen von den Werten für eine Normalverteilung ab. Auch der Kolmogorov-Smirnov-Test ($p = .001$) sowie der Shapiro-Wilk-Test ($p = .007$) sind (mit $\alpha = .05$) signifikant. Für die Q-Q-Plots sind schlichtweg zu wenige Datenpunkte vorhanden für ein aussagekräftiges Ergebnis.

In diesem Fall würde es sich also anbieten, die Ergebnisse auch mit einem robusteren Verfahren (gegen Verletzung der Normalverteilungsvoraussetzung) zu überprüfen. Dafür kann beispielsweise ein Mann-Whitney Test durchgeführt werden. Um einen solchen Test mit SPSS durchzuführen ist unter *Analyze >> Nonparametric Tests >> Legacy Dialogs >> Two-Independent-Samples Tests* die Schuhgröße im Feld „Test Variable List“ einzufügen und das Geschlecht wiederum als „Grouping Variable“, woraufhin noch die beiden Gruppen unter „Define Groups…“ zu definieren sind. Danach können die entsprechenden Kommandozeilen mittels „Paste“ wieder in die Syntax eingefügt und dort ausgeführt werden. In der Ausgabe kann in der Tabelle „Test Statistics“ unter „Asymp. Sig. (2-tailed)“ eingesehen werden, dass auch in diesem Fall ein signifikanter Unterschied für die Schuhgrößen weiblicher und männlicher Studierender erhalten wird. Eine weitere Möglichkeit wäre die Anforderung eines Bootstrap Tests mit mindestens 1000 Bootstrap-Stichproben unter Analyze >> Compare Means and Proportions >> Independent-Samples T Test… durch Auswählen von „Perform bootstrapping“ im Menü „Bootstrap…“.

Beispiel 5.9

Unter Annahme gleich vieler Testpersonen in Experimental- und Plazebogruppe muss die Gesamtstichprobe 1172 Personen umfassen, siehe Abbildung L.14.

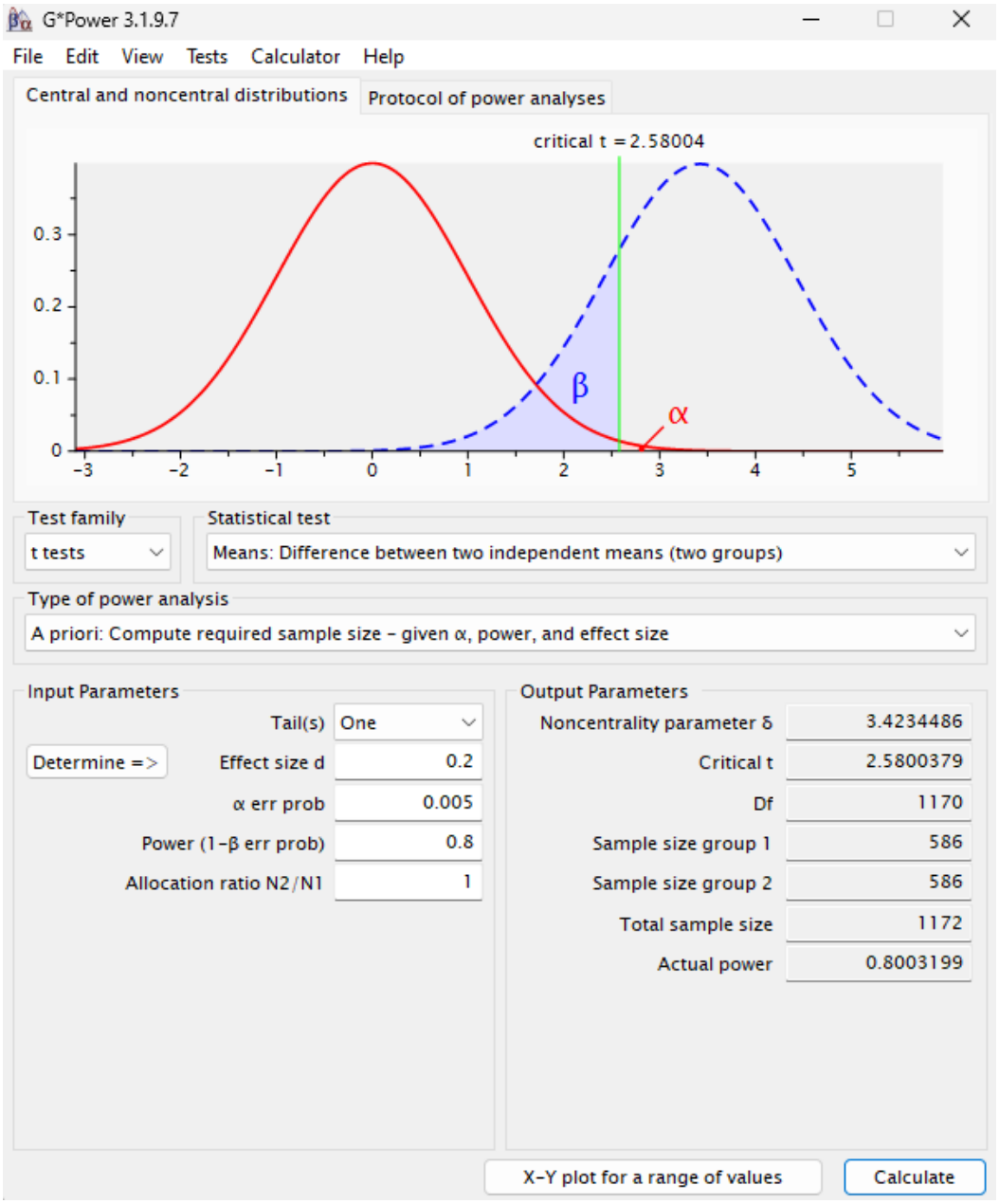


Abbildung L.14. Lösung für Beispiel 5.9.

Beispiel 5.10

Es werden mindestens 17 Personen benötigt, siehe Abbildung L.15.

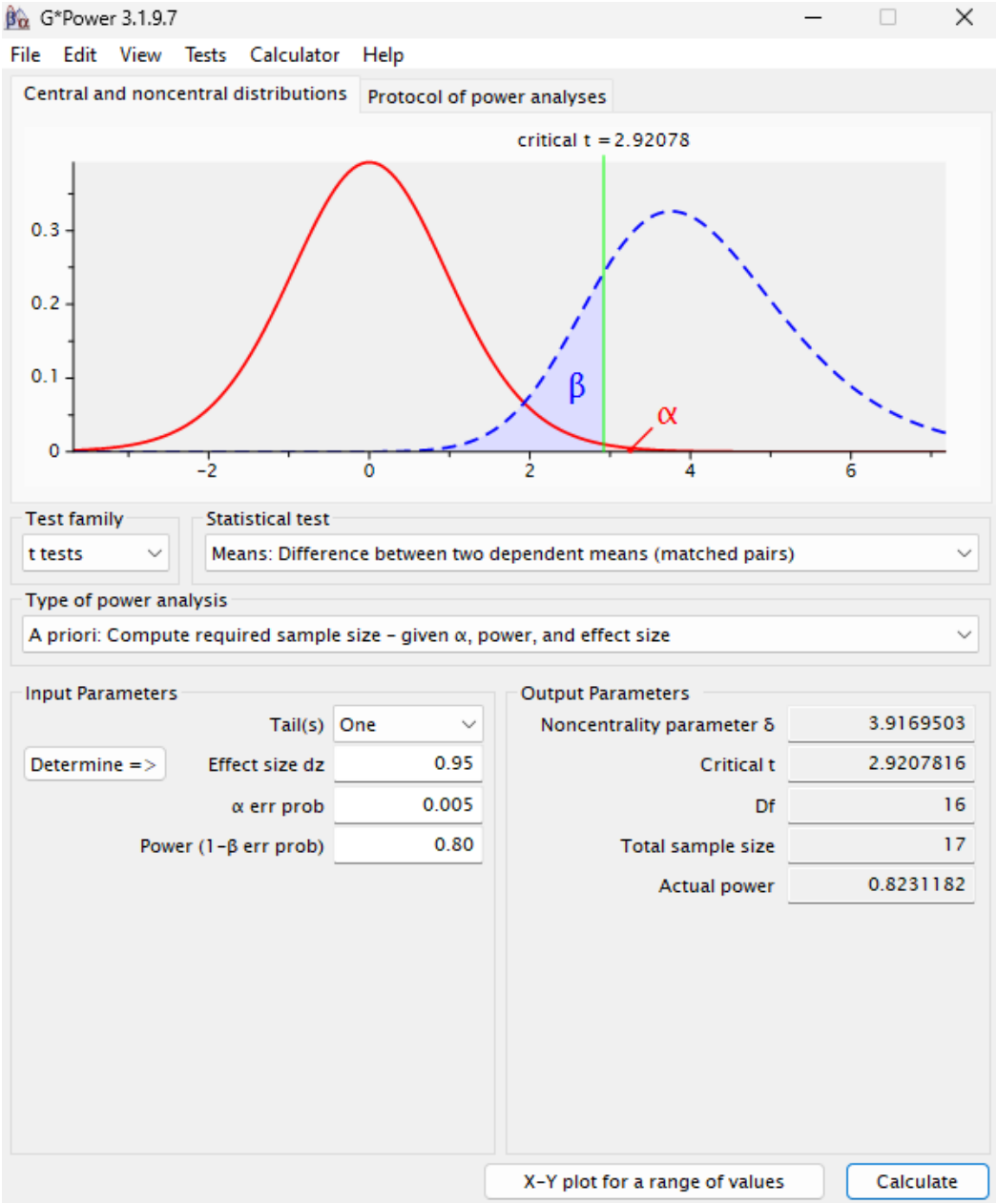


Abbildung L.15. Lösung für Beispiel 5.10.

Beispiel 5.11

Das mittlere Angstniveau unterscheidet sich (mit $\alpha = .05$) nicht signifikant zwischen Gruppe 1 (M = 5.22, SD = 1.06, n = 60) und Gruppe 2 (M = 5.66, SD = 0.98, n = 30), $t(62.28) = 1.94$, $p = 0.057$, Cohens $d = 0.42$. Gemäß Cohens Heuristik (1988) handelt es sich um einen kleinen Effekt.

Beispiel 5.12

Das mittlere Depressionsniveau unterscheidet sich (mit $\alpha = .005$) nicht signifikant zwischen Gruppe 1 (M = 4.91, SD = 1.05, n = 70) und Gruppe 2 (M = 5.25, SD = 0.78, n = 30), $t(72.97) = -1.81$, $p = .074$, Cohens $d = 0.35$. Gemäß Cohens Heuristik (1988) handelt es sich um einen kleinen Effekt.

Beispiel 5.13

Die Konzentrationsfähigkeit nach der Intervention (M = 56.30, SD = 29.33) ist (mit $\alpha = .04$) signifikant höher als die Konzentrationsfähigkeit vor der Intervention (M = 49.88, SD = 14.86), $t(72) = -1.94$, $p = .028$, Cohens $d = 0.23$. Gemäß Cohens Heuristik (1988) handelt es sich um einen kleinen Effekt.

Beispiel 5.14

Zur Beantwortung der Fragestellung wurde ein t-Test für abhängige Stichproben durchgeführt. Dieser ergibt, dass sich die Fähigkeit zur mentalen Rotation zu Zeitpunkt 1 (M = 51.00, SD = 9.80) statistisch signifikant von der Fähigkeit zur mentalen Rotation zu Zeitpunkt 2 (M = 57.30, SD = 12.91) unterscheidet, $t(92) = 6.84$, $p < .001$, Cohens $d = 0.70$. Gemäß Cohens Heuristik (1988) handelt es sich um einen mittleren Effekt.

Beispiel 5.15

Der Mittelwert der Gruppe „Schrift“ (M = 49.60, SD = 9.52) fällt niedriger aus als der Mittelwert der Gruppe „Sprache“ (M = 65.01, SD = 9.27). Zur Überprüfung der statistischen Signifikanz des Unterschieds der Mittelwerte wurde ein t-Test für unabhängige Stichproben durchgeführt. Mit einem Unterschied von 15.41 Punkten im Testergebnis unterscheiden sich die Gruppen „Schrift“ und „Sprache“ statistisch signifikant voneinander, $t(169.88) = 10.75$, $p < .001$, Cohens $d = 1.63$. Gemäß Cohens Heuristik (1988) handelt es sich um einen großen Effekt.

Beispiel 5.16

(a) Die 119 Personen mit Burnout (M = 61.48, SD = 12.82) erleben im Mittel mehr Stress am Arbeitsplatz als die 348 Personen ohne Burnout (M = 53.36, SD = 12.16). Der Unterschied der beiden Mittelwerte von 8.13 ist (mit α = .005) signifikant, $t(465) = -6.21$, $p < .001$, Cohens d = 0.66. Gemäß Cohens Heuristik (1988) handelt es sich um einen mittleren Effekt.

(b) Die Angestellten der Firma erleben weniger Stress am Arbeitsplatz (M = 55.43, SD = 12.82) als in ihrem Privatleben (M = 61.91, SD = 14.74). Der Unterschied von 6.48 ist (mit α = .005) statistisch signifikant, $t(466) = 6.59$, $p < .001$, Cohens d = 0.31. Gemäß Cohens Heuristik (1988) handelt es sich um einen kleinen Effekt.

Beispiel 5.17

Es wurde ein t-Test für abhängige Stichproben durchgeführt. Die Schmerzintensität nach Einnahme des Medikaments (M = 4.73, SD = 1.80) ist im Mittel für die n = 200 Personen (mit α = .005) signifikant geringer als die Schmerzintensität vor der Einnahme (M = 4.94, SD = 1.41), $t(199) = 2.84$, p = .003 (gerichtet), Cohens d = 0.20. Gemäß Cohen (1988) handelt es sich um einen kleinen Effekt. Die Einnahme des Medikaments scheint die Schmerzen zwar im Mittel tatsächlich ein wenig zu lindern, der Effekt ist allerdings nicht sehr stark.

Beispiel 5.18

Um die Fragestellung zu untersuchen wurde ein t-Test für unabhängige Stichproben durchgeführt. Die Bewertung des Dreigängemenüs fiel im Mittel (mit α = .005) signifikant höher in der Personengruppe aus, die Zitronensaft zu trinken bekam (M = 5.85, SD = 2.12, n = 75), als in der Gruppe, die Wasser zu trinken bekam (M = 4.81, SD = 2.13, n = 75), $t(148) = 2.98$, p = .002 (gerichtet), Cohens d = 0.49. Gemäß Cohen (1988) handelt es sich um einen kleinen Effekt.

Beispiel 5.19

Die Lernmotivation der n = 59 Schüler:innen vor der Aktivierungsübung (M = 1.95, SD = 0.35) ist im Mittel niedriger als die Lernmotivation nach der Aktivierungsübung (M = 2.15, SD = 0.36). Ein t-Test für abhängige Stichproben ergibt, dass die Lernmotivation nach der Übung (mit α = .005) signifikant höher ist als vor der Übung, $t(58) = 2.82$, p = .003 (gerichtet), Cohens d = 0.37. Gemäß Cohen (1988)

handelt es sich um einen kleinen Effekt. Die kurze Aktivierungsübung scheint sich also durchaus leicht positiv auf die Lernmotivation auszuwirken.

Beispiel 5.20

Das allgemeine Entspannungsniveau der $n = 60$ Klient:innen vor der Atemübung ($M = 1.95$, $SD = 0.35$) ist im Mittel niedriger als das Entspannungsniveau nach der Atemübung ($M = 2.24$, $SD = 0.41$). Ein t-Test für abhängige Stichproben ergibt, dass das Entspannungsniveau nach der Übung (mit $\alpha = .005$) signifikant höher ist als vor der Übung, $t(59) = 3.67$, $p < .001$ (gerichtet), Cohens $d = 0.47$. Gemäß Cohen (1988) handelt es sich um einen kleinen Effekt. Die kurze Atemübung scheint sich also durchaus leicht positiv auf das Entspannungsniveau auszuwirken.

Beispiel 5.21

Die Gesamtstichprobe muss 242 Personen umfassen.

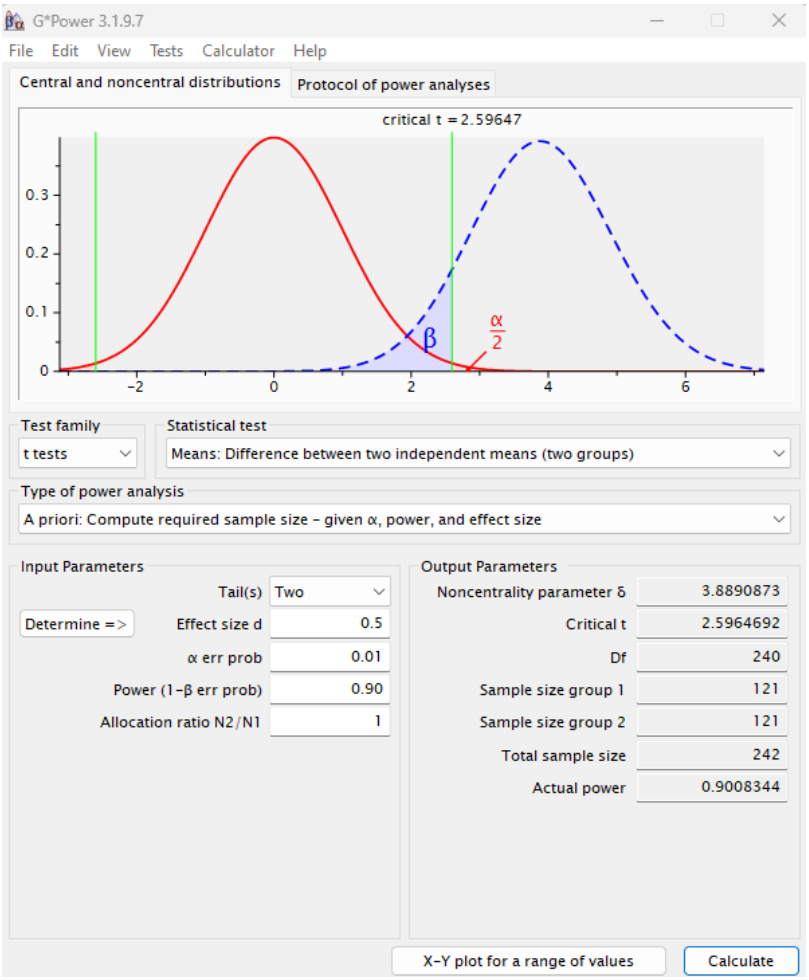


Abbildung L.16. Verlangter Screenshot für Beispiel 5.21.

Beispiel 5.22

Die Gesamtstichprobe muss 338 Personen umfassen.

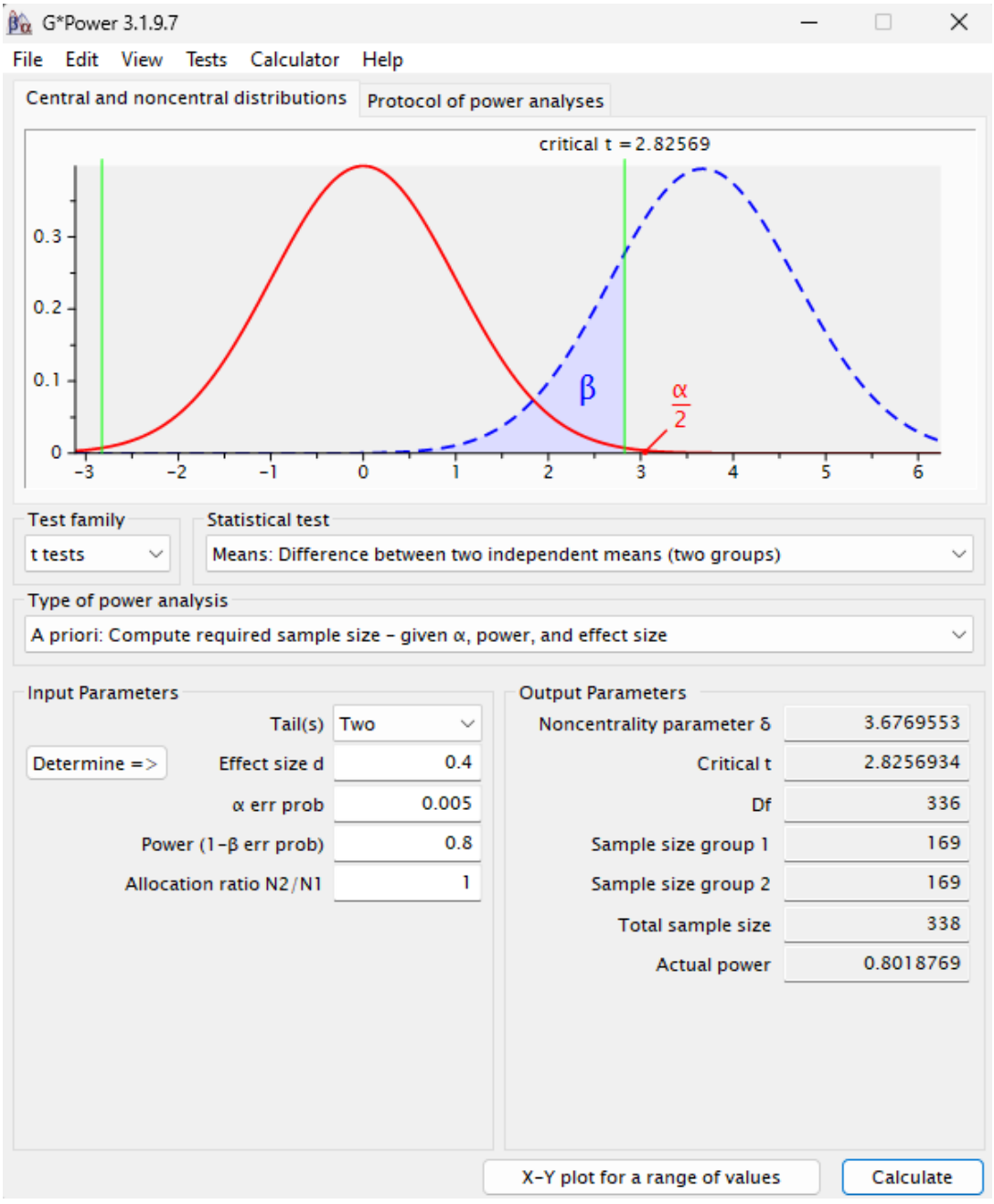


Abbildung L.17. Verlangter Screenshot für Beispiel 5.22.

Beispiel 5.23

Die Stichprobe muss 337 Personen umfassen.

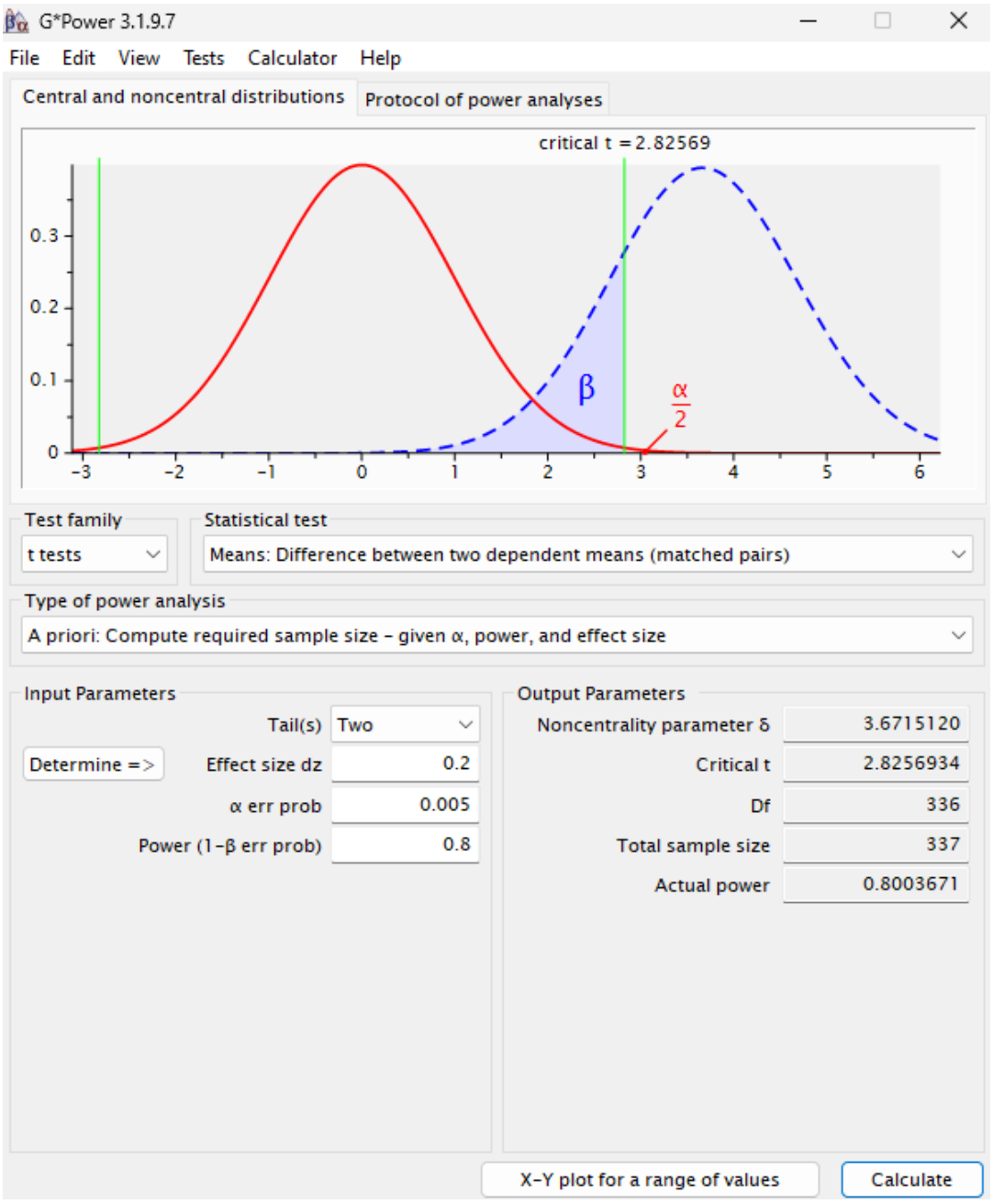


Abbildung L.18. Verlangter Screenshot für Beispiel 5.23.

Lösungen der Übungsaufgaben zu Kapitel 6

Beispiel 6.1

Richtig: (b), (c). Falsch: (a), (d).

Beispiel 6.2

Richtig: (b), (d). Falsch: (a), (c).

Beispiel 6.3

Richtig: (a), (d). Falsch: (b), (c).

Beispiel 6.4

Nr.	**Aussage**	**R/F**
1)	Gemäß Cohens Heuristik (1988) wird ein $\eta^2 = 0.4$ als kleiner Effekt bezeichnet	F
2)	Eine Effektstärke für die einfaktorielle ANOVA heißt *f* und kann aus η^2 berechnet werden kann. Diese Berechnung kann auch in G*Power durchgeführt werden.	R
3)	Fishers least-significant-difference (LSD) Test hat für den Fall einer einfaktoriellen Varianzanalyse ohne Messwiederholung für drei Gruppen eine höhere Teststärke als Tukeys honestly-significant-difference (HSD) Test und ist diesem daher vorzuziehen.	R
4)	Falls die Voraussetzung der Varianzhomogenität nicht erfüllt ist, kann anstelle einer einfaktoriellen Varianzanalyse ohne Messwiederholung eine Varianzanalyse nach Welch gerechnet werden.	R
5)	Die Voraussetzung der Normalverteilung der AV ist wichtiger als die Voraussetzung der Varianzgleichheit für einfaktorielle Varianzanalysen ohne Messwiederholung.	F
6)	Bei η^2 zwischen 0.5 und 0.8 spricht man gemäß Cohens Heuristik (1988) von einem mittleren Effekt.	F

Beispiel 6.5

Die 14 männlichen Übungsteilnehmer (M = 181.29 cm, SD = 5.99 cm) sind im Mittel größer als weibliche Übungsteilnehmerinnen (M = 166.84 cm, SD = 7.57 cm). Levenes Test zur Prüfung der Voraussetzung der Varianzhomogenität war nicht signifikant (p = .268). Der Größenunterschied ist statistisch signifikant, $F(1,49) = 41.03$, $p < .001$, und entspricht mit $\eta^2 = .46$ gemäß Cohens Heuristik (1988) einem großen Effekt.

Beispiel 6.6

Der Mittelwert der Gruppe „Schrift“ ($M = 49.60$, $SD = 9.52$, $n = 86$) fällt niedriger aus als der Mittelwert der Gruppe „Sprache“ (M = 65.01, SD = 9.27, n = 86). Zur Überprüfung der statistischen Signifikanz des Unterschieds der Mittelwerte wurde eine Varianzanalyse ohne Messwiederholung durchgeführt. Levenes Test zur Überprüfung der Varianzhomogenität war nicht signifikant (p = 0.870). Die beiden Gruppen „Schrift“ und „Sprache“ unterscheiden sich statistisch signifikant voneinander, $F(1,170) = 115.62$, $p < .001$, $\eta^2 = .41$. Der Effekt entspricht gemäß Cohens Heuristik (1988) einem großen Effekt.

Beispiel 6.7

Für insgesamt 1557 Personen, d.h. 519 pro Gruppe, siehe Abbildung L.19.

Beispiel 6.8

Für die Abneigungen gegenüber Statistikprüfungen (Skala 0-10) ergeben sich für die drei verglichenen Lieblingshauptfächer Deutsch, Englisch und Mathematik folgende Mittelwerte, Standardabweichungen sowie Stichprobengrößen:

Hauptfach	*M*	*SD*	*n*
Deutsch	3.91	1.51	11
Englisch	5.04	1.93	25
Mathematik	3.27	1.58	15

Levenes Test für die Gleichheit der Varianzen war nicht signifikant (p = .945). Die Unterschiede zwischen den Mittelwerten der drei entsprechenden Gruppen von Studierenden sind statistisch signifikant, $F(2,48) = 5.13$, p = .010. Damit ist die Abneigung gegenüber Statistikprüfungen in diesem

Sinne abhängig vom Lieblingsfach in der Schule. Mit $\eta^2 = .18$ zeigt sich gemäß Cohens Heuristik (1988) ein großer Effekt des Lieblingsfachs auf die Abneigung.

Paarweise Vergleiche mittels Fishers LSD Test ergeben einen signifikanten Unterschied zwischen den Mittelwerten für Studierende mit dem Lieblingsfach Englisch und den Mittelwerten für Studierende mit dem Lieblingsfach Mathematik, $p = .003$. Die verbleibenden beiden paarweisen Unterschiede sind nicht statistisch signifikant ($p = .080$ für den Vergleich zwischen Deutsch und Englisch, $p = .359$ für den Vergleich zwischen Deutsch und Mathematik).

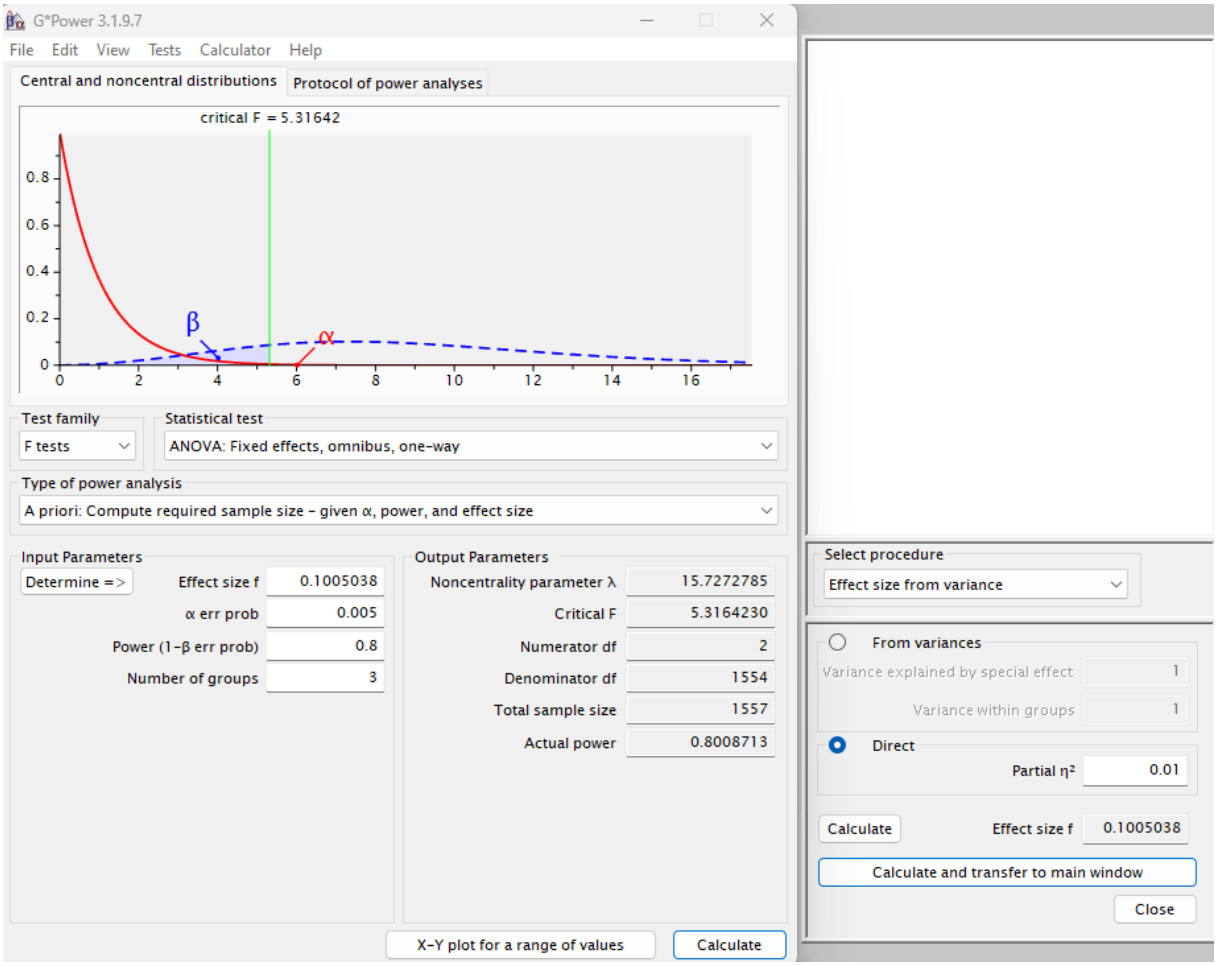


Abbildung L.19. Lösung für Beispiel 6.7.

Beispiel 6.9

Deskriptive Statistiken sind in der Tabelle in der Lösung zu Beispiel 6.8 zu finden.

Entsprechend unserer Vermutungen ergab sich mit einem Unterschied von 1.21 (95%-KI [0.17, 2.24]) im Mittel eine signifikant höhere Abneigung gegen Statistikprüfungen bei sprachaffinen Studierenden als bei Mathematik-affinen, $t(28.32) = 2.39$, $p = 0.012$, Cohens $d = 0.69$. Zudem ergab sich mit Unterschied von 1.13 (95%-KI [0.10, 2.36]) ebenfalls eine signifikant höhere Abneigung bei Englisch-affinen als bei Deutsch-affinen Studierenden, $t(24.20) = 1.89$, $p = 0.035$, $d = 0.65$. Bei beiden Effekten handelt es sich gemäß Cohens Heuristik (1988) um mittlere Effekte.

Beispiel 6.10

Die Gruppengrößen (n) sowie mittlere Anzahl an verkauften Alben (M) und deren Standardabweichungen (SD) für die drei Kategorien unterschiedlicher attraktiver Bands sind in der folgenden Tabelle zusammengefasst:

Ergebnistabelle

Deskriptive Statistiken

Attraktivität	*M*	*SD*	*n*
„ugly“	161.14	75.67	70
„average“	215.75	76.99	73
„beautiful“	203.68	80.04	57

Levenes Test zur Überprüfung der Varianzhomogenität war nicht signifikant ($p = .871$). Die Unterschiede zwischen den Mittelwerten für die drei Kategorien sind statistisch signifikant, $F(2,197) = 9.62$, $p < .001$. Mit $\eta^2 = .09$ zeigt sich gemäß Cohens Heuristik (1988) ein mittlerer Effekt der Attraktivität auf die mittleren Verkaufszahlen.

Paarweise Vergleiche gemäß Fishers LSD-Test ergeben einen signifikanten Unterschied zwischen den Mittelwerten für hässliche Bands und für durchschnittlich attraktive ($p < .001$) oder schöne Bands ($p = .002$). Die mittleren Verkaufszahlen von durchschnittlichen attraktiven und schönen Bands unterscheiden sich nicht signifikant ($p = .379$).

Beispiel 6.11

Die Ergebnisse der drei Methoden sind in Abbildung L.20 gegenübergestellt.

Multiple Comparisons

Dependent Variable: Album Sales (Thousands)

	(I) Attractiveness of the Band	(J) Attractiveness of the Band	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Tukey HSD	ugly	average	-54.61*	12.950	<.001	-85.19	-24.03
		beautiful	-42.54*	13.811	.007	-75.16	-9.92
	average	ugly	54.61*	12.950	<.001	24.03	85.19
		beautiful	12.07	13.683	.652	-20.25	44.38
	beautiful	ugly	42.54*	13.811	.007	9.92	75.16
		average	-12.07	13.683	.652	-44.38	20.25
LSD	ugly	average	-54.61*	12.950	<.001	-80.15	-29.07
		beautiful	-42.54*	13.811	.002	-69.78	-15.30
	average	ugly	54.61*	12.950	<.001	29.07	80.15
		beautiful	12.07	13.683	.379	-14.92	39.05
	beautiful	ugly	42.54*	13.811	.002	15.30	69.78
		average	-12.07	13.683	.379	-39.05	14.92
Bonferroni	ugly	average	-54.61*	12.950	<.001	-85.88	-23.34
		beautiful	-42.54*	13.811	.007	-75.89	-9.19
	average	ugly	54.61*	12.950	<.001	23.34	85.88
		beautiful	12.07	13.683	1.000	-20.97	45.11
	beautiful	ugly	42.54*	13.811	.007	9.19	75.89
		average	-12.07	13.683	1.000	-45.11	20.97

Based on observed means.
The error term is Mean Square(Error) = 5992.987.
*. The mean difference is significant at the .05 level.

Abbildung L.20. Ergebnisse für die drei verwendeten post-hoc Verfahren in Beispiel 6.11.

Man sieht: Fishers LSD-Test ergibt die kleinsten p-Werte und hat daher, da er die FWER im Fall von drei Gruppen exakt kontrolliert die höchste Teststärke von den drei Verfahren ohne Fehler 1. Art zu erhöhen. Dies gilt allerdings nur, wenn es sich genau um den Vergleich von drei Gruppen handelt.

Man sieht auch, dass p-Werte bzw. Konfidenzintervalle bei Tukeys HSD-Test kleiner sind als bei Bonferroni, d.h. höhere Teststärke, da weniger konservativ (Bonferroni kontrolliert FWER zu stark).

Beispiel 6.12

Korrektur Ergebnisbericht: Die Stichprobe umfasste insgesamt 200 Personen. Der erste Kontrast verglich die beiden Therapien mit den beiden Kontrollbedingungen. Es zeigte sich, dass die beiden Therapien zu ~~weniger~~ mehr Gewichtszunahme führten als die beiden Kontrollbedingungen (t(~~928.74~~ 192.09) = 7.13, p < .001, d = 1.01; d.h. gemäß Cohen (1988) ein großer Effekt). Zwischen den beiden Therapieformen gab es ~~keinen~~ einen signifikanten Unterschied zwischen den mittleren Gewichtszunahmen für die KVT-Gruppe (M = 5.90, SD = ~~2.66~~ 3.05) und die LKT-Gruppe (M = 4.52, SD = ~~3.05~~ 2.66; t(96.19) = 2.40, p = .018, d = 0.47; d.h. gemäß Cohen (1988) ein ~~großer~~ kleiner Effekt). Auch innerhalb der Kontrollbedingungen fand sich ein signifikanter Unterschied zwischen den mittleren Gewichtszunahmen der TAU-Gruppe (M = 3.20, SD = 2.73) und der KB-Gruppe (M = 1.37, SD = 3.15; t(96.06) = 3.12, p = ~~.020~~ .002, d = 0.63; d.h., ein mittlerer Effekt gemäß Cohen(1988)).

Beispiel 6.13

Um die Hypothese zu prüfen, wurden entsprechende a-priori Kontraste definiert, um (a) die Wirksamkeit der beiden Therapien mit der Kontrollbedingung zu vergleichen und (b) die Wirksamkeit der beiden Therapien miteinander zu vergleichen.

Bezüglich Hypothese (a) ergab ein t-Test nach Welch, dass beide Therapien (mit α = .005) signifikant besser wirken als die Kontrollbedingung, t(82.16) = 4.86, p < .001 (gerichtet), Cohens d = 0.92. Gemäß Cohen (1988) entspricht das einem großen Effekt.

Bezüglich Hypothese (b) ergab ein t-Test nach Welch, dass die Verhaltenstherapie (mit α = .005) nicht signifikant besser wirkt als die Psychoanalyse, t(77.23) = 0.87, p = .194 (gerichtet), Cohens d = 0.20. Die Effektstärke beträgt präziser d = 0.198, was gemäß Cohen (1988) knapp keinem kleinen Effekt entspricht.

Deskriptive Statistiken für die Besserung der Symptomatik sind für alle drei Therapiebedingungen in Tabelle L.3 angegeben.

Tabelle L.3

Mittelwerte und Standardabweichungen für die Besserung der Symptomatik für die drei Bedingungen (Beispiel 6.13)

Bedingung	*M*	*SD*	*n*
Treatment as usual	-1.80	13.34	40
Psychoanalyse	9.65	14.85	40
Verhaltenstherapie	12.40	13.43	40

Beispiel 6.14

Um die Hypothese zu prüfen, wurde ein entsprechender a-priori Kontrast definiert, um den Unterschied zwischen dem Mittelwert der Statistikangst von Absolvent:innen von Schulen mit Schwerpunkt Naturwissenschaft und Technik mit dem Mittelwert der Statistikangst von Absolvent:innen beider anderen Schultypen zusammen zu vergleichen. Ein t-Test nach Welch ergab, dass die Statistikangst von Absolvent:innen von Schulen mit Schwerpunkt Naturwissenschaft und Technik ($M = 5.82$, $SD = 1.58$, $n = 75$) entsprechend der Hypothese im Mittel (mit $\alpha = .005$) signifikant niedriger ist als diejenige von Absolvent:innen der beiden anderen Schultypen (für Schwerpunkt Sprache: $M = 6.44$, $SD = 1.54$, $n = 100$; für Schwerpunkt Kunst und Design: $M = 6.39$, $SD = 1.49$, $n = 50$), $t(141.81) = 2.66$, $p = .004$ (gerichtet), Cohens $d = 0.39$. Gemäß Cohen (1988) entspricht dies einem kleinen Effekt. Deskriptive Statistiken für die Statistikangst sind für alle Schultypen auch in Tabelle L.4 zusammengefasst.

Tabelle L.4

Mittelwerte und Standardabweichungen für die Statistikangst für die drei Schultypen (Beispiel 6.14)

Schwerpunkt	*M*	*SD*	*n*
Sprachen	6.44	1.54	100
Naturwissenschaft und Technik	5.82	1.58	75
Kunst und Design	6.39	1.49	50

Beispiel 6.15

Um die Hypothese zu prüfen, wurden entsprechende a-priori Kontraste definiert. Ein t-Test nach Welch ergab, dass die Statistikangst von Absolvent:innen von Schulen mit Schwerpunkt Naturwissenschaft und Technik entsprechend Hypothese (i) im Mittel (mit $\alpha = .005$) signifikant niedriger ist als diejenige von Absolvent:innen der Schultypen mit Schwerpunkten Sprachen und Kunst und Design, $t(141.81) = 2.66$, $p = .004$ (gerichtet), Cohens $d = 0.38$. Gemäß Cohen (1988) entspricht dies einem kleinen Effekt. Zudem ergab ein t-Test nach Welch, dass die Statistikangst von Absolvent:innen von Schulen mit Schwerpunkt Sport sich entgegen Hypothese (ii) im Mittel (mit $\alpha = .005$) nicht signifikant von der Statistikangst von Absolvent:innen der Schultypen mit Schwerpunkten Sprachen und Kunst und Design unterscheidet, $t(138.44) = 2.69$, $p = .008$ (ungerichtet), Cohens $d = 0.39$. Gemäß Cohen (1988) entspricht dies einem kleinen Effekt. Deskriptive Statistiken für die Statistikangst sind für alle Schultypen in Tabelle L.5 zusammengefasst.

Tabelle L.5

Mittelwerte und Standardabweichungen für die Statistikangst für die drei Schultypen (Beispiel 6.15)

Schwerpunkt	*M*	*SD*	*n*
Sprachen	6.44	1.54	100
Naturwissenschaft und Technik	5.82	1.58	75
Kunst und Design	6.39	1.49	50
Sport	7.03	1.63	75

Lösungen der Übungsaufgaben zu Kapitel 7

Beispiel 7.1

Sie unterscheiden sich nicht.

Beispiel 7.2

(i) Normalverteilung der AV in den einzelnen Populationen; (ii) Varianzgleichheit (auch als Varianzhomogenität bzw. Homoskedastizität bezeichnet); (iii) Unabhängigkeit der Beobachtungen bzw. Messungen; (iv) Intervallskalenniveau der AV.

Beispiel 7.3

Richtig: (a)-(c). Falsch: (d).

Beispiel 7.4

Richtig: (b)-(d). Falsch: (a).

Beispiel 7.5

Es wurde eine zweifaktorielle Varianzanalyse ohne Messwiederholung mit den Faktoren Geschlecht (zwei Stufen) und dem Faktor Alkoholmenge (drei Stufen) durchgeführt. Levenes Test war nicht signifikant ($p > .05$), daher wurde von Varianzhomogenität ausgegangen.

Im Mittel waren die ausgewählten Gesprächspartner:innen von Männern und Frauen nicht signifikant unterschiedlich attraktiv ($F(1,42) = 2.03$, $p = .161$, $\eta_p^2 = .05$, d.h. kleiner Effekt gemäß Cohen (1988)). Im Mittel unterscheidet sich die Attraktivität der ausgewählten Gesprächspartner:innen signifikant in Abhängigkeit der Menge getrunkenen Alkohols ($F(2,42) = 20.07$, $p < .001$, $\eta_p^2 = .49$, d.h. großer Effekt gemäß Cohen (1988)). Zwischen Geschlecht und Alkoholmenge besteht zudem eine signifikante Interaktion ($F(2,42) = 11.91$, $p < .001$, $\eta_p^2 = .36$, d.h. großer Effekt gemäß Cohen (1988)). Zur weiteren Analyse paarweiser Mittelwertsunterschiede wurden post-hoc Tests mit einer Korrektur der p-Werte für multiple Vergleiche gemäß Bonferroni durchgeführt. Im Folgenden werden lediglich korrigierte p-Werte berichtet.

Für die Menge getrunkenen Alkohols von 0 ($p = .177$) oder 2 ($p = .342$) Pint Bier unterschieden sich die ausgewählten Gesprächspartner:innen in ihrer Attraktivität nicht signifikant zwischen Männern

und Frauen. Bei einer Menge getrunkenen Alkohols von 4 Pint Bier waren die ausgewählten Gesprächspartner:innen von Männern jedoch signifikant weniger attraktiv als die ausgewählten Gesprächspartner:innen von Frauen bei derselben Menge getrunkenen Alkohols ($p < .001$).

Zudem waren die Gesprächspartner:innen von Männern bei 4 Pint Bier signifikant weniger attraktiv als bei 2 oder 0 Pint Bier (jeweils $p < .001$). Bei 2 und 0 Pint Bier bestand kein signifikanter Unterschied ($p > .999$). Gesprächspartner:innen von Frauen unterschieden sich für keinen paarweisen Vergleich signifikant in ihrer Attraktivität (0 und 2 Pint Bier: $p > .999$; 0 und 4 Pint Bier: $p > .999$; 2 und 4 Pint Bier: $p = .836$).

Punkt- und Intervallschätzungen für die Attraktivität der Gesprächspartner:innen in Abhängigkeit vom Geschlecht der Studienteilnehmer:innen und der von ihnen getrunkenen Menge Alkohols sind in Abbildung L.21 dargestellt. Mittelwerte, Standardabweichungen und Gruppengrößen sind in Tabelle L.6 zusammengefasst.

Tabelle L.6

Deskriptive Statistiken

Geschlecht	Alkoholmenge	*M*	*SD*	*n*
Männlich	0 Pints	66.88	10.33	8
	2 Pints	66.87	12.52	8
	4 Pints	35.63	10.84	8
Weiblich	0 Pints	60.62	4.96	8
	2 Pints	62.50	6.55	8
	4 Pints	57.50	7.07	8

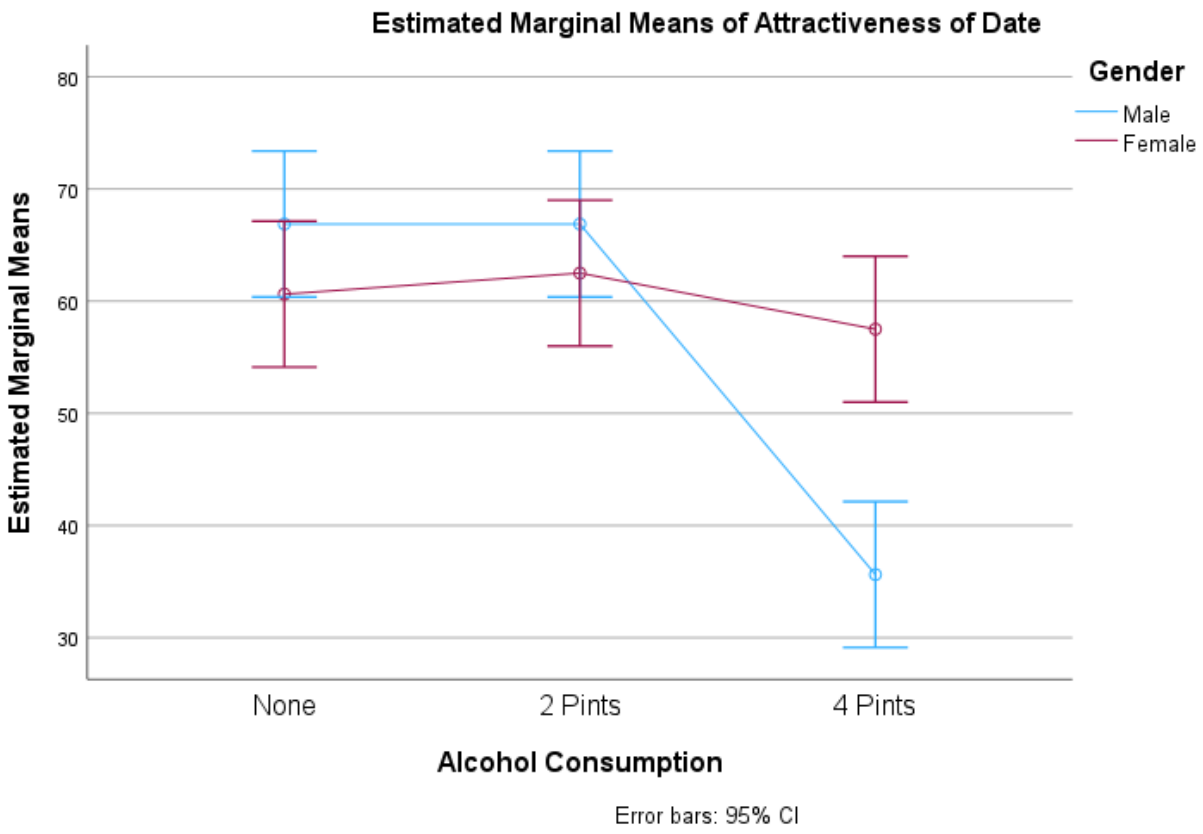


Abbildung L.21. Punkt- und Intervallschätzungen der AV in Beispiel 7.5.

Beispiel 7.6

(a) UV1: Unterrichtsmethode, UV2: Unterrichtsfach. AV: Erzielte Punkte beim Wissenstest.

(b) Jeweils 2 Stufen. Bei UV1: Tafel vs. Powerpoint. Bei UV2: Geschichte vs. Mathematik.

(c) Mit der statistischen Testung der Interaktion und des Haupteffekts für das Unterrichtsfach.

(d) Siehe unten.

Ergebnisbericht: Es wurde eine zweifaktorielle Varianzanalyse ohne Messwiederholung mit den Faktoren Unterrichtsmethode (zwei Stufen: Tafel vs. Powerpoint) und Unterrichtsfach (zwei Stufen: Geschichte vs. Mathematik) durchgeführt. Levenes Test war nicht signifikant ($p > .05$), daher wurde von Varianzhomogenität ausgegangen.

Im Mittel waren die erzielten Punkte beim Wissenstest für die beiden Unterrichtsmethoden (mit $\alpha = .005$) nicht signifikant unterschiedlich ($F(1,176) = 0.31$, $p = .581$, $\eta_p^2 < .01$). Im Mittel waren die erzielten Punkte beim Wissenstest auch für die beiden Unterrichtsfächer (mit $\alpha = .005$) nicht signifikant unterschiedlich ($F(1,176) = 3.25$, $p = .073$, $\eta_p^2 = .02$, d.h. kleiner Effekt gemäß Cohen (1988)).

Allerdings ergab sich (mit $\alpha = .005$) eine signifikante Interaktion zwischen Unterrichtsfach und Methode, $F(1,176) = 74.41, p < .001, \eta_p^2 = .30$, was einem großen Effekt gemäß Cohens Heuristik (1988) entspricht. Zur weiteren Analyse paarweiser Mittelwertsunterschiede wurden post-hoc Tests mit einer Korrektur der p-Werte für multiple Vergleiche gemäß Bonferroni durchgeführt. Im Folgenden werden lediglich korrigierte p-Werte berichtet.

Während Schüler:innen im Geschichtsunterricht signifikant mehr mit der Methode Powerpoint lernen als mit der Methode Tafel ($p < .001$), ist es im Mathematikunterricht gerade umgekehrt: dort lernen Schüler:innen signifikant mehr mit der Methode Tafel als mit der Methode Powerpoint ($p < .001$). Auch innerhalb der beiden Methoden gibt es signifikante Unterschiede zwischen den beiden Unterrichtsfächern. Mit der Methode Tafel lernen Schüler:innen signifikant mehr in Mathematik als in Geschichte ($p < .001$). Mit der Methode Powerpoint ist es wiederum gerade umgekehrt: mit dieser Methode lernen Schüler:innen signifikant mehr in Geschichte als in Mathematik ($p < .001$). Dabei scheint der Unterschied zwischen den Fächern (deskriptiv) ausgeprägter für die Methode Tafel (Punktschätzung für Betrag der Mittelwertsdifferenz: 39.16 mit plausiblen Werten gemäß 95%-KI [28.68, 49.64]; Konfidenzniveau korrigiert gemäß Bonferroni) als für die Methode Powerpoint (Punktschätzung für Betrag der Mittelwertsdifferenz: 25.62 mit plausiblen Werten gemäß 95%-KI [15.14, 36.10]; Konfidenzniveau korrigiert gemäß Bonferroni). Die deutliche Überlappung der 95%-KI zeigt aber auch an, dass die Gleichheit dieser Mittelwertsdifferenz für die beiden Methoden nicht mit hoher Konfidenz ausgeschlossen werden kann.

Punkt- und Intervallschätzungen für die erzielten Punkte beim Wissenstest in Abhängigkeit von Unterrichtsfach und -methode sind in Abbildung L.22 dargestellt. Mittelwerte, Standardabweichungen und Gruppengrößen sind in Tabelle L.7 zusammengefasst.

Tabelle L.7

Deskriptive Statistiken

Methode	Fach	*M*	*SD*	*n*
Tafel	Geschichte	30.04	23.80	45
	Mathematik	69.20	25.20	45
Powerpoint	Geschichte	60.36	26.88	45
	Mathematik	34.73	28.75	45

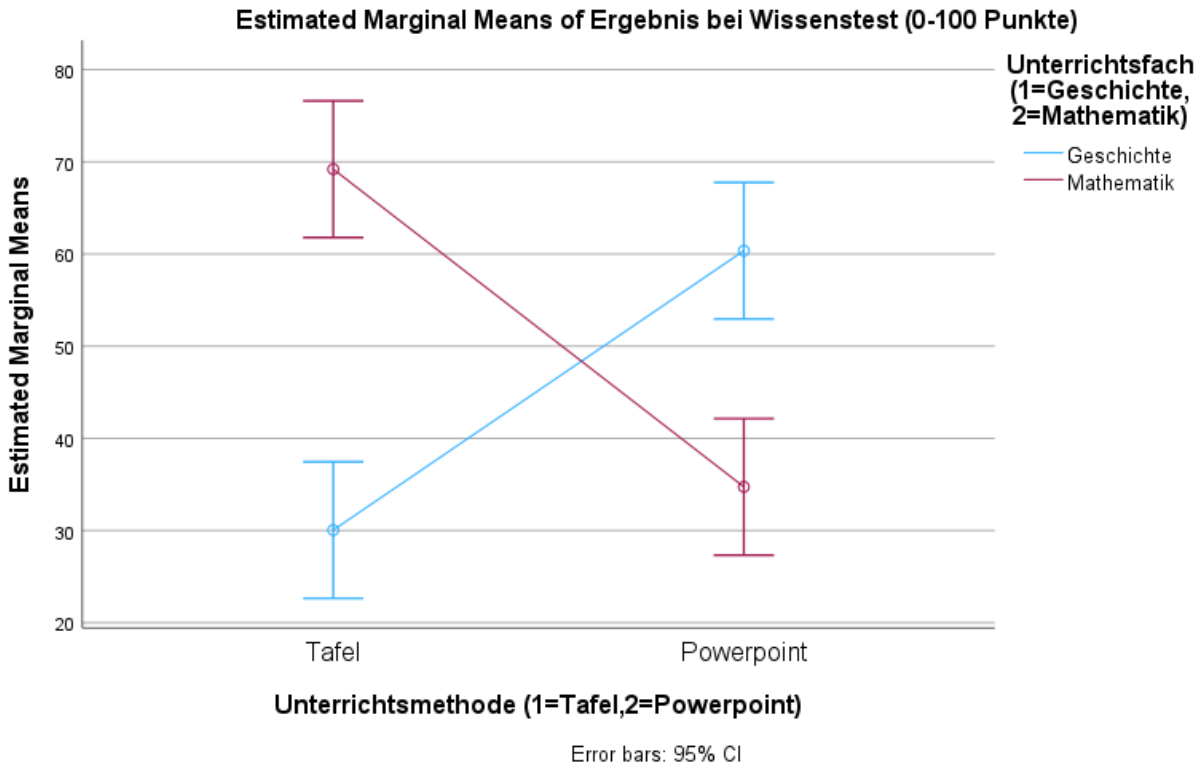


Abbildung L.22. Punkt- und Intervallschätzungen der AV in Beispiel 7.6.

Beispiel 7.7

Ergebnisbericht: Es wurde eine zweifaktorielle Varianzanalyse ohne Messwiederholung mit den Faktoren Altersgruppe (drei Stufen: jung, d.h. 18-30 Jahre, mittel, d.h. 31-50 Jahre, alt, d.h. > 50 Jahre) und Trainingsmethode (zwei Stufen: konventionelles Krafttraining mit Gewichten vs. HIIT mit eigenem Körpergewicht) durchgeführt. Das Signifikanzniveau wurde zu $\alpha = .005$ gewählt.

Insgesamt wurden Daten von 270 Personen in einem balancierten Design erhoben. Levenes Test war (mit $\alpha = .05$) nicht signifikant ($p = .831$), daher wurde von Varianzgleichheit in den einzelnen Populationen ausgegangen.

Im Mittel war die allgemeine Fitness zwischen den unterschiedlichen Altersgruppen signifikant verschieden ($F(2,264) = 54.15$, $p < .001$, $\eta_p^2 = .29$, d.h. ein großer Effekt gemäß Cohen (1988)). Im Mittel war die erzielte allgemeine Fitness auch zwischen den beiden Fitnessprogrammen signifikant unterschiedlich ($F(1,264) = 24.18$, $p < .001$, $\eta_p^2 = .08$, d.h. ein mittlerer Effekt gemäß Cohen (1988)). Die Interaktion zwischen den beiden Faktoren war nicht signifikant ($F(2,264) = 2.43$, $p = .090$, $\eta_p^2 = .02$, d.h. ein kleiner Effekt gemäß Cohen (1988)). Zur weiteren Analyse paarweiser Mittelwertsunterschiede wurden post-hoc Tests mit einer Korrektur der p-Werte für multiple Vergleiche gemäß Bonferroni durchgeführt. Im Folgenden werden lediglich korrigierte p-Werte berichtet.

Sowohl bei konventionellem Krafttraining mit Gewichten als auch bei HIIT-Programmen mit dem eigenen Körpergewicht nahm die erzielte, allgemeine Fitness mit fortschreitendem Alter ab. Bei konventionellem Krafttraining waren alle paarweisen Mittelwertsunterschiede zwischen den unterschiedlichen Altersgruppen signifikant ($p \leq .001$). Bei HIIT-Programmen war der Unterschied zwischen jungen und mittleren Erwachsenen nicht signifikant ($p > .999$), während die Unterschiede zwischen jungen und alten sowie mittleren und alten Erwachsenen jeweils signifikant waren ($p < .001$). Zudem unterschieden sich konventionelles Krafttraining und HIIT-Programme sowohl bei mittleren ($p < .001$) als auch älteren ($p = .001$) Erwachsenen signifikant, jedoch nicht bei jungen Erwachsenen ($p = .264$). Bei allen Altersgruppen war die erzielte allgemeine Fitness jedoch bei HIIT-Programmen höher als bei konventionellem Krafttraining.

Punkt- und Intervallschätzungen für die erzielte allgemeine Fitness in Abhängigkeit von Altersgruppe und verwendeter Trainingsmethode sind in Abbildung L.23 dargestellt. Mittelwerte, Standardabweichungen und Gruppengrößen sind in Tabelle L.8 zusammengefasst.

Tabelle L.8

Deskriptive Statistiken

Altersgruppe	Training	*M*	*SD*	*n*
Jung: 18-30 Jahre	Konv. Kraft	73.84	15.84	45
	HIIT	77.64	17.74	45
Mittel: 31-50 Jahre	Konv. Kraft	61.71	17.03	45
	HIIT	75.82	15.67	45
Alt: > 50 Jahre	Konv. Kraft	45.96	14.86	45
	HIIT	56.98	15.36	45

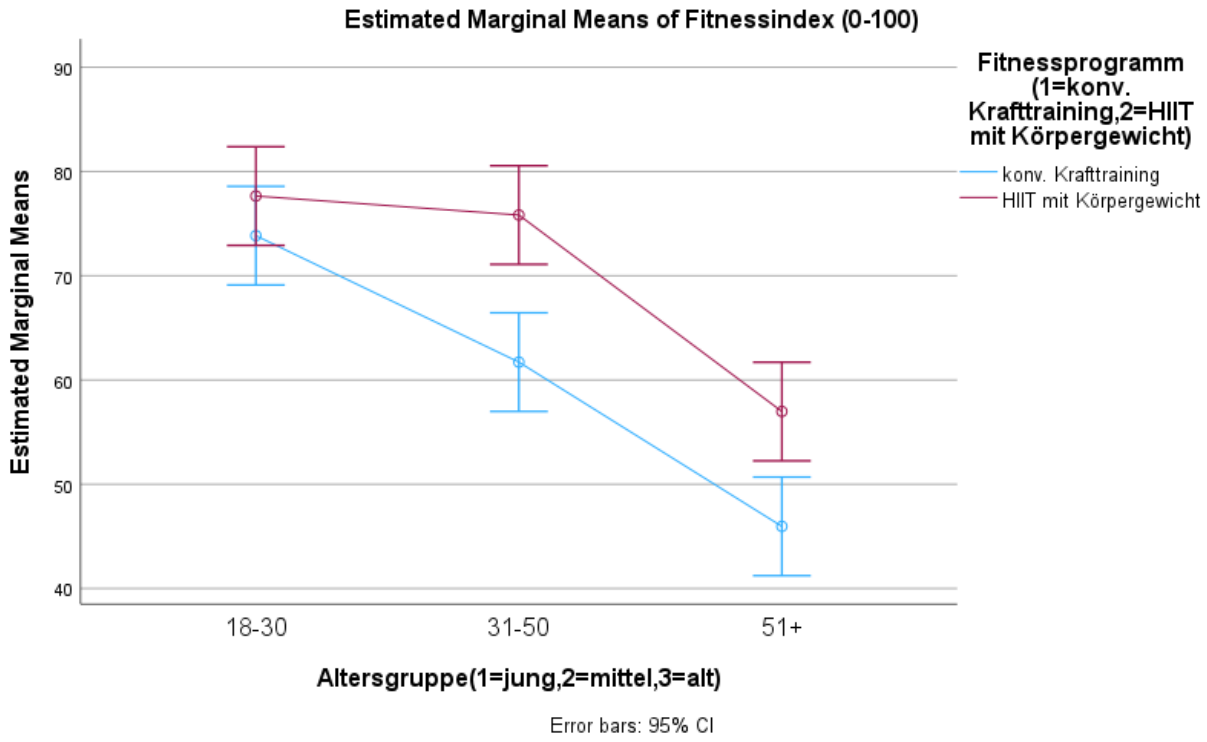


Abbildung L.23. Punkt- und Intervallschätzungen für die erzielte allgemeine Fitness in Abhängigkeit von Altersgruppe und verwendeter Trainingsmethode.

Beispiel 7.8

Um die Fragestellung zu untersuchen wurde eine zweifaktorielle Varianzanalyse ohne Messwiederholung durchgeführt. Deskriptive Statistiken für alle Kombinationen aus Faktorstufen sind in Tabelle L.9 angegeben.

Es ergibt sich ein signifikanter Haupteffekt für das Geschlecht, $F(1,116) = 18.11$, $p < .001$, $\eta_p^2 = .14$, d.h. ein großer Effekt gemäß Cohen (1988). Es ergibt sich auch ein signifikanter Haupteffekt für die Bedingung, $F(1,116) = 13.76$, $p < .001$, $\eta_p^2 = .11$, d.h. ein mittlerer Effekt gemäß Cohen (1988). Zudem ergibt sich eine signifikante Interaktion, $F(1,116) = 17.16$, $p < .001$, $\eta_p^2 = .13$, ein mittlerer Effekt gemäß Cohen (1988).

Paarweise post-hoc Vergleiche mit Bonferroni-korrigierten p-Werten zeigen, dass sich in der Kontrollbedingung Frauen und Männer nicht signifikant in der mittleren Symptomänderung unterscheiden ($p = .936$), in der Interventionsbedingung allerdings schon ($p < .001$). Während sich für Frauen Kontrollbedingung und Intervention in der mittleren Symptomänderung signifikant unterscheiden ($p < .001$), tun sie das für Männer nicht ($p = .760$).

Tabelle L.9

Mittelwerte und Standardabweichungen der Symptomstärkeänderungen sowie Stichprobenumfänge für alle Kombinationen aller Faktorstufen

Geschlecht	Bedingung	*M*	*SD*	*n*
Weiblich	tau	5.57	6.91	30
Weiblich	Intervention	17.07	7.83	30
Männlich	tau	5.40	9.01	30
Männlich	Intervention	4.77	8.19	30

Lösungen der Übungsaufgaben zu Kapitel 8

Beispiel 8.1

Richtig: (b), (c). Falsch: (a), (d).

Beispiel 8.2

Richtig: (a), (c), (d). Falsch: (b).

Beispiel 8.3

Nr.	**Aussage**	**R/F**
1)	Bei der Effektstärke η_p^2 werden Werte ab 0.01/0.06/0.14 gemäß Cohen (1988) als klein/mittel/groß bezeichnet.	R
2)	Beim Box Test handelt es sich um einen Test der Sphärizität.	F
3)	Die Greenhouse-Geisser-Korrektur ist zu konservativ, weshalb besser die Huynh-Feldt-Korrektur verwendet werden sollte.	R
4)	Die Gleichheit der Kovarianzmatrizen kann mit Mauchlys Test überprüft werden.	F

Beispiel 8.4

Zur Beantwortung der Fragestellung wurde eine einfaktorielle Varianzanalyse mit Messwiederholung durchgeführt. Beim Innersubjektfaktor handelt es sich um den Messzeitpunkt des Umweltverhaltens mit den drei Stufen (i) vor der Veranstaltung im Nationalpark, (ii) ein Monat nach der Veranstaltung und (iii) ein Jahr nach der Veranstaltung. Als Signifikanzniveau wurde α = .005 gewählt.

Da die Voraussetzung der Sphärizität verletzt war (p < .001), werden im Folgenden Huynh-Feldt-korrigierte Werte berichtet. Der Messzeitpunkt hat einen signifikanten Einfluss auf das Umweltverhalten, $F(1.58,209.11) = 7.09$, $p = .002$, $\eta_p^2 = .05$, d.h. 5% der Variabilität im Umweltverhalten können durch den Messzeitpunkt erklärt werden, was gemäß Cohen (1988) einem kleinen Effekt entspricht. Paarweise post-hoc Vergleiche mit p-Wert-Korrektur für multiple Vergleiche gemäß Fisher's LSD Methode ergaben zudem, dass sich das mittlere Umweltverhalten zu Messzeitpunkt 1 von dem zu Messzeitpunkt 2 signifikant unterscheidet (p < .001), aber nicht von dem zu Messzeitpunkt 3 (p = .076). Ferner unterscheidet sich das mittlere Umweltverhalten zu Messzeitpunkt 2 auch nicht

signifikant von dem zu Messzeitpunkt 3 (p = .081). Deskriptive Statistiken sind in Tabelle L.10 gegeben. Das Umweltverhalten ist in der Tat zu den Messzeitpunkten 2 und 3 höher als zum Messzeitpunkt 1.

Tabelle L.10

Deskriptive Statistiken

Messzeitpunkt	*M*	*SD*	*n*
1	12.61	5.79	133
2	13.36	6.08	133
3	13.05	6.31	133

Beispiel 8.5

Zur Beantwortung der Fragestellung wurde eine zweifaktorielle Varianzanalyse mit vollständiger Messwiederholung durchgeführt. Bei einem Messwiederholungsfaktor handelt es sich um den Messzeitpunkt mit den drei Stufen (i) eine halbe Stunde nach dem Lernen, (ii) ein Tag nach dem Lernen, und (iii) eine Woche nach dem Lernen. Beim anderen Messwiederholungsfaktor handelt es sich um die Bedeutung des Lernmaterials mit den zwei Stufen (i) eher niedrig (sinnlose Silbenpaare) und (ii) eher hoch (Vokabeln: Paare aus deutschen und japanischen Begriffen). Als Signifikanzniveau wurde α = .005 gewählt.

Da die Voraussetzung der Sphärizität sowohl für den Messzeitpunkt (p = .012) als auch die Interaktion zwischen Messzeitpunkt und Bedeutung (p = .036) verletzt war, werden im Folgenden Huynh-Feldt-korrigierte Werte berichtet. Es gibt einen signifikanten Haupteffekt für den Messzeitpunkt, $F(1.77,86.58) = 170.61$, $p < .001$, $\eta_p^2 = .78$, d.h. gemäß Cohen (1988) ein großer Effekt. Ebenso gibt es einen signifikanten Haupteffekt für die Bedeutung des Lernmaterials, $F(1,49) = 182.64$, $p < .001$, $\eta_p^2 = .79$, d.h. gemäß Cohen (1988) wiederum ein großer Effekt. Schließlich gibt es auch eine signifikante Interaktion zwischen den beiden Faktoren, $F(1.83,89.77) = 29.75$, $p < .001$, $\eta_p^2 = .38$, d.h. gemäß Cohen (1988) wiederum ein großer Effekt.

Um paarweise Unterschiede zu untersuchen wurden post-hoc Tests mit p-Wert-Korrektur für multiple Vergleiche gemäß Bonferroni berechnet. Es zeigt sich, dass die Behaltensleistung für beide Stufen des Faktors Bedeutung über die Zeit hinweg abnehmen. Für die sinnlosen Silben unterscheidet

sich die mittlere Behaltensleistung zwischen Zeitpunkt 1 und 2 sowie 1 und 3 signifikant (jeweils $p <$ .001), zwischen Zeitpunkt 2 und 3 jedoch nicht signifikant (p = .007). Selbiges gilt für die deutsch-japanischen Begriffspaare: auch hier unterscheidet sich die mittlere Behaltensleistung signifikant zwischen Zeitpunkt 1 und 2 sowie Zeitpunkt 1 und 3 (jeweils $p <$ .001), jedoch nicht signifikant zwischen Zeitpunkt 2 und 3 (p = .317). Die Behaltensleistung unterscheidet sich jedoch signifikant zwischen den beiden Bedeutungsstufen zu allen drei Zeitpunkten (jeweils $p <$ .001). Die Behaltensleistung ist immer höher im Fall der deutsch-japanischen Vokabeln.

Deskriptive Statistiken sind in Tabelle L.11 angeführt. Eine graphische Darstellung dieser Ergebnisse inklusive 95%-KI für die mittleren Leistungsindizes ist in Abbildung L.24 gegeben.

Tabelle L.11

Deskriptive Statistiken

Zeitpunkt	Bedeutung	*M*	*SD*	*n*
1	Sinnlose Silben	10.12	3.43	50
	Vokabeln	14.88	2.19	50
2	Sinnlose Silben	3.00	2.78	50
	Vokabeln	11.78	3.51	50
3	Sinnlose Silben	1.64	2.08	50
	Vokabeln	11.10	4.91	50

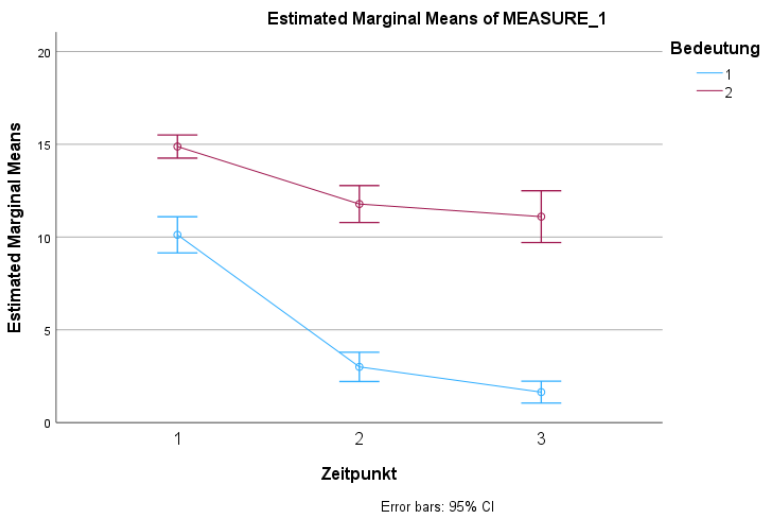


Abbildung L.24. Behaltensleistung über die drei untersuchten Zeitpunkte hinweg für jeweils 17 deutsch-japanische Vokabeln (rote Linie) und Paare aus sinnlosen Silben (blaue Linie).

Beispiel 8.6

Zur Erhellung der Fragestellung wurde eine zweifaktorielle Varianzanalyse mit einem gemischten 2x3 Design durchgeführt. Beim Faktor ohne Messwiederholung handelt es sich um die Medikation mit den drei Stufen wie in der Angabe gegeben. Beim Messwiederholungsfaktor handelt es sich um die beiden Blutdruckwerte, d.h. den systolischen und den diastolischen Blutdruck.

Die beiden Blutdruckwerte unterscheiden sich signifikant, $F(1,119) = 13953.96$, $p < .001$, $\eta_p^2 = .99$, d.h. die Art des Blutdruckwerts (d.h. systolisch oder diastolisch) erklärt 99% der Variabilität in den Werten, die nicht bereits durch andere systematische Effekte aufgeklärt werden. Auch für die Medikation gibt es einen signifikanten Haupteffekt, $F(2,119) = 98.11$, $p < .001$, $\eta_p^2 = .62$, d.h. die Medikation klärt 62% der Variabilität der Werte auf, die nicht durch andere Effekte aufgeklärt werden. Schließlich ist auch die Interaktion zwischen Art des Blutdruckwerts und Medikation signifikant, $F(2,119) = 7.49$, $p < .001$, $\eta_p^2 = .11$, was gemäß Cohen (1988) einem mittleren Effekt entspricht.

Paarweise post-hoc Vergleiche mit p-Wert-Korrektur gemäß Bonferroni zeigen, dass beide Blutdruckarten über die steigenden Medikationen hinweg abnehmen. Für beide Arten von Blutdruckwerten unterscheiden sich die Messwerte zwischen allen Medikationen signifikant ($p < .001$). Auch die beiden Blutdruckwerte unterscheiden sich für alle Stufen der Medikation signifikant voneinander ($p < .001$) mit plausiblen Bereichen für den Unterschied zwischen den Werten, die sehr gut dem Unterschied von etwa 45 mmHg zwischen den Normalbereichen für die beide Werte entsprechen.

Die plausiblen Werte für die Blutdruckwerte für die drei Stufen der Medikation sind in Tabelle L.12 numerisch gegeben und in Abbildung L.25 graphisch dargestellt. Tabelle L.12 enthält zudem deskriptive Statistiken für die Blutdruckwerte für die drei Medikationen. Man sieht, dass die plausiblen Werte für die Medikation von 16 mg Candesartan morgens und 8 mg Candesartan sowie 5 mg Amlodipin abends im Normalbereich für die beiden Arten der Blutdruckwerte liegen, während sie für die anderen beiden Medikationen teilweise zu hoch ausfallen. Die genannte Medikation scheint also von den drei überprüften Medikationen die passendste für den Patienten zu sein.

Tabelle L.12

Deskriptive Statistiken und 95%-KI

Medikation	Blutdruckwert	*M*	95%-KI	*SD*	*n*
2x8 mg Candesartan	Systolisch	148.00	[145.00,151.00]	8.13	20
	Diastolisch	102.43	[99.27,105.59]	7.54	20
16+8 mg Candasartan	Systolisch	128.44	[126.89,129.99]	7.14	75
	Diastolisch	84.14	[82.51,85.77]	7.83	75
16 mg Candesartan + 8/5 mg CandAm	Systolisch	121.67	[119.08,124.25]	4.15	27
	Diastolisch	74.20	[71.48,76.92]	4.09	27

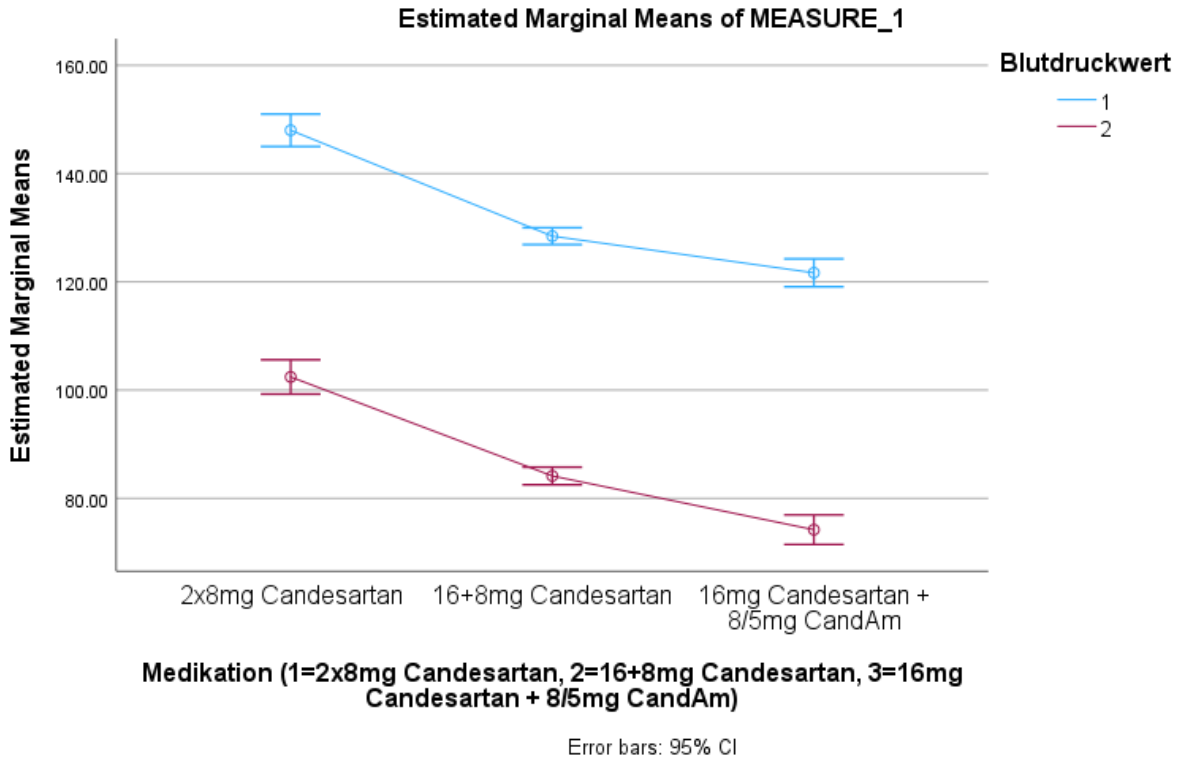


Abbildung L.25. Systolische (blaue Linie) und diastolische mittlere Blutdruckwerte und deren 95%-KI für die drei unterschiedlichen Medikationen aus Beispiel 8.6.

Beispiel 8.7

Antworten:

(a) Es ist eine zweifaktorielle Varianzanalyse mit Messwiederholung mit einem gemischten 2x3 Design durchzuführen. Bei dem Zeitpunkt des Leistungstests (Variablen t1, t2, t3) handelt es sich um einen dreistufigen Innersubjektfaktor, bei der Art der Lehrmethode (Variable Lehrmethode) um einen zweistufigen Zwischensubjektfaktor. Mit Ausnahme der Sphärizitätsannahme sind alle Voraussetzungen für die Varianzanalyse mit Messwiederholung erfüllt: (i) bei der AV handelt es sich um eine intervallskalierte Variable, (ii) es liegt ein balanciertes Design vor (selbst keine Voraussetzung, aber erhöht die Robustheit der Resultate gegenüber mehreren Voraussetzungsverletzungen), (iii) die Levene-Tests sowie der Box Test sind allesamt nicht signifikant ($p > .05$), (iv) die Normalverteilungsvoraussetzung muss gemäß Angabe nicht überprüft werden. Aufgrund der Verletzung der Sphärizitätsvoraussetzung ($W(2) = 0.62$, $p < .001$) werden im Folgenden Huynh-Feldt-korrigierte Teststatistiken berichtet.

(b) Die Ergebnisse im Leistungstest unterscheiden sich signifikant zwischen den einzelnen Testzeitpunkten, $F(1.46, 346.94) = 7768.64$, $p < .001$, $\eta_p^2 = 0.97$. Gemäß Cohen (1988) liegt damit ein großer Effekt vor.

(c) Die Ergebnisse im Leistungstest unterscheiden sich signifikant zwischen den beiden Lehrmethoden, $F(1, 238) = 15.20$, $p < .001$, $\eta_p^2 = 0.06$. Gemäß Cohen (1988) liegt damit ein mittlerer Effekt vor.

(d) Es liegt eine signifikante Interaktion zwischen Lehrmethode und Zeitpunkt vor, $F(1.46, 346.94) = 604.18$, $p < .001$, $\eta_p^2 = 0.72$. Gemäß Cohen (1988) liegt damit ein großer Effekt der Interaktion vor.

(e) Sämtliche Mittelwerte und Standardabweichungen sind in Tabelle L.13 zusammengefasst. Die Gruppengrößen waren jeweils zu $n = 120$ gegeben. Im Folgenden werden für multiple paarweise Vergleiche gemäß Bonferroni korrigierte p-Werte berichtet. Zu Zeitpunkt 1 liegt kein signifikanter Unterschied in den Ergebnissen beim Leistungstest für die beiden Lehrmethoden vor, $p = .753$. Zu Zeitpunkt 2 ist das mittlere Testergebnis signifikant höher bei der traditionellen Lehrmethode als bei Verwendung der flipped classroom Methode, $p = .022$. Zu Zeitpunkt 3 ist

das mittlere Testergebnis signifikant höher bei der flipped classroom Methode als bei der traditionellen Lehrmethode, $p < .001$. Sowohl bei der traditionellen Lehrmethode als auch bei der flipped classroom Methode ist das mittlere Testergebnis zu Zeitpunkt 1 jeweils signifikant niedriger als zu Zeitpunkt 2 (jeweils $p < .001$) und auch als zu Zeitpunkt 3 (jeweils $p < .001$), während es zu Zeitpunkt 3 jeweils signifikant niedriger ist als zu Zeitpunkt 2 (jeweils $p < .001$).

Tabelle L.13

Deskriptive Statistiken

Zeitpunkt	Lehrmethode	*M*	*SD*	*n*
1	Traditionell	29.08	11.15	120
1	Flipped classroom	29.53	11.02	120
2	Traditionell	83.00	12.69	120
2	Flipped classroom	79.26	12.49	120
3	Traditionell	42.80	16.87	120
3	Flipped classroom	65.93	16.47	120

Beispiel 8.8

Antworten:

(a) Zweifaktorielle Varianzanalyse mit Messwiederholung mit gemischtem 2x2 Design mit einem Innersubjektfaktor (Zeitpunkt mit zwei Faktorstufen: (1) vor und (2) nach der Therapie) und einem Zwischensubjektfaktor (Therapieform; ebenfalls zwei Stufen: (1) kognitive Verhaltenstherapie oder (2) achtsamkeitsbasierte Therapie).

(b) Zur Klärung der Fragestellung wurde eine Varianzanalyse mit Messwiederholung (gemischtes Design) durchgeführt. Die Depressionsschwere unterscheidet sich signifikant zwischen den beiden Messzeitpunkten (vor und nach den jeweiligen Therapien), $F(1, 98) = 129.37$, $p < .001$, $\eta_p^2 = 0.57$. Gemäß Cohen (1988) liegt damit eine große Effektstärke für den Effekt des Messzeitpunkts vor. Die Depressionsschwere unterscheidet sich nicht statistisch signifikant zwischen den beiden Therapieformen, $F(1, 98) = 0.158$, $p = .692$, $\eta_p^2 < 0.01$. Allerdings ergibt sich eine signifikante Interaktion zwischen Messzeitpunkt und Therapieform, $F(1, 98) = 10.84$,

p = .001, η_p^2 = 0.10, die gemäß Cohen (1988) einem mittleren Effekt entspricht. Paarweise post-hoc Vergleiche mit Bonferroni-Korrektur zeigen, dass sich die Depressionsschweren für beide Therapieformen weder zu Zeitpunkt 1 (p = .499) noch zu Zeitpunkt 2 (p = .206) signifikant voneinander unterscheiden. Zudem bessert sich die Depressionsschwere signifikant sowohl bei der kognitiven Verhaltenstherapie (p < .001) von Zeitpunkt 1 (M = 41.02, SD = 9.27) zu Zeitpunkt 2 (M = 30.82, SD = 11.99) als auch bei der achtsamkeitsbasierten Therapie (p < .001), ebenfalls von Zeitpunkt 1 (M = 39.60, SD = 11.53) zu Zeitpunkt 2 (M = 33.98, SD = 12.81).

(c) Es handelt sich um eine typische hybride Interaktion. Die Depressionsschwere bessert sich zwar in jedem Fall (Haupteffekt des Messzeitpunkts), unabhängig von der Therapieform, aber die Besserung ist deutlicher ausgeprägter bei der kognitiven Verhaltenstherapie (signifikanter Interaktionseffekt).

Beispiel 8.9

Ergebnisbericht: Um Unterschiede im moralischen Verhalten und dem Bedürfnis nach kognitiven Anforderungen je nach Spielegenre zu untersuchen, wurde eine zweifaktorielle Varianzanalyse ~~ohne~~mit Messwiederholung durchgeführt. Dabei wies der Zwischensubjektfaktor „Spielegenre“ ~~zwei~~drei Faktorstufen auf. Der Innersubjektfaktor berücksichtigte, um welchen der beiden Fragebögen es sich handelte. Die Interaktion zwischen den beiden Faktoren lässt darauf schließen, ob es zwischen den Spielegenres Unterschiede im Antwortverhalten auf die beiden Fragebögen gibt.

Die Varianzanalyse ergab einen signifikanten Haupteffekt für die Art des Fragebogens, F(1,297) = 9.64, p = ~~.031~~.002, η_p^2 = .03. Es ergab sich auch ein signifikanter Haupteffekt für das Spielegenre, F(2,297) = 27.21, p < .001, η_p^2 = .16. Die Interaktion zwischen Fragebogenart und Spielegenre war allerdings ~~nicht~~auch signifikant, F(2,297) = 19.42, p ~~= .116~~< .001, η_p^2 = .12, weshalb die Effekte der beiden Faktoren nicht unabhängig voneinander interpretiert werden können.

Bei Spieler:innen, die besonders gerne Egoshooter spielen, wurden sowohl beim Moralfragebogen (M = 49.90, SD = 9.09) als auch beim Kognitionsfragebogen (M = 51.34, SD = 9.52) vergleichsweise geringe Werte erreicht, die sich auch nicht signifikant voneinander unterschieden,

p = .188. Auch die Punktwerte der Spieler:innen, die besonders gerne Strategiespiele spielen, waren sehr ähnlich bei Moralfragebogen (M = ~~51.34~~58.23, SD = ~~9.52~~9.61) und bei Kognitionsfragebogen (M = 58.38, SD = 9.45), fielen aber vergleichsweise deutlich höher aus und unterschieden sich in beiden Fällen signifikant von den jeweiligen Werten der Egoshooter-Spieler:innen (p < .001). Rollenspieler:innen erzielten hingegen ganz andere Werte im Moralfragebogen (M = 59.87, SD = 9.01) als im Kognitionsfragebogen (M = 52.41, SD = 9.43), der Unterschied war ~~nicht~~auch signifikant (p < .001). Rollenspieler:innen erzielten im Kognitionsfragebogen ähnliche Werte wie Egoshooter-Spieler:innen (p = ~~.631~~> .999), während sie im Moralfragebogen ähnliche Werte wie Strategie-Spieler:innen erzielten (p > ~~.999~~=.631).

Beispiel 8.10

Ergebnisbericht: Zur statistischen Analyse wurde eine zweifaktorielle Varianzanalyse ~~ohne~~mit Messwiederholung (gemischtes Design) durchgeführt. Beim zweistufigen ~~Inner~~Zwischensubjektfaktor handelt es sich um die Variable, die angibt, ob es sich um eine Person mit besonders niedriger oder hoher Selbstwirksamkeit handelt. Beim ebenfalls zweistufigen ~~Zwischen~~Innersubjektfaktor handelt es sich um die Variable, die angibt, ob es sich um das Lernergebnis zur Spiel- oder zur Nichtspielversion der Lernaufgabe handelt.

Es ergibt sich (mit α = .05) ~~ein~~kein signifikanter Haupteffekt für die Selbstwirksamkeit (niedrig oder hoch), F(1,158) = 0.28, p = ~~.002~~.597, η_p^2 = ~~.60~~< .01. Es ergibt sich ~~ein~~kein signifikanter Haupteffekt für die Version der Lernaufgabe (Spiel oder Nichtspiel), F(1,158) = 0.10, p = ~~.001~~.753, η_p^2 = ~~.75~~< .01. Es ergibt sich eine signifikante Interaktion zwischen den beiden Faktoren, F(1,158) = 3769.10, p < .001, η_p^2 = .96.

Paarweise post-hoc Vergleiche mit gemäß Bonferroni korrigierten p-Werten ergeben, dass Personen mit niedriger Selbstwirksamkeit in der Spielversion (M = 45.56, SD = 9.81) signifikant höhere Ergebnisse als in der Nichtspielversion (M = 40.71, SD = 9.74) erzielen, p < .001. Bei Personen mit hoher Selbstwirksamkeit ist es gerade umgekehrt: Diese erzielen in der Spielversion (M = 41.44, SD = 8.01) signifikant niedrigere Ergebnisse als in der Nichtspielversion (M = 46.34, SD = 8.18), p < .001. Zudem erzielen in der Spielversion Personen mit niedriger Selbstwirksamkeit signifikant

höhere Ergebnisse als Personen mit hoher Selbstwirksamkeit, p ~~< .001~~ = .004. In der Nichtspielversion hingegen erzielen Personen mit niedriger Selbstwirksamkeit signifikant niedrigere Ergebnisse als Personen mit hoher Selbstwirksamkeit, $p < .001$.

Beispiel 8.11

Zur Erhellung der Fragestellung wurde eine einfaktorielle Varianzanalyse mit Messwiederholung durchgeführt. Deskriptive Statistiken für die Haarqualität der $n = 160$ Personen zu allen drei Zeitpunkten sind in Tabelle L.17 gegeben. Die Haarqualität unterscheidet sich im Mittel (mit $\alpha = .005$) signifikant zwischen den drei Zeitpunkten, $F(2, 318) = 163.94$, $p < .001$, $\eta_p^2 = .51$. D.h., gemäß Cohen (1988) liegt ein großer Effekt des Zeitpunkts für die resultierende Haarqualität vor. Paarweise post-hoc Vergleiche mit Bonferroni-Korrektur für die sich ergebenden p-Werte zeigen, dass die Haarqualität zu Zeitpunkt 1 zwar im Mittel mit signifikant niedrigeren Werten beurteilt wird als zu Zeitpunkt 2 ($p < .001$), aber mit signifikant höheren als zu Zeitpunkt 3 ($p < .001$). Die Haarqualität zu Zeitpunkt 2 wird zudem im Mittel auch mit signifikant höheren Werten beurteilt als zu Zeitpunkt 3 ($p < .001$).

Inhaltlich bedeutet dieses Ergebnis, dass die Haarqualität direkt nach der zweimonatigen Pflege mit den silikonhaltigen Produkten in der Tat höher ist als zu Beginn des Experiments. Das heißt insbesondere auch höher als ohne die Verwendung solcher Produkte, da lediglich Personen ausgewählt wurden, die bisher keine solchen Produkte verwendet haben. Allerdings ist die Haarqualität ein Jahr nach Beginn des Experiments, und das heißt insbesondere zehn Monate nach Absetzen der Pflege mit silikonhaltigen Produkten, geringer als zu Beginn des Experiments. Woran das genau liegen könnte, lässt sich aus diesem Experiment allerdings nicht schließen.

Tabelle L.14

Mittelwerte und Standardabweichungen für die Haarqualität zu den drei Zeitpunkten (Beispiel 8.11)

Zeitpunkt	*M*	*SD*
Beginn des Experiments	49.67	13.86
2 Monate später	58.78	15.14
1 Jahr später	39.06	14.57

Beispiel 8.12

Die mittleren Prüfungsleistungen unterscheiden sich signifikant für die beiden Messzeitpunkte, $F(1, 98) = 122.46$, $p < .001$, $\eta_p^2 = .56$, was einem großen Effekt gemäß Cohen (1988) entspricht. Die mittleren Prüfungsleistungen unterscheiden sich hingegen nicht signifikant für die beiden Lernmethoden, $F(1, 98) = 2.02$, $p = .158$, $\eta_p^2 = .02$, was einem kleinen Effekt gemäß Cohen (1988) entspricht. Zwischen Messzeitpunkt und Lernmethode besteht eine signifikante Interaktion, $F(1, 98) = 27.71$, $p < .001$, $\eta_p^2 = .22$, was gemäß Cohen (1988) einem großen Effekt entspricht.

Für paarweise post-hoc Vergleiche werden Bonferroni korrigierte p-Werte berichtet. Für beide Lernmethoden unterscheiden sich die mittleren Prüfungsleistungen signifikant voneinander zwischen beiden Messzeitpunkten (jeweils $p < .001$). Insbesondere nimmt die Prüfungsleistung für beide Lernmethoden von Messzeitpunkt 1 zu Messzeitpunkt 2 zu. Zu Messzeitpunkt 1, d.h. zu Semesterbeginn, unterscheiden sich die mittleren Prüfungsleistungen für die beiden Lernmethoden nicht signifikant voneinander ($p = .168$). Zu Messzeitpunkt 2 unterscheiden sich die mittleren Prüfungsleistungen für die beiden Lernmethoden allerdings signifikant voneinander ($p < .001$). Die Prüfungsleistung am Semesterende ist höher für das verteilte Lernen als für das massierte Lernen.

Deskriptive Statistiken sind in **Tabelle L.15** zusammengefasst.

Tabelle L.15

Mittelwerte und Standardabweichungen für die Prüfungsleistungen zu beiden Messzeitpunkte für beide Lernmethoden (Beispiel 8.12)

Zeitpunkt	Lernmethode	*M*	*SD*	*n*
Semesterbeginn	Massiertes Lernen	51.32	8.39	50
	Verteiltes Lernen	48.99	8.43	50
Semesterende	Massiertes Lernen	56.21	7.88	50
	Verteiltes Lernen	62.74	9.28	50

Beispiel 8.13

Alle folgenden inferenzstatistischen Ergebnisse beziehen sich jeweils auf ein Signifikanzniveau von $\alpha = .005$.

Die mittleren Leistungen beim Mathematiktest unterscheiden sich signifikant für die beiden Messzeitpunkte, $F(1, 98) = 181.97$, $p < .001$, $\eta_p^2 = .65$, was einem großen Effekt gemäß Cohen (1988) entspricht. Die mittleren Leistungen unterscheiden sich hingegen nicht signifikant zwischen den beiden Lehrmethoden, $F(1, 98) = 2.31$, $p = .132$, $\eta_p^2 = .02$, was einem kleinen Effekt gemäß Cohens Heuristik (1988) entspricht. Zwischen Messzeitpunkt und Lehrmethode besteht eine signifikante Interaktion, $F(1, 98) = 8.86$, $p = .004$, $\eta_p^2 = .08$, was gemäß Cohens Heuristik (1988) einem mittleren Effekt entspricht.

Für paarweise post-hoc Vergleiche werden gemäß Bonferroni korrigierte p-Werte berichtet. Für beide Lehrmethoden unterscheiden sich die mittleren Leistungen signifikant zwischen beiden Messzeitpunkten ($p < .001$). Insbesondere nimmt die Leistung im Mittel für beide Lehrmethoden von Beginn zu Ende des Semesters zu. Am Beginn des Semesters unterscheiden sich die mittleren Leistungen für die beiden Lehrmethoden nicht signifikant voneinander ($p = .849$). Auch am Ende des Semesters unterscheiden sich die mittleren Leistungen für die beiden Lehrmethoden nicht signifikant voneinander ($p = .025$).

Deskriptive Statistiken sind in Tabelle L.16 zusammengefasst. Wir sehen, dass die Mathematikleistung für beide Lehrmethoden zunimmt (Haupteffekt Messzeitpunkt). Die Leistungszunahme ist allerdings stärker ausgeprägt für die VR-Methode (Interaktion).

Tabelle L.16

Deskriptive Statistiken für beide Messzeitpunkte und Lehrmethoden (Beispiel 8.13)

Messzeitpunkt	Lehrmethode	*M*	*SD*	*n*
Semesterbeginn	VR	37.12	6.91	50
	Klassisch	40.12	6.30	50
Semesterende	VR	44.92	4.76	50
	klassisch	45.10	4.70	50

Lösungen der Übungsaufgaben zu Kapitel 9

Beispiel 9.1

Richtig: (c), (d). Falsch: (a), (b).

Beispiel 9.2

Richtig: (d). Falsch: (a), (b), (c).

Beispiel 9.3

(a) H_0: $\beta = 0$. H_1: $\beta \neq 0$.

(b) H_1.

(c) $a = 11.45$; $b = 0.34$.

(d) Bei einem Wert der Abhängigkeitskognitionen von 0 Punkten (Einheiten auf der Skala des entsprechenden Fragebogens) erwarten wir im Mittel einen Wert von 11.45 für die Depressionsschwere auf der BDI-Skala. Eine Erhöhung der Intensität der Abhängigkeitskognitionen um einen Punkt geht im Mittel mit einer Erhöhung von 0.34 Punkten auf der Skala von Becks Depressionsinventar einher.

(e) y = 11.45 + 0.34 * x.

Ergebnisbericht: Ein signifikanter Anteil der Varianz in der Depressionsschwere der untersuchten 50 Personen kann (mit $\alpha = .05$) auf die Intensität der Abhängigkeitskognitionen zurückgeführt werden, $F(1,48) = 4.32$, $p = .043$, $R^2 = 0.08$. Gemäß Cohens Heuristiken (1988) entspricht dies einem kleinen Effekt. Der Regressionskoeffizient für den Zusammenhang zwischen der Intensität der Abhängigkeitskognitionen und der Depressionsschwere unterscheidet sich signifikant von Null, $b = 0.34$ (stand. $\beta = 0.29$), $t(48) = 2.08$, $p = .043$ (zweiseitig). Der Koeffizient ist zudem positiv, d.h., je höher die Intensität der Abhängigkeitskognitionen, desto höher die Depressionsschwere. Eine Erhöhung der Intensität der Abhängigkeitskognitionen um einen Punkt geht gemäß dem einfachen Regressionsmodell im Mittel mit einer Erhöhung der Depressionsschwere um 0.34 Punkte auf der Skala von Becks Depressionsinventar einher.

Beispiel 9.4

Ergebnisbericht: Ein signifikanter Anteil der Varianz in den Verkaufszahlen kann (mit $\alpha = .005$) auf das verwendete Werbebudget zurückgeführt werden, $F(1,198) = 99.59$, $p < .001$, $R^2 = 0.34$. Gemäß Cohens Heuristiken (1988) entspricht dies einem großen Effekt. Der Regressionskoeffizient für den Zusammenhang zwischen der Anzahl verkaufter Alben und dem Werbebudget unterscheidet sich signifikant von Null, $b = 0.10$ (stand. $\beta = 0.58$), $t(198) = 9.98$, $p < .001$ (zweiseitig). Der Koeffizient ist zudem positiv, d.h., je höher das Werbebudget, desto höher die Verkaufszahlen. Eine Erhöhung des Werbebudgets um 1000 Englische Pfund geht gemäß dem einfachen Regressionsmodell im Mittel mit einer Erhöhung der Verkaufszahlen um 96 Alben einher.

Antworten auf die Fragen: Eine Erhöhung um eine Million Pfund würde im Mittel mit einer Erhöhung der Verkaufszahlen um 96000 einhergehen. Der Standardschätzfehler beträgt allerdings 65.99, d.h., die Streuung um die erwarteten Verkäufe bei einem bestimmten Werbebudget ist in derselben Größenordnung wie die Mehrverkäufe für die Steigerung des Werbebudgets selbst. Das heißt, es gibt eine beträchtliche Streuung. Das wiederum heißt, man könnte im Einzelfall sowohl weit über dem mittleren Mehrverkauf als auch weit darunter liegen.

Beispiel 9.5

Ergebnisbericht: Ein signifikanter Anteil der Varianz in den Gehältern kann (mit $\alpha = .005$) auf die Berufserfahrung zurückgeführt werden, $F(1,229) = 29.41$, $p < .001$, $R^2 = 0.11$. Gemäß Cohens Heuristiken (1988) entspricht dies einem kleinen Effekt. Der Regressionskoeffizient für den Zusammenhang zwischen dem Gehalt und der Berufserfahrung unterscheidet sich signifikant von Null, $b = 3.43$ (stand. $\beta = 0.34$), $t(229) = 5.42$, $p < .001$ (zweiseitig). Der Koeffizient ist zudem positiv, d.h., je höher die Berufserfahrung, desto höher das Gehalt. Eine Erhöhung der Berufserfahrung um ein Jahr geht gemäß dem einfachen Regressionsmodell im Mittel mit einer Erhöhung des Gehalts um 3.43 Pfund pro Tag einher.

Beispiel 9.6

Abbildung L.26 zeigt die SPSS-Ausgabe für eine lineare Regressionsanalyse mit dem Kriterium Depressionsschwere und dem Prädiktor Abhängigkeitskognitionen. Es ergibt sich eine Steigung b = 0.34 sowie ein Achsenabschnitt a = 11.49. Durch den linearen Zusammenhang mit den Abhängigkeitskognitionen können 8.3% an Varianz der Depressionsschwere der untersuchten 50 Personen aufgeklärt werden.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Abhängigkeitskognitionen[b]	.	Enter

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.287[a]	.083	.064	11.969

a. Predictors: (Constant), Abhängigkeitskognitionen

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	619.403	1	619.403	4.324	.043[b]
	Residual	6875.977	48	143.250		
	Total	7495.380	49			

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. Predictors: (Constant), Abhängigkeitskognitionen

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	11.486	3.637		3.158	.003
	Abhängigkeitskognitionen	.341	.164	.287	2.079	.043

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung L.26. SPSS-Ausgabe für eine lineare Regressionsanalyse mit dem Kriterium Depressionsschwere und dem Prädiktor Abhängigkeitskognitionen.

Daran ändert sich auch nichts, wenn die Abhängigkeitskognitionen um die mittlere Ausprägung der Abhängigkeitskognitionen in der Stichprobe zentriert werden wie Abbildung L.27 zeigt. Allerdings ist der Achsenabschnitt nun durch $a = 18.18$ gegeben. Das bedeutet, das bei mittlerer Ausprägung der Abhängigkeitskognitionen die Depressionsschwere 18.18 BDI-Punkte beträgt. An allen übrigen Bestandteilen der Ausgabe ändert sich selbstverständlich nichts.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Abhängigkeitskognitionen zentriert[b]	.	Enter

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.287[a]	.083	.064	11.969

a. Predictors: (Constant), Abhängigkeitskognitionen zentriert

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	619.403	1	619.403	4.324	.043[b]
	Residual	6875.977	48	143.250		
	Total	7495.380	49			

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. Predictors: (Constant), Abhängigkeitskognitionen zentriert

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18.180	1.693		10.741	<.001
	Abhängigkeitskognitionen zentriert	.341	.164	.287	2.079	.043

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung L.27. SPSS-Ausgabe für eine lineare Regressionsanalyse mit dem Kriterium Depressionsschwere und dem zentrierten Prädiktor Abhängigkeitskognitionen.

Wird anstelle der Abhängigkeitskognitionen ein entsprechend des Aufgabenteils (b) skalierter Prädiktor verwendet, so ergibt sich die in Abbildung L.28 gezeigte Ausgabe. Hier haben sich sowohl Achsenabschnitt und Steigung geändert. Der Achsenabschnitt a = 11.83 entspricht nun dem Depressionsniveau für die kleinste Ausprägung der Abhängigkeitskognitionen in der Stichprobe, die Steigung b = 12.95 entspricht der Änderung des Depressionsniveaus von der kleinsten zur größten Ausprägung der Abhängigkeitskognitionen in der Stichprobe.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Abhängigkeitskognitionen skaliert[b]	.	Enter

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.287[a]	.083	.064	11.969

a. Predictors: (Constant), Abhängigkeitskognitionen skaliert

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	619.403	1	619.403	4.324	.043[b]
	Residual	6875.977	48	143.250		
	Total	7495.380	49			

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. Predictors: (Constant), Abhängigkeitskognitionen skaliert

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	11.827	3.493		3.386	.001
	Abhängigkeitskognitionen skaliert	12.952	6.229	.287	2.079	.043

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung L.28. SPSS-Ausgabe für eine lineare Regressionsanalyse mit dem Kriterium Depressionsschwere und dem umskalierten Prädiktor Abhängigkeitskognitionen.

Abbildung L.29 zeigt schließlich eine entsprechende SPSS-Ausgabe für die studentisierten Abhängigkeitskognitionen als Prädiktor. Der Achsenabschnitt entspricht nun wiederum dem Depressionsniveau bei einer mittleren Ausprägung der Abhängigkeitskognitionen, die Steigung $b = 3.56$ entspricht nun der Änderung des Depressionsniveaus für eine Änderung der Abhängigkeitskognitionen um eine Standardabweichung.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Zscore: Abhängigkeitskognitionen[b]	.	Enter

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.287[a]	.083	.064	11.969

a. Predictors: (Constant), Zscore: Abhängigkeitskognitionen

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	619.403	1	619.403	4.324	.043[b]
	Residual	6875.977	48	143.250		
	Total	7495.380	49			

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

b. Predictors: (Constant), Zscore: Abhängigkeitskognitionen

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18.180	1.693		10.741	<.001
	Zscore: Abhängigkeitskognitionen	3.555	1.710	.287	2.079	.043

a. Dependent Variable: Depressionsschwere (Gesamtwert für Becks Depressionsinventar)

Abbildung L.29. SPSS-Ausgabe für eine lineare Regressionsanalyse mit dem Kriterium Depressionsschwere und dem studentisierten Prädiktor Abhängigkeitskognitionen.

Beispiel 9.7

Die Annahme eines linearen Zusammenhangs scheint berechtigt wie das Streudiagramm in Abbildung L.30 zeigt.

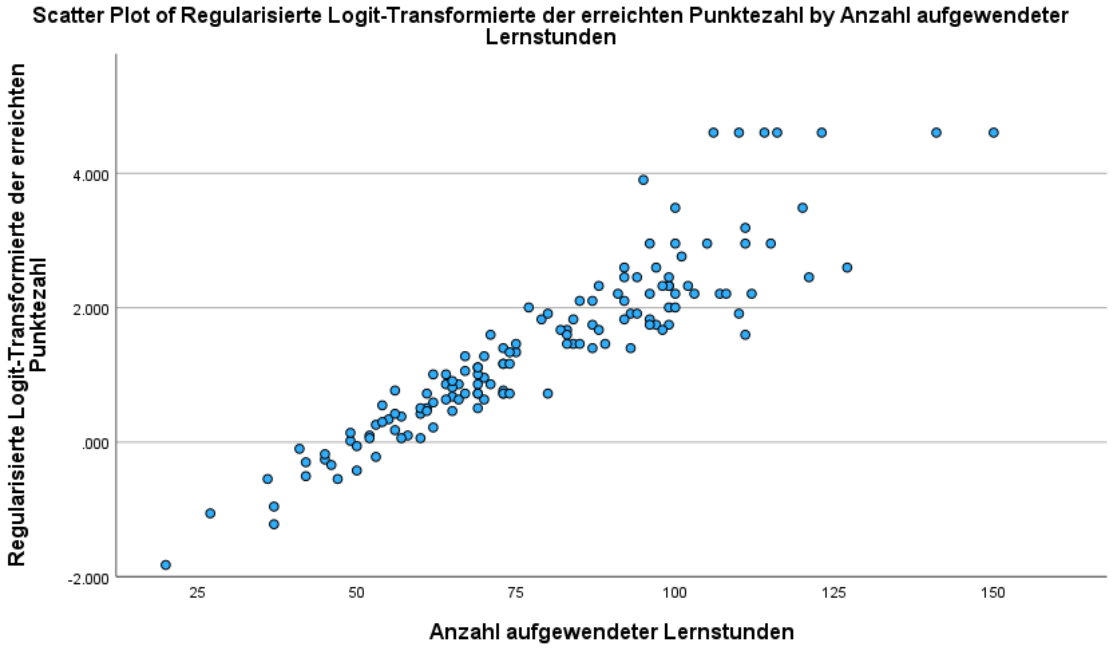


Abbildung L.30. Benötigtest Streudiagramm für Beispiel 9.7.

Eine einfache lineare Regressionsanalyse mit der Anzahl aufgewendeter Lernstunden als Prädiktor zeigt, dass das Ausmaß der aufgewendeten Lernzeit in der Tat einen signifikanten Anteil der Varianz des Kriteriums (*LogitPunkte*) erklären kann, $F(1,138) = 779.788$, $p < .001$, $R^2 = 0.85$. D.h., die aufgewendete Lernzeit kann alleine 85% der Varianz der abhängigen Variablen in der Stichprobe erklären, was gemäß Cohen (1988) einem großen Effekt entspricht. Der geschätzte Regressionskoeffizient des einzigen Prädiktors unterscheidet sich dementsprechend auch signifikant von Null, $b = 0.05$ (stand. $\beta = 0.92$), $t(138) = 27.93$, $p < .001$ (gerichtet). Der Schätzwert ist entsprechend der Hypothese positiv, d.h. nimmt die aufgewendete Lernzeit zu, so nimmt auch die abhängige Variable zu. Für einen Zuwachs der Lernzeit um 20 Stunden nimmt die abhängige Variable um 1 Einheit zu.

Beispiel 9.8

Eine einfache lineare Regressionsanalyse zeigt, dass sich im Mittel in der Tat ein signifikanter Anteil der Prüfungsleistung durch den Gemüseanteil aufklären lässt, $F(1,198) = 14.37$, $p < .001$, $R^2 = .07$, d.h., ein kleiner Effekt gemäß Cohen (1988). Der Anteil an Varianz der Prüfungsleistung, der in der Stichprobe durch den Gemüseanteil erklärt werden kann, beträgt 6.80%. Insbesondere ist der Regressionskoeffizient für den Gemüseanteil signifikant positiv, $b = 0.13$ (stand. $\beta = .26$), $t(198) = 3.79$, $p < .001$ (ungerichtet), d.h. ein höherer Gemüseanteil in der Ernährung geht mit einer höheren Prüfungsleistung einher. Eine Erhöhung des Gemüseanteils um 10% geht mit einer Erhöhung der Prüfungsleistung um 1.3% einher.

Ein Streudiagramm, in dem die Prüfungsleistung gegen den Gemüseanteil aufgetragen ist, ist in Abbildung L.31 dargestellt.

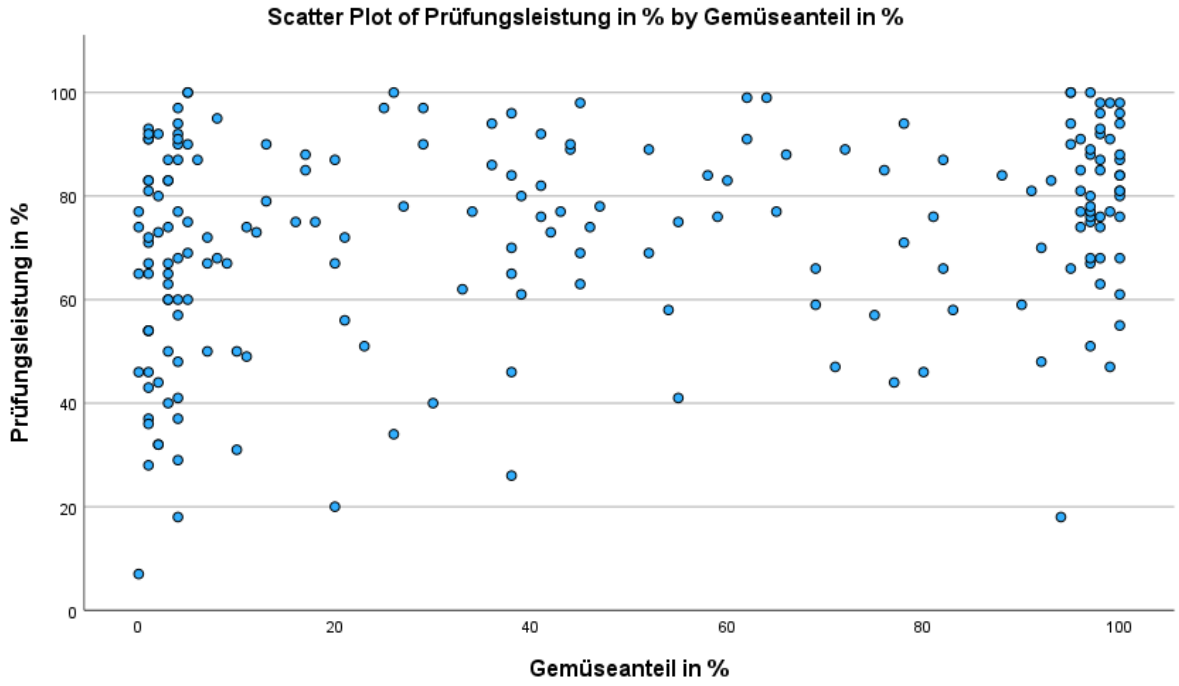


Abbildung L.31. Streudiagram für die Prüfungsleistung und den Gemüseanteil aus Beispiel 9.8.

Beispiel 9.9

(a) Ein signifikanter Anteil der Leistung bei einem Intelligenztest lässt sich auf die Menge an wöchentlich verzehrtem Brokkoli zurückführen, $F(1, 462) = 232.69$, $p < .001$, $R^2 = 0.34$. Für ein Kilogramm mehr an wöchentlich verzehrtem Brokkoli steigt die Intelligenzleistung um 39 IQ-Punkte, $b = 0.04$ (stand. $\beta = .58$), $t(462) = 15.25$, $p < .001$. Die Menge wöchentlich verzehrten Brokkolis erklärt 33.5% der Gesamtvarianz der Intelligenzleistung in der Stichprobe; es handelt sich gemäß Cohen (1988) um einen großen Effekt. Die Schätzwerte sowie Teststatistiken für die Modellparameter sind in Tabelle L.17 zusammengefasst.

(b) Das Regressionsmodell mit beiden Prädiktoren erklärt einen signifikanten Anteil der Varianz der Leistung beim Intelligenztest, $F(2, 461) = 119.28$, $p < .001$, $R^2 = 0.34$. Zusammen erklären beide Prädiktoren 34.1% der Gesamtvarianz der Intelligenzleistung in der Stichprobe; es handelt sich gemäß Cohen (1988) um einen großen Effekt. Allerdings ist (mit $\alpha = .005$) lediglich die Menge wöchentlich verzehrten Brokkolis ein signifikanter Prädiktor, $b_{\text{Brokkoli}} = 0.04$ (stand. $\beta = .54$), $t(461) = 12.59$, $p < .001$. Die Menge wöchentlich verzehrter Karotten ist hingegen (mit $\alpha = .005$) kein signifikanter Prädiktor, $b = 0.01$ ($\beta = .09$), $t(461) = 2.06$, $p = .040$. Das heißt, ist die Menge wöchentlich verzehrten Brokkolis bereits bekannt, kann mit der wöchentlich verzehrten Menge an Karotten kein signifikanter Zugewinn an nützlicher Information für die Vorhersage der Intelligenzleistung geleistet werden. Umgekehrt kann jedoch bei bekannter Menge wöchentlich verzehrter Karotten ein signifikanter Zugewinn an nützlicher Information für die Vorhersage der Intelligenzleistung durch die Berücksichtigung der wöchentlich verzehrten Menge an Brokkoli geleistet werden. Die Schätzwerte sowie Teststatistiken für die Modellparameter sind in Tabelle L.18 zusammengefasst.

Tabelle L.17

Schätzwerte und Teststatistiken für die Modellparameter des einfachen Regressionsmodells

Prädiktor	Schätzwert	Standardfehler	Stand. Koeff.	$t(462)$	p
Achsenabschnitt (a)	80.64	1.45		55.47	< .001
Brokkoliverzehr (b)	0.04	< 0.01	0.58	15.25	< .001

Tabelle L.18

Schätzwerte und Teststatistiken für die Modellparameter des multiplen Regressionsmodells

Prädiktor	Schätzwert	Standardfehler	Stand. Koeff.	$t(461)$	p
Achsenabschnitt (a)	79.41	1.57		50.68	< .001
Brokkoliverzehr (b_1)	0.04	< 0.01	0.54	12.59	< .001
Karottenverzehr (b_2)	0.01	< 0.01	0.09	2.06	.040

Beispiel 9.10

Eine multiple Regressionsanalyse zeigt, dass die Anzahl verkaufter CDs, der Ticketpreis sowie das Werbebudget einen signifikanten Anteil der Varianz der Anzahl von Konzertbesuchern erklären können, $F(3, 32) = 47.65$, $p < .001$, $R^2 = 0.82$. Insgesamt können die drei Prädiktoren 81.7% der Varianz der Anzahl der Konzertbesucher erklären, es handelt sich also um einen großen Effekt gemäß Cohen (1988). Steigt der Preis der Konzertkarten um einen Schweizer Franken, so sinkt die Besucherzahl (bei konstantem Werbebudget und CD-Verkauf) im Mittel um 43.23 Personen, $b_{\text{Preis}} = -43.23$ (stand. $\beta = -0.20$), $t(32) = -2.61$, $p = .014$. Steigt das Werbebudget um einen Schweizer Franken, nimmt die Besucheranzahl (bei konstanten Kartenpreis und CD-Verkauf) im Mittel um 0.54 Personen zu, $b_{\text{Werbung}} = 0.54$ (stand. $\beta = 0.74$), $t(32) = 9.66$, $p < .001$. Verkauft eine Band eine CD mehr, so nimmt die Besucherzahl (bei konstantem Kartenpreis und Werbebudget) im Mittel um 0.97 Personen zu, $b_{\text{CD_Verkauf}} = 0.97$ (stand. $\beta = 0.44$), $t(32) = 5.76$, $p < .001$. Die Schätzwerte sowie Teststatistiken für alle Modellparameter sind in Tabelle L.19 zusammengefasst.

Tabelle L.19

Schätzwerte und Teststatistiken für die Modellparameter des multiplen Regressionsmodells

Prädiktor	Schätzwert	Standardfehler	Stand. Koeff.	$t(32)$	p
Achsenabschnitt (a)	5091.21	1820.56		2.80	.009
Kartenpreis (b_{Preis})	-43.23	16.55	-0.20	-2.61	.014
Werbebudget (b_{Werbung})	0.54	0.06	0.74	9.66	< .001
CD-Verkauf ($b_{\text{CD_Verkauf}}$)	0.97	0.17	0.44	5.76	< .001

Beispiel 9.11

(a) Der Logarithmus der Oberflächentemperatur der untersuchten 47 Sterne kann keinen signifikanten Anteil der Varianz des Logarithmus der Leuchtkraft erklären, $F(1, 45) = 0.08$, $p = .782$. Das Ergebnis steht im Widerspruch zur theoretischen Vorhersage. Die geschätzten Modellparameter sind in Tabelle L.20 zusammengefasst.

(b) Werden die vier Sterne aus der Analyse ausgeschlossen, so kann der Logarithmus der Oberflächentemperatur der verbleibenden 43 Sterne einen signifikanten Anteil der Varianz des Logarithmus der Leuchtkraft erklären, $F(1, 41) = 30.55$, $p < .001$. Eine Erhöhung der Oberflächentemperatur um eine Größenordnung geht mit einer Erhöhung der Leuchtkraft um 1.48 Größenordnungen einher, $b = 1.48$ (stand. $\beta = 0.65$), $t(41) = 5.53$, $p < .001$. Die geschätzten Modellparameter sind in Tabelle L.21 zusammengefasst. Auf der Grundlage dieser Ergebnisse kann argumentiert werden, dass die Theorie zum Zusammenhang zwischen Leuchtkraft und Oberflächentemperatur dahingehend eventuell dahingehend präzisiert werden muss, dass der postulierte lineare Zusammenhang nur für Hauptreihensterne bzw. nicht für Sterne vom Typ Rote Riesen gilt. Diesen Eindruck erweckt auch ein Vergleich der beiden Streudiagramme, die sich für alle 47 Sterne bzw. nur für die 43 Sterne ohne den Roten Riesen ergeben, siehe Abbildung L.32 und Abbildung L.33.

Tabelle L.20

Schätzwerte und Teststatistiken für die Modellparameter des einfachen Regressionsmodells

Prädiktor	Schätzwert	Standardfehler	Stand. Koeff.	$t(45)$	p
Achsenabschnitt (a)	5.30	1.11		4.77	< .001
Oberflächentemperatur (b)	-0.07	0.25	-0.04	-0.28	.782

Tabelle L.21

Schätzwerte und Teststatistiken für die Modellparameter des einfachen Regressionsmodells

Prädiktor	Schätzwert	Standardfehler	Stand. Koeff.	$t(41)$	p
Achsenabschnitt (a)	-1.74	1.20		-1.45	.155
Oberflächentemperatur (b)	1.48	0.27	0.65	5.53	< .001

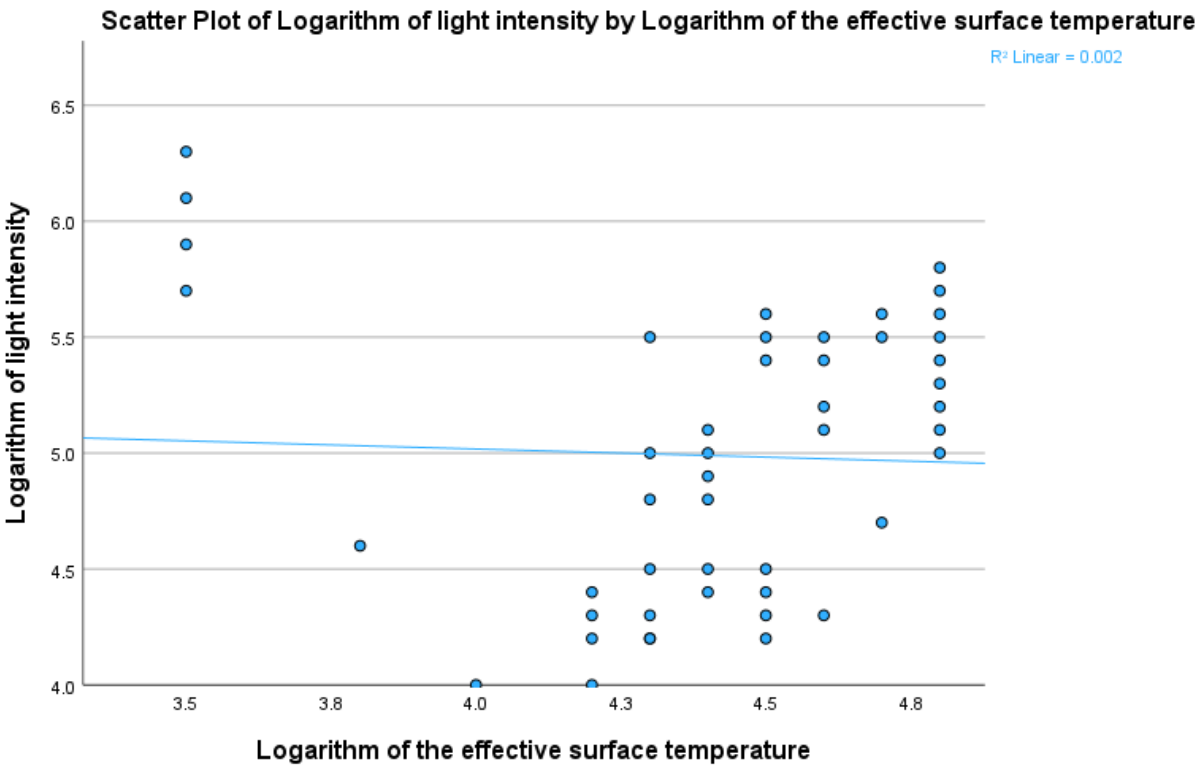


Abbildung L.32. Streudiagramm aller 47 Sterne aus Beispiel 9.11 inklusive Fitgerade (blaue Linie).

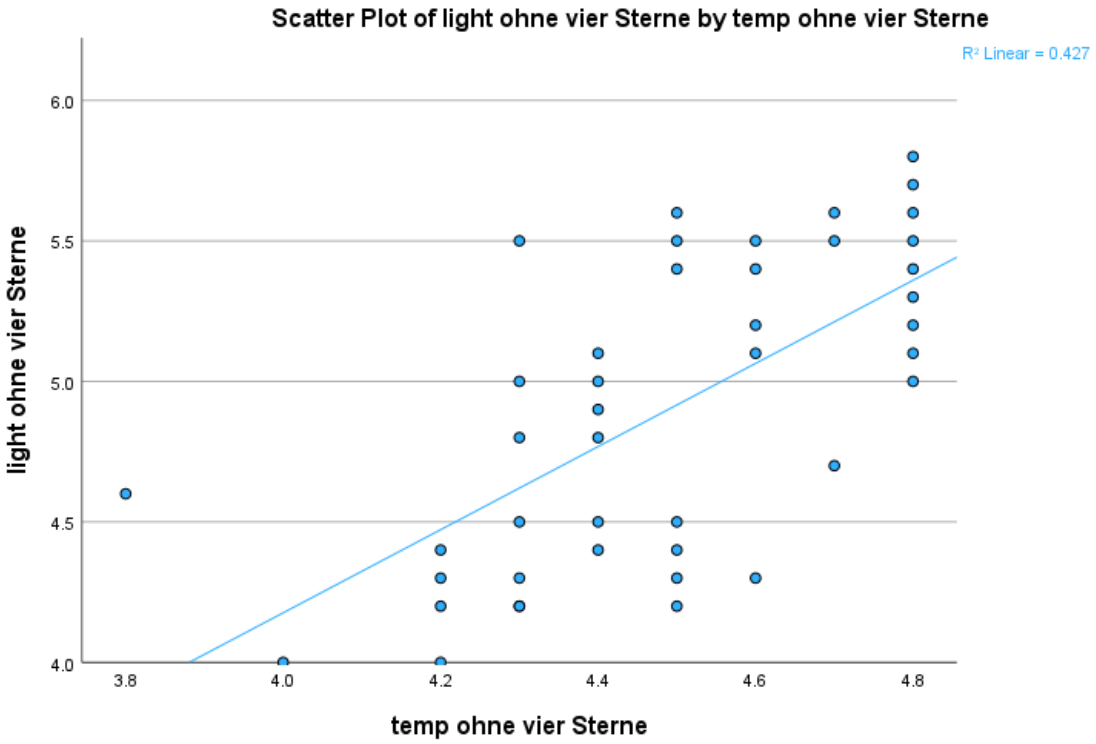


Abbildung L.33. Streudiagramm nur für die 43 Hauptreihensterne aus Beispiel 9.11 inklusive Fitgerade (blaue Linie).

Lösungen der Übungsaufgaben zu Kapitel 10

Beispiel 10.1

Richtig: (b). Falsch: (a), (c), (d).

Beispiel 10.2

Richtig: (a), (d). Falsch: (b), (c).

Beispiel 10.3

Regressionsdiagnostik für einfache Regression, d.h. für Teil (a) der Übungsaufgabe 10.3. Überprüfung der Linearitätsannahme: Keine Anzeichne für nichtlineare Verläufe, siehe Abbildung L.34. Überprüfung der Homoskedaszitätsannahme: scheint gut erfüllt (siehe Abbildung L.35). Überprüfung Normalverteilungsannahme: scheint ebenfalls gut erfüllt (siehe Abbildung L.36). Einflusswerte: 24 (= 24/464 = 5.2%) vorhanden (kritischer Wert für Cooks Distanz = 4/464 = 0.0086), aber nicht überraschend, da nahe an 5%.

Regressionsdiagnostik für multiple Regression, d.h. für Teil (b) der Übungsaufgabe 10.3. 1. Überprüfung der Linearitätsannahme: Keine Anzeichen für nichtlineare Verläufe durch Inspektion der partiellen Regressions-Plots, siehe Abbildung L.37 und Abbildung L.38. 2. Überprüfung der Homoskedaszitätsannahme: scheint gut erfüllt (siehe Abbildung L.39). 3. Überprüfung der Normalverteilungsannahme: scheint ebenfalls gut erfüllt (siehe Abbildung L.40). 4. Einflusswerte: 26 vorhanden (5.6%); allerdings nicht überraschend von erwartbaren 5% verschieden.

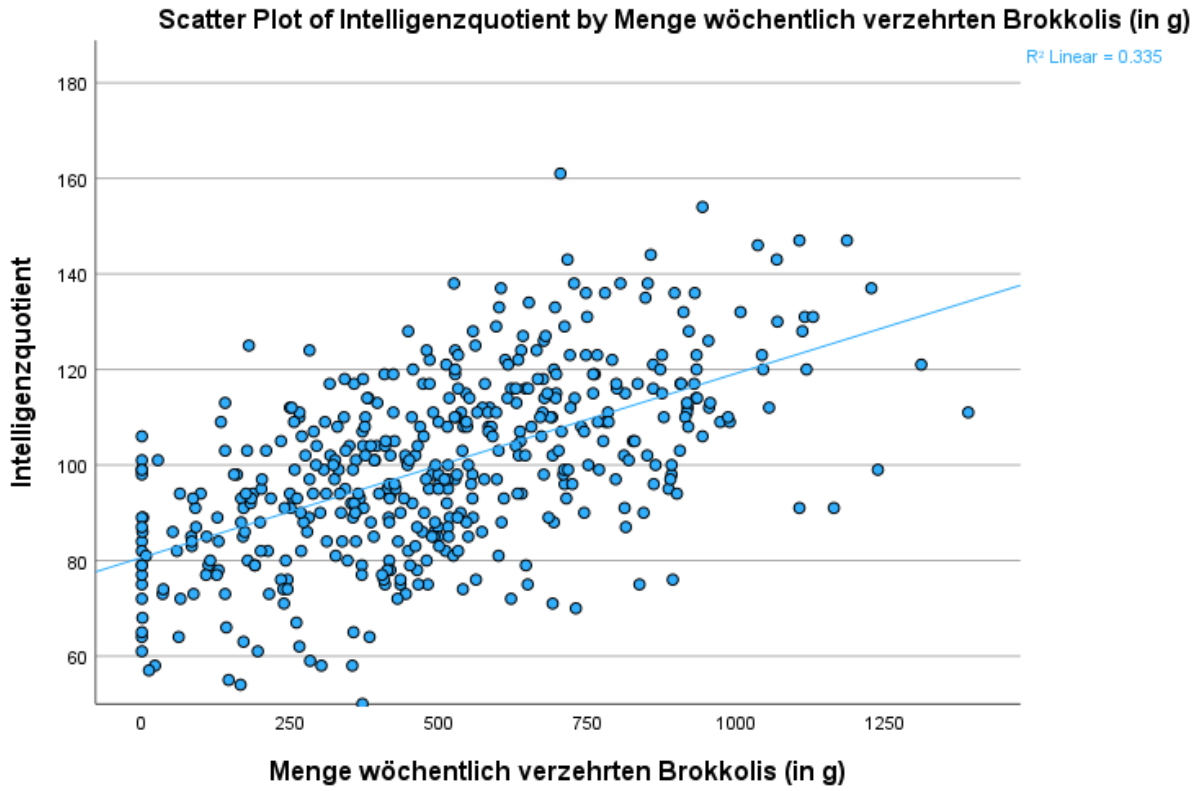


Abbildung L.34. Streudiagramm für IQ und Menge wöchentlich verzehrten Brokkolis aus Beispiel 10.3(a).

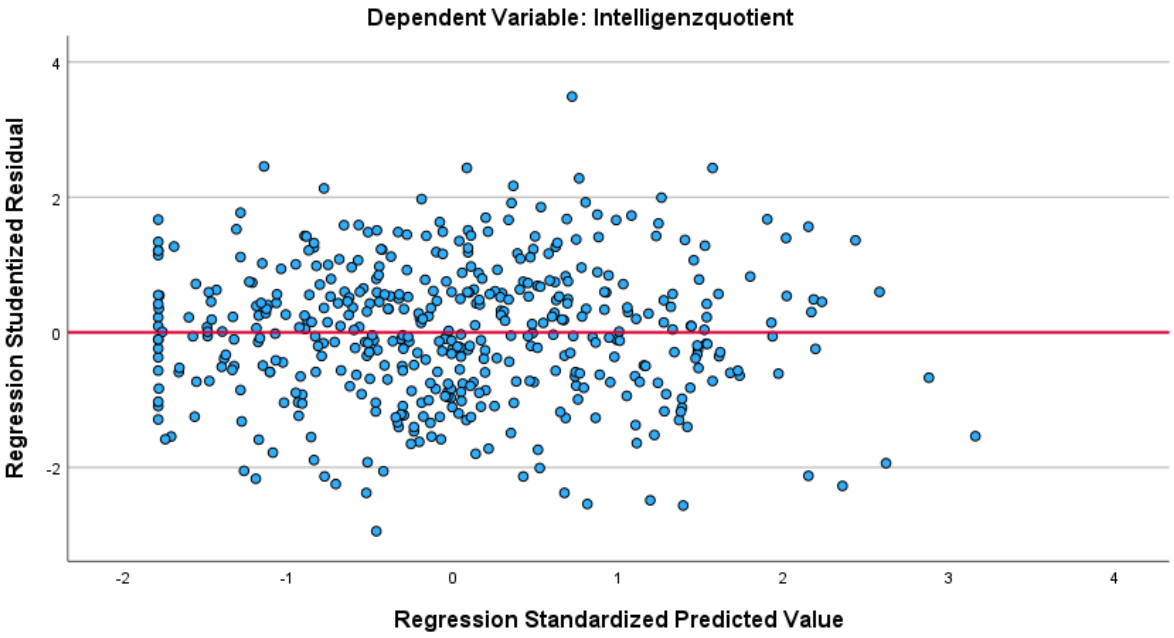


Abbildung L.35. Streudiagramm für die studentisierten Residuen aus Beispiel 10.3(a).

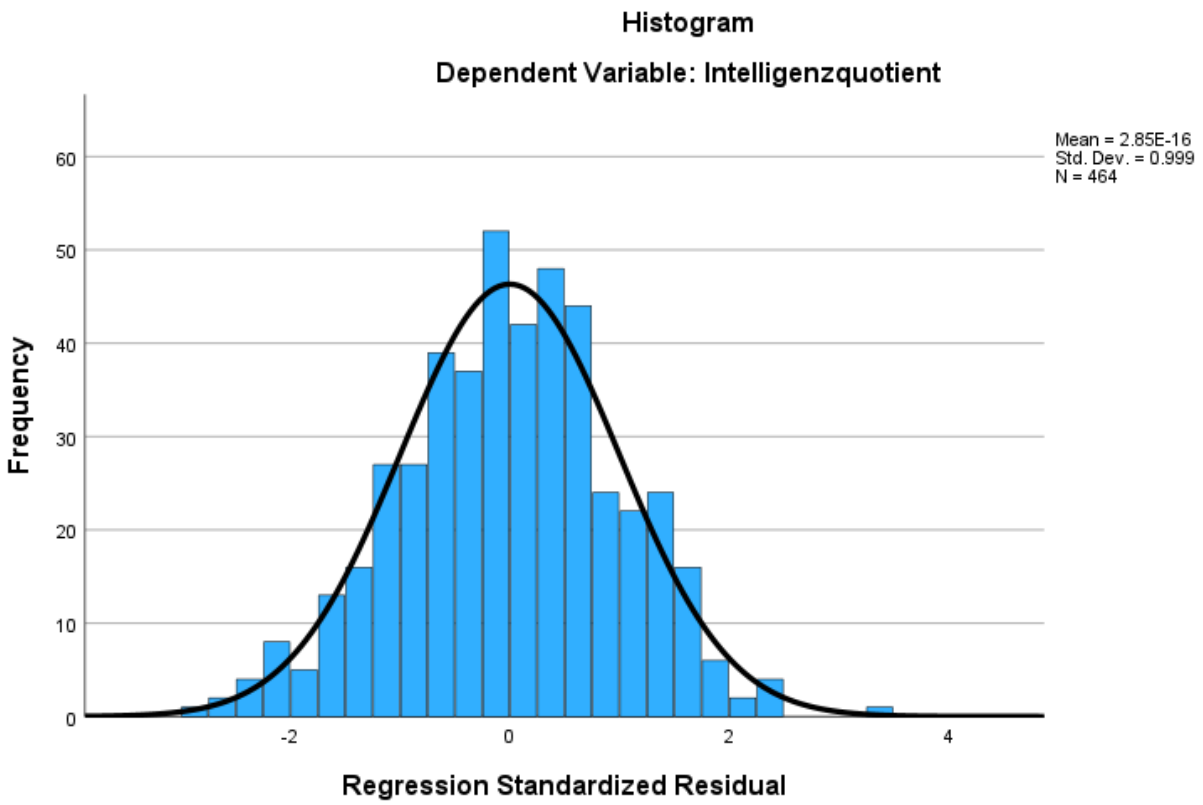


Abbildung L.36. Histogramm der standardisierten Residuen aus Beispiel 10.3(a).

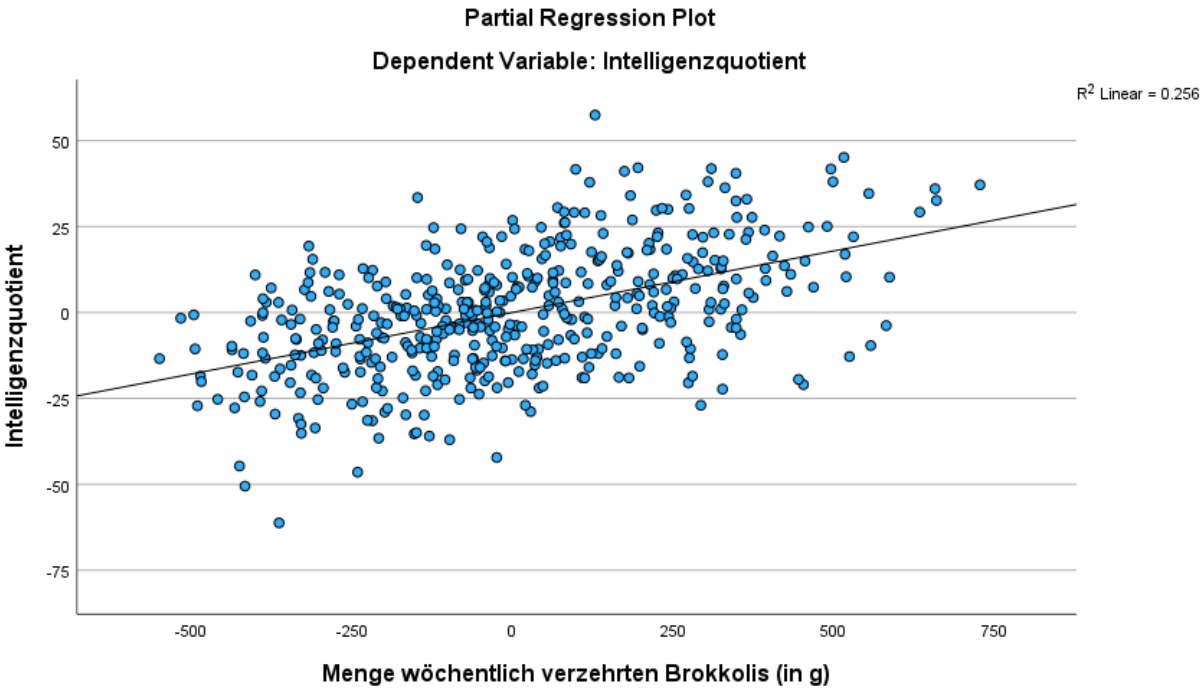


Abbildung L.37. Partieller Regressions-Plot für IQ und Menge wöchentlich verzehrten Brokkolis aus Beispiel 10.3(b).

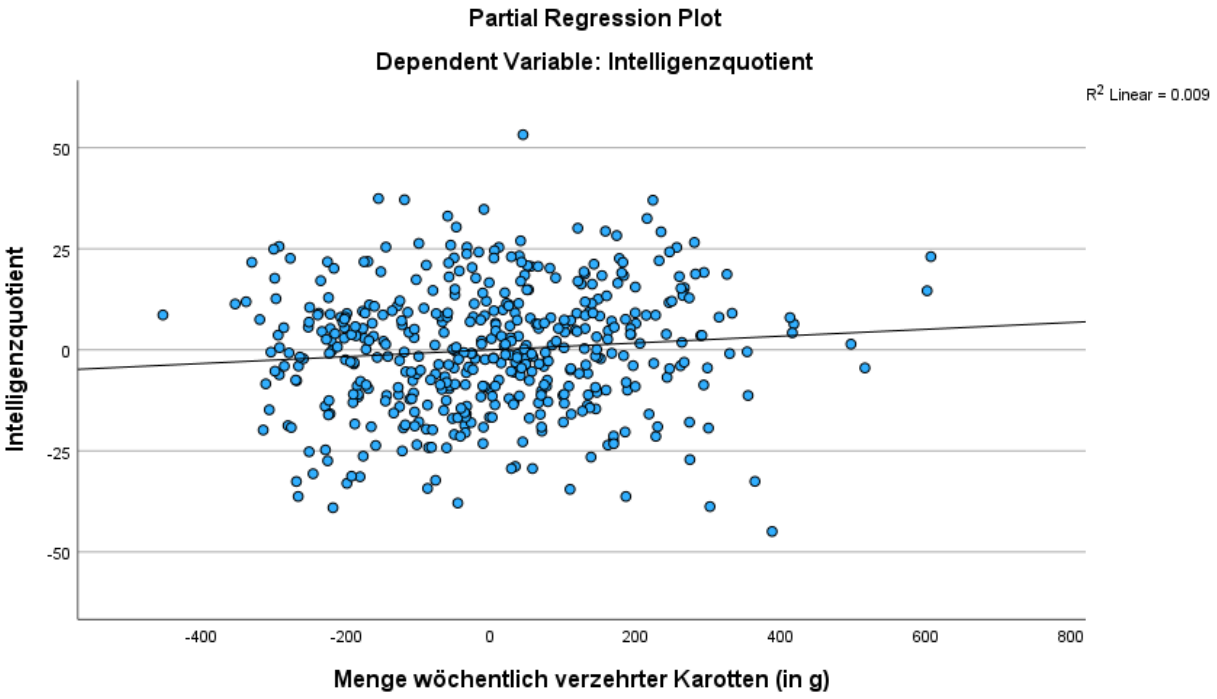


Abbildung L.38. Partieller Regressions-Plot für IQ und Menge wöchentlich verzehrter Karotten aus Beispiel 10.3(b).

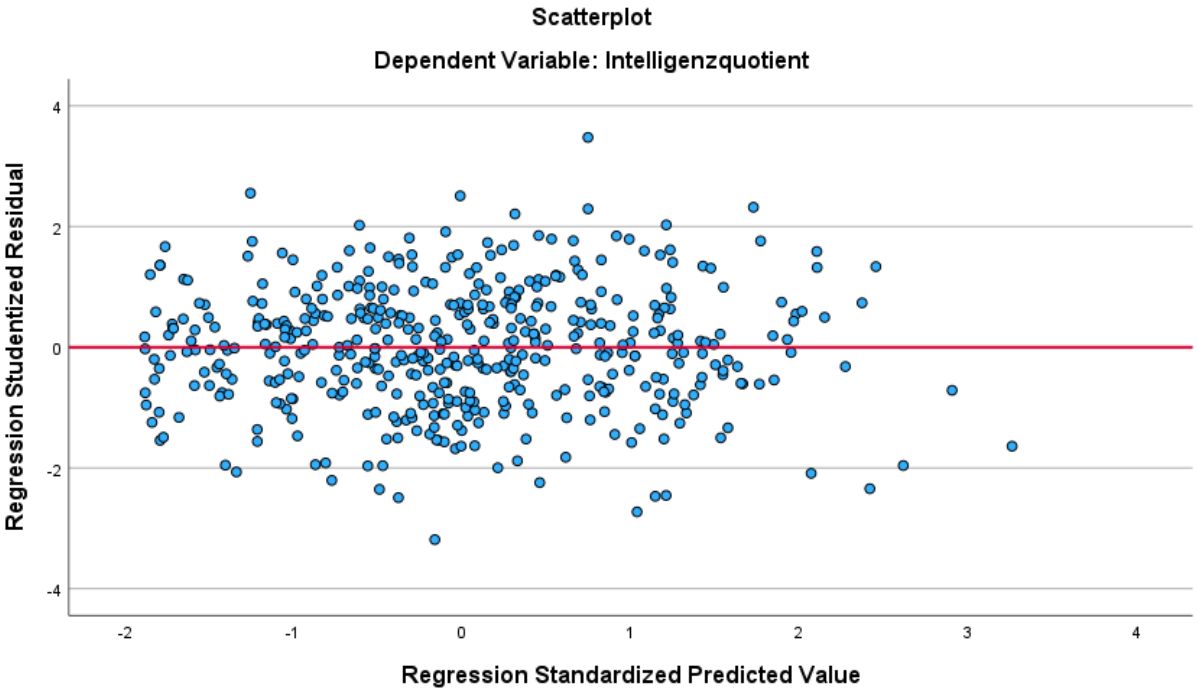


Abbildung L.39. Streudiagramm für die studentisierten Residuen aus Beispiel 10.3(b).

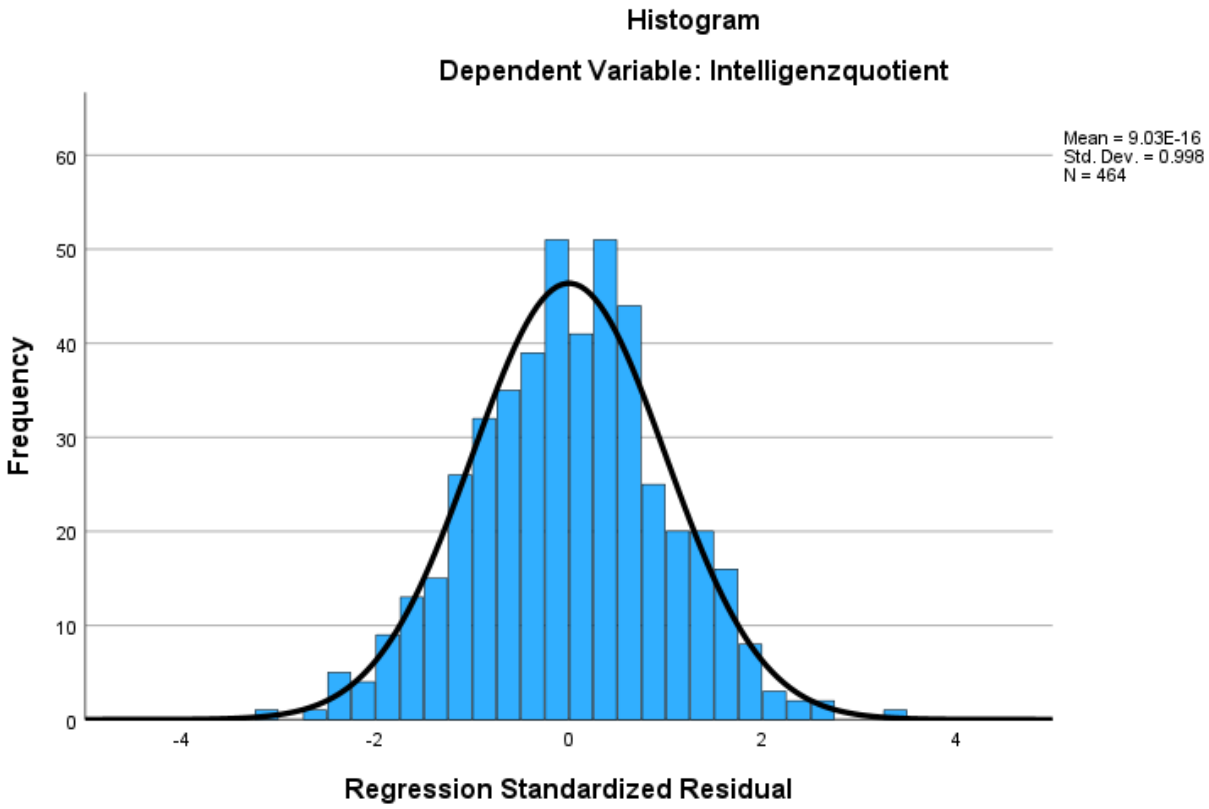


Abbildung L.40. Histogramm der standardisierten Residuen aus Beispiel 10.3(b).

Beispiel 10.4

Überprüfung der Linearitätsannahme: Keine Anzeichen für nichtlineare Verläufe durch Inspektion der partiellen Regressions-Plots, siehe „Kap10UE4.spv“.

Überprüfung der Homoskedaszitätsannahme: scheint gut erfüllt (Inspektion des Streudiagramms für die studentisierten Residuen, siehe „Kap10UE4.spv“).

Überprüfung der Normalverteilungsannahme: scheint ebenfalls gut erfüllt (Inspektion des Histogramms der standardisierten Residuen, siehe „Kap10UE4.spv“).

Drei Einflusswerte (8.3%) vorhanden (kritischer Wert für Cooks Distanz = 4/36 = 0.111); allerdings nicht überraschend von erwartbaren 5% verschieden (das wären ca. 2 von 36), siehe „konzertbesuche_inkl_cook.sav“.

Beispiel 10.5

Die Linearitätsannahme wurde mit den folgenden beiden partiellen Regressions-Plots (Abbildung L.41 und Abbildung L.42) überprüft und erscheint zumindest dem visuellen Eindruck nach gerechtfertigt.

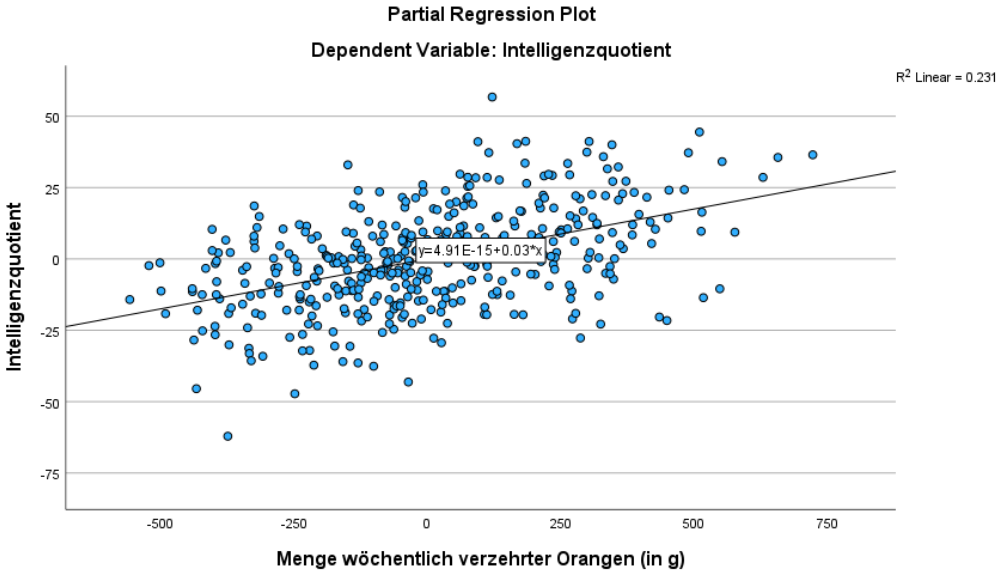


Abbildung L.41. Partieller Regressions-Plot für IQ und Menge wöchentlich verzehrter Orangen aus Beispiel 10.5.

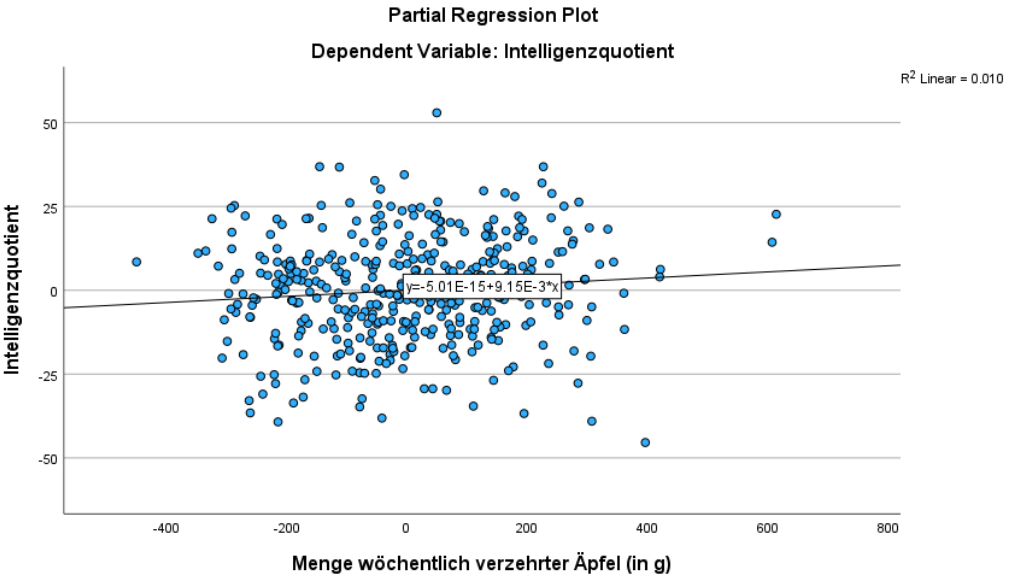


Abbildung L.42. Partieller Regressions-Plot für IQ und Menge wöchentlich verzehrter Äpfel aus Beispiel 10.5.

Die Homoskedastizitätsannahme wurde durch Inspektion des folgenden Streudiagramms (Abbildung L.43) untersucht und erscheint ebenfalls ganz gut erfüllt zu sein.

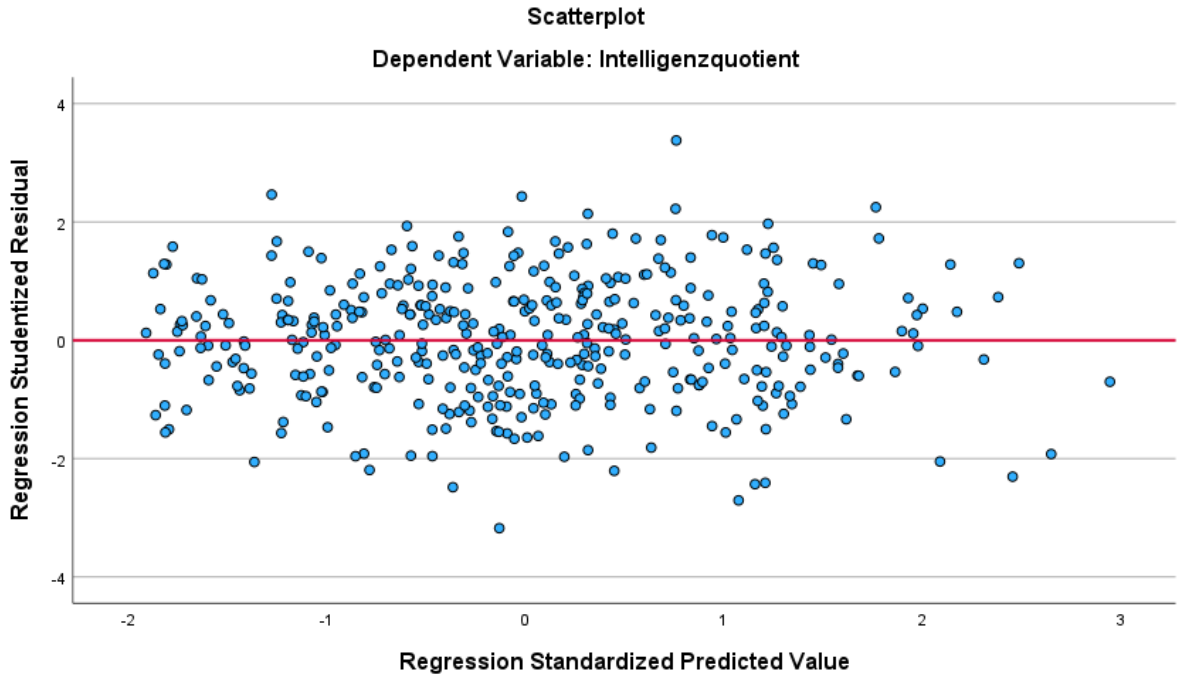


Abbildung L.43. Streudiagramm für die studentisierten Residuen aus Beispiel 10.5.

Die Normalverteilungsannahme wurde mittels Inspektion des folgenden Histogramms für die standardisierten Residuen (Abbildung L.44) überprüft und erscheint ebenfalls ganz gut erfüllt zu sein.

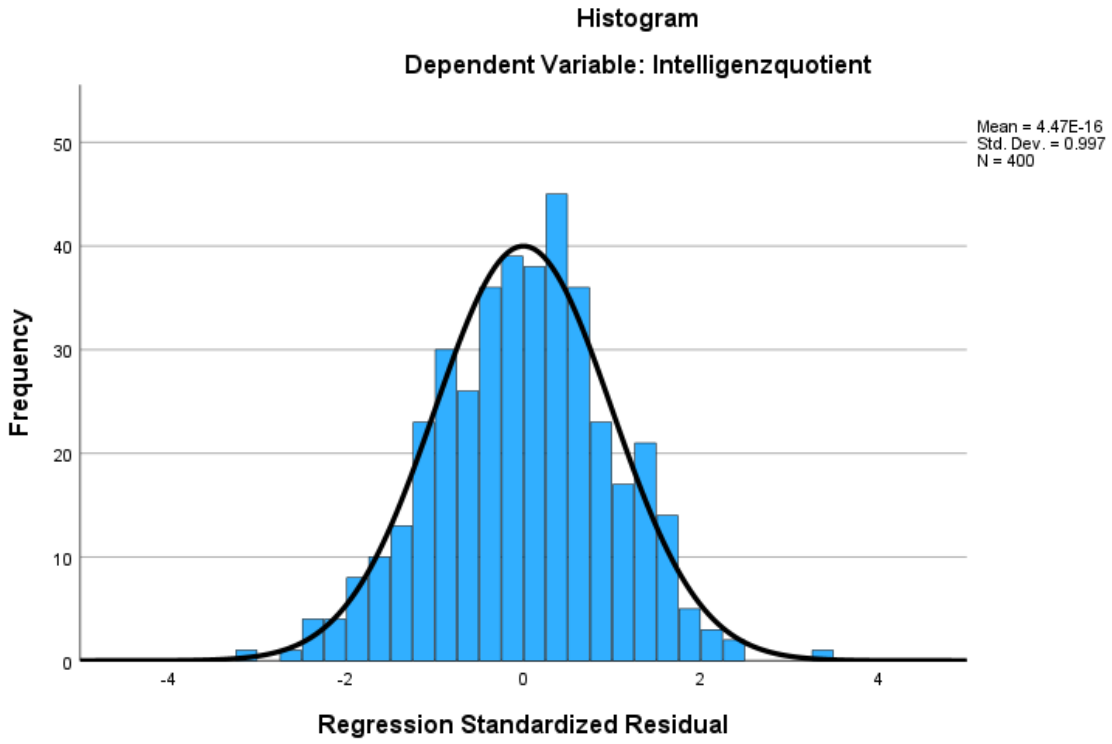


Abbildung L.44. Histogramm der standardisierten Residuen aus Beispiel 10.5.

Die Überprüfung auf Ausreißer wurde mittels Cooks Distanz durchgeführt. Es ergab sich, dass 22 Datenpunkte eine Cooks Distanz größer als $4/n = 0.01$ aufweisen, was allerdings 22/400 = 5.5% entspricht und damit im Bereich der erwarteten Anzahl an Datenpunkten mit einer solchen Cooks Distanz unter Gültigkeit aller Voraussetzungen für eine lineare Regressionsanalyse entspricht.

Beispiel 10.6

Die Linearitätsannahme wurde mit einem Streudiagramm untersucht und erscheint gerechtfertigt (siehe Abbildung L.45).

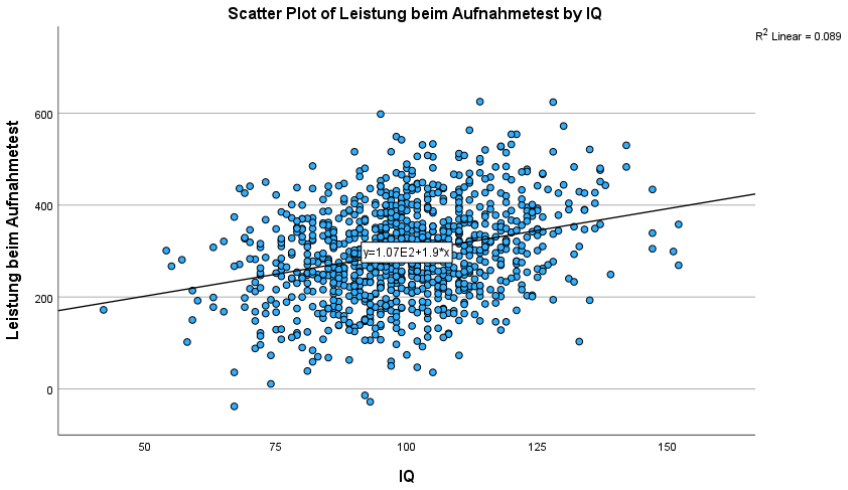


Abbildung L.45. Streudiagramm für IQ und Leistung aus Beispiel 10.6.

Die Homoskedastizitätsannahme wurde durch Inspektion des folgenden Streudiagramms (Abbildung L.46) untersucht und erscheint ebenfalls ganz gut erfüllt zu sein.

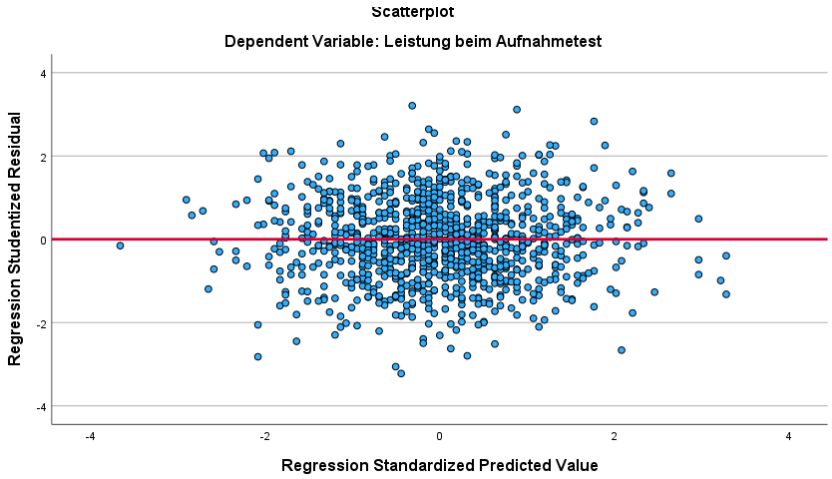


Abbildung L.46. Streudiagramm für die studentisierten Residuen aus Beispiel 10.6.

Die Normalverteilungsannahme wurde mittels Inspektion des folgenden Histogramms (Abbildung L.47) für die standardisierten Residuen überprüft und erscheint ebenfalls ganz gut erfüllt zu sein.

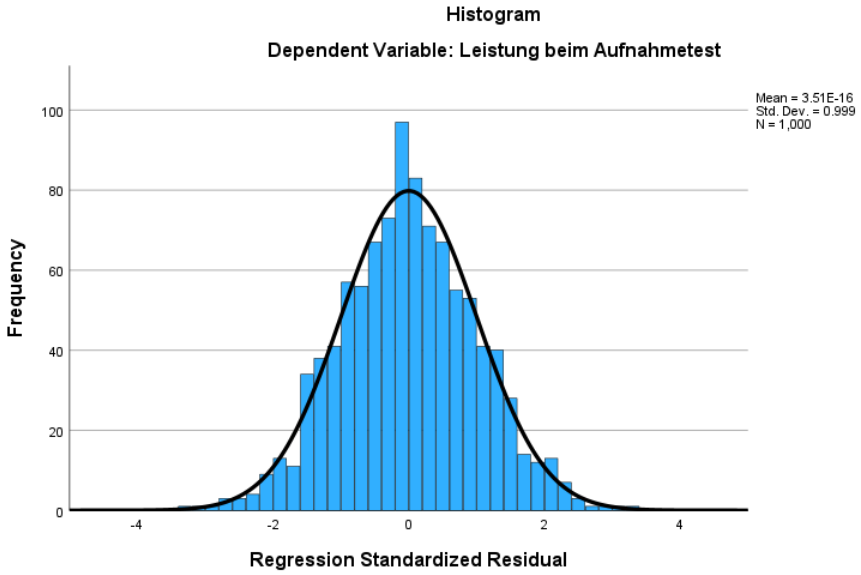


Abbildung L.47. Histogramm der standardisierten Residuen aus Beispiel 10.6.

Die Überprüfung auf Ausreißer wurde mittels Cooks Distanz durchgeführt. Es ergab sich, dass 53 Datenpunkte eine Cooks Distanz größer als $4/n = 0.004$ aufweisen, was allerdings 53/1000 = 5.3% entspricht und damit durchaus im Bereich der erwarteten Anzahl an Datenpunkten mit einer solchen Cooks Distanz unter Gültigkeit aller Voraussetzungen für eine lineare Regressionsanalyse entspricht.

Beispiel 10.7

Richtig: (a). Falsch: (b), (c), (d).

Beispiel 10.8

Richtig: (b). Falsch: (a), (c), (d).

Beispiel 10.9

Der eigenständige Beitrag zur erklärten Varianz der Menge wöchentlich verzehrten Brokkolis beträgt $0.476^2 = 0.226 = 22.6\%$. Der eigenständige erklärte Beitrag der Menge wöchentlich verzehrter Karotten beträgt $0.078^2 = 0.006 = 0.6\%$. Der Anteil der Gesamtvarianz, der nur durch beide Variablen gemeinsam erklärt werden kann, beträgt $34.1\% - 22.6\% - 0.6\% = 10.8\%$.

Beispiel 10.10

Der Preis der Konzertkarten erklärt eigenständig 3.9% der Varianz der Konzertbesucher, das Werbebudget erklärt eigenständig 53.3% der Varianz, und die Anzahl verkaufter CDs erklärt eigenständig 18.9% der Varianz. Damit verbleiben 5.7% der Varianz für die kombinierte Wirkung der drei Prädiktoren, und 18.3% der Varianz verbleiben unaufgeklärt.

Beispiel 10.11

Es werden 103 Personen benötigt.

Beispiel 10.12

Es werden 406 Personen benötigt.

Lösungen der Übungsaufgaben zu Kapitel 11

Beispiel 11.1

Richtig: (a), (d). Falsch: (b), (c).

Beispiel 11.2

Richtig: (a), (b), (c). Falsch: (d).

Beispiel 11.3

Eine einfache lineare Regressionsanalyse ergab, dass ein (mit α = .005) statistisch nicht signifikanter Anteil der Varianz im Restgeldbetrag nach einem Casinobesuch der untersuchten n = 78 Personen dadurch erklärt werden kann, ob die Personen spielsüchtig sind oder nicht, $F(1, 76) = 2.97$, $p = .089$, $R^2 = .04$; ein kleiner Effekt gemäß Cohen (1988).

Gemäß des resultierenden Regressionsmodells machen Personen ohne Spielsucht nach einem Casinobesuch im Mittel einen Nettoverlust von etwa 8 Euro (b = -8.40, $t(76) = -2.47$, $p = .016$). Dieser ist Verlust unterscheidet sich (mit α = .005) nicht statistisch signifikant unterschiedlich von Null. Personen mit Spielsucht bleiben im Mittel etwa 8 Euro weniger übrig als Personen ohne Spielsucht. Dieser Unterschied ist (mit α = .005) ebenfalls nicht statistisch signifikant (b = -8.30, stand. β = -.19, $t(76) = -1.72$, $p = .089$).

Beispiel 11.4

(a): *Transform >> Recode into Different Variables….* Die Variable *diagnosis* soll umkodiert werden in eine Dummy-Variable. Wir nennen diese hier *NOvsDIAG*, um anzudeuten, dass in dieser Variable die Referenzkategorie „keine Diagnose“ mit der Kategorie „ADHS Diagnose“ vergleichen wird.

Mit einem Klick auf den Knopf „*Old and New Values…*“ können dann die entsprechenden Werte eingegeben werden: Unter *Old Value >> Value:* geben wir no ein und unter *New Value >> Value:* geben wir 0 ein und klicken anschließend auf *Add*. Das gleiche machen wir für *Old Value >> Value:* yes und *New Value >> Value:* 1. Danach klicken wir auf *Continue* und *Paste*, um den Code in die Syntax zu bekommen. Dort können wir ihn dann ausführen und erhalten im Datensatz eine neue Variable namens *NOvsDIAG*.

(b): *Transform >> Recode into Different Variables…*. Die Variable *medication* soll umkodiert werden in zwei Dummy-Variablen. Wir nennen diese hier *NONEvsRIT* und *NONEvsADD*, um anzudeuten, dass hier „keine Medikation" die Referenzkategorie ist und jeweils mit der Kategorie „ritalin" und „adderall" vergleichen wird.

Mit einem Klick auf den Knopf „*Old and New Values…*" können dann die entsprechenden Werte eingegeben werden. Zuerst bauen wir die Dummy-Variable für *NONEvsRIT*: Unter *Old Value >> Value:* geben wir none ein und unter *New Value >> Value:* geben wir 0 ein und klicken anschließend auf *Add*. Das gleiche machen wir für *Old Value >> Value: adderall* und *New Value >> Value:* 0 und *Old Value >> Value: ritalin* und *New Value >> Value:* 1. Danach klicken wir auf *Continue* und *Paste*, um den Code in die Syntax zu bekommen. Dort können wir ihn dann ausführen und erhalten im Datensatz eine neue Variable namens *NONEvsRIT*.

Danach bauen wir die Dummy-Variable für *NONEvsADD*: Unter *Old Value >> Value:* geben wir none ein und unter *New Value >> Value:* geben wir 0 ein und klicken anschließend auf *Add*. Das gleiche machen wir für *Old Value >> Value: ritalin* und *New Value >> Value:* 0 und *Old Value >> Value: adderall* und *New Value >> Value:* 1. Danach klicken wir auf *Continue* und *Paste*, um den Code in die Syntax zu bekommen. Dort können wir ihn dann ausführen und erhalten im Datensatz eine neue Variable namens *NONEvsADD*.

(c): Zuerst lassen wir uns den Mittelwert der Variable *tolerance* über *Analyze >> Descriptive Statistics >> Frequencies…* ausgeben. Mit Doppelklick in die Tabelle in der Ausgabe und dann einem Doppelklick auf den Mittelwert (21.385), wird der Mittelwert mit > 3 Nachkommastellen angezeigt und kann so mit höherer Genauigkeit rauskopiert werden (21.384920634920633).

Danach wird über *Transform >> Compute Variable…* die neue zentrierte Variable berechnet. Unter „Target Variable" schreiben wir den neuen Variablennamen c_tolerance (= centered tolerance) hinein. In das Feld „Numeric Expression" schreiben wir tolerance - 21.384920634920633. Danach klicken wir auf *Paste* und führen den Code in der Syntax aus.

(d): Über *Transform >> Compute Variable…* können wir schlussendlich auch die Interaktionsvariable zwischen dem Dummy für ADHS Diagnose (*NOvsDIAG*) und der zentrierten Frustrationstoleranz-

Variable berechnen. Unter „Target Variable“ schreiben wir den neuen Variablennamen NOvsDIAG_X_c_tol hinein. In das Feld „Numeric Expression“ schreiben wir NOvsDIAG * c_tolerance. Danach klicken wir auf *Paste* und führen den Code in der Syntax aus.

Beispiel 11.5

Eine multiple lineare Regressionsanalyse ergab, dass ein (mit $\alpha = .005$) statistisch signifikanter Anteil der Varianz im Restgeldbetrag nach einem Casinobesuch der untersuchten $n = 78$ Personen durch die Prädiktoren Spielsucht, Restgeldbetrag eines vorangegangenen Casinobesuchs und deren Interaktion aufgeklärt werden kann, $F(3, 74) = 51.93$, $p < .001$, $R^2 = .68$; ein großer Effekt gemäß Cohen (1988).

Betrachtet man die einzelnen Regressionsparameter, machen Personen ohne Spielsucht, die bei einem vorangegangenen Casinobesuch mit 0 Euro Restgeldbetrag ausgestiegen sind, nach einem erneuten Casinobesuch im Mittel einen Nettogewinn von etwa 8 Euro ($b = 7.78$, $t(74) = 1.52$, $p = .133$). Dieser ist Gewinn unterscheidet sich (mit $\alpha = .005$) nicht statistisch signifikant unterschiedlich von Null.

Personen mit Spielsucht, die bei einem vorangegangenen Casinobesuch mit 0 Euro Restgeldbetrag ausgestiegen sind, bleiben im Mittel etwa 17 weniger übrig als Personen ohne Spielsucht ($b = -16.58$, stand. $\beta = -.17$, $t(74) = -2.17$, $p = .033$). Dieser Unterschied ist (mit $\alpha = .005$) nicht statistisch signifikant.

Wenn bei Personen ohne Spielsucht der Restbetrag beim Casinobesuch vor der Intervention um 1 Euro höher ist, erwartet man einen um im Durchschnitt etwa 60 Cent höheren Restbetrag beim Casinobesuch nach der Intervention ($b = 0.66$, stand. $\beta = .28$, $t(74) = 2.55$, $p = .013$). Dieser Anstieg ist (mit $\alpha = .005$) nicht statistisch signifikant.

Wenn bei Personen mit Spielsucht der Restbetrag beim Casinobesuch vor der Intervention um 1 Euro höher ist, erwartet man einen um im Durchschnitt etwa 2 Euro (= 0.656 + 1.343) höheren Restbetrag beim Casinobesuch nach der Intervention.

Der Unterschied zwischen den beiden Effekten der Spielsüchtigen und Nicht-spielsüchtigen ist (mit $\alpha = .005$) statistisch signifikant ($b = 1.34$, stand. $\beta = .32$, $t(74) = 4.15$, $p < .001$).

Beispiel 11.6

Eine multiple lineare Regressionsanalyse ergab, dass ein (mit $\alpha = .005$) statistisch signifikanter Anteil der Varianz im ADHS-Wert der untersuchten $n = 126$ Personen durch die Prädiktoren Frustrationstoleranz, ADHS-Diagnose und deren Interaktion aufgeklärt werden kann, $F(3, 122) = 21.19$, $p < .001$, $R^2 = .34$; ein großer Effekt gemäß Cohen (1988).

Betrachtet man die einzelnen Regressionsparameter, haben Personen ohne ADHS Diagnose und einer mittleren Frustrationstoleranz im Mittel einen ADHS-Wert von etwa 31 Punkten ($b = 30.50$, $t(122) = 20.11$, $p < .001$). Im Vergleich dazu, haben Personen mit ADHS-Diagnose (bei mittlerer Frustrationstoleranz) im Mittel etwa 5 Punkte mehr im ADHS Fragebogen ($b = 4.90$, stand. $\beta = .20$, $t(122) = 2.67$, $p = .009$). Dieser Unterschied im ADHS-Wert zwischen Personen mit und ohne ADHS-Diagnose ist (mit $\alpha = .005$) nicht statistisch signifikant.

Weiters zeigt sich bei Personen ohne ADHS-Diagnose ein negativer Zusammenhang zwischen Frustrationstoleranz und ADHS-Wert: Bei einem Anstieg von 1 Punkt im Frustrationstoleranz-Fragebogen, erwartet man eine Verringerung des ADHS-Werts um etwa 0.5 Punkte ($b = -0.52$, stand. $\beta = - .34$, $t(122) = -2.91$, $p = .004$). Dieser Zusammenhang ist (mit $\alpha = .005$) statistisch signifikant.

Bei Personen mit ADHS-Diagnose findet sich ebenfalls ein negativer Zusammenhang zwischen Frustrationstoleranz und ADHS-Wert: Bei einem Anstieg von 1 Punkt im Frustrationstoleranz-Fragebogen, erwartet man eine Verringerung des ADHS-Werts um etwa 0.9 Punkte (= -0.52 – 0.42 = - 0.94). Der Unterschied im Zusammenhang von Frustrationstoleranz und ADHS-Wert zwischen Personen mit und ohne ADHS-Diagnose ist (mit $\alpha = .005$) nicht statistisch signifikant ($b = -0.42$, stand. $\beta = -.20$, $t(122) = -1.78$, $p = .078$).

Beispiel 11.7

Eine multiple lineare Regressionsanalyse ergab, dass ein (mit $\alpha = .05$) statistisch signifikanter Anteil der Varianz im ADHS-Wert der untersuchten $n = 126$ Personen durch die Prädiktoren Frustrationstoleranz, ADHS-Medikation und deren Interaktion aufgeklärt werden kann, $F(5, 120) = 21.39$, $p < .001$, $R^2 = .47$; ein großer Effekt gemäß Cohen (1988).

Betrachtet man die einzelnen Regressionsparameter, haben Personen, die kein ADHS-Medikament einnehmen, und eine mittlere Frustrationstoleranz haben im Mittel einen ADHS-Wert von etwa 31 Punkten ($b = 30.50$, $t(120) = 22.24$, $p < .001$).

Im Vergleich dazu, haben Personen (mit mittlerer Frustrationstoleranz), die Ritalin einnehmen, im Mittel etwa 4 Punkte mehr im ADHS-Fragebogen ($b = 3.96$, stand. $\beta = .16$, $t(122) = 1.99$, $p = .048$). Dieser Unterschied im ADHS-Wert zwischen Personen, die kein ADHS-Medikament einnehmen, im Vergleich zu Personen, die Ritalin einnehmen, ist (mit $\alpha = .05$) statistisch signifikant. Personen (mit mittlerer Frustrationstoleranz), die Adderall einnehmen, haben im Vergleich zu Personen, die kein ADHS-Medikament einnehmen, etwa 3.5 Punkte mehr im ADHS-Fragebogen ($b = 3.50$, stand. $\beta = .15$, $t(122) = 1.85$, $p = .068$). Dieser Unterschied ist (mit $\alpha = .05$) statistisch nicht signifikant.

Weiters zeigt sich bei Personen, die kein ADHS-Medikament einnehmen, ein negativer Zusammenhang zwischen Frustrationstoleranz und ADHS-Wert: Bei einem Anstieg von 1 Punkt im Frustrationstoleranz-Fragebogen, erwartet man eine Verringerung des ADHS-Werts um etwa 0.5 Punkte ($b = -0.52$, stand. $\beta = -.34$, $t(120) = -3.21$, $p = .002$). Dieser Zusammenhang ist (mit $\alpha = .05$) statistisch signifikant.

Bei Personen, die Ritalin einnehmen, ist der negative Zusammenhang zwischen Frustrationstoleranz und ADHS-Wert noch stärker negativ: Bei einem Anstieg von 1 Punkt im Frustrationstoleranz-Fragebogen, erwartet man eine Verringerung des ADHS-Werts um etwa 1.7 Punkte. Der Unterschied im Zusammenhang zwischen Frustrationstoleranz und ADHS-Wert ist bei Personen, die kein ADHS-Medikament einnehmen, und Personen, die Ritalin einnehmen, (mit $\alpha = .05$) statistisch signifikant ($b = -1.13$, stand. $\beta = -.39$, $t(120) = -4.34$, $p < .001$).

Bei Personen, die Adderall einnehmen, ist der Zusammenhang zwischen Frustrationstoleranz und ADHS-Wert leicht negativ: Bei einem Anstieg von 1 Punkt im Frustrationstoleranz-Fragebogen, erwartet man eine Verringerung des ADHS-Werts um etwa 0.1 Punkte. Der Unterschied im Zusammenhang zwischen Frustrationstoleranz und ADHS-Wert ist bei Personen, die kein ADHS-Medikament einnehmen, und Personen, die Adderall einnehmen, (mit $\alpha = .05$) statistisch nicht signifikant ($b = 0.39$, stand. $\beta = .13$, $t(120) = 1.48$, $p = .141$).

Beispiel 11.8

(a) Die vervollständigte Tabelle sieht wie folgt aus:

	Unstandardized B	*t*	Sig.
(Constant)	108.89	53.67	<.001
Gruppe	-8.94	-2.83	.007

(b) Eine einfache lineare Regressionsanalyse ergab, dass ein (mit $\alpha = .05$) statistisch signifikanter Anteil der Varianz im IQ der untersuchten $n = 46$ Personen dadurch erklärt werden kann, ob die Personen BWL oder Psychologie studieren, $F(1, 44) = 8.02$, $p = .007$, $R^2 = .15$; ein mittlerer Effekt gemäß Cohen (1988).

Gemäß des resultierenden Regressionsmodells haben Psychologiestudierende im Mittel einen IQ von zirka 109 ($b = 108.89$, $t(44) = 53.67$, $p < .001$). BWL-Studierende haben im Mittel einen IQ der etwa 9 IQ-Punkte niedriger ist als der von Psychologiestudierenden ($b = -8.94$, $t(44) = -2.83$, $p = .007$). Dieser Unterschied ist (mit $\alpha = .005$) statistisch signifikant.

Beispiel 11.9

(a) Zuerst lassen wir uns den Mittelwert der Variablen *water* und *shade* über *Analyze >> Descriptive Statistics >> Frequencies…* ausgeben. Hier in diesem Beispiel sieht man auch mit Blick in die Daten, dass der Mittelwert beider Variablen genau 2 ist.

Danach wird über *Transform >> Compute Variable…* die neue zentrierte Variable zuerst für *water*, dann für *shade* berechnet. Unter „Target Variable“ schreiben wir den neuen Variablennamen c_water (= centered water) hinein. In das Feld „Numeric Expression“ schreiben wir „water – 2“. Danach klicken wir auf Paste, führen den Code in der Syntax aus und wiederholen das Ganze für die Variable *shade*.

(b) Über *Transform >> Compute Variable…* können wir dann die Interaktionsvariable zwischen den beiden zentrierten Variablen *c_water* und *c_shade* berechnen. Unter „Target Variable“ schreiben wir den neuen Variablennamen *c_water_X_c_shade* hinein. In das Feld „Numeric Expression“ schreiben wir *c_water*c_shade*. Danach klicken wir auf *Paste* und führen den Code in der Syntax aus.

(c) Eine multiple lineare Regressionsanalyse ergab, dass ein (mit α = .005) statistisch signifikanter Anteil der Varianz in der Größe der Tulpenblüten von n = 27 Tulpen durch die Prädiktoren Feuchtigkeit, Beschattung und deren Interaktion aufgeklärt werden kann, $F(3, 23) = 23.33$, $p < .001$, $R^2 = .75$; ein großer Effekt gemäß Cohen (1988).

Betrachtet man die einzelnen Regressionsparameter, haben Tulpenblüten bei mittlerer Feuchtigkeit und mittlerer Beschattung eine durchschnittliche Größe von 129 (b = 128.99, $t(23) = 12.68$, $p < .001$). Für eine mittlere Beschattung, erwartet man bei einem Anstieg der Feuchtigkeit um 1 einen Anstieg in der Blütengröße um 75.8 (b = 75.80, stand. β = .68, $t(23) = 6.56$, $p < .001$). Dieser Anstieg ist (mit α = .005) statistisch signifikant. Bei einer mittleren Feuchtigkeit der Blumenerde, erwartet man bei einer Erhöhung der Beschattung um 1 eine Reduktion der Blütengröße um 41.6 (b = -41.60, stand. β = -.37, $t(23) = -3.60$, p = .002). Dieser Anstieg ist (mit α = .005) ebenfalls statistisch signifikant.

Weiters verändert sich der Zusammenhang zwischen Feuchtigkeit und Blütengröße je nach Beschattungsniveau (und umgekehrt): Bei einer Erhöhung der Beschattung um 1 sinkt die Steigung der Regressionsgerade des Effekts von Feuchtigkeit auf Blütengröße um 52.9 (b = -52.85, stand. β = -.39, $t(23) = -3.74$, p = .001). Diese Interaktion bzw. Veränderung des Zusammenhangs ist (mit α = .005) statistisch signifikant.

Beispiel 11.10

Um in SPSS ein Streudiagramm zu erstellen, bei dem die Datenpunkte andere Farben je nach Ausprägung einer dritten Variablen haben, muss diese in der SPSS-Datendatei als nominal klassifiziert sein. Im Beispiel 11.9 hat die Variable *shade* genau 3 Ausprägungen (1, 2 und 3), obwohl sie eine stetige Variable ist. Indem wir die Variable über *Transform >> Compute Variable…* einfach kopieren (shade_diskret = shade), legt SPSS automatisch eine als nominal-skalierte neue Variable namens *shade_diskret* an. Optional kann auch händisch einfach das Skalenniveau in der Variablenansicht in der Spalte „Measure“ umgestellt werden, obwohl es sich auch hier empfiehlt, zuerst eine Kopie der originalen *shade* Variable zu machen.

Im Anschluss kann über *Graphs >> Chart Builder…* das erwünschte Streudiagramm ausgegeben werden. Unter *Scatter/Dot* wird ein Streudiagramm ausgewählt (Scatter Plot). Auf die x-

Achse setzen wir die zentrierte Variable *c_water* und auf die y-Achse das Kriterium *blooms*. Rechts oben im Diagramm bei „Set color?“ ziehen wir die neue diskrete Variable *shade_diskret* rein. Die Diagrammvorschau sollte dann gleich unterschiedlich färbige Punkte zeigen. In der rechten Spalte des Chart Builder Fensters wählen wir ganz unten noch bei den „Linear Fit Lines“ „Subgroups“ aus, damit uns für jede Gruppe der Belichtung (1 = niedrige Beschattung, 2 = mittlere Beschattung, 3 = hohe Beschattung) eine eigene Regressionsgerade ausgegeben wird. Danach klicken wir auf *Paste* und führen den Code in der Syntax aus.

Nachwort zur ersten überarbeiteten Fassung dieses Manuskript

Wenn nach einem Monat bereits die erste überarbeitete Fassung eines Manuskripts vorliegt (die jetzt auch über eine Versionsnummer verfügt, um zukünftige Aktualisierungen voneinander und von früheren Fassungen unterscheiden zu können), heißt das mindestens zweierlei. Erstens, dass in der ursprünglichen Fassung Mängel vorhanden waren, die zügiger Korrektur bedurften. Zweitens, dass es Personen gibt, deren kritische Durchsicht die Entdeckung dieser Mängel überhaupt ermöglichte, da Verfasser:innen selbst für diese bekanntlich bei der Erstellung der entsprechenden Texte zunehmend blind werden und jedenfalls für einen gewissen Zeitraum nach der Fertigstellung auch blind bleiben.

Ich kann mich jedenfalls glücklich schätzen, solch kritische Leser:innen für die erste Fassung diese Manuskripts gefunden zu haben, welchen ich mich zutiefst zu Dank verpflichtet fühle. In erster Linie geht dabei mein Dank an Prof. H. Harald Freudenthaler, der neben seinen Verpflichtungen und noch dazu mitten im Semester die Zeit fand, die erste Fassung dieses Manuskripts in nicht einmal ganz zwei Wochen zur Gänze durchzusehen. Ihm ist es zu verdanken, dass diese erste überarbeitete Fassung nun auch einen umfassenderen Abschnitt zur Korrelation, d.h. zu Maßen des Zusammenhangs zwischen Variablen, enthält. Letztere wurden in der ersten Fassung lediglich in einem Exkurs im Rahmen der linearen Regression abgehandelt. Dies war allerdings mehr der Eile geschuldet, mit der die erste Fassung im Verlauf eines bereits begonnen Semesters fertiggestellt wurde, als dem Raum, dem eine grundlegende Einführung in solche Maße des Zusammenhangs gebühren sollte, auch wenn hier nach wie vor nur die wesentlichen Anwendungsaspekte derselben berührt werden. Erwähnung finden nun jedenfalls neben Pearsons Korrelationskoeffizient auch Spearmans Rangkorrelationskoeffizient sowie Kendalls tau. Die unterschiedlichen Arten wie diese Koeffizienten Zusammenhänge zwischen Variablen erfassen werden dabei in einer Reihe von Beispielen gegenübergestellt. Ganz dem Credo der ersten Fassung dieses Manuskripts treu bleibend werden dafür auch die Übungsaufgaben genutzt, die zu diesem Zweck um wesentliche Beispiele erweitert wurden. So hat Anscombes berühmt berüchtigtes Quartett nun Eingang in diese Sammlung aus Übungen gefunden. Zudem wurde ein Beispiel, das auf Rand R. Wilcox zurückgeht, in die Übungsaufgaben eingearbeitet, um zu veranschaulichen, dass alle drei behandelten Maße des Zusammenhangs empfindlich auf einige ungewöhnliche Datenpunkte (sog. Ausreißer) reagieren können. Interessierten Leser:innen sind damit hoffentlich auch genug Hinweise

gegeben, wie sie, wenn Notwendigkeit oder Neugier es verlangen, über diese grundlegenden Möglichkeiten Zusammenhänge zwischen Variablen zu beschreiben hinausgelangen können.

Zum Dank verpflichtet bin ich ein weiteres Mal auch meinen Studierenden, die mich unablässig beim gemeinsamen Üben der vielfältigen Beispiele auf die eigenen Versehen in den vorbereiteten oder ad-hoc demonstrierten Lösungen hinweisen. Als Lehrenden erfreut mich das mindestens in zwei Hinsichten. Erstens heißt es, dass einige sich so gewissenhaft und gleichzeitig inhaltlich mit den Lernmaterialien auseinandersetzen, dass ihnen diese Mängel überhaupt ins Bewusstsein treten, und diese Tatsache allein muss dem Lernerfolg bereits zuträglich sein. Zweitens bedient das Lernmaterial damit auch einen Lernansatz, den ich, gerade, wenn es um komplexere Lerninhalte geht, für äußerst effektiv halte: das Lernen durch kritisches Prüfen. In Zeiten, in denen Hilfsangebote aller möglichen künstlichen Intelligenzen zu allen möglichen Inhalten immer weiter um sich greifen, erscheint mir das kritische Überprüfen angebotener Informationen sowohl als geeignete als auch willkommene Versicherung gegen ein vielleicht effizientes, aber unmündiges Übernehmen scheinbarer Wahrheiten. Wer sich der Informationsflut unserer Zeiten nicht ohnmächtig ausliefern will, muss sich zwangsläufig in der Kultivierung der eigenen Urteilsfähigkeit üben. Das kritische Überprüfen von Lernmaterial mit der Intention der Aneignung praktischen, anwendungsbezogenen Wissens scheint mir dafür nicht der schlechteste Anlass zu sein.

Wenn mich dann Studierende etwa darauf hinweisen, dass bei der „Varianz σ“ doch irgendwo ein Quadrat (im Exponenten) verloren gegangen sein muss, dann geht mir als Lehrendem ebenso wie als selbst unablässig Lernendem natürlich im wahrsten Sinne des Wortes das Herz auf. Denn in diesen Momenten wird mir demonstriert: hier wird den Inhalten nach-gedacht, über die Inhalte reflektiert, ohne die Inhalte für bare Münze zu nehmen, und schließlich auf dem Fundament des eigenen Verstehens geurteilt; und wo das geschieht, findet Lernen, Einsicht und Verstehen statt. Und ein angemessenerer und schönerer Sinn kann diesem Manuskript zum selbständigen Üben, Lernen und Verstehen wohl kaum gegeben werden. Aber dem vorliegenden Manuskript solchen Sinn zu verleihen, ist nicht meine, sondern die Errungenschaft eben jener Studierenden, für die ich allein aus diesem Grunde schon nicht anders als in tiefster Dankbarkeit und Hochachtung verbleiben kann.

Stefan E. Huber, Graz am 26. Mai 2025

Nachwort zur ersten Fassung dieses Manuskripts

Die Idee zu diesem Manuskript war ebenso schlicht wie naiv: einen Text verfassen, den jemand im Selbststudium Stück für Stück durcharbeiten könnte, um dann am Ende ein besseres Verständnis für jene Inhalte erlangt zu haben, die in der Zeit, als ich an der Universität Graz Übungen in der Anwendung statistischer Verfahren unterrichten durfte, in entsprechenden Vorlesungen zur Statistik für Psycholog:innen gelehrt wurden.

Die Idee war schlicht, weil genau das das Ziel der Lehrveranstaltung war, die mir (unter vielen anderen) anvertraut worden war: Studierenden ein Verständnis für jene grundlegenden statistischen Verfahren zu vermitteln, die im zu dieser Zeit üblichen Lehrkanon zur statistischen Grundbildung im Bachelorstudium Psychologie vorgesehen waren. Die inhaltlichen Grenzen, sowohl was statistische Verfahren als auch Werkzeuge zu ihrer Anwendung (Computerprogramme, Software) anging, waren mit diesen Rahmenbedingungen relativ klar abgesteckt. Auch wenn diese nicht unbedingt meinen persönlichen Präferenzen entsprachen (Wo ist das Prinzip maximaler Entropie? Wo ist das Gesetz inverser Wahrscheinlichkeiten? Wo sind die Diskussionen eines unhaltbaren Begriffs der „Äquivalenz“ von Zufallsexperimenten? Wo sind die Diskussionen über andere Zugänge zum Begriff der Wahrscheinlichkeit? Und wo sind schließlich die erkenntnistheoretischen Bezüge und die Bezüge zum Prozess der Forschung selbst – die Achillesferse jedes frequentistischen Zugangs, da selbst Fisher, nachdem er ein Leben lang darüber nachgedacht hat, erkennen musste, dass ein solcher Zugang den Forschungsprozess, das, was Wissenschaft und Erkenntnis überhaupt ist, nicht konsistent beschreiben kann?), so war mir der Gedanke daran, Studierende, die noch am Anfang ihres Studiums der Wissenschaft vom Menschen stünden und dieses Studium nicht unbedingt aufgrund ihrer Liebe zu Wahrscheinlichkeitstheorie und Statistik wählten, dabei unterstützen zu können, sich einen Reim auf das machen zu können, was sich da eben über das letzte Jahrhundert als statistische Methodologie in der Psychologie etabliert hatte, doch von Anfang an ein erfreulicher. Denn die Verwirrung über dieses Thema ist groß. Unter Anfänger:innen ebenso wie unter sogenannten Expert:innen. Und nicht minder beim Verfasser dieser Zeilen.

Zum Teil liegt das daran, weil vieles, wenn nicht das Meiste, aus den Bereichen der Wahrscheinlichkeitstheorie oder der Statistik nicht intuitiv, nicht leicht zugänglich ist. Manche Aspekte scheinen vielmehr so entgegen jeder Zugänglichkeit, dass man für den Hauch eines Verständnisses dafür das tun muss, was andernorts als Verrücktheit gilt: sich immer und immer wieder – scheinbar vergeblich – Gedanken darüber zu machen, mit zweifelhaftem Ausgang. Die historischen Figuren der betreffenden Fachgebiete können völlig zurecht nicht deshalb Expert:innen genannt werden, weil sie – um nur ein Beispiel zu nennen – etwa wissen, was ein p-Wert ist, sondern weil sie sich jahrzehntelang mit der Problematik, die mit dem Konzept eines p-Werts einhergeht, beschäftigt haben und deshalb besser verstehen als die meisten, wie problematisch dieses Konzept im Grunde ist. Die Grundlagen der Statistik besser zu verstehen, soll einen oder eine nicht in die Lage versetzen, wie etwas scheinbar *richtig* gemacht wird; vielmehr soll es in die Lage versetzen, an der Diskussion darüber überhaupt teilnehmen zu können.

Naiv war die Idee zu diesem Manuskript deshalb, weil ein Text allein diese Ziele vermutlich immer verfehlen muss. Statistische Grundlagen besser zu verstehen, ist keine Sache des Lesens, es ist eine Sache des Denkens, aber auch des Tuns, des mentalen und praktischen Arbeitens mit den Inhalten. Lesen kann aber einen Anstoß zu diesem aktiven Bearbeiten geben. Und mehr noch: ein Text kann immerhin ein Gerüst, ein bisschen Anleitung bieten, wie man mit bestimmten Inhalten überhaupt arbeiten *kann*. Und diesem Ziel kann ein Text durchaus gerecht werden. Ob dieses Manuskript das tut, daran habe ich so meine Zweifel. Seine Intention ist es aber immer gewesen.

In der Tat war mein Leitgedanke derjenige, der schon im ersten Kapitel in aller Breite erläutert worden ist. Eine Person liest sich ein Kapitel dieses Manuskripts durch, wiederholt dabei einige wesentliche theoretische Konzepte, versucht sich dann an deren Anwendung für einige der gegebenen Übungsaufgaben, stellt sich in diesem Prozess – wie jede:r Lernende zwangsläufig – neue Fragen, bringt diese Fragen in die gemeinsamen Lehrveranstaltungsstunden mit, wo wir sie alle diskutieren, uns daran üben, dabei lernen, und schließlich daran wachsen können.

Dieser Gedanke ist ebenfalls überaus naiv. Denn die viel realistischere Annahme ist, dass niemand, außer dem Verfasser selbst wahrscheinlich, freiwillig jemals dieses Manuskript von vorn bis hinten durchlesen wird, ganz zu schweigen von *durcharbeiten*. Einzelne Kapitel, vielleicht; wenn man

es soll. Die Übungsaufgaben? Wenn es sich vermeiden lässt, eher nicht. Der Realismus dieser Annahme – der jeder Person, die selbst einmal Studierende gewesen ist, offenkundig sein muss – entbehrt nicht einer gewissen Ironie. Denn selbst wenn das Einzige, was zur Teilnahme an einer meiner Lehrveranstaltungen motiviert, deren positiver Abschluss ist, gibt es vermutlich keinen gleichzeitig bequemeren als auch nützlicheren Weg zu diesem positiven Abschluss zu kommen als sich regelmäßig, Kapitel für Kapitel, Woche für Woche, mit diesem Manuskript zu befassen. Was immer man für einen positiven Abschluss der Lehrveranstaltung benötigt, es steht hier. Und das nicht nur angedeutet, in Form von Stichworten, auf Vortragsfolien, die mehr durch ihr Design als ihren Inhalt bestechen, nein: ausformuliert, in ganzen Sätzen, und wichtiger: in von vorn bis hinten durchexerzierten Beispielen.

Ich war selbst einmal Student, bin bis heute einer, habe viele Lehrveranstaltungen besucht, viele Lehrbücher konsultiert. Die besten Lehrer und Lehrer:innen (sowohl solche, die ich sprechen und vortragen gehört habe, als auch jene, die ich nur lesend kennengelernt habe) waren mir dabei jene, die nicht darüber gesprochen haben, wie „etwas“ zu machen sei, sondern, die mir *gezeigt* haben, wie sie dieses „etwas“ im konkreten Fall machen. Die beste Vorlesung, die ich jemals besucht habe, war in der Tat eine *Vorlesung*: ein Philosoph, der aus einem Buch vorlas, und nach einem Absatz oder zwei, das Buch zur Seite legte und laut über das Gelesene nachdachte. In dieser Vorlesung habe ich zwei Dinge gelernt: *Lesen* und *Denken*. Und ich lernte es, weil mir jemand zeigte, wie man das machen kann. Nicht, wie man es machen soll. Sondern wie jemand mit sehr viel mehr Erfahrung in diesen Dingen das machen kann (inklusive der Fehler, die dabei passieren, der Irrwege, die dabei eingeschlagen werden können).

Ich bin froh, dass die Intention dieses Manuskripts keine so erhabene war wie die Illustration des Philosophierens selbst. Aber auch für die Anwendung statistischer Verfahren war mir völlig klar, was mein Manuskript leisten können sollte: es sollte Personen, die daran – aus welchen Gründen auch immer – interessiert sind, *zeigen*, wie man grundlegende statistische Verfahren auf konkrete Fragestellungen anwenden kann. Das heißt insbesondere, es darf nicht abstrakt bleiben, es muss *zeigen*, d.h. konkret machen, Sätze ausformulieren, nicht sagen, was – abstrakt – in einem Ergebnisbericht stehen sollte, sondern konkrete Ergebnisberichte in ganzen Sätzen ausformuliert beinhalten, nicht die Durchführung einer Methode andeuten, sondern an einem konkreten Beispiel vorführen, was eine Methode durchführen oder anwenden alles beinhaltet, welche Entscheidungen dabei getroffen werden

müssen – und wie selten diese klar und deutlich sind, vom oft gar nicht so simplen Einlesen eines Datensatzes bis zur oft alles anderen als eindeutigen Interpretation der Ergebnisse, die einen nicht selten mit mehr Fragen als Antworten zurücklässt. Und es muss viele Möglichkeiten bieten, all das selbst an weiteren konkreten Fällen ausprobieren zu können, für die es ebenfalls wenigstens einen gut ausgeleuchteten, illustrierten Lösungsweg bereits gibt, mit dem der eigene am Ende verglichen werden kann. Denn wie viel durfte ich daraus lernen, weil ich andere dabei beobachten konnte, wie sie an schwierige, herausfordernde Probleme herangingen (nicht an die einfachen, symmetrischen, die man gerne in Vorlesungen an der Tafel vorführt)? Alles – würde ich behaupten, was mir überhaupt erst die Mittel an die Hand gab, später selbst Probleme lösen zu können, für die es bis dahin noch keine bekannte Lösung gab. Ohne diese konkreten, leibhaftigen Beispiele wäre aus mir kein Forscher geworden; das heißt, keiner, der das Gewinnen von Erkenntnis aus dem Unbekannten praktiziert.

Das bedeutete, das Manuskript musste die Anwendung grundlegender statistischer Verfahren in ihrer ganzen Konkretheit vorzeigen, sie für den Nachvollzug, das Nach-Denken und Nach-Machen verfügbar machen, und das in einer Mannigfaltigkeit an Beispielen. Wer diese Beispiele dann nachverfolgt, nachbearbeitet, und dabei immer wieder *tut*, was eben jemand tut, der grundlegende statistische Verfahren anwendet, der oder die wird von Beispiel zu Beispiel vertrauter damit, übt sich in diese eigentümliche Tätigkeit ein, und versteht besser und besser, worin sie eigentlich besteht.

Wie gut es mir gelungen ist, diesen Leitgedanken gerecht zu werden, können nur jene Studierenden beurteilen, die sich kühn daran wagen, ihre Fähigkeiten in der Anwendung grundlegender statistischer Verfahren durch die Beschäftigung mit diesem Manuskript (hoffentlich zum Positiven) zu verändern. Dafür, dass es diese Möglichkeit nun überhaupt geben kann, bin ich umso dankbarer all diesen, die auszugsweise manche Teile dieses Manuskripts im wahrsten Sinne des Wortes ausprobiert und erprobt haben, lange bevor es auch nur annähernd mehr war als ein bloßes Hirngespinst.

Allen voran danke ich dafür Nadine Schmer, die nicht nur einen großen Teil des Manuskripts korrekturgelesen und zahlreiche der Übungsaufgaben und ihrer Lösungen überprüft hat, sondern auch die erste Fassung des zweiten Kapitels verfasst hat. Danach danke ich meinen Studierenden an der Universität Graz aus den Kursen „Anwendung statistischer Verfahren am Computer“ der Jahrgänge

2023-2025, ohne die es dieses Manuskript schlichtweg nicht geben würde. Neben zahlreichen inhaltlichen Rückmeldungen und Anmerkungen zu Schwierigkeiten und Herausforderungen beim Lernen statistischer Inhalte im Allgemeinen, aber insbesondere auch bei der Anwendung der im Kurs verwendeten Software, der von mir angebotenen Hilfestellungen und Erklärungsversuche, oder der Übertragung von konzeptuellen Vorlesungsinhalten in konkrete Problemstellungen, waren es vor allem der beispielhafte Wille sich auch mit schwierigen Inhalten wiederholt auseinanderzusetzen und das Engagement, die Neugier und die Freude, mit der sie mir in den einzelnen Lehrveranstaltungseinheiten immer wieder begegnet sind, die mich inspiriert und motiviert haben, mich stets aufs Neue der Arbeit dieser Niederschrift zuzuwenden. Ich hoffe, dass das auf dieser Grundlage gewachsene Manuskript zur Lehrveranstaltung nun auch den Studierenden im voraussichtlich letzten Jahr meiner Lehrzeit an dieser Universität dazu dienen kann, nicht nur das Erreichen der konkreten Lernziele, sondern auch das Verstehen statistischer Konzepte und ihrer Anwendung im Allgemeinen zu erleichtern.

Stefan E. Huber, Graz am 26. April 2025

Literatur

Adesope, O.O., Trevisan, D.A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*, 659–701. https://doi.org/10.3102/0034654316689306

American Psychological Association. (2019). *Publication manual of the American psychological association* (7. Aufl.). American Psychological Association.

Anscombe, F. J. (1973). Graphs in Statistical Analysis. American Statistician, 27, 17-21. https://doi.org/10.1080/00031305.1973.10478966

Bahrick, H.P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*, 296–308. https://doi.org/10.1037/0096-3445.108.3.296

Bammes, G. (1998). *Studien zur Gestalt des Menschen*. Ravensburg.

Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current Practices in Data Analysis Procedures in Psychology: What Has Changed? *Frontiers in Psychology, 9*, 2558. https://doi.org/10.3389/fpsyg.2018.02558

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin, 145*, 1029–1052. https://doi.org/10.1037/bul0000209

Bühner, M., Pargent, F., Schönbrodt, F., Sckopke, P., Zygar-Hoffmann, C., Schoedel, R., Schiestel, L., & Sust, L. (2025, 18. Februar). Offene Lehrmaterialien des Lehrstuhls für Psychologische Methodenlehre und Diagnostik der Ludwig-Maximilians-Universität München. URL: https://osf.io/c59hv

Bühner, M., & Ziegler, M. (2017). *Statistik für Psychologen und Sozialwissenschaftler* (2. Aufl.). Pearson.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Duran, D. (2017). Learning-by-teaching: Evidence and implications as a pedagogical mechanism. *Innovations in Education and Teaching International, 54*(5), 476–484. https://doi.org/10.1080/14703297.2016.1156011

Ebersbach, M., Lachner, A., Scheiter, K., & Richter, T. (2022). Using spacing to promote lasting learning in educational contexts: Promises and challenges. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 54*, 151–163. https://doi.org/10.1026/0049-8637/a000259

Elvira, Q., Imants, J., Dankbaar, B., & Segers, M. (2017). Designing education for professional expertise development. *Scandinavian Journal of Educational Research, 61*, 187-204. https://doi.org/10.1080/00313831.2015.1119729

Field, A. (2024). *Discovering Statistics Using IBM SPSS Statistics* (6. Aufl.). Sage.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606. https://doi.org/10.1016/j.socec.2004.09.033

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*, 337-350. https://doi.org/10.1007/s10654-016-0149-3

Heitmann, S., Grund, A., Berthold, K., Fries, S., & Roelle, J. (2018). Testing is more desirable when it is adaptive and still desirable when compared to note-taking. *Frontiers in Psychology, 9*, 2596. https://doi.org/10.3389/fpsyg.2018.02596

Heitmann, S., Grund, A., Fries, S., Berthold, K., & Roelle, J. (2022). The quizzing effect depends on hope of success and can be optimized by cognitive load-based adaptation. *Learning and Instruction, 77*, 101526. https://doi.org/10.1016/j.learninstruc.2021.101526

Higham, P.A., Zengel, B., Bartlett, L.K., & Hadwin, J.A. (2022). The benefits of successive relearning on multiple learning outcomes. *Journal of Educational Psychology, 114*, 928–944. https://doi.org/10.1037/edu0000693

Huber, L. (2014). Forschungsbasiertes, Forschungsorientiertes, Forschendes Lernen: Alles dasselbe? Ein Plädoyer für eine Verständigung über Begriffe und Entscheidungen im Feld forschungsnahen Lehrens und Lernens. *Das Hochschulwesen*, *62*(1+2). https://pub.uni-bielefeld.de/record/2905797

Huber, O. (2019). *Das psychologische Experiment*. Springer.

Kendall, M. G. (1945). The Treatment of Ties in Ranking Problems. *Biometrika, 33*, 239–251. https://doi.org/10.2307/2332303

Lachner, A., Hoogerheide, V., van Gog, T., & Renkl, A. (2022). Learning-by-teaching without audience presence or interaction: When and why does it work? Educational Psychology Review, 34, 575-607. https://doi.org/10.1007/s10648-021-09643-4

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*, 464-488. https://doi.org/10.3758/s13428-019-01246-w

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika, 63*(3), 655-660. https://doi.org/10.1093/biomet/63.3.655

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2. Aufl.). Chapman and Hall/CRC.

Meier, U. (2006). A note on the power of Fisher's least significant difference procedure. *Pharmaceutical Statistics, 5*, 253-263. https://doi.org/10.1002/pst.210

Metcalfe, J. (2017). Learning from Errors. Annual Review of Psychology, 68, 465-489. https://doi.org/10.1146/annurev-psych-010416-044022

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychological Bulletin, 144*, 710–756. https://doi.org/10.1037/bul0000151

Rajh-Weber, H., Huber, S. E., Arendasy, M. (2025, 17. April). A practice-oriented guide to statistical inference in linear modeling for non-normal or heteroskedastic error distributions. URL: https://osf.io/preprints/osf/cdj46_v1

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140*, 283–302. https://doi.org/10.1037/a0023956

Rawson, K.A., & Dunlosky, J. (2022). Successive relearning: An underexplored put potent technique for obtaining and maintaining knowledge. *Current Directions in Psychological Science, 31*, 362–368. https://doi.org/10.1177/09637214221100484

Rawson, K.A., Dunlosky, J., & Sciartelli, S.M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*, 523–548. https://doi.org/10.1007/s10648-013-9240-4

Roediger III, H.L., & Karpicke, J.D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255. https://doi.org/10.1111/j.1467-9280.200601693.x

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction, 49*, 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008

Roelle, J., & Richter, T. (2025). Üben aus der Sicht der Lernpsychologie: Wie kann Üben das nachhaltige Lernen fördern? Unterrichtswissenschaft. https://doi.org/10.1007/s42010-025-00227-7

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley.

Rueß, J., Gess, C., & Deicke, W. (2016). Forschendes Lernen und forschungsbezogene Lehre–empirisch gestützte Systematisierung des Forschungsbezugs hochschulischer Lehre. *Zeitschrift für Hochschulentwicklung*, *11*(2), 23-44.

Rummer, R., Schweppe, J., & Schwede, A. (2019). Open-Book Versus Closed-Book Tests in University Classes: A Field Experiment. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00463

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology, 17*, 688–690. https://doi.org/10.1093/beheco/ark016

Sonntag, M., Rueß, J., Ebert, C., Friederici, K., Schilow, L., & Deicke, W. (2017). *Forschendes Lernen im Seminar*. Humboldt-Universität zu Berlin.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science, 333*, 776-778. https://doi.org/10.1126/science.1207745

Student (1908). The probable error of a mean. *Biometrika, 6*(1), 1-25. https://doi.org/10.2307/2331554

Wilcox, R. R. (2017). *Understanding and Applying Basic Statistical Methods Using R*. John Wiley & Sons.

Wilcox, R. R. (2022). *Introduction to Robust Estimation and Hypothesis Testing* (5. Aufl.). Academic Press.

Wittgenstein, L. (2003). *Tractatus logico-philosophicus, Logisch-philosophische Abhandlung*. Suhrkamp.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*, 173–181. https://doi.org/10.1348/000711004849222