Die Hypothesenprüfung – Eine schrittweise Anleitung

Gundula Wagner

Zur Autorin

Gundula Wagner, Dr., Pädagogische Hochschule Wien

Kontakt: gundula.wagner@phwien.ac.at

1 Einleitung

Rund um die Prüfung von Hypothesen gibt es einige Missverständnisse, die auf der Verwechslung deskriptiv explorativer Studien mit quantitativ explanativer Forschung beruhen[1]. Zielstellung deskriptiver Studien ist die quantitative Beschreibung von z. B. soziodemografischen Merkmalen in der Bevölkerung (vgl. Bortz & Döring 2006, S. 356). In populärwissenschaftlichen Publikationen werden deskriptive Studien jedoch, entgegen jeglicher guten wissenschaftlichen Praxis, regelmäßig zur Ursachenerklärung herangezogen. Die Meinungsforschung ist hierfür ein klassisches Beispiel, weshalb Schnell et al. (2011, S. 39) auch eine klare Abgrenzung seitens der empirischen Sozialforschung fordern. Vor diesem Hintergrund möchte der vorliegende Beitrag das Thema der vermeintlichen Hypothesenprüfung in deskriptiven Studien gezielt aufgreifen, um anschließend die forschungsmethodisch korrekte Vorgehensweise zu erklären.

2 Sonderfall deskriptive Studien

Deskriptive Studien dienen laut Bortz und Döring (2006, S. 393ff.) primär dazu, Phänomene in der jeweiligen Stichprobe zu *beschreiben,* um eine genauere

[1] Der Begriff Exploration steht für die Entwicklung von Hypothesen, der Begriff Explanation für die Prüfung von Hypothesen (vgl. Bortz & Döring, 2006, S. 356)

In: C. Fridrich, B. Herzog-Punzenberger, H. Knecht, N. Kraker, P. Riegler, G. Wagner (Hg.) (2023): Forschungsperspektiven 15. Wien: Lit. S. 197-210. DOI: 10.52038/978364351139_14

Schätzung von Populationsparametern vornehmen zu können. Die quantitative Datenerhebung erfolgt hier zumeist über standardisierte Fragebögen und die so erhobenen Daten werden mit Hilfe deskriptiver Statistik ausgewertet und in ansprechenden Grafiken aufbereitet. Sowohl die Verwendung eines Fragebogens als auch die Darstellung von quantitativen Daten in Form von Häufigkeiten und Mittelwerten lässt Laien immer wieder irrtümliche annehmen, hier würde es sich um quantitativ explanative Forschung handeln. Und so sehen sich Autor*innen populärwissenschaftlicher Publikationen regelmäßig veranlasst, mit ihren deskriptiven Studienergebnissen Hypothesen prüfen zu wollen und Kausalzusammenhänge erklären zu können. Tatsächlich ist die Zielstellung deskriptiver Studien aber eher mit jener von qualitativen Studien vergleichbar, wo ebenfalls Phänomene beschrieben und daraus explorativ Hypothesen abgeleitet werden können[2].

Es ist anzunehmen, dass zur Verwechslung deskriptiver mit quantitativ explanativen Studien in erster Linie die *Statistik* beiträgt, die in beiden Fällen zur Anwendung kommt. Dabei wird aber übersehen, dass es sich in deskriptiven Studien eben nur um die beschreibende und nicht um die, von der Stichprobe auf die Population rückschließende, Inferenzstatistik handelt (siehe Kap. 3.4). Zudem wird in populärwissenschaftlichen Studien oftmals der aus der quantitativen Forschung bekannte Begriff des Mittelwertsvergleichs zweckentfremdet. Anstelle der Berechnung einer Differenz (siehe Kap. 3.3) werden die Mittelwerte einander in Tabellen oder Diagrammen gegenübergestellt und *optisch* miteinander *verglichen*. Aus der Relation der Zahlen zueinander wird dann das Muster eines Unterschieds oder auch eines Zusammenhangs *herausgelesen.* Indem nach Mustern im Datenmaterial gesucht wird, bedient man sich bekanntermaßen der Vorgehensweise des Kodierens, vergleichbar mit dem methodischen Vorgehen in der qualitativen Forschung (siehe dazu Bortz & Döring 2006, S. 330). Da sich die auf diese Weise entdeckten Unterschiede oder Zusammenhänge aber nur auf die untersuchte Stichprobe beziehen, können in deskriptiven Studien keine Verallgemeinerungen angestellt werden.

Deskriptive Studien nehmen demnach eine Sonderstellung in der Forschungsmethodik ein: Obwohl das Datenmaterial in Zahlenform vorliegt, sind sie eher dem qualitativen Paradigma zuzuordnen. Vergleichbar mit qualitati-

[2] Für eine anschauliche Erklärung des qualitativen und quantitativen Paradigmas siehe den Beitrag „Wissenschaftstheoretische Spielregeln der Bildungsforschung verständlich erklärt" in diesem Band.

ven Studien beantworten sie Fragen nach dem Wie, nicht nach dem Warum (vgl. Bortz & Döring 2006, S. 340). Dies ist der quantitativen Forschung vorbehalten, die explanativ also hypothesenprüfend vorgeht. Der Unterschied zwischen deskriptiven und quantitativen Studien wird aber wohl erst dann eindeutig, wenn man sich die methodische Begründung einer Kausalbeziehung im Detail ansieht[3].

3 Ursachenforschung in quantitativen Studien

Aufgabe des quantitativen Paradigmas ist es, Wirkungen bzw. Einflüsse, die aus der Theorie abgeleitet werden, zu prüfen. In Kurzfassung bedeutet dies: Die angenommenen Kausalbeziehungen werden in Hypothesen formuliert, anschließend werden im Untersuchungsdesign entsprechende empirische Werte gesammelt und diese im Signifikanztest bestätigt. Kann keine Signifikanz ausgewiesen werden, bedeutet dies, dass die angenommene Kausalbeziehung nicht nachgewiesen werden konnte. Die Hypothese muss entsprechend verworfen werden.

3.1 Annahme einer Kausalbeziehung

Als erster Schritt wird zumeist eine allgemeine Forschungsfrage formuliert, die nach der Ursache von Sachverhalten fragt (Warum-Frage). Sobald in der Literatur eine mögliche theoretische Antwort auf dieser Frage gefunden wurde, kann eine Hypothese formuliert werden. Dabei handelt es sich um eine wissenschaftliche Aussage, in der sich aufgrund bestehender Erkenntnisse (= Theorie) bestimmte Erwartungen ausdrücken, die aber auch über den aktuellen Kenntnisstand hinausgeht (vgl. Bortz 1999, S. 108). Folgende Kriterien sind bei der Formulierung von Hypothesen zu bedenken, andernfalls kann nicht von einer Hypothese gesprochen werden:

- „Eine wissenschaftliche Hypothese bezieht sich auf *reale* Sachverhalte, die *empirisch untersuchbar* sind.
- Eine wissenschaftliche Hypothese ist eine *allgemeingültige*, über den Einzelfall oder ein singuläres Ereignis hinausgehende *Behauptung (All-Satz).*

[3] Eine grafische Übersicht zum Unterschied zwischen deskriptiven und quantitativen Studien findet sich unter https://phwien.ac.at/wp-content/uploads/2023/05/Uebersicht-ueber-quantitative-Unterrichtsdesigns_wagnergundula.pdf

– Einer wissenschaftlichen Hypothese muss zumindest implizit die *Formalstruktur* eines sinnvollen Konditionalsatzes („Wenn-dann-Satz“ oder „Jedesto-Satz“) zugrunde liegen.
– Der Konditionalsatz muss potentiell *falsifizierbar* sein, d. h., es müssen Ereignisse denkbar sein, die dem Konditionalsatz widersprechen“ (Bortz & Döring 2006, S. 4, Hervorhebung d. Autorin).

Wissenschaftliche Hypothesen müssen also in eine bestimmte Satzform gebracht werden, damit sie mathematisch überprüfbar sind. Dazu müssen sich die in der Hypothese vorkommenden Begriffe (z. B. Motivation, Schulleistung) beobachtbaren Daten zuordnen lassen. Diesen Vorgang nennt man *Operationalisierung*, die wiederum die Voraussetzung für die Prüfung von Hypothesen ist.

Grundsätzlich unterscheidet die Ursachenforschung zwischen zwei Hypothesenarten, der Zusammenhangs- und der Unterschiedshypothese. Dies ist für Forschungsanfänger*innen durchaus hilfreich, weil es die Auswahlmöglichkeiten unter den verschiedenen Hypothesenarten sehr beschränkt. Beginnen wir mit der leichter verständlichen Form von Hypothesen, der *Zusammenhangshypothese*. Hier wird ein Zusammenhang zwischen zwei Merkmalen angenommen. Eine *ungerichtete Hypothese* würde folgendermaßen lauten: Es wird angenommen, dass zwischen der Motivation und der Schulleistung ein Zusammenhang besteht. In einer *gerichteten Hypothese* wird theoriegleitet eine Aussage über die Art des Zusammenhangs gemacht. Hier werden Formulierungen wie „wenn… dann“ oder „je… desto“ verwendet. Eine gerichtete Hypothese würde daher lauten: Je höher die Motivation ist, desto höher steigt auch die Schulleistung.

Daneben gibt es die schon erwähnten *Unterschiedshypothesen*, für die *Vergleiche* zweier (oder auch mehrerer) Stichproben bzw. Unterstichproben (z. B. Mädchen und Buben) charakteristisch sind. Schwieriger zu verstehen ist hier die Unterscheidung zwischen einer unabhängigen Variable (uaV), von der eine Wirkung ausgehen soll, und der abhängigen Variable (aV), an der die Wirkung geprüft werden soll. Als Beispiel für eine abhängige Variable dient uns wiederum die Lernmotivation, die zwischen Mädchen und Buben oder zwischen Klasse A und B (uaV) unterschiedlich sein kann. Eine entsprechende Unterschiedshypothese würde demnach lauten: Es wird angenommen, dass es einen Unterschied zwischen Mädchen und Buben hinsichtlich Leistungsmoti-

vation gibt. Abhängig von der jeweiligen theoretischen Annahme könnte auch formuliert werden, dass Mädchen eine höhere Leistungsmotivation als Buben zeigen.

3.2 Konstruktion der Kausalbeziehung

Im Untersuchungsdesign wird nun versucht, die in der Hypothese formulierte theoretische Annahme quasi in der Realität nachzubauen, um sie so an den Erfahrungen der Proband*innen in der Stichprobe empirisch prüfen zu können. Zu diesem Zweck gibt es eine Vielzahl unterschiedlicher Forschungsdesigns, die in diesem Beitrag nicht erschöpfend thematisiert werden können. Für einen schnellen Überblick bietet sich aber die Unterscheidung in *Ein-Gruppen-* bzw. *Zwei-Gruppen-Pläne* an (vgl. Rost 2022, S. 158).

Der klassische Zwei-Gruppen-Untersuchungsplan ist das *Experiment*, das aus einer Experimental- und einer Kontrollgruppe besteht. Je nachdem wie zufällig die Zuordnung der Untersuchungsteilnehmer*innen zu den zwei Gruppen erfolgt (= Randomisierung), spricht man von einem *echten-* oder einem *quasi-experimentellen Design*. Echte Experimente kommen eher unter Laborbedingungen zustande, während in Felduntersuchungen vielfach nur quasi-experimentelle Versuchspläne mit vorgegebenen Versuchsgruppen (z. B. Schulklassen) möglich sind. Ziel der experimentellen Designs ist, eine möglichst *direkte Kontrolle der unabhängigen Variablen* vorzunehmen und dadurch mögliche andere Einflüsse des Treatments bzw. der Intervention (z. B. Unterrichtsmethode, Lernprogramm) auf die abhängige Variable auszuschließen.

Nun ist es in den Sozialwissenschaften aber vielfach nicht möglich, die uns interessierenden Variablen zu kontrollieren. Dazu gehören Variablen wie das Geschlecht oder auch von außen wirkende Variablen wie beispielsweise das soziale Umfeld oder das Herkunftsland. Die kontrollierte Konstruktion von Ursache-Wirkungs-Zusammenhängen, die ein experimentelles Untersuchungsdesign ausmacht, fällt demzufolge weg. Folge dessen muss man sich in den Sozialwissenschaften oft mit vorexperimentellen *Ein-Gruppen-Designs* begnügen. Das bedeutet, dass nicht zwischen Versuchs- und Vergleichsgruppe unterschieden wird, sehr wohl aber ein Prä- und Posttest vor bzw. nach einer Intervention durchgeführt wird. Bei der *klassischen Fragebogenuntersuchung*, als gängigem Instrument in den Sozialwissenschaften, wird aber oftmals nur zu einem Zeitpunkt gemessen. Bortz und Döring (2006, S. 506) sprechen hier

von einfachen Querschnittsuntersuchungen in Kombination mit Zusammenhangsanalysen. Diese werden angewendet, wenn keine Trennung in abhängige und unabhängige Variablen möglich ist bzw. die Kausalität des Zusammenhangs nicht eindeutig ist. Aus Zusammenhangsanalysen lassen sich *keine echten Kausalaussagen* ableiten, sie zeigen nur auf, wie zwei oder mehrere Merkmale sich miteinander verändern bzw. variieren (vgl. Rost 2022, S. 102; Reinders & Gniewosz 2015).

Zwei-Gruppen-Plan mit Vorher-Nachher Messung	Experiment mit Experimental- und Kontrollgruppe	Unterschiedshypothese
	Quasi-Experiment mit Versuchs- und Vergleichsgruppe	Unterschiedshypothese
Ein-Gruppen-Plan mit Vorher-Nachher Messung	Vorexperimentelles Design ohne Vergleichsgruppe	Unterschiedshypothese
Ein-Gruppen-Plan mit nur einmaliger Messung	Querschnittsuntersuchung (klassische Fragebogenuntersuchung)	vorwiegend Zusammenhangshypothesen; Unterschiedshypothesen bei Vergleich zwischen Unterstichproben möglich

Übersicht 1: Übersicht über Untersuchungsdesigns und dazugehörige Hypothesen (in Anlehnung an Rost 2022, S. 158)

Wie in Übersicht 1 ersichtlich, ergibt sich eine Hierarchie der Untersuchungsdesigns betreffend ihrer Qualität, Ursache-Wirkungs-Zusammenhänge zu prüfen. An der Spitze dieser Hierarchie befindet sich unbestritten das Experiment. Die in den Sozialwissenschaften häufig durchgeführte klassische Fragebogenuntersuchung als Ein-Gruppen-Plan mit einmaliger Messung lässt echte Kausalaussagen nicht zu. Die Forschungsergebnisse von Einzelstudien sind daher dementsprechend vorsichtig zu formulieren. Bevor Kausalaussagen über Zusammenhänge getätigt werden können, muss die gleiche Zusammenhangshypothese in einer Reihe von Untersuchungen bestätigt werden (vgl. Bortz & Döring 2006, S. 522). Metastudien leisten hier einen wertvollen Beitrag.

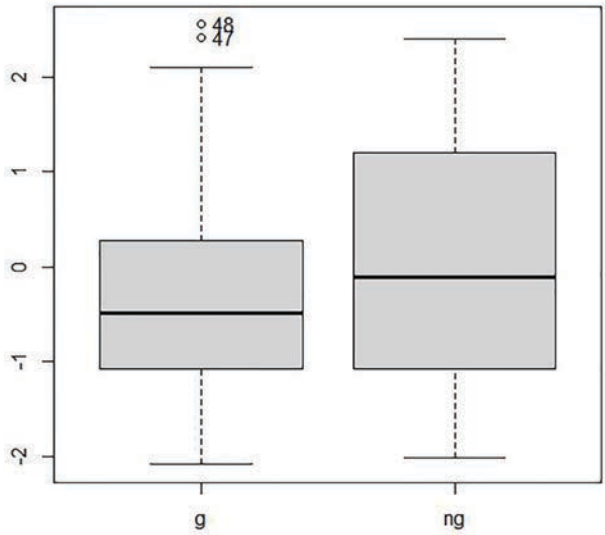
3.3 Berechnung der Kausalbeziehung

Nachdem heutzutage vorwiegend per Computerprogramm ausgewertet wird, ist vielen Nutzer*innen die mathematische Grundlage der Kausalbeziehungen im Signifikanztest nicht mehr geläufig. Bevor wir uns also im nächsten Kapitel mit dem Grundprinzip des Signifikanztests auseinandersetzen, gilt es, zunächst das mathematische Grundprinzip eines Mittelwertsvergleichs und eines Zusammenhangs zu verstehen.

Ein Mittelwertsvergleich wird angestellt, sobald eine aus der Theorie abgeleitete Unterschiedshypothese vorliegt, d. h. die Hypothese gibt das Verfahren vor (vgl. Reinders & Gniewosz 2015, S. 137). In einer korrekt formulierte Unterschiedshypothese wird deutlich, dass es sich hier um einen *Mittelwertsunterschied zweier* (oder auch mehrerer) *Gruppen* handelt (z. B. Es wird angenommen, dass Mädchen eine höhere Leistungsmotivation als Buben zeigen). Um diesen zu berechnen, wird eine *Subtraktion* angestellt. Vom Mittelwert der einen Stichprobe wird der Mittelwert der anderen Stichprobe abgezogen, das Ergebnis ist eine *Mittelwertsdifferenz* (vgl. Bortz 1999, S. 137). Grafisch darstellen lässt sich ein Mittelwertsunterschied in einem Boxplot, wobei es sich aber genau genommen um einen Medianvergleich handelt. Der Median[4] wird durch die fett-gedruckte Linie in Übersicht 2 markiert, die Box um den Median stellt 50 % der Verteilung dar und die strichlierten Linien markieren die Spannweite der Verteilung. Ob der zu beobachtende Unterschied zwischen den zwei Medianen der beiden Gruppen signifikant ist, lässt sich aus dieser Grafik nicht herauslesen. Dazu benötigt es einen Signifikanztest (siehe Kap. 3.4).

Bei Vorliegen einer Zusammenhangshypothese werden Korrelationsanalysen angestellt. Dahinter verbirgt sich als Berechnungsgrundlage eine *Funktion,* in der zwei Variablen, eine davon auf der x-Achse die andere auf der y-Achse, zueinander in Beziehung gesetzt werden. Funktionen sind der mathematische Ausdruck einer *Wenn-dann*-Beziehung, in der beide Variablen systematisch miteinander variieren (vgl. Bortz 1999, S. 173). Eine Korrelation kann positiv wie negativ sein. Ein Beispiel für einen positiven Zusammenhang wäre die Hypothese: Je höher die Leistungsmotivation ist, desto höher ist die Schulleistung. Erkennbar ist dies an einer nach rechts oben steigenden Punktwolke, die eine monoton steigende Funktion darstellt (siehe Übersicht 3). Ein Beispiel

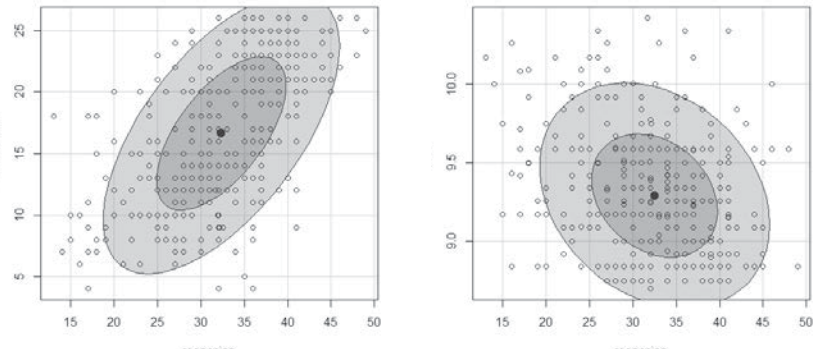
[4] Der Median teilt eine Verteilung von Daten genau in zwei Hälften.



Übersicht 2: Boxplot mit grafischer Darstellung eines Mittelwertsunterschieds (Quelle: eigene Darstellung)

für einen negativen Zusammenhang ist die Hypothese: Je länger der Unterrichtstag dauert, desto niedriger ist die Konzentrationsfähigkeit. Diesmal fällt die Punktewolke nach rechts unten ab (siehe Übersicht 3). Es handelt sich um eine monoton fallende Funktion. Zudem kann ein Zusammenhang unterschiedlich stark oder schwach sein. Dies äußert sich visuell in einer mehr oder weniger dichten Punktewolke. Eine exakte Aussage liefert der Korrelationskoeffizient r, der sich zwischen -1 (negative Korrelation) und +1 (positive Korrelation) bewegt.

Eine bei Forschungsanfänger*innen immer wieder anzutreffende Fehlvorstellung ist, dass ein in der Stichprobe nachgewiesener Zusammenhang oder Unterschied automatisch signifikant sein muss, d. h. auf die Population verallgemeinert werden kann. Dies kommt daher, da in den gängigen Statistikprogrammen per Knopfdruck gleichzeitig sowohl die Berechnung der Kausalbeziehung als auch die Signifikanzprüfung ausgewiesen werden. Tatsächlich handelt es sich hier aber um zwei getrennt voneinander zu betrachtende Verfahren. Herauszufinden, ob der in einem ersten Schritt ermittelte Unterschied oder Zusammenhang systematisch und nicht etwa zufällig entstanden ist, ist



Übersicht 3: Beispiel für einen positiven Zusammenhang (links) und für einen negativen Zusammenhang (rechts). (Quelle: eigene Darstellung)

erst im zweiten Schritt Aufgabe des Signifikanztests (vgl. Reinders & Gniewosz 2015, S. 135).

3.4 Verallgemeinerung der Kausalbeziehung

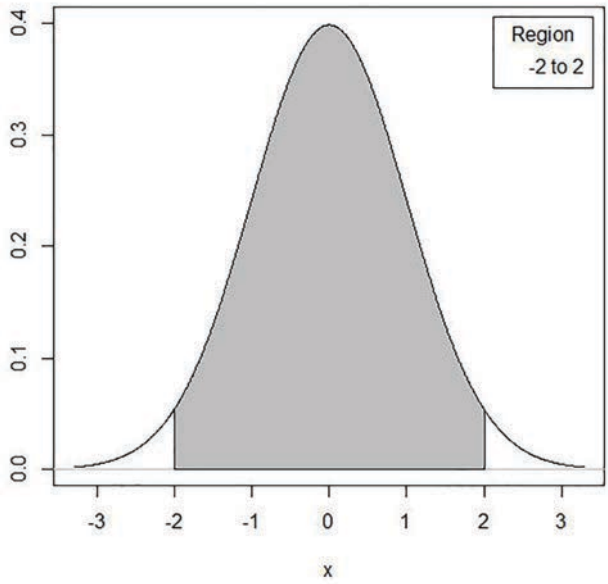
Bevor wir uns das Verfahren des Signifikanztests genauer ansehen, sollten wir uns nochmals in Erinnerung rufen, wofür wir diesen eigentlich brauchen. Ziel der Hypothesenprüfung ist es, allgemeingültige Aussagen für die Gesamtpopulation zu tätigen. Da es aber nur in wenigen Fällen möglich ist, die Gesamtpopulation zu untersuchen, bedient man sich in Untersuchungen stattdessen einer Teilmenge, der Stichprobe, und zieht anschließend mittels Signifikanztest *Rückschlüsse* auf die Gesamtpopulation. Man spricht daher auch von der *schließenden Statistik* oder der *Inferenzstatistik*. Im Umkehrschluss heißt das aber auch, dass wir im Falle einer Gesamterhebung der Population keinen Signifikanztest bräuchten.

Ausgangspunkt des Signifikanztests sind die in der Hypothese vermuteten systematischen Ursache-Wirkungs-Zusammenhänge. Man spricht hier auch von der *Forschungs- oder Alternativhypothese (H_1)* (siehe Kap. 3.1). Die *Nullhypothese (H_0)* hingegen besagt, dass die in der Stichprobe gefundenen Zusammenhänge oder Unterschiede nur zufällig zustande gekommen sind und es diese in der Population nicht gibt. Forschungshypothese und Nullhypothese werden zusammen als statistisches Hypothesenpaar bezeichnet und korrekterweise immer gemeinsam formuliert. Grundlage für den Signifikanztest ist die

Nullhypothese und dieser bewertet nun – in der Annahme, dass in der Population keine systematischen Mittelwertsunterschiede oder Zusammenhänge vorliegen – die Wahrscheinlichkeit der in der Stichprobe empirisch aufgetretenen Mittelwertsunterschiede oder Zusammenhänge (vgl. Leonhard 2009, S. 163). *Zweck* des *Signifikanztests* ist es also, die in der Stichprobe ermittelten Zusammenhänge oder Unterschiede *gegenüber Zufallsfunden abzusichern* (vgl. Reinders & Gniewosz 2015, S. 135). Mit etwas anderen Worten erklärt: Im Signifikanztest wird entschieden, ob die aus den Daten errechnete Mittelwertsdifferenz oder der ermittelte Zusammenhang eher zur Verteilung der Population (H_0) gehören und damit zufällig sind, oder eher zur Verteilung der Stichprobe (H_1) passen und somit systematisch auftreten (vgl. Beller 2008, S. 103).

Da die wahren Werte der Population zumeist nicht bekannt sind, müssen sie aus den Stichprobenwerten geschätzt werden (= Parameterschätzung). Dies ist möglich, da mittels Computersimulationen nachweisbar ist, dass die meisten Merkmale (z. B. Intelligenz, Motivation, Schulangst) in der Population normalverteilt sind (vgl. Rasch et al. 2010, S. 29). Ebenso bekannt ist, dass sich die Daten einer Normalverteilung um den Mittelwert gruppieren, was der Verteilung die typische Glockenform verleiht. Mit Hilfe der obigen Überlegungen lässt sich daher die Bandbreite angeben, innerhalb derer sich der Wert in der Population sehr *wahrscheinlich* bewegt. Zu 95 % befindet er sich in der grauen Fläche unter der Kurve (siehe Übersicht 4). Wir können also mit 95 %iger Wahrscheinlichkeit darauf vertrauen, dass unsere empirischen Daten um den Mittelwert der Population innerhalb der grauen Fläche liegen (vgl. Rasch et al. 2010, S. 32). Der graue Bereich unter der Kurve wird daher auch als *Vertrauens- oder Konfidenzintervall* bezeichnet.

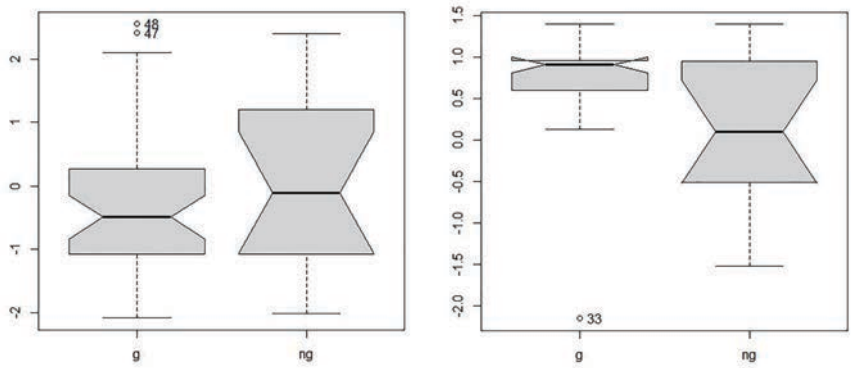
Dennoch muss man sich immer bewusst sein, dass die wahren Populationswerte nur geschätzt werden können und niemals ein 100 %iges Ergebnis vorliegt. Es wird daher immer ein gewisses Maß an Irrtumswahrscheinlichkeit bleiben. Sie ist jene Wahrscheinlichkeit, mit der wir uns irren würden, wenn wir unsere Forschungshypothese H_1 annehmen, obwohl die Nullhypothese H_0 gilt (= α-Fehler oder Fehler 1. Ordnung). Um den α-Fehler nach Möglichkeit gering zu halten, wird vorab ein Ablehnungsbereich vereinbart. Als Grenze hat sich in den Sozialwissenschaften eine tolerierte Irrtumswahrscheinlichkeit (p) von 5 % eingebürgert, was als Signifikanzniveau $p < 0{,}05$ bekannt ist. Dieser Ablehnungsbereich – in der Grafik an den weißen Flächen rechts



Übersicht 4: Normalverteilungskurve mit Konfidenzintervall von 95 % (grauer Bereich) (Quelle: eigene Darstellung)

und links der Kurve zu erkennen (siehe Übersicht 4) – ist im Grunde nichts anderes ist als die Kehrseite des Konfidenzintervalls (vgl. Hauser & Humpert 2009, S. 122). Nur wenn der Signifikanztest für einen Zusammenhang oder einen Unterschied einen Wert ermittelt, der unter $p < 0{,}05$ liegt, können wir unsere Forschungshypothese annehmen und von einem signifikanten Ergebnis sprechen (vgl. Reinders & Gniewosz 2015, S. 138). Gleichzeitig müssen wir aber auch mitbedenken, wir könnten uns zu 5 % irren.

Soweit die allgemeine Erklärung zum Signifikanztest. Liegt eine Unterschiedshypothese vor und ist die Stichprobe normalverteilt, kommt im einfachsten Fall ein *t-Test* zum Einsatz. Der t-Test untersucht, ob sich zwei empirisch gefundene Stichprobenmittelwerte systematisch voneinander unterscheiden oder ob der Unterschied zufällig ist. Zur Erinnerung: Die Nullhypothese besagt, dass der ermittelte Unterschied zwischen zwei Stichproben (z. B. Mädchen und Buben) zufällig zustande gekommen ist. Unter dieser Annahme errechnet der t-Test die Wahrscheinlichkeit für das systematische Auftreten der gefundenen Mittelwertsdifferenz oder einer Differenz, die noch größer ist (vgl. Rasch et al. 2010, S. 55).



Übersicht 5: Boxplots mit statistisch nicht bedeutsamen Unterschied (links) und mit statistisch bedeutsamen Unterschied (rechts) (Quelle: eigene Darstellung)

Das Statistikprogramm R bietet mit der Option der gekerbten Boxplots die Besonderheit einer optisch darstellbaren Signifikanzprüfung. Die Kerben symbolisieren das Konfidenzintervall von 95 %. Sind die Mittelwerte gleich oder ähnlich groß, dann überlappen die Kerben mit einer 95 %igen Wahrscheinlichkeit. Sofern sich die Kerben *nicht überlappen*, kann von einem statistisch bedeutsamen Unterschied ausgegangen werden, der nicht zufällig vorliegt (vgl. Reisinger & Wagner 2017, S. 150). In Übersicht 5 ist daher beim linken Boxplot kein signifikanter Unterschied zwischen den Gruppen zu erwarten, sehr wohl jedoch beim rechten Boxplot, auch wenn die Kerben als ein nicht sehr strenges Prüfkriterium gelten und ein Signifikanztest zur Absicherung anzuraten ist. Die Signifikanzprüfung einer Zusammenhangshypothese erfolgt bei normalverteilten Daten mit dem *Pearson-Korrelations*-Test und weitgehend analog zum t-Test[5].

Zum Abschluss dieses Kapitels noch ein Hinweis zu einer möglichen Fehlerquelle, der Kumulierung des α-Fehlers. Im Rahmen populärwissenschaftlicher Studien ist immer wieder zu beobachten, dass eine Vielzahl an Signifikanztests – zumeist auch ohne jede theoretische Begründung – gerechnet wird. Daher sei an dieser Stelle darauf hingewiesen, dass mit Daten aus einer Stichprobe *nicht beliebig* viele Signifikanztests gerechnet werden können. Der α-Fehler, der dann auftritt, wenn wir irrtümlich die Alternativhypothese H_1

[5] Für eine nähere Erklärung der unterschiedlichen Korrelationsverfahren siehe den Beitrag von Roßnagl in diesem Band.

annehmen, obwohl die Nullhypothese H_0 gilt, erhöht sich mit jedem Signifikanztest. Bei einer höheren Zahl an Signifikanztests wäre daher als Gegenmaßnahme eine sog. Bonferoni Korrektur durchzuführen. Dabei wird das Signifikanzniveau durch die Anzahl der Tests dividiert und entsprechend abgesenkt (vgl. Bortz 1999, S. 261). Bei fünf Signifikanztests und einem ursprünglichen Signifikanzniveau von $p < 0{,}05$ müsste dieses folglich auf $p < 0{,}01$ angeglichen werden.

4 Zusammenfassung

Lässt man sich von einem Statistikprogramm ausschließlich Mittelwerte oder Häufigkeiten auswerfen, um anschließend einen rein optischen Zahlenvergleich anzustellen, findet keine quantitative Forschung statt. Als Daumenregel gilt: Werden aus den Kennwerten deskriptiver Statistik interpretativ Zusammenhänge oder Unterschiede abgeleitet, handelt es sich um eine deskriptiv explorative Studie. Aus der Beschreibung der Stichprobe können, ähnlich der Vorgehensweise in qualitativen Studien, Hypothesen abgeleitet, durch die Beschreibung aber keine Hypothesen geprüft werden. Allgemeine Aussagen über Ursache-Wirkungs-Zusammenhänge wären in diesem Fall nicht seriös, denn die im Datenmaterial entdeckten Muster könnten auch *zufällig* entstanden sein. Fällt also die Entscheidung zur Durchführung einer deskriptiven Studie, sollte man sich nicht zu einfachen populärwissenschaftlichen Ursachenerklärungen verleiten lassen.

Zur Hypothesenprüfung bedarf es mathematisch, statistischer Modelle, die eine *Absicherung gegenüber Zufallsfunden* gewährleisten. Von großer Bedeutung für die Erklärung von Kausalbeziehungen ist jedoch auch das Untersuchungsdesign. Da in den Sozialwissenschaften experimentelle Designs oftmals nicht möglich sind, wird häufig auf Querschnittsuntersuchungen per Fragebogen zurückgegriffen. Diese sind für Aussagen über Ursache-Wirkungs-Zusammenhänge leider wenig geeignet. Ergebnisse aus Einzelstudien müssen daher entsprechend vorsichtig interpretiert werden und es bedarf regelmäßiger Metastudien.

Literatur

Beller, S. (2008). *Empirisch forschen lernen. Konzepte, Methoden, Fallbeispiele, Tipps.* Bern: Huber.

Bortz, J. (1999). *Statistik für Sozialwissenschafter*. Berlin: Springer.
Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschafter*. Berlin: Springer.
Hauser, B. & Humpert, W. (2009). *Signifikant? Einführung in statistische Methoden für Lehrkräfte.* Seelze-Velber: Klett-Kallmeyer.
Leonhart, R. (2009). *Lehrbuch Statistik. Einstieg und Vertiefung*. Bern: Huber.
Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2010). *Quantitative Methoden Band 1. Einführung in die Statistik für Psychologen und Sozialwissenschafter.* Berlin: Springer.
Reinders, H. & Gniewosz, B. (2015). Quantitative Auswertungsverfahren. In H. Reinders, H. Ditton, Gräsel, C. & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung. Strukturen und Methoden* (S. 131–140). Wiesbaden: Springer.
Reisinger, C. M. & Wagner, G. (2017). *Aller Anfang ist leicht. Datenanalyse mit dem R Commander.* Wien: Facultas.
Rost, D. H. (2022). *Interpretation und Bewertung pädagogischer und psychologischer Studien*. Bad Heilbrunn: Julius Klinkhardt.
Schnell, R., Hill, P. B. & Esser, E. (2011). *Methoden der empirischen Sozialforschung*. München: Oldenburg.