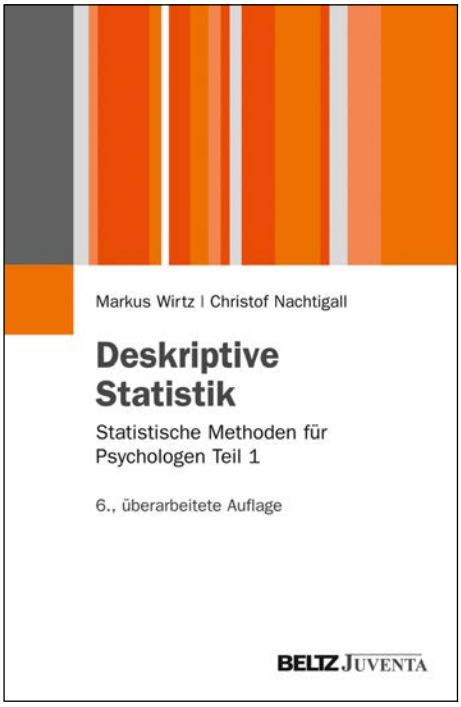

Leseprobe aus Wirtz und Nachtigall, Deskriptive Statistik, ISBN 978-3-7799-2252-0

http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-2252-0

I Einleitung

In der Psychologie werden statistische Methoden angewendet um empirische Fragestellungen zu entscheiden oder um Modelle psychologischer Sachverhalte zu bilden und zu prüfen. Die Psychologie versteht sich als *empirische Wissenschaft*, d.h. in der Psychologie werden z.B. theoriegeleitet Hypothesen über Zusammenhänge zwischen Merkmalen aufgestellt und anschließend wird überprüft, ob sich diese Zusammenhänge aufgrund von Erfahrungen oder allgemeiner aufgrund empirischer Daten bestätigen lassen. Die theoretische Überlegung, dass bestimmte Sachverhalte gelten müssten, wird erst akzeptiert, wenn diese sich tatsächlich für eine beliebige Personengruppe vorhersagen lassen.

In der Statistik und der Methodenlehre hat man sich auf bestimmte standardisierte Vorgehensweisen geeinigt, die gewährleisten sollen, dass Untersuchungsergebnisse eindeutig interpretierbar (Kriterium der Eindeutigkeit) und unabhängig vom jeweiligen Untersucher (Kriterium der Objektivität) gültig sind.

Das *Kriterium der Eindeutigkeit* besagt, dass Schlussfolgerungen aus empirischen Befunden nur dann zwingend oder eindeutig sind, wenn keine alternativen Interpretationsmöglichkeiten existieren, die diese Befunde in gleicher Weise vorhersagen.
Existieren alternative Interpretationsmöglichkeiten, so kann der empirische Befund lediglich als Hinweis auf die Richtigkeit der Schlussfolgerung gelten.

Angenommen, es wurde in einer Studie[3] festgestellt, dass sozial eher isolierte Menschen auch eher an depressiven Gefühlen leiden. Dann würde die Schlussfolgerung, dass die soziale Isolation die Entstehung von Depressionen zur Folge hat, das Kriterium der Eindeutigkeit verletzen. Denn es könnte auch sein, dass eher depressive Menschen dazu neigen, sich sozial zurückzuziehen. In diesem Fall würde die Depressivität die soziale Isolation bedingen und nicht umgekehrt. Es existieren also mindestens zwei Wirkmodelle, die den beobachteten Zusammenhang der beiden Merkmale ‚erklären' können. Deshalb ist es nicht zwingend oder eindeutig, aufgrund der Beobachtung eines Zusammenhanges eine bestimmte kausale Wirkbeziehung als bewiesen anzusehen.

Eine eindeutige Aussage über die Ursache eines Zusammenhanges kann auch im folgenden Beispiel nicht gemacht werden: In einer Untersuchung hat man

[3] In diesem Buch werden viele fiktive Untersuchungsbeispiele angeführt, die zur Verdeutlichung typischer Problemstellungen dienen. Wenn die empirische Nachweisbarkeit von Ergebnissen nicht explizit betont wird, so entstammen die berichteten Daten und Ergebnisse nicht tatsächlich durchgeführten empirischen Untersuchungen.

nachgewiesen, dass in einer Gruppe etwa 40-jähriger Raucher die Sauerstoffaufnahmefähigkeit des Blutes deutlich geringer ist als in einer Gruppe etwa 20-jähriger Nichtraucher. Aus diesem Ergebnis ist nicht erkennbar, ob der Effekt auf das höhere Alter, auf die Tatsache, dass man raucht, oder auf beide Merkmale zurückzuführen ist.

Wenn man einen Zusammenhang zwischen Merkmalen gefunden hat, so gibt es meist mehrere mögliche Modelle, die diesen Zusammenhang vorhersagen würden. Man läuft häufig Gefahr, nur das als Erklärungsmodell ins Auge zu fassen, wovon man schon vor Kenntnis des Untersuchungsergebnisses überzeugt war. Ein Ergebnis einer Untersuchung sagt dann nicht nur etwas über die beobachtete Empirie aus, sondern auch über die Konzepte, die der Untersucher aus seinem Menschen- oder Weltbild heraus bereit ist zu akzeptieren.
Ein Beispiel: Ein überzeugter Astrologe verfasst ein ‚Steinbock'-Horoskop. Er stellt fest, dass Angehörige des Sternkreiszeichens ‚Steinbock' angeben, dass dieses Horoskop sehr gut für sie zutrifft. Der Astrologe wird u.U. hieraus folgern, dass von ihm ersonnene Horoskope Kenntnisse vermitteln, die für eine bestimmte Menschengruppe spezifisch gültig sind. Ein Skeptiker könnte jedoch einwenden, dass zur Bestätigung dieser Schlussfolgerung erst ein alternatives Erklärungsmodell getestet werden muss: Menschen sind vielleicht grundsätzlich geneigt, allgemein formulierte Aussagen, die für die meisten Menschen zutreffen, als für sich spezifisch zu empfinden, wenn ihnen gesagt wird, dass diese Aussage speziell für sie bzw. ihr Sternzeichen erstellt wurde. Er könnte den Astrologen in arge Argumentationsschwierigkeiten bringen, wenn Personen aus dem ‚Widder'-Sternkreiszeichen das gleich lautende Horoskop ebenfalls als für sich selbst sehr gut zutreffend empfinden, wenn ihnen mitgeteilt wurde, dass es sich um ein ‚Widder'-Horoskop handelt.

Das *Kriterium der Objektivität* fordert, dass Untersuchungsbefunde nicht durch die Person des Untersuchers beeinflusst werden. Zudem müssen verschiedene Forscher unabhängig voneinander dasselbe Datenmaterial ähnlich beurteilen.

Angenommen, ein Psychoanalytiker stellt fest, dass die Befindlichkeit von Personen, die von ihm therapiert wurden, besser ist als von Personen, die nicht therapiert wurden. Damit die Aussage über die Effektivität der Therapie als objektiv gelten kann, müsste ein außen stehender Experte, der nicht weiß, welche Personen therapiert wurden, für die einzelnen Klienten aufgrund ihrer Befindlichkeit erkennen können, ob diese therapiert wurden oder nicht. Ansonsten kann nicht ausgeschlossen werden, dass der Therapeut – bewusst oder unbewusst – aufgrund des Wissens über die Teilnahme an der Therapie die Effektivität der Behandlung nicht mehr unvoreingenommen beurteilen kann.
Außerdem sollte gewährleistet sein, dass das Erkennen einer Befindlichkeitsbesserung nicht davon abhängt, welcher theoretischen Schule der Beurteiler angehört. Demnach müsste es egal sein, ob der Beurteiler psychoanalytisch ‚denkt', wenn er eine Befindlichkeitsbesserung feststellt. Ist ein Klient wegen

einer Angststörung in Behandlung, so wird ein Psychoanalytiker dies u.U. als Indikator für das Vorliegen ungelöster Konflikte in der frühen Kindheit ansehen. Er würde eine therapeutische Intervention auch dann schon als effektiv ansehen, wenn der Klient erkennt, dass er wichtige frühkindliche Erfahrungen noch nicht bewältigt hat. Ein Verhaltenstherapeut würde dies allerdings nicht als Therapieerfolg werten: Er würde bei der Beurteilung eines Klienten darauf achten, dass sich sein Verhalten normalisiert, ohne dass dabei notwendigerweise eine Einsicht in die Ursache der Entstehungsproblematik stattgefunden haben muss. Kurzum: Ein Psychoanalytiker wird den Therapieerfolg völlig anders definieren als ein Verhaltenstheoretiker. Patienten, die dem einen erfolgreich therapiert erscheinen, gelten für den anderen vielleicht als unverändert behandlungsbedürftig.
Ist eine Therapie wirksam, so sollte diese Wirksamkeit jedoch erkennbar sein, ohne dass man eine theoretische ‚Brille' aufsetzen muss. Kommen Anhänger unterschiedlicher Überzeugungen beim Vorliegen desselben Datenmaterials zu unterschiedlichen Schlüssen, so sagen diese Schlüsse u.U. mehr über die Überzeugungen der Beurteiler aus als über die Eigenschaften der Beurteilten: Dies ist aber nicht zulässig, wenn in der Schlussfolgerung ‚Psychoanalytische Therapie führt zur Befindlichkeitsbesserung' implizit Allgemeingültigkeit und Objektivität der Erkenntnis behauptet wird. Objektivität ist in diesem Beispiel deshalb gefordert, weil wir ja etwas über die Patienten erfahren wollen und nicht über die Therapeuten.

Befragt man Bundestagsabgeordnete, ob die soziale und wirtschaftliche Entwicklung in der Regierungszeit Angela Merkels positiv verlaufen ist, so kann man aus der Antwort sehr gut erkennen, ob der betreffende Politiker der Regierungskoalition angehört oder nicht. Was wir aber darüber erfahren, was wir eigentlich wissen wollten, bleibt ungeklärt. Hierfür müssten wir erst entscheiden, ob wir eher Angehörigen der Koalition oder der Opposition vertrauen: Also müssten wir uns die ursprünglich gestellte Frage implizit selbst beantworten.

Glaubt man, dass Schulnoten objektive Einschätzungen der Leistungen von Schülern sind, so sollte es sich nicht auf die Noten auswirken, wenn neue Lehrer Vorinformationen über die bisherigen Leistungen der Schüler erhalten. Angenommen, man gibt neuen Lehrern zufällig erstellte ‚Zeugnisse' vor und lässt sie in dem Glauben, dass dies ernsthafte Einschätzungen der Schüler durch ihre Vorgänger seien. Dann sollte sich bei Objektivität kein Zusammenhang zwischen den Zufallsnoten und der Einschätzung ergeben, die aus dem tatsächlichen Umgang mit den Schülern resultiert. Dies ist, wie sich in empirischen Studien gezeigt hat, mitnichten der Fall (Pygmalion-Effekt).

Für eine empirisch gesicherte These müssen die Beobachtungen, die diese These bestätigen sollen, des Weiteren *wiederholbar* und *zuverlässig* sein.

Das *Kriterium der Wiederholbarkeit* fordert, dass ein Ergebnis nur dann zu akzeptieren ist, wenn es sich bei wiederholten Beobachtungen immer wieder in ähnlicher Weise zeigen wird.

> Eine wissenschaftliche Aussage gewinnt man nicht dadurch, dass man in einer *bestimmten* Gruppe von Personen eine Struktur oder einen Zusammenhang zwischen Merkmalen *erkennt*. Erst wenn sich diese Struktur für eine *beliebige* Gruppe *vorhersagen* lässt, kann von der Allgemeingültigkeit der Aussage ausgegangen werden.

Würfelt man etwa achtmal nacheinander, so wird man in der gewürfelten Zahlenreihenfolge immer irgendeine Systematik erkennen können. Ein Beispiel:

1. Wurffolge: 5 3 6 4 2 4 3 1
2. Wurffolge: 2 6 3 3 1 4 5 3

Beide Wurffolgen sind Ergebnisse eines Zufallsprozesses. Trotzdem könnte man jeweils eine Systematik unterstellen: Für die erste Wurffolge könnte man z.B. behaupten, dass zu Beginn eher größere Augenzahlen gefallen sind oder dass jeweils die Summe aller aufeinander folgenden Zweierpaare [(5+3), (6+4), (2+4), (3+1)] stets eine gerade Zahl ergibt. Außerdem unterscheiden sich zwei aufeinander folgende Ziffern nie um mehr als 3 Einheiten. Dies sind wahre Behauptungen für die erste Folge, jedoch keine allgemeingültigen Erkenntnisse über die Systematik des Würfelwerfens. Bei Wiederholung des Zufallsprozesses zeigen sich diese Strukturen nämlich nicht wieder, d.h. diese Strukturen sind nicht systematisch vorhersagbar. Keine der drei Aussagen lässt sich in der zweiten Wurffolge bestätigen. Sicherlich könnten die Strukturen aus der ersten Wurfreihe zufallsbedingt wieder auftreten, jedoch würde sich bei oftmaliger Wiederholung zeigen, dass dies nicht häufiger der Fall ist als man es aufgrund von Zufall erwarten würde. Übertragen auf eine psychologische Fragestellung würde dies z.B. bedeuten, dass es bei einer Studie zur Wirksamkeit von Therapien nicht ausreichen würde, bei einer Gruppe von Patienten *im Nachhinein (a posteriori)* festzustellen, dass die Therapie um so wirksamer war, je größer der Leidensdruck zu Beginn war. Erst wenn man *im Voraus (a priori)* vorhersagen kann, dass man diesen Zusammenhang im Allgemeinen auch bei anderen Gruppen wieder finden wird, kann man von einer Erkenntnis sprechen, die allgemeingültig ist und in dieser Gruppe nicht nur durch Zufall zustande kam.

Beobachten Sie eine Menge von Merkmalen in einer Gruppe von wenigen, z.B. 20 Personen, so werden Sie tendenzielle Zusammenhänge von beliebigem Unsinnigkeitsgrad feststellen können: z.B. ‚Diejenigen mit eher dunklen Haaren sind auch die eher größeren', ‚Diejenigen, die studieren, haben eher helle Pullover an', ‚Die Nachnamen der Brillenträger beginnen mit Buchstaben, die eher am Anfang des Alphabetes stehen' usw. All das wird Ihnen normalerweise nicht bewusst werden, weil Sie wissen, dass es unsinnig ist, diese Aussagen zu verallgemeinern. Uns springt gewöhnlich nur etwas ins Auge, wenn wir diesem einen gewissen inhaltlichen Sinn unterstellen können. Ähnlich wie es schon beim Kriterium der Objektivität beschrieben wurde, laufen wir hier Gefahr, nur das als auffällig zu erkennen, was unsere Theorien von der Welt und den Menschen zu bestätigen scheint.

Zufällige Strukturen sehen manchmal sehr unzufällig aus: wenn man nur genügend viele zufällige Strukturen untersucht, werden auch einige auffällige darunter sein. Aber nur wenn man prüft, ob diese Strukturen systematisch wieder auftreten, lässt sich ihr Informationsgehalt testen.

Das *Kriterium der Zuverlässigkeit oder der Reliabilität* fordert, dass das Ergebnis einer Untersuchung möglichst genau sein sollte. Ein Befund ist dann zuverlässig oder genau, wenn Störquellen oder Zufallskomponenten einen geringen Einfluss auf das Ergebnis haben.

Das Kriterium der Reliabilität ist eng verwandt mit den oben beschriebenen Kriterien der Objektivität und der Wiederholbarkeit: Je objektiver und wiederholbarer ein Befund ist, desto zuverlässiger ist er im Allgemeinen. Zusätzlich wird bei der Reliabilität gefordert, dass ein Ergebnis möglichst präzise oder genau sein sollte. Weiß man z.B., dass die Müdigkeit das Ergebnis eines Intelligenztests beeinflusst, sollte man dafür sorgen, dass sich Personen in der Testsituation hinsichtlich der ‚Wachheit' nicht unterscheiden. Wenn man Intelligenz messen möchte, sollte das Ergebnis ja möglichst ausschließlich von der Intelligenz bestimmt werden. Die Störquelle ‚Wachheit' würde die Reliabilität oder Genauigkeit des Tests systematisch beeinträchtigen.

Möchte man den Zusammenhang zwischen Schnelligkeit und Sprungkraft bestimmen, so wird man ein unzuverlässigeres oder ungenaueres Ergebnis erhalten, wenn man die beiden Merkmale nur in groben Ordnungskategorien fasst (eher langsam bzw. geringe Weite bis eher schnell bzw. große Weite), als wenn man die beiden Merkmale genau in Sekunden bzw. in Zentimetern bestimmt.

Bei den meisten Fragestellungen sollte man ein Ergebnis erst dann akzeptieren, wenn ein Mindestmaß an Reliabilität gegeben ist: Entscheiden Sie z.B. aufgrund eines Testergebnisses, dass ein Schüler nicht das Gymnasium besuchen sollte, so ist diese Entscheidung nur dann zu vertreten, wenn das Testergebnis möglichst genau die betreffenden Eigenschaften des Schülers widerspiegelt.

Nach dem Kriterium der Wiederholbarkeit sollte ein Effekt also immer wieder auffindbar sein und nach dem Kriterium der Reliabilität sollte der gefundene Effekt in seiner Stärke möglichst genau bestimmbar sein.

Die Statistik und die Methodenlehre – die mehr oder weniger fließend ineinander übergehen bzw. häufig nur zwei verschiedene Betrachtungsweisen derselben Sachverhalte darstellen – bieten Vorgehensweisen an, die gewährleisten sollen, dass die oben beschriebenen vier Kriterien möglichst gut erfüllt bleiben. Die Statistik hat sich dabei als eine Art stark formalisierte Sprache in der Psychologie etabliert, die den Forscher dazu zwingt, das, was er tut, möglichst genau auszudrücken. Hierdurch wird transparent gemacht, wie man tatsächlich zu seinen Erkenntnissen gelangt. Dies ermöglicht es dem Rezipienten, der die ‚Sprache der Statistik' versteht, einen genauen Eindruck von der Bedeutung der Befunde zu erhalten.

Möchte man in der Psychologie also ein inhaltliches Problem lösen, so geschieht dies meist, indem man eine empirische Untersuchung durchführt. Hierfür muss ein Untersuchungsdesign entwickelt werden, mit Hilfe dessen die Fragestellung optimal beantwortet werden kann: Ein psychologisches Problem wird also erst formal übersetzt und dann durch statistische Verfahren analysiert. Die Aussage des statistischen Ergebnisses ist festgelegt, da dieses mathematisch herleitbar ist. Anschließend wird das Ergebnis der statistischen Analyse wieder in inhaltlich-psychologische Begriffe rückübersetzt. Bei der Interpretation des Ergebnisses bezieht man sich dann wieder auf die inhaltliche Bedeutung der Begriffe und bindet das Ergebnis an den bestehenden Forschungsstand an. Durch dieses gestufte Vorgehen lässt sich identifizieren bzw. trennen, was empirisch tatsächlich nachweisbar ist und was nur aufgrund des theoretischen Rahmens aus den Ergebnissen geschlussfolgert werden kann.

Mit Hilfe eines konkreten Beispiels soll dieser Prozess nun veranschaulicht werden. Dabei werden einige zentrale Probleme, auf die man normalerweise bei empirischer Forschung stößt, kurz dargestellt. Die skizzenhaften Lösungsansätze sollen lediglich grobe Orientierungen bieten, die in den folgenden Kapiteln der beiden Bände der ‚Statistischen Methoden für Psychologen' genauer ausgeführt werden.

Angenommen, ein Sozialwissenschaftler glaubt – aufgrund theoretischer Vorüberlegungen oder durch eigene Beobachtungen begründet – dass die Beurteilung des eigenen Verhaltens wesentlich durch die Reaktion der Umwelt mitbestimmt wird. Er möchte folgende Hypothese testen: ‚Je größer der Anteil von Personen in der sozialen Umwelt ist, die eine bestimmte Verhaltensweise missbilligen, desto negativer wird das Verhalten durch den Ausführenden selbst eingeschätzt.'

In diesem Beispiel sollen folgende Fragen skizzenartig diskutiert werden:

1. Wie lassen sich inhaltlich formulierte Merkmale messen?
2. Was muss für eine statistische Kennzahl gelten, damit sie geeignet ist, uns Aufschluss über die Richtigkeit der Hypothese zu geben?
3. Wie muss eine Stichprobe ausgewählt werden, damit unser Ergebnis verallgemeinert werden kann?
4. Worauf muss außerdem geachtet werden, damit unsere Beobachtung auch für Personen gilt, die wir nicht direkt untersucht haben?
5. Wie muss eine Untersuchung geplant werden, damit wir Aufschluss über die Ursache eines Zusammenhanges erhalten?

Zu 1: Wie lassen sich inhaltlich formulierte Merkmale messen?
(s. Band 1: Abschnitt II.A)

Um die Fragestellung des Forschers mit statistischen Methoden entscheiden zu können, muss festgelegt werden, welche Merkmale oder Eigenschaften unter-

sucht werden sollen und wie diese Merkmale in Zahlen gefasst werden können. Der Forscher möchte das Merkmal ‚Anteil der Personen, die das Verhalten missbilligen' untersuchen. Dieses Merkmal kann z.B. quantifiziert als prozentualer Anteil bestimmt werden: Man müsste hierzu einfach erheben, wie viel Prozent der Personen in der Umwelt ablehnend reagieren. Das zweite Merkmal ‚Bewertung einer Verhaltensweise' wird im Allgemeinen nicht als ‚in eine Zahl gefasst' oder quantifiziert empfunden. Die *Messtheorie (II.A)* beschäftigt sich z.B. mit der Frage, wie für Merkmalsausprägungen, die normalerweise in sprachlichen ‚mehr oder weniger'-Abstufungen vorliegen, ein Zahlensystem gefunden werden kann, das diese Ausprägungsgrade gut repräsentiert. In unserem Beispiel wäre es z.B. möglich, eine 9-stufige Skala zu konstruieren, auf der dem Wert ‚1' die Bedeutung ‚stark negative Einschätzung', dem Wert ‚5' die Bedeutung ‚neutrale Einschätzung' und dem Wert ‚9' die Bedeutung ‚stark positive Einschätzung' zugeordnet wird. Nach dem Kriterium der Eindeutigkeit sollte die Skala semantisch genau das beinhalten, was man unter ‚Bewertung des eigenen Verhaltens' versteht. Wenn eine Skala diese Bedingung erfüllt, bezeichnet man sie als gültig oder valide.

Die inhaltliche Fragestellung ließe sich wie folgt statistisch umformulieren:

Merkmal X = Anteil der Personen, die ein Verhalten missbilligen
Merkmal Y = Wert auf der 9-stufigen Schätzskala zur Verhaltensbewertung
Hypothese: Je größer der Wert in X ist, desto kleiner wird im Allgemeinen der Wert in Y sein.

Zu 2: Was muss für eine statistische Kennzahl gelten, damit sie geeignet ist, uns Aufschluss über die Richtigkeit der Hypothese zu geben? (s. Band 1:Abschnitt II.C und II.D)

Es soll eine Maßzahl gefunden werden, die Aufschluss über die Stärke des postulierten Zusammenhanges (Je größer X ist, desto größer/kleiner ist Y) gibt. Eine Kennzahl oder ein Koeffizient für die Stärke des Zusammenhangs zwischen Merkmalen soll die Eigenschaft besitzen, dass sein Wert umso höher ist, je stärker der Zusammenhang zwischen den beiden Merkmalen ist. Angemessen wäre hier z.B. eine Prozentzahl.

Ein Beispiel: Zwischen Haarlänge und Körpergröße besteht im Allgemeinen kein Zusammenhang. Die Maßzahl für die Stärke des Zusammenhanges sollte etwa den Wert 0% annehmen. Zwischen Körpergewicht und Körpergröße besteht ein deutlicher aber nicht unbedingt sehr starker Zusammenhang: Obwohl große Menschen im Allgemeinen schwerer sind, lässt sich die Körpergröße nicht perfekt aus dem Gewicht vorhersagen. Unser Koeffizient sollte dann irgendwo im mittleren Bereich bei 50% liegen. Hingegen ist der Zusammenhang zwischen Körpergröße in cm und Körpergröße in m perfekt: Es besteht ein 100%iger Zusammenhang, da man von einer Person der Größe 180 cm unmittelbar weiß, dass sie 1.80 m groß ist.

Für die Hypothese gilt also: Je näher die Zusammenhangsmaßzahl an 100% liegt, desto besser trifft die Hypothese zu. Bestünde ein 100%iger Zusammenhang, so ließe sich die Bewertung des eigenen Verhaltens perfekt aus der Anzahl der Personen, die das Verhalten missbilligen, vorhersagen. Bestünde ein 0%iger Zusammenhang, so wäre es für den Einzelnen unbedeutsam, wie viele Personen das eigene Verhalten missbilligen. Mit der Frage, wie eine solche Maßzahl gefunden werden kann, werden wir uns vor allem in dem Kapitel *Korrelations- und Regressionsrechnung (Abschnitt II.C)* beschäftigen.

Zu 3: Wie muss eine Stichprobe ausgewählt werden, damit unser Ergebnis verallgemeinert werden kann? (s. Band 2 – II. Inferenzstatistik)

Um die Hypothese zu prüfen, wählt man eine Gruppe von Personen (Stichprobe) zufällig aus und schaut, ob der obige Zusammenhang für diese Personengruppe besteht. Dabei ist es wichtig, dass diese Gruppe sich hinsichtlich wichtiger Eigenschaften nicht systematisch von der Menge aller Personen (Population), für die die Hypothese gelten soll, unterscheidet. Untersucht man z.B. nur Frauen, so ist es nicht zulässig, die Gültigkeit für Männer zu behaupten. Untersucht man nur junge Menschen, so ist nach der Untersuchung weiterhin fraglich, ob die Aussagen für ältere Menschen zutreffen. Bei einer *zufälligen Auswahl* von Personen kann man bei einer genügend großen Stichprobe davon ausgehen, dass die Stichprobe die Population annähernd repräsentativ, d.h. ähnlich hinsichtlich unkontrollierter Merkmale wie Alter, Geschlecht usw., abbildet.

Zu 4: Worauf muss außerdem geachtet werden, damit unsere Beobachtung auch für Personen gilt, die wir nicht direkt untersucht haben? (s. Band 2 – II. Inferenzstatistik)

Die Inferenzstatistik beschäftigt sich mit dem Problem der Verallgemeinerbarkeit von Untersuchungsergebnissen: Angenommen, in einer Zufallsstichprobe von 40 Personen findet der Forscher, dass ein Zusammenhang zwischen den beiden Merkmalen X und Y für genau diese 40 Personen existiert. Damit ist jedoch noch nicht sichergestellt, dass dieser Zusammenhang für jede beliebige andere Stichprobe auch gilt, obwohl unsere Hypothese dies behauptet.

Beispiel: Wählen wir eine Gruppe von 25 Dortmundern und 25 Münsteranern zufällig aus und betrachten die durchschnittlichen Körpergrößen, so werden wir immer einen Unterschied feststellen. Unterschiede treten also in Stichproben auf, obwohl es keinen ‚wahren', systematischen Unterschied zwischen diesen beiden Gruppen gibt. D.h. Dortmunder und Münsteraner sind im Allgemeinen im Durchschnitt gleich groß. Es stellt sich dann die Frage, welchen Sinn es macht, Stichproben zu untersuchen, wenn man weiß, dass sich je nach Stichprobe unterschiedliche Zusammenhänge oder Effekte zeigen, obwohl in der Allgemeinheit diese Effekte u.U. nicht gelten. Je nach Stichprobe werden mal die Münsteraner und mal die Dortmunder größer sein, und man müsste sehr viele Stichproben untersuchen, um ein angemessenes Bild zu erhalten.

Eine Lösung dieses Problems liefert der so genannte *Signifikanztest*: Ist ein Zusammenhang signifikant, so weiß man mit einer bestimmten Sicherheit, dass der Zusammenhang in einer untersuchten Stichprobe so groß ist, dass man diesen Zusammenhang in jeder anderen beliebigen Stichprobe ebenfalls erwarten kann. Anders ausgedrückt: Von der Gültigkeit eines Zusammenhanges in der Stichprobe kann auf die Gültigkeit in der Allgemeinheit geschlossen werden.

Ein Signifikanztest würde in unserem Beispiel also zu folgender Aussage führen: Für unsere Stichprobe ist der Zusammenhang zwischen der Selbstbeurteilung und dem Anteil der Personen, die das Verhalten missbilligen, so stark, dass mit einer gewissen Sicherheit davon ausgegangen werden kann, dass sich dieser Zusammenhang stets finden lässt. Der Zusammenhang hat sich nicht nur gezeigt, weil wir genau diese Stichprobe untersucht haben.

Zu 5: Wie muss eine Untersuchung geplant werden, damit wir Aufschluss über die Ursache eines Zusammenhanges erhalten? (s. Band I – Abs. II.C.3):

Problem der Kausalität und der Konfundierung: Die zu testende Hypothese lautet: ‚Je mehr Personen aus der sozialen Umwelt eine bestimmte Verhaltensweise missbilligen, desto negativer wird das Verhalten durch den Ausführenden selbst eingeschätzt'. Man beachte hierbei, dass der Forscher eigentlich die Theorie hat (s. S. 22), dass die Umweltreaktion einen *verursachenden, kausalen Einfluss* auf die Selbstbewertung hat. Diese kausale Wirkrichtung ist in der Hypothese jedoch nicht explizit berücksichtigt worden. Dies stellt in diesem Zusammenhang ein wichtiges Problem dar. Man betrachte folgendes Beispiel: Ein Arzt, der durch eine schwierige Operation einem Menschen das Leben gerettet hat, wird genau wie seine Umwelt seine Tat als uneingeschränkt positiv bewerten. Dies liegt jedoch daran, dass das Verhalten des Arztes an sich positiv zu bewerten ist. Somit kommt es zu einer positiven Einschätzung durch den Arzt und durch die Umwelt, ohne dass die Umwelteinschätzung die Selbsteinschätzung des Arztes beeinflusst. Also: Wenn ein Zusammenhang zu erkennen ist, ist noch nicht unbedingt klar, welche sachlogische Beziehung zwischen den Merkmalen besteht.

Auch wenn die Lebenserwartung in den letzten einhundert Jahren mit der Luftverschmutzung angestiegen ist, ist davon abzuraten, durch das Bevorzugen luftverschmutzter Lebensräume seine Lebenserwartung steigern zu wollen. Die zunehmende Industrialisierung führte zum einen zu erhöhter Luftverschmutzung, wirkte sich aber auch positiv aus, indem die medizinischen Bedingungen sowie der allgemeine Lebensstandard verbessert wurden. In einem solchen Fall sagt man, dass die Variable ‚Luftverschmutzung' und die Variable ‚Industrialisierung' bzw. ‚Lebensstandard' gekoppelt oder konfundiert sind und somit u.U. nicht mehr unabhängig voneinander interpretiert werden können.

Auf unsere Ausgangsfragestellung bezogen bedeutet dies, dass ein Zusammenhang zwischen dem ‚Anteil der missbilligenden Personen (X)' und der ‚Selbsteinschätzung (Y)' auf unterschiedliche kausale Wirkungen hindeuten kann.

(i) $X \Rightarrow Y$: X verursacht Y, d.h. die Umwelt beurteilt ein Verhalten negativ und in Folge dessen verändert sich die Selbsteinschätzung des Ausführenden.

(ii) $Y \Rightarrow X$: Y verursacht X, d.h. dadurch dass der Ausführende selbst sein Handeln negativ beurteilt, missbilligt auch die Umwelt das Verhalten.

(iii) $X \Leftrightarrow Y$: X und Y beeinflussen sich wechselseitig, d.h. es bestehen beide Wirkrichtungen gleichzeitig. Durch Kommunikation zwischen Individuum und Umwelt finden beide Seiten zu ähnlichen Einschätzungen.

(iv): $Z \Rightarrow X$; $Z \Rightarrow Y$: Es existiert eine dritte Variable Z, die die Ausprägung in X und Y bestimmt. Die oben angesprochene Eigenschaft ‚Valenz der Tat unabhängig von der Beurteilungssituation' wäre ein Beispiel für eine solche Wirkbeziehung. Da eine bestimmte Tat (z.B. die des Arztes) an sich immer positiv (negativ) bewertet wird, wird sie sowohl vom Ausführenden als auch von der Umwelt eher als positiv (negativ) eingeschätzt.

Es gibt noch viele andere Wirkgefüge, die zu einem Zusammenhang von X und Y führen könnten. Falls der Forscher aber wirklich daran interessiert ist, die Wirkbeziehung in (i) nachzuweisen, kann ihm mit der Durchführung eines *Experiments* geholfen werden. Einem Experiment liegt der Gedanke zugrunde, dass ein verursachender Effekt dann bewiesen werden kann, wenn zwei Situationen sich nur hinsichtlich *eines* Merkmals (unabhängige Variable) unterscheiden und im Folgenden eine Veränderung eines anderen Merkmals (abhängige Variable) beobachtet wird. Alle Merkmale, die die abhängige Variable beeinflussen könnten, außer der unabhängigen Variablen werden konstant gehalten. So können diese Störvariablen in der Experimentalsituation keinen Einfluss auf die abhängige Variable mehr haben. In unserem Beispiel würde also gelten:

- Unabhängige Variable: Anzahl der Personen, die ein Verhalten missbilligen (X)
- Abhängige Variable: Einschätzung des eigenen Verhaltens (Y)
- Eventuelle Störvariable: Valenz der Tat unabhängig von der Beurteilungssituation (Z)

Wir würden also beachten müssen, dass alle zu testenden Personen dasselbe Verhalten ausführen müssen, und können damit garantieren, dass sich für alle Situationen die Beurteilung der Tat (Z) an sich nicht verändert. Des Weiteren müssen wir dafür sorgen, dass sich die Anzahl der missbilligenden Personen (X) für verschiedene Situationen ändert, obwohl die Tat dieselbe bleibt. Dieses Problem kann man mehr oder weniger geschickt gelöst werden, indem man sogenannte Konföderierte (vom Experimentator eingeweihte Personen) einsetzt. Die Experimentalsituation könnte z.B. so arrangiert sein: Immer dann, wenn die zu testende Person ihr Verhalten zeigen soll, sind 10 Personen anwesend. Die-

sen ist vorgegeben, ob sie das Verhalten missbilligen sollen oder nicht. Dass die Konföderierten nicht echt reagieren, darf der ausführenden Person natürlich nicht bekannt sein. Nachdem die Person, die getestet wird, die Reaktion der Anwesenden wahrgenommen hat, wird ihre Selbsteinschätzung erhoben. Durch dieses Vorgehen wird insbesondere gewährleistet, dass die Umweltreaktion nicht von der Selbsteinschätzung des Ausführenden beeinflusst werden kann (s. Fälle (ii) und (iii)).

Die dargestellte ‚Lösung' des behandelten Problems ist natürlich sehr skizzenhaft und in den Details verbesserungsbedürftig. Aber auch wenn dieses Experiment sehr gekünstelt und unnatürlich wirkt – was es im Übrigen nicht unbedingt von vielen tatsächlich durchgeführten psychologischen Experimenten unterscheidet –, so ist doch die Kontrolle möglicher Faktoren, die die Eindeutigkeit der Interpretation erschweren, häufig unabdingbar. Es besteht im Allgemeinen eine Art Unschärfebeziehung: Je besser man das Problem der Interpretierbarkeit löst, desto gravierender werden andere problematische Einflüsse. Je eindeutiger ein Verhalten zu interpretieren ist, desto unnatürlicher ist das Verhalten, desto mehr stellt sich die Frage, ob das Verhalten in der Experimentalsituation irgendeinen Aufschluss über das Verhalten in natürlichen Umgebungen im Alltag zulässt. Die Tatsache, dass man sich bewusst ist, gerade an einem Experiment teilzunehmen, kann dazu führen, dass Menschen sich nicht mehr natürlich verhalten. Soll die untersuchte Person ihrem psychischen Befinden durch Ankreuzen einer Zahl in einem Fragebogen Ausdruck verleihen, stellt sich ebenfalls die Frage, welchen Aufschluss diese Messung über das Erleben im Alltag gibt. Das Ankreuzen einer Zahl auf einer Einschätzungsskala in einem Fragebogen ist nicht ohne Weiteres mit dem Entwickeln von Gefühlen oder Verhaltensweisen in konkreten Situationen gleichzusetzen.

Wenn in einer Experimentalsituation Verhalten erzeugt oder angegeben wird, das im normalen ökologischen Umfeld so nicht gezeigt wird, so sind die Ergebnisse genauso wenig interpretierbar, wie wenn nicht klar ist, worauf das Verhalten denn nun tatsächlich zurückzuführen ist. Eine ‚gute' psychologische Untersuchung muss also sowohl die Interpretierbarkeit als auch die Übertragbarkeit (ökologische Gültigkeit oder Validität) gewährleisten.

Leseempfehlungen

In den ‚Statistischen Methoden für Psychologen' wird ein Überblick und eine Einführung in grundlegende statistische Verfahren gegeben. Gerade in der Statistik ist es aber anzuraten, sich zu denselben Inhalten die Darstellungen in verschiedenen Lehrbüchern anzuschauen. Insbesondere Texte und Bücher, die die Anwendung in den Vordergrund stellen, sind geeignet, den Sinn und Zweck der Statistik in der Psychologie und in der Forschung allgemein zu klären. Wenn Ihnen also Texte zu Beginn des Studiums begegnen, die statistische Methoden behandeln oder verwenden, versuchen Sie für sich selbst oder mit Kommilitonen/Dozenten explizit zu klären, weshalb diese angewendet werden und welcher Nutzen dadurch entsteht.

An dieser Stelle sei lediglich auf einige wichtige Statistikbücher hingewiesen. In den Leseempfehlungen der folgenden Kapitel werden detailliertere Informationen gegeben. Im deutschsprachigen Raum sind die Lehrbücher von Eid, Gollwitzer und Schmidt (2010) sowie Sedlmeier und Renkewitz (2008) besonders zu empfehlen, da sie neben den ausführlichen Einführungen der statistischen Verfahren auch sehr gut die Verbindung zu allgemeinen Forschungsmethoden verdeutlichen. Die Lehrbücher von Bortz und Schuster (2010) sowie Fahrmeier, Pigeot, Künstler und Tutz (2011) stellen die statistichen Aspekte in den Mittelpunkt. Bortz und Döring (2006) ermöglichen einen sehr umfassenden Einstieg in die psychologische Forschungsmethodik und -methodologie.

Zur Einführung in komplexere statistische Analyseverfahren eignen sich vor allem die Lehrbücher von Rudolf und Müller (2004) sowie Backhaus, Erichson, Plinke und Weiber (2010). Auch wenn in letzterem die angewandten Beispiele der Betriebswirtschaft entstammen, so ist der Transfer auf psychologische Problemstellungen meist sehr naheliegend. Das englischsprachige Buch von Hair, Anderson, Tatham und Black (2009) ist zwar nicht unbedingt für den Einstieg geeignet: Wenn Sie aber im Verlaufe des Studiums (z.B. bei der Erstellung der Diplomarbeit) einfache oder komplexe statistische Verfahren anwenden, so gibt es unseres Wissens nach keinen Text, der in ähnlich vollständiger Weise Standards bei der Durchführung des Verfahren und Lösungen für potentielle Probleme liefert. Das Buch von Tabachnik und Fidell (2007) ist in diesem Zusammenhang ebenfalls zu empfehlen, insbesondere weil auf der zugehörigen homepage Beispieldatensätze zur Verfügung gestellt werden.

Die für Einsteiger besonders zu empfehlende Bücher von Beller (2008) und Huber (2002) verdeutlichen für das empirische Forschen allgemein und das Experiment, wie eng verknüpft inhaltliches Vorgehen und die Anwendung statistischer Methoden in der Psychologie sind. Wenn Sie selbst Ihren ersten Versuch oder Ihr erstes Experiment durchführen, finden Sie im Text von Hager, Spieß und Heise (2001) eine prägnante Einführung in Standards bei der Durchführung und Darstellung von Untersuchungsergebnissen. Die Methode der Beobachtung wird von Greve und Wentura (1997) und die Methode der Test- und Fragebogenkonstruktion wird von Bühner (2010) einführend dargestellt: In beiden Büchern wird die Bedeutung statistischer Methoden gut verständlich heraus gearbeitet.

Einen sehr unterhaltsamen Zugang zum Umgang mit statistischen Fragen, erhält man durch die Bücher von Krämer (2000, 2001), Beck-Bornholt und Dubben (2001) sowie Gonick und Smith (1993). Insbesondere die Darstellung von Fehlern in der Anwendung und bei der Interpretation, zeigt besonders plakativ, wie wichtig ein angemessener Umgang mit statistischen Methoden in der Forschung allgemein ist. Das Buch von Hall (1998) gibt einen sehr guten und unterhaltsamen Einblick in gängige Forschungs- und Veröffentlichungspraxis.