Statistik leicht gemacht

Guido Strunk

Statistik leicht gemacht

[Arbeitstitel]

Complexity-Research, Forschung & Lehre, Verlag

Statistik leicht gemacht
[Arbeitstitel]

ISBN 978-3-903291-[\ \]

1050 Wien, Schönbrunner Str. 32 / 20, www.complexity-research.com

Umschlaggestaltung: Sofie Strunk
Druck: Books on Demand GmbH, D-22848 Norderstedt, In de Tarpen 42

Guido Strunk, Technische Universität Dortmund, Deutschland, Complexity-Research Wien, Österreich,
FH Campus Wien, Österreich

Inhalt

1 Einleitung

Bereits in der griechischen Antike – insbesondere bei den Pythagoreern – findet sich die Überzeugung, dass die Gesetze der Natur ihrem Wesen nach mathematisch seien (Herrmann, 2014, S. 46 ff.). Aristoteles (2019/4. Jhdt. v. Chr., S. 38, Buch I, B) schreibt in seiner Metaphysik über die Pythagoreer:

> Da sie nun auch darauf aufmerksam wurden, dass die Verhältnisse und Gesetze der musikalischen Harmonie sich in Zahlen darstellen lassen, und da auch alle anderen Erscheinungen eine natürliche Verwandtschaft mit den Zahlen zeigten, die Zahlen aber das Erste in der gesamten Natur sind, so kamen sie zu der Vorstellung, die Elemente der Zahlen seien die Elemente alles Seienden und das gesamte Weltall sei eine Harmonie und eine Zahl.

Galilei (1953/1623) formuliert anschaulich und aus voller Überzeugung, dass das Buch der Natur in der Sprache der Mathematik geschrieben sei, wohingegen Albert Einstein verwundert nachfragt:

> Wie ist es möglich, dass die Mathematik, die doch ein von aller Erfahrung unabhängiges Produkt des menschlichen Denkens ist, auf die Gegenstände der Wirklichkeit so vortrefflich passt? (Einstein, 2002/1918-1921, S. 385, zitiert nach Krey, 2012)

Statistik ist angewandte Mathematik, die sich vor allem auf empirische Gegebenheiten bezieht.

Vor diesem Hintergrund versteht sich die Statistik als eine Art angewandte Mathematik. Im Gegensatz zur reinen Mathematik beschäftigt sie sich mit empirischen Sachverhalten. Sie verwendet Zahlen zur Abbildung dieser empirischen Sachverhalte und mathematisch fundierte Methoden zur Interpretation dieser Zahlen. Grob lässt sich die Statistik in drei Bereiche unterteilen: Die so genannte *deskriptive* Statistik beschreibt die Welt mit Hilfe von Zahlen. Merkmale in der empirischen

Welt werden in Zahlen abgebildet und die deskriptive Statistik versucht, die Beschreibung der Welt durch Zahlen zu verbessern, z. B. Beobachtungen zu objektivieren oder vergleichbar zu machen. Die sogenannte *explorative, entdeckende* Statistik geht einen Schritt weiter und sucht nach Mustern in Daten. Diese Muster werden mit Hilfe der Methoden erst entdeckt und können daher neu und überraschend sein. Diese Form der Statistik kann daher helfen, bisher unbekannte Phänomene oder neuartige Erklärungen für Phänomene zu entdecken. Die so genannte *prüfende* (*Inferenz-*) Statistik bewertet das Ausmaß der Übereinstimmung oder der Abweichung empirischer Befunde von zuvor formulierten Hypothesen oder Prognosen.

2 Messung, Mess- und Skalenniveau

Statistik basiert auf Zahlen, die durch Messungen gewonnen werden. Da sich die Statistik in der Regel auf die Empirie bezieht, ist sie in der einen oder anderen Form auf Messdaten angewiesen, die Sachverhalte der empirischen Welt in Zahlen abbilden. Diese Abbildung der empirischen Welt in die Zahlenwelt wird als Messung bezeichnet. In Bezug auf die Physik formuliert Lord Kelvin es wie folgt:

> In physical science a first essential step in the direction of learning any subject is to find principles of numerical reckoning and methods for practicably measuring some quality connected with it. I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, … (Thomson, 1889, S. 73)

Messungen können auf sehr unterschiedliche Weise durchgeführt werden. Bei ganz einfachen Messungen werden Ereignisse, Merkmale, Objekte usw. lediglich gezählt. Oder es werden Unterschiede zwischen Objekten durch unterschiedliche Zahlen ausgedrückt, z. B. indem größere Objekte größere Zahlen erhalten als kleinere Objekte. Nach einer Messung liegen Zahlen vor. Diese Zahlen lassen dann nicht mehr erkennen, wie sie zustande gekommen sind und welche Messverfahren sie hervorgebracht haben. Das ist ein Kernproblem der Statistik. Denn in der Mathematik kann man mit Zahlen alles machen, was man mit Zahlen machen kann. Man kann die Grundrechenarten anwenden, aber auch die Wurzel ziehen, den Logarithmus bestimmen usw. Wenn die Zahlen aber nur ein Code für z. B. die Sozialversicherungsnummer sind, dann sind solche Rechenkunststücke mathematisch zwar durchaus ebenfalls durchführbar, aber inhaltlich wahrscheinlich sinnlos. Da man den Zahlen selbst nicht ansieht, wofür sie stehen und was daher eine sinnvolle mathematische Verwendung sein könnte, ergeben sich hier einige gravierende Schwierigkeiten.

Zusammenfassend kann gesagt werden, dass bei einer Messung empirische Gegebenheiten durch Zahlen repräsentiert werden. Ziel ist es, die Un-

terschiede, Ähnlichkeiten oder Beziehungen, in denen die empirischen Sachverhalte zueinander stehen, möglichst gut durch die Zahlen wiederzugeben. Nach der Messung liegen nur die Zahlen vor und es muss kommuniziert werden und bekannt sein, wie die Messzuordnung erfolgt ist und welche Eigenschaften der Zahlen interpretiert werden können und was aufgrund der Art der Messung nicht interpretiert werden kann.

Formal betrachtet lässt sich eine Messung wie folgt definieren:

Definition Eine Messung ist die homomorphe Abbildung eines empirischen Relativs in ein numerisches Relativ.

Empirische Sachverhalte werden hier formal als „empirisches Relativ" bezeichnet. Damit wird betont, dass es sich um Strukturen mit relativen Unterschieden, Ähnlichkeiten usw. in der empirischen Welt handelt. Dieses empirische Relativ wird auf die Zahlenwelt abgebildet, die ebenfalls Unterschiede, Ähnlichkeiten usw. aufweist und daher als „numerisches Relativ" bezeichnet wird. Die Abbildung des empirischen Relativs (empirische Strukturen) in das numerische Relativ (numerische Strukturen) soll möglichst strukturerhaltend erfolgen. Eine strukturerhaltende Abbildung wird als „homomorphe Abbildung" bezeichnet.

Die Strukturen, die die Zahlenwelt im Angebot hat, sind in der Regel vielfältiger als die Strukturen, die die Messung tatsächlich nutzt. Messungen werden daher danach klassifiziert, welche Strukturen der Zahlenwelt mit der empirischen Welt übereinstimmen und welche nicht interpretiert werden können. Zwei Aspekte sind für die Klassifikation von Messungen von Bedeutung. Zum einen geht es um die sogenannte Stetigkeit (d.h. die Frage, ob es beliebige Zwischenwerte geben kann). Darauf wird weiter unten eingegangen. Der andere Aspekt wird als Skalenniveau bezeichnet, und dieses Niveau kann unterschiedlich hoch sein (Stevens, 1946). Ein höheres Niveau ist statistisch gesehen immer besser, da bei einer Messung mit einem hohen Niveau viele Informationen aus den Zahlen herausgelesen und interpretiert werden können.

2.1 Skalenniveaus

Die Skalenniveaus werden vor allem durch die Veränderungen definiert, die man an den Zahlen nach der Messung noch vornehmen darf, ohne die Messung kaputt zu machen (vgl. Tabelle 1). Das klingt etwas abstrakt. Ge-

meint ist, dass Messdaten nach der Messung durchaus noch verändert werden können. So ist es üblich, Messdaten aus einer Maßeinheit (z. B. Kilometer) in eine andere Maßeinheit (z. B. Seemeilen) umzurechnen. Dafür gibt es viele Beispiele. So haben verschiedene Länder z. B. unterschiedliche Maßeinheiten für die Temperatur (Grad Celsius, Grad Fahrenheit, Kelvin) oder für Längenmaße (Meter, Elle, Meile, Seemeile, Astronomische Einheit usw.) oder für Geld (Euro, Jen, Dollar usw.). Es ist also durchaus üblich, dass Zahlen nach der Messung in andere Einheiten oder Zahlensysteme usw. umgerechnet werden.

Skalenniveau	**Das darf eine Transformation nicht verändern ...**	**Zulässige Interpretation, Beispiel**
Nominal	Ein-eindeutig Zuordnung	Code, Bezeichnung, Beispiel: Berufe
Ordinal	Reihenfolge	Rangordnung Beispiel: Schulbildung
Intervall	Intervalle zwischen den Zahlen (erlaubt ist die Addition/Subtraktion von Konstanten, sowie die Multiplikation/Division mit Konstanten)	Abstände (Intervalle) zwischen den Zahlen Beispiel: Alter
Verhältnis	Verhältnisse zwischen den Zahlen (erlaubt ist die Multiplikation/Division mit Konstanten)	Verhältnisse zwischen den Zahlen Beispiel: Gehalt
Absolut	Nichts darf verändert werden.	Verhältnisse zwischen den Zahlen, Kardinalzahl Beispiel: Häufigkeiten

Tabelle 1: Skalenniveaus
Die Skalenniveaus – ohne Absolutskala – wurden zuerst von Stevens (z. B. 1946) beschrieben.

Diese nachträgliche Veränderung von Messdaten hat allerdings Grenzen. Da eine Messung Strukturen der empirischen Welt im Zahlenraum abbilden möchte, darf eine nachträgliche Veränderung der Messdaten diese Strukturen nicht wieder zerstören. Wenn es beispielsweise das Ziel ist, Objekte nach ihrer Größe zu sortieren, und man misst die Größe auf die eine oder andere Weise, dann kann man die Messdaten zwar nachträglich in

andere Maßeinheiten umrechnen, aber wenn dadurch die angestrebte Sortierung zerstört wird (man also nicht mehr weiß, welches Objekt das größte und welches das kleinste ist), dann hat man etwas falsch gemacht. Nicht jede Form der nachträglichen Manipulation von Messdaten ist sinnvoll. Je besser die Messdaten erfasst wurden, desto weniger sind nachträgliche Veränderungen zulässig. Wenn ich z. B. frage, wie viele Buchstaben dieser Satz hat, dann ist die Antwort eine ganz bestimmte Zahl, und diese Zahl kann nicht mehr sinnvoll in andere Einheiten umgerechnet werden. Dies Zahl ist absolut. Sie heißt auch Kardinalzahl.

Nominalskala

Das geringste Skalenniveau hat die *Nominalskala*: Zahlen werden ein-eindeutig verwendet und den Objekten oder Merkmalen so zugeordnet, dass man diese empirischen Gegebenheiten anhand der Zahl erkennen kann. Die Zahlen sind also Bezeichnungen für die empirischen Objekte, Merkmale usw. Sie ermöglichen es, diese Objekte, Merkmale etc. zu identifizieren. Beispiel: Zahlencode für Berufe, Bäcker:in = 234, Professor:in = 43, ... Die Höhe der Zahlen hat keinerlei Bedeutung. Daher kann jede Zahl verwendet werden, um z. B. eine Bäcker:in zu codieren, solange diese Zahl eben nur für diesen Beruf verwendet wird. Bei der Nominalskala ist also jede nachträgliche Veränderung der Zahlen erlaubt, solange dadurch die ein-eindeutige Zuordnung nicht zerstört wird. Statistisch lässt sich mit dieser Skala nicht viel berechnen, ein Mittelwert z. B. kann zwar für die Zahlencodes ermittelt werden – inhaltlich macht er aber keinen Sinn. Die Tabelle 2 (S. 26) stellt die Methoden vor, die üblicherweise verwendet werden, um diese Daten deskriptiv (beschreibend) darzustellen.

Praktische Hinweise

Die Nominalskala kommt häufig zum Einsatz, um das Vorliegen eines bestimmten Merkmals festzustellen. Dabei wird erfasst, ob ein Merkmal (z. B. ein bestimmter Beruf) vorliegt oder nicht. Daher lassen sich nominale Daten in einer Datenbank auf zwei Arten abspeichern. Man könnte beispielsweise jeden Beruf in einer eigenen Variable speichern. Es gäbe dann eine Variable „Bäcker:in“ und dort würde gespeichert werden, ob eine befragte Person diesen Beruf hat oder nicht. In einem solchen Fall sollte man die Zahlen 0 für „trifft nicht zu“ und 1 für „zutreffend“ verwenden. Diese Kodierung mit 0 und 1 wird auch als Boolsche Codierung bezeichnet und für weiterführende statistische Testungen ist diese häufig Voraussetzung. Da es sich um eine Nominalskala handelt, könnten auch andere Kodierungen gewählt werden. In der Praxis hat die Kodierung mit 0 vs. 1 jedoch viele Vorteile, weshalb man sich angewöhnen

sollte, alle Daten, die nur zwei Zustände kennen (sog. dichotome Daten), immer einheitlich mit 1 und 0 zu kodieren. Leider machen das viele Online-Fragebögen falsch und vergeben häufig die Zahlen 1 und 2, wobei mitunter sogar 1 für „ja" und 2 für „nein" verwendet wird. Diese Daten müssen vor der Benutzung umkodiert werden. Wenn jeder Beruf beispielsweise mit einer eigenen Variable erfasst wird, können Mehrfachantworten problemlos miterfasst werden. So können Menschen ja gleichzeitig mehrere Berufe haben. Oder sie können gleichzeitig sowohl unselbstständig als auch selbstständig tätig sein. Bei Untersuchungen mit Papierfragebögen kommt es immer wieder vor, dass Personen zwei oder mehr Antworten ankreuzen, obwohl nur eine Antwort erwartet wird. Bei Online-Fragebögen wird die Möglichkeit einer Mehrfachantwort häufig verhindert. Das kann hilfreich sein, führt aber zu hohen Abbruchraten, wenn Menschen daran gehindert werden, ihre Lebensrealität in den Fragebogen einzutragen. In einigen Bereichen gehört es zur Lebensrealität, mehrere berufliche Tätigkeiten gleichzeitig auszuüben. Ein Fragebogen, der dazu zwingt, nur eine Tätigkeit anzugeben, kann Ärger auslösen. Wenn für eine Variable wirklich nur eine Antwort infrage kommt, kann es einfacher sein, diese eine Frage als Variable zu erfassen und dort für jeden Beruf einen Code einzutragen. Falls sich in der Literatur – beispielsweise bei statistischen Ämtern – bereits bewährte Kodierungen finden, sollten diese verwendet werden. Beispielsweise gibt es eine internationale Standardklassifikation für Berufe (ISCO). Für viele statistische Auswertungen müssen die in einer einzigen Variable erfassten Codes später dennoch in neue Variablen umgruppiert werden, die für jeden erfassten Beruf das Vorliegen oder Nichtvorliegen einzeln abbilden. Das liegt daran, dass mit nominalen Daten nur Codes erfasst werden, die sich statistisch kaum auswerten lassen. Beispielsweise kann die Variable „Beruf" in der Regel nicht mit Erkrankungen korreliert werden, wenn die Erkrankungen und die Berufe mit beliebigen Zahlencodes in jeweils einer Variable erfasst werden. Wird ein Beruf aber z. B. mit „Pflege" ja vs. nein – also 0 vs. 1 – kodiert und eine Erkrankung wie „unspezifische Rückenschmerzen" ebenfalls mit 1 und 0, können beide Variablen leicht miteinander korreliert werden.

Ordinalskala

Das nächst höhere Skalenniveau ist das der *Ordinalskala*: Die Anordnung der Zahlen nach ihrer Größe entspricht einer Ordnung der empirischen Gegebenheiten. Diese wird aber nur grob wiedergegeben bzw. ist tatsächlich nur grob vorhanden. Der Abstand der Zahlen zueinander kann daher nicht als Abstand der empiri-

schen Sachverhalte zueinander interpretiert werden. Ein Beispiel dafür ist der höchste Bildungsabschluss: Pflichtschule = 1, Matura/Abitur = 2, Studium = 3, ... Höhere Bildungsabschlüsse erhalten höhere Zahlen. Da aber der konkrete Zahlenwert in diesem Fall keine Bedeutung besitzt, können statt 1, 2, 3, ... auch die Zahlen -30, 44, 102 verwendet werden. Während bei der Nominalskala die Höhe der Zahlen gar keine Rolle spielt, nutzt die Ordinalskala die Reihenfolge von Zahlen, die man ja leicht nach Größe ordnen kann. Alle Transformationen sind erlaubt, solange sie die Reihenfolge nicht verändern. Statistisch lässt sich auch mit dieser Skala nicht viel berechnen, ein Mittelwert z. B. kann aus den Ordinalzahlen ebenfalls nicht sinnvoll ermittelt und interpretiert werden. Aber die Mitte einer sortierten Anordnung (*Median*) erlaubt eine sinnvolle statistische Kennzeichnung einer zentralen Tendenz der Daten (vgl. Tabelle 2, S. 26).

Praktische Hinweise

Für Ordinalskalen könnte es in der Literatur ebenfalls übliche Kodierungen geben, denen idealerweise gefolgt werden sollte. Oft muss jedoch selbst überlegt werden, wie sich Angaben im Fragebogen am besten in Datenbanken eintragen lassen. Die Empfehlung lautet, nach Möglichkeit aufsteigende Zahlen zu verwenden, die immer bei 1 beginnen. Das macht die spätere Datenanalyse bedeutend einfacher. Während bei der Nominalskala beliebige Merkmale ohne eine Richtung von „mehr“ oder „weniger“, „größer“ oder „kleiner“ etc. kodiert werden, gibt es bei der Ordinalskala eine Antwortrichtung. Große Probleme entstehen, wenn Zahlen vergeben werden, die nicht zur Antwortrichtung passen. Beispielsweise könnte die Zufriedenheit mit drei möglichen Antworten – „gar nicht zufrieden“, „weder noch“ und „sehr zufrieden“ – abgefragt worden sein. Die Zahlen 1, 2 und 3 sind eine gute Wahl. Sie sollten dann aber so vergeben werden, dass hohe Zahlen eine hohe Zufriedenheit ausdrücken. Wichtig ist, dass der Name der Variable – hier beispielsweise Zufriedenheit – die Messrichtung angibt und hohe Zahlen dieser Messrichtung entsprechen. Verwirrend sind Notenskalen, bei denen hohe Zahlen eine schlechte Note bedeuten. Notenskalen wie „sehr gut“ (1), „gut“ (2) usw. bilden nicht die Leistung, sondern das Unvermögen ab. Das muss bei der Interpretation von Korrelationen berücksichtigt werden. Eine positive Korrelation zwischen Note und Alter bedeutet beispielsweise, dass hohe Zahlen in der Note regelmäßig bei hohen Zahlen für das Alter vorkommen.

Intervallskala

Das nächst höhere Skalenniveau ist das der *Intervallskala*: Die Abstände zwischen den Zahlen können bei einer Intervallskala sinnvoll interpretiert werden. Zahlenverhältnisse sind hingegen nicht definiert und können nicht verwendet, also z. B. nicht sinnvoll interpretiert werden. Das Alter gemessen in Jahren ist ein gutes Beispiel für eine Intervallskala. Ein Jahr ist ein klar definiertes Zeitmaß und wenn eine Person zwei Jahre älter ist als eine andere Person ist dieser zeitliche Abstand, also das Intervall zwischen den beiden inhaltlich gut verstehbar. Wichtig ist auch festzuhalten, dass sich dieser Abstand nicht verändert, wenn Zeit vergeht. Der Abstand ist invariant (unveränderlich) gegenüber der Zeit. Eine Messung sollte so vorgenommen werden, dass sie Messwerte liefert, die mit den Besonderheiten der vorliegenden inhaltlichen Problemstellung umgehen kann. Es ist nützlich, dass zeitliche Abstände sich nicht verändern, wenn weiterhin Zeit vergeht. Das Beispiel wird so ausführlich behandelt, weil daran gezeigt werden kann, dass weitere, höhere Skaleneigenschaften für Zeitmessungen nicht invariant sind und auch bei noch so präzise gemessenen Zeiten nicht interpretiert werden können. Wenn eine Person in einem gegebenen Moment exakt doppelt so alt ist wie eine andere, ist sie es am nächsten Tag oder in der nächsten Stunde oder Minute, im nächsten Augenblick nicht mehr. Das Zahlenverhältnis aus Altersangaben ist nicht sinnvoll interpretierbar. Es ändert sich mit jeder Sekunde, die vergeht. Das Alter ist (nur) eine Intervallskala. Die Intervalle sind invariant gegenüber weiterhin vergehender Zeit. Die Zahlenverhältnisse sind hingegen in jeder Sekunde die vergeht andere. Das Alter ist also keine Verhältnisskala.

Nachträglich kann eine Intervallskala durchaus noch in andere Maßzahlen umgerechnet werden, wenn die Abstände dabei nicht zerstört werden und interpretierbar bleiben. So kann ein Intervall von 2 Jahren auch als 24 Monate angegeben werden. Dies wird durch Multiplikation mit 12 Monaten pro Jahr berechnet. Jede Multiplikation oder Division sowie jede Addition oder Subtraktion (im Beispiel vergeht Zeit und das wäre eine Addition auf das Alter der beiden Personen) ist erlaubt. Die Intervalle werden dadurch nicht zerstört. Statistisch kann hier erstmals sinnvoll der Mittelwert berechnet werden, da der Mittelwert die Abstände (Intervalle) zwischen den Zahlen berücksichtigt (vgl. Tabelle 2, S. 26).

Praktische Hinweise

Viele naturwissenschaftlich relevante Variablen besitzen konkrete Maßeinheiten, denen eine Kodierung folgen sollte. Wenn Maßeinheiten beispielsweise von Land zu Land variieren,

ist das häufig ein Hinweis darauf, dass nur Intervallskalen vorliegen. Ein häufiger Fehler entsteht, wenn naturwissenschaftliche Präzision zum Anlass genommen wird, anzunehmen, dass die Zahlen der Skala die Berechnung von Zahlenverhältnissen erlauben würden. Eine prozentuale Steigerung oder Verringerung kann für Intervalldaten jedoch nicht sinnvoll interpretiert werden. Ein besonderes Problem in den Sozialwissenschaften stellt die Ratingskala dar. Sie könnte ordinal- oder intervallskaliert sein. Viele statistische Auswertungen lassen sich erst für Intervallskalen durchführen. Eine Statistik ist schnell beendet, wenn Ratingskalen als Ordinalskalen betrachtet werden. Das führt dazu, dass viele Auswertungen stillschweigend Ratingskalen als Intervallskalen betrachten. Darüber wird in den Wissenschaften immer wieder gestritten. Praktisch gesehen wird man in sehr vielen Fällen nicht umhin können, Ratingskalen als Intervallskalen zu behandeln. Siehe dazu auch unten. Häufig ist zudem die Interpretation von Messwerten problematisch, für die es keine genormten Maßeinheiten gibt. Liegt beispielsweise eine Norm für eine Temperatur vor, dann lässt sich leicht interpretieren, dass eine Lufttemperatur von 40 °C sehr heiß ist. Wenn in einem Fragebogen mit Ratingskalen aber 40 Punkte erreicht werden, kann man nicht ablesen, ob das viele oder wenige Punkte sind. Das bereitet auch dann Schwierigkeiten, wenn man die maximal mögliche Punktzahl als Vergleich heranzieht. Denn es kann an der Art der Fragestellung liegen, dass 40 Punkte erreicht werden. Für viele Skalen in den Sozialwissenschaften sind häufig keine Maßeinheiten bekannt. Daher können zwar Vergleiche zwischen Gruppen interpretiert werden – beispielsweise wenn die eine Gruppe 30 Punkte und die andere 40 Punkte erreicht –, aber ohne Vergleichsmöglichkeit ist die ermittelte Zahl in ihrer Höhe nicht interpretierbar. Das gilt auch, wenn um Benotungen nach dem Schulnotensystem gebeten wird. Denn den befragten Personen ist eine echte Normierung ja gar nicht bekannt. Für „echte“ Schulnoten würde man vorher das Lernziel festlegen. Eine Person, die das Lernziel gerade eben erreicht, bekommt eine 4 (ausreichend/genügend). Die Normierung „echter” Schulnoten erfolgt am vorher festgelegten Lernziel. Wenn man jedoch fragt, wie zufrieden Kund:innen mit einem Mittagessen sind, dann hat jede:r eigene Ansprüche daran, was er oder sie als gerade ausreichend empfindet.

Einige Statistik-Lehrbücher enden mit der Intervallskala die gerne auch als *metrische* Skala bezeichnet wird. Es gibt aber noch zwei höhere Skalen, die

ebenfalls zu den metrischen Skalen gezählt werden können (vgl. hierzu die Einteilung von Bortz et al., 2000, S. 62).

Verhältnisskala

Die nächsthöhere Skala nach der Intervallskala ist die *Verhältnisskala*: Zahlenverhältnisse können hier erstmals sinnvoll interpretiert werden. Ein gutes Beispiel für eine Verhältnisskala ist das Gehalt. Während man in Bezug auf das Alter nicht sinnvoll davon sprechen kann, dass eine Person doppelt so alt ist wie eine andere (da sich dies durch die Addition der weiterhin verstreichenden Zeit ständig ändert), ist ein doppelt so hohes Gehalt eine gut interpretierbare Relation. Messgrößen dieser Skala können aber durchaus noch sinnvoll verändert werden. Beispielsweise kann das Gehalt in andere Währungen umgerechnet werden. Durch die Umrechnung des Gehalts in unterschiedliche Währungen würde ein doppelt so hohes Gehalt weiterhin doppelt so hoch bleiben. Die Währung spielt bei der Interpretation der Zahlenverhältnisse keine Rolle. Zulässige Transformationen der Messwerte sind Multiplikation oder Division (tatsächlich werden bei der Währungsumrechnung keine anderen mathematischen Operationen verwendet). Addition oder Subtraktion hingegen zerstören die Zahlenverhältnisse und sind daher nicht zulässig. Messungen mit einem klar definierten Nullpunkt, der inhaltlich nicht sinnvoll verschoben werden kann, sind Verhältnisskalen (denn Addition oder Subtraktion wären Verschiebungen des Nullpunktes). Beim Gehalt ist ein Gehalt von Null Euro ein solcher Nullpunkt. Da Verhältnisskalen Verhältnisse abbilden können, ist es manchmal auch sinnvoll, Mittelwerte solcher Verhältnisse zu berechnen. Je nachdem, ob Multiplikationen oder Divisionen im Vordergrund stehen, ist dann entweder ein geometrisches oder ein harmonisches Mittel zu bilden (vgl. Tabelle 2, S. 26).

Praktische Hinweise

In vielen Statistiklehrbüchern werden Skalenniveaus, die über der Intervallskala liegen, unterschlagen. Das kann daran liegen, dass solche Skalen in dem Bereich, mit dem sich das Buch beschäftigt, selten vorkommen. Die Unterscheidung der Skalenniveaus wurde jedoch in die Statistik eingeführt, um das passende statistische Verfahren für die gegebenen Daten zu ermitteln. So kann beispielsweise ein Mittelwert erst ab dem Intervallskalenniveau sinnvoll interpretiert werden und ist bei Nominalskalen wahrscheinlich nicht sinnvoll. Tatsächlich kann es zu erheblichen Fehlern kommen, wenn man davon ausgeht, dass die Intervallskala das Nonplusultra ist und nur den sogenannten arithmetischen Mittelwert verwendet. Da Verhältnisskalen hingegen Zahlenverhältnisse ausdrücken können, kann es manchmal wichtig sein, ein

mittleres Zahlenverhältnis zu ermitteln. Dazu muss das geometrische Mittel verwendet werden, denn Zahlenverhältnisse sind Multiplikationen und der gesuchte Faktor ist der mittlere Multiplikationsfaktor. Das arithmetische Mittel würde fälschlicherweise die additiven Zuwächse berücksichtigen, nicht die Multiplikation. Der Kontostand am Monatsende könnte beispielsweise wie folgt aussehen:

Kontostand	Absolute Differenz zum Vormonat	% Wachstum	Multiplikationsfaktor
1500			
2000	500	33,33%	1,3333
1750	-250	-12,50%	0,8750
1500	-250	-14,29%	0,8571
2000	500	33,33%	1,3333
Arithmetisches Mittel			1,09970
Geometrisches Mittel			1,07457

Wenn man den richtigen durchschnittlichen Multiplikationsfaktor verwendet, müsste man, ausgehend vom ersten Kontostand, viermal mit diesem Faktor multiplizieren und käme dann auf 2.000 Euro.

Für das arithmetische Mittel käme aber heraus:

$$1500 * 1{,}09970 * 1{,}09970* 1{,}09970* 1{,}09970 = 2193{,}77$$

Das ist offensichtlich falsch.

Für das geometrische Mittel kommt heraus:

$$1500 * 1{,}07457 * 1{,}07457* 1{,}07457* 1{,}07457 = 2000{,}00$$

Das trifft zu. Das geometrische Mittel ist korrekt.

Wie wird es berechnet? Man multipliziert alle vier Multiplikationsfaktoren miteinander und zieht aus dem Ergebnis die vierte Wurzel. Die Wurzel die gezogen wird orientiert sich an der Zahl der Werte. Hier sind es vier Multiplikationsfaktoren.

$$GM = \sqrt[n]{\prod \prod x_i}$$

GM: Geometrisches Mittel, n: Zahl der Messwerte, x(i): Messwert der Nummer i, ∏: Produkt aller Messwerte.

Ebenfalls bei Verhältnisskalen von Relevanz ist die Mittelwertbildung über den Nenner von Brüchen. Eine Geschwindigkeit wird beispielsweise als Weg geteilt durch die Zeit berechnet. Wenn ein PKW die ersten 10 km in 5 Minuten, die zweiten 10 km in 7 Minuten und die dritten 10 km in 6 Minuten fährt, dann war er mit den folgenden Geschwindigkeiten unterwegs:

10 km / 5 Min *60 = 120 km/h.

10 km / 7 Min *60 = 85,71 km/h.

10 km / 6 Min *60 = 100 km/h.

Das arithmetische Mittel bestimmt die Durchschnittsgeschwindigkeit, indem es die drei Geschwindigkeiten addiert und diese Summe durch drei teilt: (120 km/h + 85,71 km/h + 100 km/h) / 3 = 101,90 km/h. Dies Ergebnis ist jedoch falsch, was sich leicht nachprüfen lässt. Die insgesamt 30 km wurden in 18 Minuten zurückgelegt, also: 30 km / 18 Min *60 = 100 km/h.

Das harmonische Mittel summiert den Kehrwert der Geschwindigkeiten also (1 / 120 km/h + 1 / 85,71 km/h + 1 / 100 km/h) = 0,03 h/km. Anschließend wird die Zahl der Messwerte (hier 3) durch diese Summe geteilt, also: 3 / 0,03 h/km = 100 km/h. Das harmonische Mittel ist der korrekte Wert.

$$HM = \frac{n}{\sum \frac{1}{x_i}}$$

HM: Harmonisches Mittel, n: Zahl der Messwerte, x(i): Messwert der Nummer i, Σ: Summe der Kehrwerte aller Messwerte.

Absolutskala

Das höchste Skalenniveau ist das der *Absolutskala*: Bei einer Absolutskala ist eine nachträgliche Umrechnung der Zahlen in andere Maßeinheiten nicht sinnvoll (Klein, 2004). Ihre Werte sind absolut. So ist z. B. die Anzahl der Personen in einem Raum eine Zahl, die genau diese Anzahl angibt und nicht mehr sinnvoll verändert werden kann. Zählungen sind immer Absolutskalen.

Praktische Hinweise Skalen mit einem geringen Skalenniveau können verbessert werden, indem bestimmte interessante Aspekte gezählt werden. Berufe sind beispielsweise eine sehr offene nominale Skala. Wenn man sich jedoch auf ganz bestimmte Berufe konzentriert und diese jeweils einzeln als zutreffend oder unzutreffend erfasst, erhält man mehrere absolute Skalen, also eine eigene Skala für jeden Beruf. Auch ordinale Skalen kann man in eine Zählung umwandeln, indem beispielsweise das Erreichen einer bestimmten Zahlenhöhe als zutreffend/unzutreffend gezählt wird. So könnte der höchste Bildungsabschluss über vielleicht fünf verschiedene Stufen abgefragt werden. Gezählt werden könnte dann, wie viele Personen mindestens die dritte Stufe erreicht haben. Die ordinale Variable enthält beispielsweise die Werte: Pflichtschule = 1, Matura/Abitur = 2, Bachelor-Studium = 3, Master = 4, Promotion = 5. Die absolute Skala enthielte dann den Wert „mindestens Bachelor“ und würde nur „zutreffend/unzutreffend“ erfassen.

2.2 Interpretationsprobleme und Konventionen

Je nach Skalenniveau können also unterschiedliche Eigenschaften der Zahlen interpretiert werden. Das Skalenniveau entscheidet somit über die Art der zulässigen statistischen Weiterverarbeitung. Die Skalenniveaus sind daher der Ausgangspunkt für alle weiteren statistischen Entscheidungen. Letztlich hängt alles davon ab, was man mit den erhobenen Zahlen sinnvoller Weise anfangen kann. Viele statistische Fehlinterpretationen resultieren aus der Anwendung von Verfahren, die für das jeweilige Skalenniveau ungeeignet sind. Beispielsweise berücksichtigt ein Mittelwert die Intervalle zwischen den Zahlen. Skalenniveaus unterhalb der Intervallskala führen daher zu Mittelwerten, die wahrscheinlich nicht interpretiert werden können. Es wird daher davon abgeraten, Kennwerte zu ermitteln, deren Interpretation Schwierigkeiten bereitet. Es ist jedoch nicht verboten, einen Mittelwert für numerische Codes zu berechnen. Der Computer stürzt auch nicht ab, wenn man dies tut. Es macht allerdings wenig Sinn, Kennwerte zu berechnen, die nachher nicht interpretiert werden können.

Die Beurteilung des Skalenniveaus kann anhand der bereits vorgestellten Kriterien in drei Schritten erfolgen:

1. In einem ersten Schritt kann entschieden werden, ob zumindest Intervallskalenniveau vorliegt. Denn in der Regel ist schnell er-

kennbar, ob es sich um eine Nominalskala (die Zahlen sind Bezeichnungen für Objekte, Merkmale, Kategorien) oder um eine Ordinalskala handelt. Liegt beides nicht vor, dann handelt es sich um eine der drei metrischen Skalen.

2. In einem zweiten Schritt stellt sich daher die Frage, ob ein Nullpunkt vorhanden ist und inhaltlich sinnvoll interpretiert werden kann. Beispielsweise ist der Nullpunkt der Grad-Celsius-Skala durch das Gefrieren von Wasser gegeben. Das ist zwar sinnvoll, aber man hätte ebenso gut einen ganz anderen Nullpunkt wählen können. Besonders problematisch ist hier der Umstand, dass es tatsächlich verschiedene Temperaturskalen mit unterschiedlichen Nullpunkten gibt. Wenn – wie bei der Temperaturskala – keine Einigkeit über einen Nullpunkt erzielt werden kann, dann liegt kein Verhältnisskalenniveau vor. Im Vergleich dazu ist ein Kontostand von genau Null Euro ein eindeutiger Nullpunkt, der nicht sinnvoll verschoben werden kann. Liegt ein eindeutig interpretierbarer Nullpunkt vor, so ist dies ein Hinweis darauf, dass zumindest ein Verhältnisskalenniveau vorliegt.

3. Ein dritter Schritt kann klären, ob es weitere sinnvolle Transformationen oder Umrechnungen in andere Maßeinheiten geben könnte. Ist dies nicht der Fall, liegt eine Absolutskala vor.

Mitunter erfordern diese drei Schritte Fachwissen über den jeweiligen Gegenstand. Probleme entstehen entweder dadurch, dass nicht klar ist, wie die „echte" empirische Welt tatsächlich strukturiert ist, oder dadurch, dass nicht nachvollzogen werden kann, wie die Messung als Abbildung in den Zahlenraum konkret funktioniert:

- **Unklarheiten auf Seiten des empirischen Relativs.** Ist es wirklich sinnvoll, von einem absoluten Nullpunkt der Temperatur auszugehen? Um das zu entscheiden, bedarf es einiger theoretischer Annahmen und auch empirischer Experimente. Nicht immer ist von vornherein klar, ob die „wahre Natur" einer Verhältnisskala oder einer anderen Skala ähnelt.
- **Unklarheiten bei der Abbildung in den Zahlenraum.** Auf der anderen Seite wird beim Messen mit Hilfe von Messvorschriften versucht, eine – vielleicht teilweise noch unbekannte – Realität in den Zahlenraum abzubilden. Die dabei verwendeten Abbildungsvorschriften und -regeln können ihrerseits die Güte der Skalierung einschränken. Es nützt nichts,

wenn die Temperatur in „Wirklichkeit“ einen echten Nullpunkt hat, wir aber – um gut interpretierbare Zahlen zu erhalten – einen Nullpunkt am Gefrierpunkt festlegen.

Es ist wichtig zu verstehen, dass das Skalenniveau erst durch den konkreten Messvorgang und die Überprüfung erlaubter Transformationen festgelegt wird. Es geht nicht darum, wie der Messgegenstand bzw. -inhalt in „Wahrheit“ existiert, sondern wie er nach der Messung interpretiert werden kann. Eine Messung kann gut erfolgen und der „wahren“ Natur sehr nahekommen oder eher schlecht erfolgen und sich stark von ihr entfernen. Auch Moden und Gewohnheiten spielen eine Rolle. Man kann Menschen schließlich nicht verbieten, eine Temperatur in das Maß umzurechnen, das sie gewohnt sind. Wenn durch eine erlaubte Transformation ein Skalenniveau zerstört wird, ist es wenig sinnvoll, dieses Skalenniveau im Rahmen einer Studie als gegeben anzunehmen, nur weil in der vorgelegten Studie diese Transformation nicht durchgeführt wird.

Beispiele	Art der Erhebung (Beispiele)	Niveau
Alter	Alter:_____ Geburtsjahr:_____ ❍ 1 10-15 Jahre ❍ 2 16-25 Jahre ❍ 3 26-35 Jahre ❍ 4 älter als 35 Jahre	Intervall Ordinal
Geschlecht	❍ 0 männlich ❍ 1 weiblich ❍ 2 divers	Nominal
Beruf	❍ 1 Arbeitslosigkeit ❍ 2 Arbeitsunfähigkeit ❍ 4 Unselbstständig erwerbstätig ❍ 5 Selbstständig erwerbstätig ❍ 6 Studium ❍ 7 Ausbildung ❍ 8 Hausfrau / Hausmann ❍ 9 Rente / Pension ❍ 10 Wehr- / Zivildienst ❍ –1 Sonstiges/Unbekannt	Nominal
Gerade Anzahl, Abstufungen	❍ ❍ ❍ ❍ gut schlecht	Intervall / Ordinal
Ungerade Anzahl, Abstufungen	❍ ❍ ❍ ❍ ❍ gut schlecht	Intervall / Ordinal
Keine Abstufungen (Visuelle-Analog-Skala VAS)	|——————————| gut schlecht	Intervall
Ungerade Anzahl, Abstufungen, Beschriftet	❍ ❍ ❍ ❍ ❍ sehr gut gut mittel schlecht sehr schlecht	Intervall / Ordinal

Abbildung 1: Beispiele für Skalenniveaus in Fragebögen

Die Zahlen neben den Ankreuzmöglichkeiten geben die Codierung an, mit der die Daten in die Rohdatentabelle eingetragen werden. Weitere Beispiele für Ratingskalen finden sich bei Bortz und Döring (2002, S. 176-177). Ratingskalen werden üblicherweise mit Mittelwerten zusammengefasst. In dem Fall werden sie als Intervallskalen aufgefasst. Möglicherweise handelt es sich aber um Ordinalskalen. Der Fehler den man begeht, wenn sie als Intervallskalen behandelt werden, scheint gering zu sein. In einigen Studien kann die Intervallskalenqualität von Ratingskalen sogar empirisch nachgewiesen werden (vgl. die Vorschläge für die Überprüfung von Ratingskalen in Westermann, 1985).

Das Skalenniveau hängt also von beiden Seiten ab: Auf der einen Seite steht das empirische Relativ, auf der anderen das numerische. Die Abbildungsvorschrift des einen Relativs in das andere bestimmt den Messprozess und letztlich ist das Skalenniveau das Endergebnis dieses gesamten Prozesses. Vergleiche hierzu auch die Beispiele in Abbildung 1.

Probleme bei der Abschätzung des Skalenniveaus treten z. B. dann auf, wenn der Einfluss der Messung selbst unterschätzt wird. So wird z. B. für physikalische Größen gerne ein sehr hohes Skalenniveau erwartet, einfach weil man annimmt, dass die Natur selbst exakt und wohlgeordnet ist und die Naturwissenschaften Messungen grundsätzlich auf einem hohen Skalenniveau durchführen. Dies mag auch der Fall sein, aber wenn sich die Menschen nicht auf ein geeignetes Messverfahren einigen können, wird die Messung kein hohes Skalenniveau erreichen.

Beispielsweise wären Temperaturangaben in Kelvin in einem Fachartikel über den Klimawandel wenig intuitiv, und es ist nicht ungewöhnlich, eine Temperaturskala zu verwenden, die in der Meteorologie der Region, über die berichtet wird, üblich ist. Diese Skalen haben einen anderen Nullpunkt als die Kelvin-Skala. Die Personen, die den Artikel verfassen, könnten sich dafür entscheiden, Vergleichswerte für Wien im Januar als Beispiel anzugeben und dafür Grad Celsius zu verwenden. Das bedeutet, dass die Temperatur auf der Intervallskala und nicht auf der Verhältnisskala angegeben wird. Eine prozentuale Temperaturänderung ist die Änderung eines Verhältnisses und setzt daher eine Verhältnisskala voraus. Wer prozentuale Temperaturänderungen in Grad Celsius angibt, bekommt Interpretationsprobleme, wenn andere Personen dies in Fahrenheit umrechnen und damit zwangsläufig zu ganz anderen prozentualen Temperaturänderungen kommen.

Gelegentlich wird in Fachartikeln versucht, deutlich zu machen, dass in dieser Arbeit beispielsweise alles nur in Kelvin gerechnet wird und daher Verhältnisskalenniveau vorliegt und daher auch prozentuale Änderungen angegeben werden können. Ja, man kann versuchen, einen solchen Artikel zu veröffentlichen, aber man muss trotzdem damit rechnen, dass z. B. in einem Medienbericht über die Ergebnisse die üblichen Umrechnungen in Fahrenheit oder Celsius zu finden sind. Verwirrung ist vorprogrammiert. Obwohl die Temperatur eine physikalische Größe ist und es in der Physik auch eine perfekte Verhältnisskala gibt, wird die Temperatur im alltägli-

chen Gebrauch anders verwendet. Das muss also bei Publikationen berücksichtigt werden.

Aufwerten

Es gibt auch Beispiele, in denen den Beteiligten bewusst ist, dass das verwendete Maß z. B. nur eine Ordinalskala ist, sie aber trotzdem den Mittelwert berechnen. Sie werten ihre etwas unsaubere Ordinalskala auf und verwenden sie als Intervallskala. Das ist nicht verboten, kann aber zu Fehlinterpretationen führen. Die Frage ist, ob man mit diesem Fehler leben kann. Eine Note in Englisch ist sicher keine Intervallskala. Dazu müsste der Unterschied zwischen den Noten 1 und 2 genau so groß sein wie der zwischen 2 und 3 und zwischen 3 und 4 ... Vieles spricht dagegen. Trotzdem werden die Noten gemittelt. Bei der Mittelwertbildung wird stillschweigend davon ausgegangen, dass die Abstände zwischen den Noten gleich sind. Das ist zwar unsauber und wird in einigen Statistik-Lehrbüchern als grundfalsch angeprangert, kann aber bei der Lösung praktischer Probleme helfen. Es ist möglich, dass die Durchschnittsnote in Englisch in der Schule ein guter Prädiktor für die Studienleistung im Fach Anglistik ist. In diesem Fall ist es praktikabel, so vorzugehen, und wenn es funktioniert, könnte es sogar sein, dass die Noten nicht so dramatisch von der Intervallskala abweichen wie erwartet.

Ratingskalen

Eine ähnliche Diskussion wird seit Jahrzehnten über die Bewertung von Ratingskalen (manchmal fälschlicherweise als Likert-Skalen bezeichnet) geführt (vgl. Bortz & Döring, 2002, S. 180 f.). Die einen sind der Meinung, dass es sich bei Ratingskalen um Ordinalskalen handelt, die anderen gehen davon aus, dass es Intervallskalen sind. Das Problem besteht darin, dass man nicht einmal sagen kann, wie Menschen eine Empfindung mittels einer Ratingskala ausdrücken. Haben Menschen eine Intervallskala „im Kopf“ oder sind sie nur dazu in der Lage etwas ordinal zu bewerten? Es kann durchaus sein, dass Menschen bei der Beurteilung bestimmter Sachverhalte in der innerpsychischen Urteilsbildung einer Skala mit einem sehr hohen Skalenniveau folgen. Aber auch das Gegenteil ist möglich. Hinzu kommt, dass das Messinstrument selbst Verzerrungen aufweisen kann. So ist z. B. eine Ratingskala an den Polen begrenzt und dies entspricht möglicherweise nicht dem Eindruck der Urteilenden, die davon ausgehen, dass es noch viel extremere Werte geben könnte.

Die Diskussion um die Grenzen von Ratingskalen wurde zum Teil recht heftig geführt und hat Untersuchungen angeregt, die für bestimmte Frage-

stellungen anhand von Ratingskalen unterschiedlicher Breite recht eindeutig gezeigt haben, dass es sich hier um Intervallskalen handelt – es gibt aber auch Studien, die dies nicht zeigen konnten (vgl. die Vorschläge für die Überprüfung von Ratingskalen in Westermann, 1985). Heute gelten Ratingskalen als das zentrale Forschungsinstrument in den Sozialwissenschaften, und es ist durchaus üblich, mehrere Ratings durch Mittelwerte zusammenzufassen (Bortz & Döring, 2002, S. 180). Damit wird implizit automatisch von einem Intervallskalenniveau ausgegangen. Auch hier stellt sich die Frage, wie groß der Fehler ist, wenn kein Intervallskalenniveau vorliegen würde. In der Regel wird heute davon ausgegangen, dass dieser Fehler vernachlässigbar ist. Dies liegt aber auch daran, dass für ordinale Daten keine tiefgehenden statistischen Verfahren zur Verfügung stehen. Man ist dann häufig auf Methoden angewiesen, die Intervallskalenniveau voraussetzen und geht in Ermangelung anderer Methoden eben davon aus, dass es nicht so dramatisch ist, wenn kein perfektes Intervallskalenniveau vorliegt.

Abwerten

Umgekehrt werden Durchschnittsgehälter häufig mit Hilfe des Medians angegeben. Der Median ist das Gehalt, das die Stichprobe in genau zwei gleich große Hälften teilt. Die eine Hälfte hat ein höheres, die andere ein niedrigeres Gehalt. Wie viel höher oder niedriger ein Gehalt ist, wird bei diesem einfachen statistischen Verfahren nicht berücksichtigt. Der Median berücksichtigt die Abstände zwischen den Zahlen nicht. Gehälter haben aber durchaus interpretierbare Abstände. Tatsächlich wären sogar darüber hinaus die Zahlenverhältnisse interpretierbar.

Man könnte also Mittelwerte bestimmen, die diese Informationen nutzen und ein konkretes detailliertes Bild zeichnen. Wenn man stattdessen auf den Median zurückgreift, behandelt man die qualitativ hochwertigen Daten so, als läge nur eine Ordinalskala vor. Das ist statistisch nicht gerechtfertigt. Warum tut man es dann? Löhne sind nach oben offen. Es gibt Leute, die extrem hohe Gehälter haben. Aber nach unten sind die Gehälter begrenzt. Die seltenen, aber sehr hohen Gehälter führen zu einem Mittelwert, der ebenfalls sehr hoch ist. Viele Menschen wären schockiert, wenn sie den tatsächlichen Mittelwert in der Zeitung lesen würden. Der Median interessiert sich nicht dafür, wie viel diejenigen haben, die über der Mitte liegen, und wie wenig diejenigen haben, die darunter liegen. Der Median teilt die Stichprobe in der Mitte, und das ist eine viel „schönere“ Zahl, und deshalb wird diese Zahl gerne genommen.

2.3 Passende deskriptive Statistiken

Zusammenfassend kann gesagt werden, dass das Skalenniveau für die weitere statistische Verarbeitung der Daten von großer Bedeutung ist. Werden statistische Verfahren gewählt, die nicht dem Skalenniveau entsprechen, werden entweder Informationen vernachlässigt, die eigentlich vorhanden wären, oder es werden Informationen künstlich hinzu-„erfunden", die nicht wirklich erwartet werden können. Beides kann in Ausnahmefällen sinnvoll sein. Wichtig ist, dass man sich bewusst ist, was man tut. Dies zeigt, dass Statistik viel mehr als von vielen erwartet eine Wissenschaft ist, die nicht wahr oder falsch als Leitdifferenz hat, sondern sich an der Bewertung nützlich vs. nicht nützlich orientiert.

Skalenniveau	**zentrale Tendenz**	**Abweichungsmaß**	**Anmerkung**
Nominal	**Modalwert** (häufigster Wert)	**Prozent**, seltenster Wert	Werden z. B. in der Medizin auch gerne als kategoriale Daten bezeichnet.
Ordinal	**Median** (Modalwert)	**(Inter)-Quartilsabstand** (Prozent, seltenster Wert)	
Intervall	**Mittelwert** (Median) (Modalwert)	**Standardabweichung, Varianz (Stichprobe)** (Inter)-Quartilsabstand (Prozent, seltenster Wert)	Median und Inter-(Quartilsabstand) sind mitunter anschaulicher, weil Ausreißer wenig ins Gewicht fallen.
Verhältnis	Wie Intervall, aber zudem auch geometrisches Mittel (z. B. Zinsen) oder harmonisches Mittel (z .B. Geschwindigkeiten)	wie Intervall	

Tabelle 2: Skalenniveaus und deskriptive Statistik

Inwieweit Skalenniveau und statistische Methode zwingend übereinstimmen müssen, ist in der Statistik umstritten (vgl. z. B. Klein, 2004, Thomas, 2019). Abweichungen sind durchaus möglich und teilweise auch üblich. Die Tabelle versteht sich daher als Liste von Empfehlungen, die methodisch und mathematisch gut begründet sind. Wird davon abgewichen, muss dies ebenfalls gut begründet werden.

Viele Regeln der Statistik sind zunächst und vor allem Empfehlungen. Je besser diese in ihrer Begründung verstanden werden, desto besser kann man mit diesen Empfehlungen umgehen. Wer zum ersten Mal selbst eine Statistik berechnet, wird sich eher an die Empfehlungen halten. Etwas verwirrend erscheint vielen Neulingen aber dann, dass in guten Fachzeitschriften recht häufig von diesen Empfehlungen abgewichen wird, ohne dass diese Abweichung groß erklärt wird. Norman (2010) verweist hier auf die hohe Robustheit statistischen Handelns:

> One of the beauties of statistical methods is that, although they often involve heroic assumptions about the data, it seems to matter very little even when those are violated. (Norman, 2010, S. 627)

Verbote, so heißt es in neueren Lehrbüchern, gibt es in der Statistik nur dort, wo es um die Grenzen der Mathematik geht. So ist eine Division durch null unzulässig (Wenn bei der Berechnung einer Korrelation eine Varianz von null auftritt, wird durch null dividiert und das Computerprogramm stürzt ab oder verweigert hier die weitere Analyse.). Von diesen harten und klaren Verboten der Mathematik abgesehen, versucht eine gute Statistik, die getroffenen Entscheidungen zu begründen und inhaltlich zu verstehen. Dies hilft bei der Interpretation. Dementsprechend sollte man vorsichtig sein, Verfahren zu verwenden, die schwer zu interpretieren sind.

In neueren Statistik-Lehrbüchern wird nicht mehr davon ausgegangen, dass es „Verbote“ für die Berechnung von Mittelwerten gibt, sondern dass man wissen muss, was man tut, wenn man von den Empfehlungen abweicht.

2.4 Stetigkeit

Neben dem Skalenniveau spielt auch die Stetigkeit eine wichtige Rolle. Diskrete Messgrößen haben nur vorgegebene Stufen, aber keine Zwischenwerte zwischen den Stufen. Ein Würfel mit sechs Seiten zeigt immer eine der sechs Zahlen, aber nie Zwischenwerte. Ein Würfel liefert also diskrete Zahlen. Eine Besonderheit stellen diskrete Messgrößen mit nur zwei Stufen dar. Dies ist z. B. der Fall, wenn die Antwort ja oder nein sein kann, oder aber wenn keine Zwischenstufe vorgesehen ist obwohl diese inhaltlich denkbar wäre. Wenn es nur zwei diskrete Möglichkeiten gibt, werden diese Messungen auch als dichotome (griechisch) oder binäre (lateinisch) Mes-

sungen bezeichnet. Kontinuierliche Messungen können dagegen Zwischenwerte haben. Sie lassen immer mehr Zwischenwerte zu, je genauer die Messung durchgeführt wird.

2.5 Übungsaufgaben

[Dieses Arbeitsblatt wird Ihnen auch als Word-Dokument zur Verfügung gestellt. Link im Syllabus.]

Name: _____

Bitte versuchen Sie, die Liste der „Phänomene“ zu bewerten. Welche Skalenniveaus nehmen Sie an? Wenn Sie sich nicht sicher sind, kann es hilfreich sein, nach entsprechenden Informationen zu suchen, z. B. im Internet. Sie können Ihre Antworten gerne begründen oder darauf hinweisen, dass unter bestimmten Bedingungen, die Sie dann nennen sollten, auch andere Lösungen möglich wären.

Phänomen	**Skalenniveau**	**Diskret oder stetig (nach der Messung)**	**Kodierung: Wie werden Zahlen vergeben?**
1. Geschlecht	*Nominal*	*Diskret*	*0: männlich* *1: weiblich* *2: divers*
2. Höhe eines Berges			
3. Reiseziele (Urlaub)			
4. Güteklassen von Gemüse			
5. Inflationsraten			

6. Bruttojahresgehalt			
7. Anzahl Verkehrstote (März)			
8. Berufe			
9. Höchster Bildungsabschluss			
10. Berufsjahre im Job			
11. IQ			
12. Anzahl belegter Betten pro Monat			
13. Systolischer Blutdruck			
14. Schulnote in Englisch			
15. Schmerzskala (VAS)			
16. Lottozahlen			

17. Welchen Sinn kann es machen, für Lottozahlen den Mittelwert zu berechnen?

3 Messung und Testtheorie

Wie im vorangegangenen Abschnitt erläutert, ist eine Messung eine strukturerhaltende Abbildung eines empirischen Relativs in die Zahlenwelt. Das Ergebnis einer solchen Messung wird als Skala bezeichnet. Das Skalenniveau gibt an, welche Eigenschaften der Zahlen interpretiert werden dürfen und welche sich der Interpretation entziehen. Da die Skala das Ergebnis des Messprozesses ist, hängt viel davon ab, wie die Messung tatsächlich durchgeführt wird. In der Aufgabe aus 2.5 befindet sich ganz rechts eine Spalte, die fragt, wie die Zahlen vergeben werden. Binet et al. (2003/1905) haben den Intelligenzquotienten beispielsweise als das Verhältnis des sogenannten Intelligenzalters zum Lebensalter, multipliziert mit 100, definiert: Wenn ein Kleinkind etwas kann, was Kinder üblicherweise mit drei Jahren können, dann hat es ein Intelligenzalter von drei Jahren. Ist das Kind tatsächlich jünger, beispielsweise zwei Jahre, dann ist es seinem Alter voraus. Das Zahlenverhältnis 3/2 = 1,5 wird mit 100 multipliziert, wodurch sich ein IQ von 150 ergibt. Da der so konstruierte IQ über ein Zahlenverhältnis definiert ist, liegt ein Verhältnisskalenniveau vor.

Die Messung kann jedoch auch auf eine andere Weise vorgenommen werden. So könnten beispielsweise 100 Aufgaben vorgelegt und gezählt werden, wie viele davon gelöst werden. Als Vergleichsmaßstab dient eine repräsentative Stichprobe, die z. B. zeigt, dass in der Bevölkerung im Durchschnitt 40 Aufgaben gelöst werden. Anschließend wird geprüft, ob sich eine Person von diesem Durchschnitt unterscheidet und wie groß der Unterschied ist. Der Bevölkerungsdurchschnitt, der zur Normierung verwendet wird, wird in regelmäßigen Abständen neu erhoben, um die Normierung an aktuelle Entwicklungen in der Bevölkerung anzupassen. In diesem Fall werden der Durchschnitt und damit auch der mögliche Nullpunkt verschoben. Es liegt daher kein Verhältnisskalenniveau, sondern ein Intervallskalenniveau vor. (Auch diese Skala lässt sich auf einen Durchschnitt von 100 verschieben. Die Zahl der von einer Person i gelösten Aufgaben sei x_i. Davon wird der Bevölkerungsdurchschnitt $\bar{x}$ abgezogen und 100 addiert. Die Verschiebung der Skala auf Zahlenwerte, die um den Durchschnitt von 100 schwanken können, zeigt erneut das Intervallskalenniveau. Verschiebungen durch Additionen oder Subtraktionen sind bei Intervallskalen erlaubte Transformationen.)

Das Skalenniveau wird also durch den gewählten Messalgorithmus bestimmt. Dabei sind in der Regel verschiedene Algorithmen denkbar. Beispielsweise kann der IQ, wie gesehen, einmal als Quotient und ein anderes Mal als Abstand zum Bevölkerungsdurchschnitt berechnet werden.

3.1 Operationale Definition hypothetischer Konstrukte

In den Sozialwissenschaften und der Psychologie stellt sich vor jeder Messung die Frage nach der Operationalisierung: Wie und mit welchen Methoden kann man etwas messen, das im Verborgenen liegt, das man nicht anfassen und greifen kann und von dem man nicht einmal sicher weiß, ob es überhaupt existiert?

Während man in der physikalischen Welt einige Eigenschaften von Objekten direkt mit den Sinnen erfassen kann, zeigen sich andere Eigenschaften erst, wenn man sie durch sogenannte Operationen, also konkrete Messvorschriften, hervorlockt. So sind Größe und Gewicht eines Gegenstandes beispielsweise bereits mit den Sinnen wahrnehmbar und lassen sich dann ohne große Umstände mit geeigneten Maßstäben feststellen. Andere Eigenschaften, wie beispielsweise die Wasserlöslichkeit eines Stoffes, sind dagegen erst prüfbar, wenn man den Stoff im Wasser löst. Die Eigenschaft der Wasserlöslichkeit zeigt sich also erst, wenn eine sogenannte Operationalisierungsvorschrift befolgt wird. Eine solche Operationalisierungsvorschrift gibt an, was man tun muss, um die betreffende Eigenschaft hervorzulocken. In der Regel ist eine Operationalisierungsvorschrift eine Art Rezept oder Algorithmus, der genau befolgt werden sollte. In Bezug auf die Wasserlöslichkeit könnten die Menge des Wassers und die Menge des Stoffes, die Temperatur des Wassers, die Dauer des Verrührens des Stoffes im Wasser, die Prüfung der Trübung des Wassers durch den Stoff und möglicherweise noch weitere Rahmenbedingungen bedeutsam sein und müssten dann auch berücksichtigt werden. Der konkrete Algorithmus stellt insgesamt eine *Definition* dar, die exakt festlegt, wann ein Stoff als wasserlöslich gilt und wann nicht.

Wir unterscheiden also zwischen direkt gegebenen bzw. leicht zugänglichen Eigenschaften bzw. Merkmalen und solchen Eigenschaften bzw. Merkmalen, die erst durch eine Operationalisierungsvorschrift hervorgebracht werden müssen. Im Folgenden geht es darum, diese Unterscheidungen etwas genauer zu betrachten. Das ist notwendig, da viele Aspekte, die

in den Naturwissenschaften als selbstverständlich angesehen werden, in der Psychologie oder den Sozialwissenschaften nicht einmal geprüft werden können.

Zunächst wurde hier zwischen Eigenschaften bzw. Merkmalen unterschieden, die bereits vorhanden sind und möglicherweise direkt mit den Sinnen erfasst werden können, und solchen, die erst durch eine konkret definierte Operationalisierungsvorschrift hervorgelockt werden müssen. Im zweiten Fall handelt es sich also um Eigenschaften, die nicht immer sichtbar sind und daher erst sichtbar gemacht werden müssen. Diese Unterscheidung ist nicht immer leicht und eindeutig. Denn selbst die Vermessung direkt erfahrbarer Größen wie der Länge eines Körpers mit einem Maßstab ist eine Tätigkeit, die man ungeschickt oder geschickt ausführen kann. Auch hierfür sollte ein Algorithmus vorgegeben werden, bei dem der Vergleichsmaßstab, beispielsweise die Messung in Metern, festgelegt und definiert wird.

Die Unterscheidung zwischen Merkmalen, die direkt erkennbar sind, und solchen, die erst durch eine umfangreiche Operationalisierungsvorschrift hervorgelockt werden müssen, ist insbesondere im Hinblick auf Konstrukte der Sozialwissenschaften und der Psychologie von Bedeutung. Dies wird verständlich, wenn man sich vor Augen führt, dass in der Psychologie und nahestehenden Forschungsrichtungen in der Regel Eigenschaften oder Merkmale gemessen werden sollen, die nicht nur erst hervorgelockt werden müssen, sondern darüber hinaus rein hypothetischer Natur sind und möglicherweise gar nicht existieren. Demgegenüber kann die Wasserlöslichkeit eines Stoffes eindeutig empirisch geprüft werden. In der Psychologie liegt die Sache ganz anders: Gibt es Intelligenz, emotionale Intelligenz, Resilienz oder Persönlichkeitseigenschaften wie emotionale Stabilität überhaupt? Es handelt sich um Konstrukte, die angenommen werden, um bestimmte Phänomene zu erklären. Diese Konstrukte sind Annahmen, also Hypothesen – es könnten auch ganz andere Erklärungen zutreffen.

Strunk (2024) fasst die Literatur zum Gegenstand der Psychologie zusammen und definiert die Psychologie als die Wissenschaft, die sich mit dem inneren Erleben beschäftigt. Dieses innere Erleben kann bei anderen Menschen niemals direkt beobachtet werden. Messvorschriften in der Psychologie und verwandten Forschungsrichtungen bringen eine Eigenschaft nie tatsächlich zum Vorschein, sondern liefern nur Hinweise darauf, dass im Inneren eines Menschen möglicherweise etwas vorliegt, das die Wissenschaft z. B. als Intelligenz bezeichnen würde. Ob dies tatsächlich der Fall

ist, bleibt unklar. Bei der Wasserlöslichkeit ist das anders. Sie wird durch die Operationalisierungsvorschrift konkret geprüft und tatsächlich hervorgelockt und sichtbar gemacht.

Zusammenfassend lässt sich folgende Unterscheidung treffen:

a) Direkt erfahrbare Merkmale oder Eigenschaften der empirischen Welt.

b) Merkmale, die durch Operationalisierungsvorschriften hervorgebracht werden und dann direkt erfahrbar sowie empirisch prüfbar sind.

c) Hypothetische Konstrukte, die sinnvolle Annahmen über verborgene Merkmale und Eigenschaften darstellen und mit einer Operationalisierungsvorschrift indirekt erschlossen werden sollen. Diese werden durch die Operationalisierung nicht direkt erfahrbar, sondern bleiben weiterhin verborgen. In neueren statistischen Ansätzen, die sich mit sogenannten linearen Strukturgleichungsmodellen beschäftigen, werden hypothetische Konstrukte auch als „latente Variablen" bezeichnet.

Aus dem Konzept der hypothetischen Konstrukte lässt sich eine bedeutsame Folgerung ableiten. Oben wurde eine Operationalisierungsvorschrift als Definition für ein Konstrukt eingeführt. Es ist sinnvoll, sich den Begriff der „Definition" genauer anzusehen, um zu prüfen, ob sich auch hypothetische Konstrukte operational definieren lassen. Einerseits sind Definitionen wichtige Bestandteile wissenschaftlicher Theorien, andererseits sind sie für die Durchführung von Messungen als Operationalisierungsvorschrift von Bedeutung. Ganz allgemein betrachtet sind Definitionen Identitäten, die die beiden Seiten eines Gleichheitszeichens gleichsetzen. Das Definiendum (der definierte Begriff) ist demnach identisch mit dem Definiens (der definierenden Erklärung, Beschreibung oder Kennzeichnung). Durch die Definition wird eine Art Abkürzung angeboten. Das, was definiert werden soll, wird in der Regel ausführlich, klar und präzise beschrieben und dann benannt. Die Benennung ist häufig ein einzelnes Wort, beispielsweise „Intelligenz", während die Definition selbst ein längerer Text ist. Nach der Definition sollte klar sein, was mit dem kurzen Wort gemeint ist. In der Folge kann das Wort daher verwendet werden, ohne den langen Text erneut anführen zu müssen. Wichtig ist dabei, dass das Wort und der lange Text exakt dasselbe bezeichnen. Wenn dies der Fall ist, dann sind sie

identisch. Dies ist auch der zentrale Anspruch, den die Wissenschaften an eine Definition haben. Der definierte Begriff bzw. das definierte Konstrukt ist identisch mit der ausführlichen Darstellung.

Ein Beispiel: Strunk (2025) fasst in einem Artikel über künstliche Intelligenz die Definition von Intelligenz in der Psychologie wie folgt zusammen:

> An dieser Stelle ist es sinnvoll, den Begriff der Intelligenz näher zu betrachten. Die Psychologie versteht darunter die Fähigkeit, potenzielle Hindernisse, die das Erreichen von Zielen beeinträchtigen könnten, eigenständig zu identifizieren und zu überwinden. Sie erfordert Wissen, schlussfolgerndes Denken, die Reflexion der eigenen Stellung in der Welt, die Fähigkeit zu abstrahieren und aus Erfahrungen zu lernen. Dabei geht es nicht um Wissen im Sinne von Bücherwissen, sondern um ein umfassenderes Verständnis und die Fähigkeit, mit den Veränderungen in der Umwelt Schritt zu halten, daraus einen Sinn abzuleiten und herauszufinden, was zu tun ist (Nisbett et al., 2012).

Kriterium der Eliminierbarkeit und Nicht-Kreativität

Man kann darüber streiten, ob damit präzise gesagt wird, was Intelligenz ist. Deutlich wird jedoch, dass der Begriff „Intelligenz" (sog. Definiendum) und seine längere Begriffsbestimmung (sog. Definiens) gleichgesetzt werden. In der Folge kann der kürzere Begriff verwendet werden, ohne dass er erneut erklärt werden muss. Durch die Gleichsetzung ist eine Definition jederzeit wieder eliminierbar und kann rückgängig gemacht werden. Es muss jederzeit möglich sein, das Definiendum im gesamten Text durch das Definiens (und umgekehrt) zu ersetzen. Der Sinn des Textes darf sich dadurch nicht ändern. Denn eine Definition ist nichts anderes als eine Ersetzungsvorschrift. Das bedeutet auch die Forderung nach Nicht-Kreativität. Es darf nicht erst durch eine Definition eine Wahrheit erzeugt oder bewiesen werden, die ohne sie nicht beweisbar wäre (Suppes, 1957).

Auf die Gefahr hin, vom Thema abzuschweifen, soll an dieser Stelle noch einmal betont werden, dass Definitionen nichts anderes als Ersetzungsvorschriften sind. Das Ziel darf nicht darin bestehen, sich etwas auszudenken und es durch eine Definition zum Leben zu erwecken. Viele populärwissenschaftlich verwendete psychologische Konstrukte wirken jedoch so, als wären sie aus der Luft gegriffen: neue, bislang unbekannte Persönlichkeitseigenschaften, bislang unbekannte psychische Erkrankungen usw. werden „erfunden" und erhalten einen möglichst kreativen Namen. Die Tatsache, dass etwas definiert werden kann, bedeutet eben noch nicht, dass es existiert. Denn eine Definition ist „nur" eine sprachliche Ersetzungsvorschrift.

Im Rahmen einer wissenschaftlichen Arbeit ist es möglich, eine beliebige Ersetzungsvorschrift zu wählen, sofern sie im Kontext der Arbeit sinnvoll erscheint. Eine auf die konkrete Arbeit beschränkte Definition wird als Realdefinition bezeichnet. Dadurch gibt es unzählige Definitionen, beispielsweise für Intelligenz. Wer jetzt fragt, welche Definition denn die richtige sei, hat vielleicht noch nicht ganz verstanden, was Definitionen sind. Jede Definition, die ein eindeutiges Gleichheitszeichen zwischen Begriff und Erklärung des Begriffs setzt, ist korrekt. Eine Definition hat keinen direkten Bezug zu „Wahrheiten" in der empirischen Welt. Sie ist ein Sprachspiel, aber kein Instrument der Wahrheitsfindung. Das ist auch der Grund dafür, dass Intelligenz, Resilienz usw. so uneinheitlich definiert sind. Es ist möglich, erlaubt und niemals falsch, eine beliebige Definition zu nutzen. Es kann jedoch verwirrend sein, wenn Definitionen gewählt werden, die inhaltlich stark von den üblichen Definitionen abweichen. Das ist jedoch weder verboten noch falsch. Versuche, einen Begriff ein für alle Mal festzulegen, heißen *Nominaldefinitionen*. Ob ein solches abschließendes letztes Wort für eine sich beständig verändernde Wissenschaft möglich und sinnvoll ist, kann bezweifelt werden. Wenn psychische Erkrankungen in der Klinischen Psychologie oder der Psychiatrie präzise und mit dem Anspruch einer Norm definiert werden, dann erfüllt das neben der wissenschaftlichen Aufgabe auch einen normierenden Zweck. Dieser soll die Kommunikation mit Betroffenen, deren Familien, Krankenkassen etc. vereinheitlichen. Dies sind politische, ökonomische und kommunikative Funktionen, die zudem der Behandlungsplanung dienen sollen. Diese Funktionen normierender Nominaldefinitionen sollten nicht mit den wissenschaftlichen Zielen von Definitionen im Allgemeinen verwechselt werden.

Zurück zum Thema. In Bezug auf die Messung von Eigenschaften oder Merkmalen wurden oben drei Konstellationen unterschieden. Sowohl direkt gegebene Merkmale als auch nach Durchführung einer Operationalisierung hervorgebrachte Eigenschaften sind echte, das heißt eindeutige, jederzeit wieder eliminierbare und nicht-kreative Definitionen. Kurz: Es sind Definitionen. Hypothetische Konstrukte wie die Intelligenz verletzen die Kriterien für eine Definition immer und in jedem Fall. Da sie niemals – auch durch eine Operationalisierung nicht – direkt erfahrbar sind, können sie niemals mit Sicherheit erfasst werden. Mehr noch: Das hypothetische Konstrukt ist immer größer als das, was die Operationalisierung zu zeigen vermag. Zwischen einer operationalen Vorschrift und einem hypothetischen Konstrukt kann niemals ein Gleichheitszeichen stehen. Es handelt sich also

nicht um Definitionen im eigentlichen Sinne. Das folgende Beispiel soll dies verdeutlichen.

Man könnte beispielsweise sagen, dass sich Intelligenz darin zeigt, dass jemand Kopfrechenaufgaben gut lösen kann. In diesem Fall wären Kopfrechenaufgaben eine Operationalisierung für Intelligenz. Die Leistung im Kopfrechnen zeigt zunächst einmal, ob jemand Kopfrechenaufgaben gut lösen kann. Nur indirekt kann dies zudem als Hinweis auf Intelligenz gewertet werden, etwa indem man annimmt, dass Intelligenz die Leistung im Kopfrechnen beeinflusst.

Wir könnten auch andere Operationalisierungen versuchen (z. B. Satzergänzungen, Schlussfolgerndes Denken etc.). Im Idealfall würden wir jeweils Hinweise darauf erhalten, dass Intelligenz vorliegt. Denn wir verstehen Intelligenz als hypothetische Erklärung für bestimmte Leistungen, von denen wir annehmen, dass sie durch Intelligenz zustande kommen. Intelligenz selbst ist jedoch niemals direkt beobachtbar. Sichtbar werden lediglich Hinweise, die so interpretiert werden, als ob Intelligenz vorläge.

Deutlicher wird dieses Problem, wenn man sich die Vielzahl verschiedener Kopfrechenaufgaben vor Augen führt. Kopfrechnen ist sicherlich nicht der einzige Faktor, der auf Intelligenz hindeutet, aber es sind bereits unendlich viele verschiedene Kopfrechenaufgaben denkbar. Was ist mit Satzergänzungstests, bei denen Intelligenz unterstellt wird, wenn jemand einen Satz korrekt vervollständigen kann? Auch hierfür gibt es unendlich viele mögliche Aufgaben. Und wie sieht es mit Aufgaben zum logischen Schlussfolgern aus? Es gibt unendlich viele Möglichkeiten, Intelligenz mit Aufgaben zu operationalisieren. Aber keine dieser Aufgaben ist die einzig richtige und keine macht Intelligenz selbst sichtbar oder greifbar. Es werden immer nur Hinweise geliefert, die darauf hindeuten könnten, dass Intelligenz vorliegt.

Zwischen dem Begriff der Intelligenz und seiner Operationalisierung durch Aufgaben besteht kein Gleichheitszeichen. Die Zahl der möglichen Aufgaben ist unendlich groß und immer viel größer als die Zahl der Aufgaben, die in einem Intelligenztest tatsächlich gestellt werden können. Aus praktischen Gründen muss jede Operationalisierung beschränkt bleiben und kann das hypothetische Konstrukt niemals vollständig erfassen. Das ist schlicht und einfach nicht möglich.

Schlussfolgerung Hypothetische Konstrukte sind stets größer und umfassender als jede denkbare Operationalisierung. Zwischen dem Begriff und der gewählten Operationalisierung kann niemals ein Gleichheitszeichen stehen. Operationalisierungen von hypothetischen Konstrukten sind somit keine Definitionen. Sie sollten daher auch nicht als solche bezeichnet werden.

Praktische Konsequenzen Praktisch gesehen bedeutet das für die Messung hypothetischer Konstrukte, dass möglichst viele Aufgaben verwendet werden sollten. Da Intelligenz viele Aspekte umfasst, sollte das Messinstrument zumindest versuchen, möglichst viele davon zu berücksichtigen. Konkret bedeutet das, dass hypothetische Konstrukte niemals mit nur einer Frage gemessen werden sollten.

Eine weitere Folgerung betrifft die Frage, ob es Aufgaben, Items oder Operationalisierungen gibt, die besser geeignet sind als andere, um ein hypothetisches Konstrukt zu erfassen. Denn wenn es unendlich viele Möglichkeiten gibt, könnte es auch welche geben, die weniger gut geeignet sind. Wenn Kopfrechenaufgaben beispielsweise so leicht sind, dass jedes Kind sie lösen kann, kann man sich diese sehr leichten Aufgaben sparen und stattdessen Aufgaben verwenden, die besser zwischen Personen, die gut Kopfrechnen können, und Personen, die das nicht so gut können, unterscheiden.

Wie im folgenden Abschnitt noch gezeigt wird, ist es nie verkehrt, Konstrukte mit mehreren Items zu messen. Das hat auch messtheoretische Gründe. Allerdings soll damit nicht gemeint sein, dass direkt gegebene Konstrukte, wie das Alter, ebenfalls mit mehreren Items abgefragt werden sollten. Erst in den letzten Jahren formiert sich eine Diskussion, die die Vorteile mehrfacher Messungen anerkennt, aber auch von diesem Vorgehen abweichende Single-Item-Messungen befürwortet, sofern dies begründet werden kann:

> Rossiter (2002) proposes that if the object can be conceptualized as concrete and singular, it does not require multiple items to represent it in the measure, and if the attribute can be conceptualized as concrete, it does not require multiple items either. However, […] it is virtually impossible to get a journal article accepted […] unless it includes multiple-item measures of the main constructs. (Bergkvist & Rossiter, 2007, S. 175)

3.2 Klassische Testtheorie

In der Psychologie ist man sich schon früh bewusst gewesen, dass man es mit hypothetischen Konstrukten zu tun hat und es daher unendlich viele Möglichkeiten gibt, diese zu messen. Da die Messung immer indirekt ist und ein einzelnes Item niemals ausreichen kann, stellt sich die Frage, wie aus der unendlichen Menge möglicher Items diejenigen ausgewählt werden, die besonders gut geeignet sind.

Gesucht sind also mehrere, in der Regel ähnlich gestaltete Operationalisierungen, die geeignet sind, das hypothetische Konstrukt möglichst umfassend zu messen. Das hypothetische Konstrukt wird also mittels einer Skala gemessen, die sich aus mehreren Operationalisierungsmöglichkeiten zusammensetzt, die üblicherweise als Items bezeichnet werden. Viele Items ergeben dabei insgesamt die Skala, die das Endergebnis der Messung repräsentiert. In der Praxis werden die Items in der Regel durch Mittelwerte zusammengefasst oder es wird die Summe der gelösten Aufgaben gebildet (was dem gleichen Prinzip entspricht, nur dass nicht durch die Zahl der gestellten Aufgaben dividiert wird).

Die ersten Ideen, wie man viele Items zu einer Skala zusammenfasst und dabei geeignete von ungeeigneten unterscheidet, stammen von Likert (1932) und Thurstone (1927). Ihre Vorschläge waren bahnbrechend. Likert schlägt für die Messung von Einstellungen Items mit fünfstufigen Ratings vor. Tatsächlich haben seine Arbeiten die Verwendung von Ratingskalen stark befördert, weshalb diese heute noch immer gerne als Likert-Skalen bezeichnet werden. Dies ist jedoch nicht ganz korrekt, denn eine „echte“ Likert-Skala ist laut Likert die Zusammenfassung mehrerer besonders gut geeigneter Items (also mehrere Ratings) zu einer Skala. Eine Skala ist also etwas anderes als ein einzelnes Rating.

Die Suche nach den besten Items für eine Skala wird als Skalenkonstruktion oder Itemselektion bezeichnet. Likerts Vorschläge zur Identifikation geeigneter Items für die Skalenkonstruktion sind inzwischen veraltet, sodass Likert-Skalen heute nicht mehr in der von Likert vorgeschlagenen Form verwendet werden. In der Regel erfolgt die Itemselektion und Skalenkonstruktion heute entweder auf der Grundlage der sogenannten *klassischen Testtheorie* oder mithilfe der *probabilistischen Testtheorie* (auch als Item-Response-Theorie bezeichnet).

Die Bezeichnung „Testtheorie" führt einen weiteren Begriff in die Diskussion ein. Ein Test ist in der Psychologie ein umfassend geprüftes Messinstrument, mit dem ein hypothetisches Konstrukt der Psychologie operationalisiert wird. Diese Bezeichnung findet sich auch in Begriffen wie „Persönlichkeitstest" oder „Intelligenztest". Die zentralen Annahmen und Schlussfolgerungen der klassischen Testtheorie gelten jedoch nicht nur für die Psychologie. Sie können auch auf Messungen in der Physik oder auf anderen Naturwissenschaften übertragen werden. Vereinfachend kann ein Test aus einer oder mehreren Skalen bestehen. Die Skalen sind Zusammenfassungen von Items. Die Idee der Zusammenfassung der Items zu einer Skala ist die zentrale Folgerung aus der klassischen Testtheorie.

Die klassische Testtheorie geht davon aus, dass jede Messung einen Messwert produziert, der nicht unbedingt mit dem tatsächlichen Wert übereinstimmen muss. Vereinfacht ausgedrückt können Messungen also auch danebenliegen. Es wäre zwar durchaus wünschenswert und auch das Ziel der Messung den wahren Wert durch die Messung festzustellen, aber Messung und wahrer Wert stimmen nicht immer überein. Messfehler sind durchaus zu erwarten. Gleichzeitig sieht man einem Messwert nicht an, ob er einen Fehler enthält und wie groß dieser ist.

Die klassische Testtheorie gilt nicht nur für hypothetische Konstrukte, sondern ganz allgemein für jede Art der Messung. Wenn etwa die Länge eines Tisches gemessen wird und 1 Meter und 8 Zentimeter herauskommen, wäre es schön, wenn der Tisch tatsächlich 1,08 Meter lang wäre. Auf der einen Seite haben wir die wahre Länge des Tisches, auf der anderen Seite den Messwert. Idealerweise stimmt der gemessene Wert mit dem wahren Wert überein. Wenn jedoch ein Messfehler auftritt, gehen der wahrer Wert und der Messwerte aufgrund des Messfehlers auseinander. Die klassische Testtheorie geht davon aus, dass der wahre Wert (mit dem griechischen Buchstaben μ bezeichnet) und der Messfehler (mit e für *error* bezeichnet) zusammen den Messwert (mit x bezeichnet) ergeben:

$$x = \mu + e.$$

Das Ziel ist μ zu ermitteln, aber eine Messung liefert x. Wie kann man also μ bestimmen, wenn man nur x kennt? Wenn ich das in Seminaren frage, stellen einige Studierende die Gleichung um und sagen, ich müsse nur den Fehler von x abziehen, um μ zu erhalten. Aber e ist ja auch nicht bekannt. Wir kennen immer nur x und haben keine Informationen über μ oder e.

Was kann man tun?

Wer schon einmal etwas ausgemessen hat, beispielsweise, um eine neue Tischplatte zu bestellen, hat vielleicht bereits intuitiv etwas getan, das den Weg weist. Man misst mehrfach. Unterscheiden sich die Messungen – vorausgesetzt, der Tisch ist zwischen den Messungen weder gewachsen noch geschrumpft (und welcher Tisch würde das schon tun?) –, zeigen die Unterschiede den Fehler an. Wenn eine Messung fehlerfrei ist, sollte bei einer Wiederholung das gleiche Ergebnis herauskommen. Jeder Unterschied zwischen den Messungen zeigt also, dass ein Messfehler vorliegt. Mehrfache Messungen helfen also, den Messfehler einzuschätzen.

Diese Zusammenhänge sollen im Folgenden etwas vertieft werden. Zunächst soll davon ausgegangen werden, dass die Messung auf die bestmögliche Weise erfolgt, d. h., es werden keine systematischen Fehler begangen, indem das Messinstrument falsch angewendet wird oder Zahlen systematisch falsch abgelesen werden. Aber selbst bei einem korrekten Vorgehen ist keine Messung völlig fehlerfrei. Wenn keine systematischen Fehler vorliegen, dann sind die immer noch vorhandenen und nicht vermeidbaren Fehler unsystematisch und wahrscheinlich zufällig. Sie schwanken mal in die eine und mal in die andere Richtung; die Messwerte sind mal zu hoch und mal zu gering. Das bietet einen Ansatz, den Fehler in den Griff zu bekommen und μ einzugrenzen. Führt man viele Messungen durch und mittelt diese, so mittelt sich der Fehler zu null heraus und der Mittelwert entspricht μ. Die folgenden Gleichungen zeigen dieses Verhalten recht deutlich.

Durch die wiederholten Messungen werden mehrere Messwerte, also mehrere x-Werte, ermittelt. Die verschiedenen Messungen werden mit einem Index i durchnummeriert. Das i steht tiefgestellt beim x und zeigt an, welchen Messwert das x repräsentiert. Der Index i zählt von 1 aufwärts bis zur letzten Messung, die mit n bezeichnet wird. Es werden also n Messungen durchgeführt. Um den Mittelwert über alle Messungen zu berechnen, werden alle x_i zusammengezählt und dann durch n geteilt. In der Statistik zeigt das Summenzeichen Sigma (Σ), dass die Größen, die rechts neben dem Sigma stehen, zusammengezählt werden. Im Folgenden steht rechts neben dem Sigma x_i, was bedeutet, dass alle x_i zusammengezählt werden.

$$\overline{x} = \frac{1}{n}\sum x_i$$

Vor dem Summenzeichen steht $1/n$, was bedeutet, dass die Summe mit $1/n$ multipliziert wird. Das ist dasselbe wie eine Division durch n. Das ist dann der Mittelwert, der in der Gleichung als $\overline{x}$ (gelesen: x-quer) bezeichnet wird. Da zudem bekannt ist, dass jedes $x_i = \mu + e_i$ ist, lässt sich für x_i auch $\mu + e_i$ schreiben:

$$\overline{x} = \frac{1}{n}\sum(\mu + e_i)$$

Wenn immer das Gleiche gemessen wird und sich während der Messungen nichts verändert, ist μ bei allen Messungen gleich. Es hat daher auch keinen Index i, aber der Fehler e kann sich von Messung zu Messung ändern. Er kann höher oder niedriger sein, positiv oder negativ. Der Fehler e ist daher mit dem Index i versehen. Im Folgenden wird die Klammer aufgelöst. Dabei darf nicht vergessen werden, dass beide Größen in der Klammer zusammengezählt und jeweils mit $1/n$ multipliziert werden müssen:

$$\overline{x} = \frac{1}{n}\sum\mu + \frac{1}{n}\sum e_i$$

Doch was ist der Mittelwert von μ? Die Antwort lautet μ. Wenn man beispielsweise 15 Messungen durchgeführt hat, dann werden 15-mal das gleiche μ zusammengezählt. Anschließend wird durch 15 geteilt. 15 kürzt sich heraus und es bleibt nur das μ stehen:

$$\overline{x} = \mu + \frac{1}{n}\sum e_i$$

Der Mittelwert aller durchgeführten Messungen ist somit μ plus den Mittelwert des Fehlers. Ist der Fehler unsystematisch und zufällig, geht der Mittelwert des Fehlers gegen null, da sich die positiven und negativen Fehler dann ausgleichen:

Wenn

$$\frac{1}{n}\sum e_i \sim 0$$

dann folgt daraus

$$\overline{x} \sim \mu.$$

Der Mittelwert vieler Messungen minimiert einen unsystematischen, zufälligen additiven Fehler auf null, sodass der Mittelwert dieser vielen Messungen ungefähr dem Wert von μ entspricht.

Anstelle des Gleichheitszeichens habe ich das „ungefähr-gleich"-Zeichen (~) verwendet. Denn es ist nicht sicher, ob der Mittelwert der Fehler tatsächlich null ergibt und somit der Mittelwert der Messwerte dem wahren Wert entspricht. Es gilt jedoch, dass mehr Messungen die Wahrscheinlichkeit für eine Fehlerminimierung erhöhen. Daher fordert auch die klassische Testtheorie dazu auf, viele Messungen durchzuführen und zu mitteln.

Diese Überlegungen lassen sich noch erweitern. Die Messwerte x_i zeigen unterschiedliche Werte, die mal nach oben und mal nach unten vom Mittelwert für x_i abweichen. Diese Schwankungen oder Abweichungen vom Mittelwert werden in der Statistik Varianz oder Standardabweichung genannt.

In Seminaren frage ich gerne, was die Standardabweichung eigentlich ist, und eine häufige Antwort lautet: „Die Standardabweichung ist die mittlere Abweichung vom Mittelwert." Die Studierenden vermuten, dass man zunächst den Mittelwert berechnen und müsse dann man danach alle Messwerte einzeln vom Mittelwert abziehen könne. Das wäre dann die Abweichung vom Mittelwert.

Zählt man dann aber alle Abweichungen vom Mittelwert zusammen, so erhält man immer null. Das ist vielen Studierenden nicht bewusst und stellt ein Hindernis bei der Berechnung der Standardabweichung dar.

$$\frac{1}{n}\sum(\overline{x} - x_i) = 0$$

Wie der Name schon sagt, liegt der Mittelwert exakt in der Mitte. Einige Messwerte weichen nach oben, andere nach unten ab. Die Summe der Abweichungen nach oben und die Summe der Abweichungen nach unten sind immer gleich groß. Immer. Der Mittelwert ist so definiert, dass dies erreicht wird.

Um die Schwankungen um den Mittelwert zu beziffern, müssen die Vorzeichen der positiven und negativen Abweichungen eliminiert werden. Das funktioniert, indem man die Differenz quadriert. Das Quadrat ist im-

mer eine positive Zahl. Wird diese Summe der quadrierten Abweichungen durch n geteilt, erhält man die Varianz, die häufig als s^2 bezeichnet wird. Die hochgestellte 2 zeigt das Quadrat an, welches der Berechnung zugrunde liegt. Zieht man aus der Varianz die Wurzel, verschwindet das Quadrat und man erhält die Standardabweichung s, die also die Wurzel aus der mittleren quadrierten Abweichung vom Mittelwert ist. Das ist nichts anderes als die mittlere Abweichung vom Mittelwert, die aber aus den gezeigten Gründen etwas aufwändiger berechnet werden muss:

$$s_x^2 = \frac{1}{n}\sum(\overline{x} - x_i)^2 \qquad \text{[Varianz]}$$

$$s_x = \sqrt{\frac{1}{n}\sum(\overline{x} - x_i)^2} = \sqrt{s_x^2} \qquad \text{[Standardabweichung]}$$

Das kleine, tiefgestellte x beim s zeigt, dass es sich um die Standardabweichung bzw. Varianz der x-Werte handelt. Man könnte auch die Varianz der Fehler berechnen, das wäre s_e^2. Und wenn verschiedene μ vorlägen, etwa weil verschiedene Tische vermessen würden, dann könnte man auch die Varianz der μ mit s_μ^2 angeben. Da die Summe aus dem wahren Wert und dem Fehler den Messwert ergibt, gilt diese Beziehung auch für die Varianzen. Es gilt also immer:

$$s_x^2 = s_\mu^2 + s_e^2$$

Wir sind nun fast am Ziel unserer Reise angelangt. Welchen maximalen Wert kann der folgende Bruch erreichen und was ist dafür notwendig?

$$r = \frac{s_\mu^2}{s_x^2}$$

Richtig, mehr als eins kann nicht herauskommen. Und eins kommt nur heraus, wenn es keine Fehlervarianz gibt. Das ist der Fall, wenn die Fehlerwerte nicht nur im Durchschnitt null sind, sondern bei jeder einzelnen Messung null betragen. Die Messung wäre perfekt und vollkommen fehlerfrei. Die Varianzen von μ und x würden dann übereinstimmen und der Bruch wäre 1. Der Bruch wird als Definition für die *Reliabilität* bezeichnet. Die Reliabilität ist das Verhältnis der Varianz der wahren Werte zur Varianz der

gemessenen Werte. Eine Reliabilität von $r = 1$ bedeutet eine perfekte Messung ohne jeden Messfehler. In der Praxis gilt ein Wert von 0,7 als akzeptable Reliabilität für Gruppenstudien und ein Wert von 0,8 als geeignet für die Messung bei Einzelpersonen. (Damit ist gemeint, dass in einer Gruppenstudie erneut mehrere Messungen durchgeführt und die Ergebnisse gemittelt werden, wodurch eine Verbesserung der Messung erreicht wird. Wird hingegen nur einmal gemessen, beispielsweise der IQ einer Person, dann muss die Reliabilität höher sein.)

Wer den Ausführungen gefolgt ist, hat vielleicht bereits eine Unstimmigkeit bemerkt: Durch die Messung erhalten wir immer nur x-Werte, aber niemals μ oder den Fehler. Wie soll man dann feststellen, ob die Fehlervarianz null ist und die Varianz der wahren Werte mit der der gemessenen Werte übereinstimmt? Tatsächlich sind die Ausführungen bis zu dieser Stelle Theorie. Sie sind – hoffentlich – gut nachvollziehbar. Auf der Grundlage dieser Theorie kann man jedoch nicht sofort losrechnen, denn tatsächlich sind nur die Messwerte bekannt.

In der Praxis lassen sich verschiedene Methoden nutzen, um die Reliabilität dennoch zu schätzen. So wurde oben bereits gezeigt, dass bei Messfehlerfreiheit wiederholte Messungen perfekt übereinstimmen müssen. Möchte man also die Reliabilität (also das Ausmaß der Messfehlerfreiheit) prüfen, könnte man eine Messung an vielen Personen durchführen und die Messung an denselben Personen noch einmal wiederholen. So könnte beispielsweise der IQ von 100 Personen zwei Mal im Abstand von einem Monat mit einem IQ-Test bestimmt werden. Da sich der IQ innerhalb eines Monats wahrscheinlich nicht verändert, sollten die Messungen übereinstimmen. In der Praxis werden die Messungen der beiden Zeitpunkte miteinander korreliert. Die Korrelation kann Werte von maximal 1 erreichen, wenn die beiden Messungen exakt übereinstimmen. Die Korrelation zwischen den beiden Messungen wird als Retest-Reliabilität bezeichnet, wobei „Retest“ die erneute Testung bezeichnet. Problematisch bei der erneuten Testung ist jedoch, dass sich die getesteten Personen erinnern und aufgrund dieser Erinnerung ähnliche Antworten geben. Dadurch wird die Reliabilität künstlich erhöht, also überschätzt. Daher sollte etwas Zeit zwischen den Messungen vergehen. Allerdings darf die Zeit auch nicht zu lang sein, da es sonst zu Veränderungen kommen könnte, die die Korrelation beeinträchtigen. Die Reliabilität würde dann unterschätzt werden.

Daher ist ein anderer Ansatz interessant. Wie oben dargestellt, sollten hypothetische Konstrukte mit vielen Items gemessen werden. Das liegt daran, dass das hypothetische Konstrukt tatsächlich über unendlich viele Items erfasst werden könnte und ein einziges Item daher niemals ausreichen kann. Wenn man mehrere Items verwendet, versucht jedes einzelne Item, das Konstrukt zu messen. Die Messungen dieser Items sollten also alle das gleiche Ergebnis liefern. Man könnte die Items durchnummerieren und die Items mit geraden Nummern (2, 4, 6, …) zu einem Messwert zusammenfassen und die Items mit ungeraden Nummern (1, 3, 5, …) zu einem anderen Messwert zusammenfassen. Beide Messwerte sollten korrelieren. Diese Form der Reliabilitätsberechnung wird Odd-Even-Reliabilität genannt. Allerdings sind Messungen, die jeweils nur die Hälfte der Items benutzen durch diese Reduktion der Items ohnehin schlechter.

Noch eleganter ist daher die von Cronbach vorgeschlagene Methode. Er nennt seine Methode „Alpha“. Cronbachs Alpha korreliert – vereinfacht gesprochen – jedes Item mit jedem anderen Item und liefert eine Art Gesamtkorrelation. Diese Form der Korrelation wird auch „interne Konsistenz“ genannt, da sie nur die im Messinstrument vorhandenen Items verwendet und nicht auf andere Testdurchführungen (Retest) setzt. Zudem zeigt der Begriff der Konsistenz, dass diese Methode prüft, ob die Items zueinander passen und konsistent das gleiche Konstrukt erfassen. Denn Cronbachs Alpha kann nur dann hoch sein, wenn die Messungen der Items übereinstimmen, was wiederum nur der Fall sein kann, wenn sie das gleiche Konstrukt erfassen.

Der Punkt ist wichtig: Wenn wir ein Messinstrument verwenden, das Items enthält, die verschiedene Konstrukte erfassen, und die interne Konsistenz dafür berechnen, dann wird diese gering ausfallen. Das Messinstrument ist dann nicht reliabel. Beispielsweise könnten fünf Aufgaben gestellt werden, die vorgeben, die Kopfrechenfähigkeit zu testen. Vier davon tun dies auch, aber eine der Aufgaben fragt nach der Schuhgröße. Da die Schuhgröße nicht mit der Kopfrechenfähigkeit korreliert, wird die interne Konsistenz dieser fünf Aufgaben gering ausfallen.

Wahrscheinlich wird niemand derartige unpassende Aufgaben in einen IQ-Test aufnehmen, aber es kann durchaus Aufgaben geben, die besser geeignet sind, und andere, die irgendwie nicht passen. Bei der sogenannten Skalenkonstruktion wird ein großer Pool von Items verwendet und dann geprüft, wie sich Cronbachs Alpha ändert, wenn das erste Item gar nicht für

die Berechnung verwendet wird. Anschließend wird geprüft, was geschieht, wenn das zweite Item nicht verwendet wird, und so weiter. Cronbachs Alpha wird besser, wenn ein nicht konsistentes Item gelöscht wird. So findet man die Items, die nicht so gut zur Skala passen, und kann sie aussortieren.

Wenn man unpassende Items Schritt für Schritt aussortiert, stellt man irgendwann fest, dass sich Cronbachs Alpha nicht mehr durch das Löschen von Items verbessern lässt. Werden weitere Items gelöscht, nimmt die Reliabilität deutlich ab. Wir haben bereits gesehen, dass der Mittelwert der Messwerte die beste Schätzung für den wahren Wert darstellt. Je mehr Messungen durchgeführt werden, desto höher wird die Reliabilität. Es gibt also eine Mindestanzahl an Items, die nötig ist, um eine hohe Reliabilität überhaupt erreichen zu können. Diese Mindestanzahl hängt jedoch davon ab, wie gut die Items geeignet sind. Sind alle Items ungeeignet, helfen auch hunderte davon nicht. Sind die Items ausreichend geeignet, können 10 bis 15 Items genügen, um insgesamt eine reliable Skala mit $r > 0{,}7$ zu erreichen.

Zusammenfassend kann festgehalten werden, dass es keine Messung ohne Messfehler gibt. Dieser sollte im Idealfall gering sein. Die Reliabilität ist ein Maß für die Messfehlerfreiheit und weist Zahlenwerte zwischen 0 und 1 auf, wobei 1 ein völliges Fehlen des Messfehlers anzeigt. 0,7 gilt als Mindestanforderung für Gruppenstudien. Da sich die Messung verbessert, wenn man sie wiederholt durchführt, werden insbesondere für hypothetische Konstrukte mehrere Items verwendet. Dabei zeigt sich, dass eine Mindestanzahl an Items notwendig ist, um eine zuverlässige Skala zu erhalten. Es ist möglich, die Eignung einzelner Items zu prüfen und so eine Itemauswahl vorzunehmen.

3.3 Schlussfolgerungen aus der klassischen Testtheorie

Die Grundannahmen der klassischen Testtheorie werden als Axiome bezeichnet. Es handelt sich dabei um Annahmen, die notwendig sind, damit sich die Schlussfolgerungen aus der Theorie ergeben können. Diese Annahmen können im Rahmen der Theorie selbst nicht geprüft werden, sondern werden vorausgesetzt. Plausibel ist die Annahme, dass jede Messung einen Fehler enthalten kann und der Messwert selbst nicht verrät, wie hoch der

wahre Wert und der Fehler sind. Dieses Dilemma lässt sich recht elegant durch Mittelwertbildung lösen. Dies ist jedoch nur dann eine passende Lösung, wenn der Fehler unabhängig vom wahren Wert auftritt und daher als additiver Summand behandelt werden kann. Wenn der Fehler jedoch z. B. mit dem wahren Wert wächst oder sonstwie mit dessen Höhe zusammenhängt, bricht die Herleitung zusammen. Zudem setzt das Konzept das Intervallskalenniveau voraus. Dieses ist möglicherweise bei Verwendung von Ratings als Items verletzt. In Ermangelung anderer, besser geeigneter Ansätze wird hier stillschweigend immer von Intervallskalen ausgegangen. Dies scheint auch dadurch gerechtfertigt zu sein, dass die klassische Testtheorie selbst Methoden zur Prüfung der Qualität einer Skala anbietet. Sollte diese Skala aus gänzlich ungeeigneten Items zusammengesetzt sein, würde sie – so kann man jedenfalls vermuten – schlecht funktionieren und die Probleme würden z. B. durch eine geringes Cronbach Alpha auffallen.

Obwohl die klassische Testtheorie für jede Form der Messung Gültigkeit beansprucht, ist sie insbesondere in den Sozialwissenschaften von Bedeutung, wenn es um die Messung hypothetischer Konstrukte geht. Auch die Messung der Länge eines Tisches wird genauer, wenn man sie wiederholt durchführt und Mittelwerte heranzieht. Galileo Galilei hat Bewegungs- und Pendelgesetze mit sehr unzulänglichen Zeitmessern (tropfende Eimer oder Pulsschlagzählung) untersucht. Erst durch viele Wiederholungen konnten die Zusammenhänge insgesamt erfasst werden. In der Sozialforschung wird man jedoch in der Regel darauf verzichten, einfache Tatsachen mehrfach in einem Fragebogen abzufragen, um den Messfehler zu begrenzen. Wenn man mehrfach nach dem Alter oder dem Geschlecht fragt, können die Interviewten das nur schwer nachvollziehen, obwohl schnelle Antworten auf die Altersfrage immer wieder auch zu unpräzisen Antworten führen. („Welches Jahr haben wir denn?“ „Im Kopf bin ich schon ein Jahr weiter,“ …).

Bei einfachen Tatsachenfragen sind Abfragen mit nur einem Item üblich und gelten statistisch auch als ausreichend reliabel. Dennoch stellt man bei Studien, die die gleichen persönlichen Daten ein zweites Mal erheben, nicht selten Unterschiede fest, die die Retest-Reliabilität als nicht so hoch erscheinen lassen, wie gemeinhin angenommen wird.

Die klassische Testtheorie ist vor allem in Bezug auf hypothetische Konstrukte relevant. Aus inhaltlicher und definitorischer Sicht wurde oben bereits darauf hingewiesen, dass ein Item nicht ausreicht, um ein Konstrukt

zu erfassen, das sich mit unendlich vielen unterschiedlichen Items erfassen ließe. Auch wenn es schwer ist einzuschätzen, wie viele Items eigentlich genügen können, ist klar, dass eines allein nicht ausreicht. Aus Sicht der klassischen Testtheorie wird dieses Argument erweitert. Denn während man bei der Abfrage des Alters gemeinhin davon ausgeht, dass dieses bekannt ist und zuverlässig mitgeteilt werden kann, ist bei weniger klaren Konstrukten nicht bekannt, wie zuverlässig ein Item ist. Sobald mehrere Items verwendet werden, kann Cronbachs Alpha berechnet werden, wodurch eine Schätzung der Reliabilität ermöglicht wird. Das ist ein entscheidender Vorteil. In guten wissenschaftlichen Zeitschriften wird deshalb verlangt, dass Skalen für hypothetische Konstrukte mehrere Items aufweisen und dass standardmäßig Cronbachs Alpha berechnet und angegeben wird. Da eine Skala mit einem Alpha kleiner als 0,7 als unbrauchbar gilt, können Studien, die diese Mindestanforderung nicht erfüllen, nicht publiziert werden.

Die klassische Testtheorie geht davon aus, dass es für das zu messende Konstrukt einen wahren Wert gibt. Dies ist bereits eine interessante Annahme, denn sie setzt voraus, dass das hypothetische Konstrukt existiert und real ist. Wenn in der Freudschen Tiefenpsychologie vom Ich, Es und Über-Ich die Rede ist oder in der neueren Zeit von Resilienz, dann sind das jeweils hypothetische Konstrukte. Es handelt sich also um theoretische Begriffe, die als Erklärung für ein Phänomen herangezogen werden, ohne dass direkt empirisch geprüft werden könnte, ob diese Begriffe etwas benennen, das tatsächlich existiert.

Dabei ist es wichtig, zwischen Fakten und Phänomenen auf der einen Seite sowie der Erklärung dieser Phänomene durch hypothetische Konstrukte auf der anderen Seite zu unterscheiden. Wenn in einem IQ-Test Aufgaben gelöst werden, dann ist die gelöste Aufgabe ein Faktum. Ob die Lösung jedoch durch Intelligenz befördert wurde, ist eine theoretische Annahme, die allein durch die Fakten nicht geprüft werden kann. Lange Zeit wurde der Psychoanalyse vorgeworfen, dass ihre Vorhersagen wenig zwingend seien. Es wurde von Menschen berichtet, die ähnliche Traumata erlebt hatten wie die von Freud beschriebenen Fallbeispiele, aber psychisch nicht erkrankten oder eine ganz andere Störung ausbildeten. Später stellte man fest, dass es Menschen gibt, die in ihrer Kindheit objektiv betrachtet schwersten Bedrohungen an Leib und Seele ausgesetzt waren, sich aber zu ausgeglichenen, glücklichen und gesunden Menschen entwickelten. Diese Tatsachenberichte sind Fakten. Wie lassen sie sich erklären? Möglicher-

weise waren hier glückliche Fügungen hilfreich oder bereits aus der Literatur bekannte Persönlichkeitseigenschaften, Intelligenz etc. Die Überzeugung, dass es sich dabei um eine bislang unbekannte Fähigkeit oder Eigenschaft der Person handelt, die heute als Resilienz bezeichnet wird, ist nur eine von sehr vielen möglichen Erklärungen. Sie benennt ein hypothetisches Konstrukt. Gemessen werden kann jedoch nur das, was es „wirklich" gibt und wofür die Existenz eines „wahren" Wertes sinnvoll angenommen werden kann. Die klassische Testtheorie liefert jedoch keine Beweise für die Existenz solcher Werte. Im Gegenteil: Sie setzt voraus, dass wir als Forschende wissen, was wir tun, und dass wir gute Argumente dafür haben, dass unsere Behauptungen zur Existenz hypothetischer Konstrukte nicht aus der Luft gegriffen sind.

Nun könnte man annehmen, dass, wenn eine zuverlässige Messung gelingt, auch das operationalisierte, hypothetische Konstrukt existiert. Das ist allerdings ein Fehlschluss. Die Reliabilität betrifft allein die Messfehlerfreiheit der Messung. Es ist jedoch möglich, dass eine reliable Skala gar nicht das angezielte hypothetische Konstrukt misst, sondern etwas anderes (z. B. messen einige psychologische Fragebögen die soziale Erwünschtheit oder die Fähigkeit, Intelligenz und Empathie zu erraten, welche Antwort nach außen hin gut aussieht). Die Frage, ob eine Skala tatsächlich das misst, was sie zu messen vorgibt, wird nicht durch die Reliabilität, sondern durch die sogenannte Validität bestimmt.

Die mehrfache Messung desselben Konstrukts ist die zentrale Idee der klassischen Testtheorie. In der Praxis bedeutet dies für die Skalenkonstruktion, dass für ein hypothetisches Konstrukt mehrere Items formuliert werden, die versuchen, das Konstrukt in seiner gesamten Breite zu erfassen. Wenn man beispielsweise die Zufriedenheit mit einem Kinobesuch erfassen möchte, würde man sich fragen, woran man diese Zufriedenheit erkennen könnte. Man könnte die Person auch direkt fragen, ob sie mit dem Kinobesuch zufrieden ist (Bist du zufrieden mit dem Kinobesuch?), was jedoch voraussetzt, dass die befragte Person den Begriff „Zufriedenheit" kennt und beurteilen kann. Wenn man beispielsweise jemanden fragt: „Wie sehr warst du heute im Einklang mit deinem Über-Ich-Ideal?", kann dies nur eine Person beantworten, die mit der Psychoanalyse sehr vertraut ist. Treffsicherer ist es, wenn einfache, möglicherweise verhaltensnahe Aspekte abgefragt werden, die Indikatoren dafür sind, dass das hypothetische Konstrukt vorlag. In Bezug auf den Kinobesuch könnten dies die folgenden Aspekte sein: „Ich habe den Film bereits weiterempfohlen", „Ich hatte

nachher gute Laune", „Ich gehe morgen noch einmal in den Film", „Ich habe nachher sehr gut geschlafen" oder „Am nächsten Morgen habe ich noch über den Film geschmunzelt" …

Es ist also durchaus üblich, nicht direkt das hypothetische Konstrukt zu erfragen („Sind Sie resilient?"), sondern es im Rahmen des Theorieteils einer wissenschaftlichen Arbeit durch geeignete Indikatoren abzubilden und somit zu operationalisieren. Dabei ist es sinnvoll, theoretisch schlüssig zu argumentieren, warum „gute Laune" nach dem Film beispielsweise ein Zeichen für Zufriedenheit mit dem Film sein könnte. Ebenso könnte man begründen, warum Kognitionen, Emotionen und Verhalten in Bezug auf den Film sowie unspezifische körperliche Empfindungen wie Unruhe, Bauchschmerzen, guter Schlaf oder Schmetterlingsgefühle berücksichtigt werden sollen. Ein inhaltlich valides Messinstrument (die Validität wird später noch behandelt, siehe unten) überlässt es nicht dem Zufall, was andere unter Zufriedenheit mit einem Kinofilm verstehen. Vielmehr werden die Indikatoren, die aus theoretischer Sicht für das Vorliegen des hypothetischen Konstrukts sprechen, mit Theorie und Literatur begründet.

Die Anwendung der klassischen Testtheorie wird problematisch, wenn eine Frage in einem Fragebogen mehrfach gestellt wird. Denn wenn es sich um dieselbe Frage handelt, wird sie in der Regel auch gleich beantwortet. Die Reliabilität ist dann zwar hoch, jedoch nicht, weil ein hypothetisches Konstrukt durch einen ganzen Strauß von Indikatoren umfassend überprüft worden wäre, sondern weil exakt das Gleiche – vielleicht mit unterschiedlicher Wortstellung oder mal als Verneinung und mal als Zustimmung formuliert – mehrfach gefragt wurde. Die Befragten empfinden ein solches Vorgehen als ähnlich irritierend wie eine dreifache Frage nach dem Alter. Besser ist es also, breit und umfassend nach Indikatoren zu suchen, diese theoretisch zu begründen und im Rahmen der Skalenkonstruktion zu erproben.

Cronbachs Alpha beziffert die Reliabilität einer Skala und gilt als Goldstandard für die Messung der internen Konsistenz eines Fragebogens. Neuere Verfahren, wie sie beispielsweise in Strukturgleichungsmodellen genutzt werden, um die sogenannte Faktorreliabilität einer Skala zu bestimmen, sind statistisch nicht besser, genauer oder dem Alpha von Cronbach irgendwie überlegen.

Cronbachs Alpha lässt sich bei gegebener Itemqualität verbessern, wenn mehr Items gleicher Qualität herangezogen werden. Dies folgt aus der be-

reits bekannten Verbesserung der Messqualität bei Messwiederholung und anschließender Mittelwertbildung. In der Literatur ist immer wieder zu lesen, dass diese Form der Reliabilität ein Hinweis auf die Homogenität einer Skala darstellt. Das ist jedoch nicht der Fall. In den vorhergehenden Absätzen wurde ausgeführt, dass mitunter versucht wird, durch möglichst identische Fragen das Alpha zu erhöhen. Eine Skala aus fast gleichlautenden Items wird als homogen bezeichnet. Tatsächlich gelingt es einer solchen Skala, ein hohes Alpha zu erreichen. Aber auch heterogene Skalen mit vielen sehr unterschiedlichen Indikatoren können ein hohes Alpha aufweisen, sofern alle diese Indikatoren auf dasselbe Konstrukt verweisen. Alpha misst die Reliabilität und nicht die Homogenität. Die Homogenität einer Skala kann bestimmt werden, indem alle Items der Skala miteinander korreliert und dann die mittlere Korrelation berechnet wird. (Vorsicht, das ist komplizierter, als es klingt, denn Korrelationen dürfen nicht einfach gemittelt werden. Zunächst muss eine Fischer-Z-Transformation durchgeführt werden.) Items, die sehr ähnlich formuliert sind, korrelieren sehr hoch und sollten vermieden werden. Eine mittlere Item-Interkorrelation zwischen 0,3 und 0,7 gilt als ideal.

Prinzipiell ist es auch möglich, eine zuverlässige Skala zu erreichen, indem man sehr viele weniger gut geeignete Items verwendet. Da die Messung besser wird, wenn man mehrfach misst, kann man die Anzahl der Items einfach erhöhen. Allerdings könnte das auch Fragen nach der Passung der Items aufwerfen. Gebräuchliche Persönlichkeitstests der Psychologie weisen beispielsweise 10 bis 15 Fragen pro Skala auf. Wenn man hingegen 30 Items benötigt, um gerade eben ein Alpha über 0,7 zu erreichen, bedeutet das, dass die eigenen Items nicht besonders gut geeignet sind, um das Konstrukt zu erfassen. Dies kann am Konstrukt oder an den Items liegen. Die Passung eines Items zu einer Skala kann durch die sogenannte Item-Trennschärfe bestimmt werden. Items mit einem Wert unter 0,4 sollten nicht verwendet werden. Die Item-Trennschärfe korreliert das Item mit der Skala aus allen anderen Items. Die Skala wird aus den Items gebildet. Wenn alle Items etwas messen, was zur Skala passt, dann sollten sie auch alle mit der Skala korrelieren. Bei der Berechnung der Korrelation muss das fragliche Item allerdings aus der Skala herausgelassen werden. Wäre es in der Skala enthalten, wäre es ja auch kein Wunder, wenn es mit dieser korreliert. Die Korrelation eines Items mit der Skala die ohne dieses Item gebildet wurde heißt Item-Trennschärfe. Mithilfe der Item-Trennschärfe lassen sich ungeeignete Items schnell identifizieren. Diese können dann im Prozess der

Itemselektion gelöscht werden. Es sollten allerdings immer nur einzelne Items gelöscht und geprüft werden, wie sich dies auf das Cronbachs Alpha und die Trennschärfe der übrigen Items auswirkt. Wenn nur mit Blick auf die Kennwerte (Alpha oder Trennschärfe) optimiert wird, besteht die Gefahr, dass inhaltlich spannende Items vorschnell gelöscht werden. Es besteht auch die Tendenz, bei der Optimierung der Kennwerte sehr ähnliche Items zu behalten und heterogene Items zu löschen. Es ist also nicht ganz leicht, eine Skala zu optimieren.

Besondere Probleme ergeben sich, wenn keine oder nur eine sehr geringe Varianz vorliegt. Wie oben dargestellt, zeigt die Mathematik harte Grenzen für das auf, was möglich ist und was nicht. So ist es beispielsweise nicht möglich, durch null zu dividieren. Gerade bei Zufriedenheitsbefragungen zeigt sich häufig, dass die Messwerte eine sehr hohe Zufriedenheit anzeigen. Wenn man von flüchtigen Bekannten gefragt wird, wie es einem geht, antworten die meisten Menschen mit „gut". Das ist nicht ungewöhnlich. Wenn in einer Befragung jedoch alle Befragten die gleiche Note für die Zufriedenheit mit einem Kinofilm angeben, dann ist die Varianz der x-Werte null. In der Gleichung für die Reliabilität (siehe oben) würde dann durch Null geteilt werden. Die klassische Testtheorie hat große mathematische Probleme bei fehlender oder sehr geringer Varianz. Was keine Varianz hat, kann auch nicht korrelieren. Dabei ist es durchaus denkbar, dass von 40 befragten Personen tatsächlich alle den Film gut fanden und die gleichen Antworten gaben. Das liegt auch im Interesse der Filmschaffenden. Sie möchten, dass sich alle zufrieden äußern. Genau dann aber versagt die Mathematik hinter der klassischen Testtheorie.

Die probabilistische Testtheorie versucht, diese Probleme zu lösen, führt dafür aber zu anderen Schwierigkeiten. Da die Varianz für die klassische Testtheorie eine wichtige Rolle spielt, kann es sinnvoll sein, sich die Items genauer anzusehen, um zu prüfen, wie viele Menschen dem Item zustimmen, es ablehnen und wie groß seine Varianz ist. Ein Item, das von niemandem zustimmend beantwortet wird, wäre beispielsweise unnütz. Wahrscheinlich ist es zu voraussetzungsreich. In der Statistik spricht man von der Schwierigkeit eines Items. Damit ist nicht gemeint, wie herausfordernd die Lösung einer schwierigen Aufgabe ist, sondern wie schwer es Menschen fällt, dem Item zuzustimmen. Die Behauptung „Dies ist auf jeden Fall der beste Film aller Zeiten" könnte die Latte beispielsweise einfach zu hoch legen. Möglicherweise wird dem niemand zustimmen. Das Item gilt statistisch als zu schwer. Wenn tatsächlich niemand zustimmt, beträgt die

Varianz des Items null. Es ist dann für die klassische Testtheorie unbrauchbar. Gleiches gilt mit umgekehrten Vorzeichen für zu leichte Items. Ein Beispiel hierfür ist, wenn alle der Behauptung zustimmen, dass der Film „zumindest nicht schlecht“ war. Hier liegt die Latte vielleicht zu niedrig.

4 Gütekriterien

Im Rahmen eines Handbuches über das Schreiben wissenschaftlicher Abschlussarbeiten definiert Strunk (2022) Wissenschaft wie folgt:

Definition *Wissenschaft versucht – auf nachvollziehbare, transparente und überprüfbare Art und Weise – zutreffende Antworten auf bislang unbeantwortete Fragen zu liefern.*

Dazu führt Strunk (2022) weiter aus:

> Am Anfang steht die Forschungsfrage. Diese kann der Neugier eines Menschen entspringen oder die Fragen kommen aus der Wissenschaft selbst. Das ist zum Beispiel dann der Fall, wenn mit einer Erkenntnis in dem einen Feld in einem anderen eine Forschungslücke aufgeworfen wird. Oder die Frage wird aufgeworfen durch gesellschaftliche Herausforderungen, etwa für das Zusammenleben, die Gesundheit, die Suche nach Selbstverwirklichung, die Optimierung von Versorgungsstrukturen und so weiter. Wo auch immer die Fragen herkommen – ob sie große gesellschaftliche Probleme zu lösen versuchen oder eine private Neugier in Worte kleiden – sie sind der Ausgangspunkt und Zielpunkt, um den es in wissenschaftlichen Arbeiten geht.

Wichtig ist jedoch, dass Wissenschaft eine Forschungslücke adressiert, also Fragen beantwortet, die bislang entweder gar nicht oder nicht aus dieser Perspektive oder mit dieser Methode etc. beantwortet wurde. Wissenschaft baut auf bereits bekanntem Wissen auf und versucht darüber hinauszugehen („auf die Schulter von Riesen steigen“). Sie strebt mit ihren Forschungsfragen also darauf ab weiter zu sehen, tiefer zu verstehen, zutreffender zu prognostizieren, Probleme mit neueren Technologien effizienter zu lösen, altbekannte Verfahrensweisen auf der Grundlage neuer Erkenntnisse zu überprüfen etc. Das jedenfalls ist gemeint, wenn von „bislang unbeantwortete Fragen“ die Rede ist.

> Bei der Beantwortung von Fragen wendet die Wissenschaft Methoden an, deren zentrale Eigenschaften die Nachvollziehbarkeit, Transparenz und Überprüfbarkeit sind. Wissenschaftliche Methoden haben das Ziel Antworten auf Fragen zu generieren. Methoden wie Experimente, Fragebögen, Beobachtungsinstrumente oder Signifikanztests, dienen der Generierung, der Auswahl und der Erprobung möglicherweise geeigneter Antworten auf eine Forschungsfrage. Aber einsames Meditieren, ein Gebet oder ein Rauschzustand

> können auch dazu dienen Antworten zu finden. Die Anwendung solcher Methoden wäre aber problematisch, wenn es transparentere und leichter nachprüfbare Erkenntniswege für die gleichen Fragen gäbe. Wissenschaftliche Methoden unterscheiden sich von allen anderen Methoden der Antwortsuche dadurch, dass sie nachvollziehbar und transparent dokumentieren und demonstrieren, wie sie zu den Erkenntnissen gelangt sind. Dies erlaubt die Nachprüfbarkeit der Erkenntnisse. Wissenschaftliche Erkenntnis ist nachvollziehbar, so dass sie kritisiert und von anderen korrigiert werden kann. Beispielsweise [..] zeigen [Studien], wie viele Fehler in der medizinischen Forschung aus Unwissenheit über die statistische Methodik entstehen (z. B. Goodman, 2008). Das Besondere am wissenschaftlichen Arbeiten ist, dass diese Fehler überhaupt auffallen und später korrigiert werden können. Denn die Forderung zur transparenten Dokumentation des Vorgehens führt eben dazu, dass andere die Ergebnisse prüfen können. Intransparenz würde eine nachträgliche Prüfung verhindern. (Strunk, 2022, S. 8 f.)

Diese Überlegungen führen zur Objektivität als zentrales Gütekriterium wissenschaftlichen Arbeitens:

> Ein wichtiges Gütekriterium wissenschaftlicher Tätigkeit und wissenschaftlicher Ergebnisse ist daher die Objektivität. Die Objektivität kann auch als Fundament der wissenschaftlichen Tätigkeit verstanden werden. Denn andere Gütekriterien bauen auf der Objektivität auf. Fehlt diese, dann können auch andere Erfordernisse nicht erfüllt [werden]. Mit der Objektivität ist gemeint, dass wissenschaftliche Erkenntnisse frei sein sollen von subjektiven Einflüssen und von individuellen, persönlichen blinden Flecken. Die Antworten auf Forschungsfragen sollten nicht persönlich gefärbt und von unwissenschaftlichen Interessen gelenkte sein, denn das wäre dann persönliches Geltungsstreben, Politik, Egoismus, Gewinnstreben etc. – aber keine Wissenschaft. Die Transparenz hilft anderen die blinden Flecken zu sehen, sie hilft zu hinterfragen ob eine Erkenntnis subjektiv gefärbt ist oder einer Prüfung durch Andere Stand hält. Es kann darüber gestritten werden, ob Objektivität überhaupt erreicht werden kann. Denn Menschen sind nun einmal subjektiv. Vollkommene Objektivität ist illusorisch, aber sie ist als Ideal erstrebenswert. Das Ideal der Objektivität kann erreicht werden, wenn Wissenschaft transparent, nachvollziehbar und überprüfbar ist. Dazu gehört es dann z. B. auch, mögliche subjektive Einflüsse in der Arbeit zu diskutieren und damit transparent zu machen. Der Begriff der Objektivität führt mitunter zu Missverständnissen. Gemeint ist nicht der Inhalt der Forschung, sondern die Methode mit der die Inhalte untersucht werden. Es gibt Inhalte, etwa in den Naturwissenschaften die von

> Haus aus objektiv erscheinen, etwa das Volumen eines Körpers, sein Gewicht etc. Inhalte in den Sozialwissenschaften gelten hingegen als wenig objektiv. Das heißt aber nicht, dass es dort nicht möglich ist objektive Forschung durchzuführen. Wenn z. B. in der Psychologie über Träume geforscht wird, dann sind diese Träume […] subjektive Erfahrungen und zunächst nur den Träumenden selbst zugänglich. Der Forschungsgegenstand der Psychologie ist daher natürlicherweise durch und durch subjektiv. Wissenschaftliche Methoden, die sich mit Träumen beschäftigen, können aber nichts desto trotz objektiv sein. Wird ein Traum im Rahmen eines leitfadengestützten Interviews erzählt und diese Erzählung aufgezeichnet und später transkribiert, dann ist es eine nachweisbare Tatsache, dass dieser Traum tatsächlich so und nicht anders berichtet wurde. Objektivität bezieht sich in diesen Fall auf die wissenschaftliche Methode und nicht auf den Inhalt. (Strunk, 2022, S. 9 f.)

Zusammenfassend lässt sich sagen, dass die Objektivität das Fundament der drei Hauptgütekriterien wissenschaftlichen Arbeitens darstellt. Sie ist bei der Datenerhebung, der Auswertung und der Interpretation zu berücksichtigen und anzustreben. Unterschiede in der Ansprache, variierende Hilfestellungen und Unterschiede im Auftreten gegenüber den Studienteilnehmenden können die Ergebnisse beeinflussen. Auch die Kodierung von Antworten sollte frei von subjektiven Einflüssen sein. Werden Fragebögen oder Testbögen elektronisch präsentiert, entfallen viele subjektive Einflüsse, da Instruktionen und Fragen immer identisch angezeigt werden. Auch Antwortvorgaben, die nur angekreuzt werden müssen, können hilfreich sein. Bei papiergestützten Befragungen kommt es dagegen immer wieder zu Abweichungen vom vorgesehenen Antwortschema. So wird beispielsweise ein Rating nicht beantwortet, sondern es wird angemerkt, dass man eigentlich 12 Punkte vergeben würde, es aber nur 10 Kästchen gibt. Solche Probleme treten bei computergestützten Befragungen nicht auf. Dafür wirken diese Befragungen recht unpersönlich und werden zu unterschiedlichen Zeiten an möglicherweise wenig geeigneten Orten bearbeitet (z. B. schnell in der U-Bahn). Auch bei der Auswertung können subjektive Einflüsse eine Rolle spielen, etwa wenn Punktwerte erst vergeben werden müssen, das Zusammenzählen von Punkten nach einem bestimmten Schema erforderlich ist oder jede Frage von Hand ausgewertet werden muss. Schließlich werden Daten, Messungen und Ergebnisse interpretiert, wobei es auch hier möglicherweise zu subjektiven Einflüssen kommen kann, die einige Ergebnisse hervorheben, während andere außer Acht gelassen werden.

Die hier angeführten Beispiele sind nicht vollständig, sondern sollen lediglich andeuten, welche Faktoren möglicherweise einen Einfluss auf die Objektivität einer Untersuchung haben könnten. Objektivität bzw. Subjektivität spielen im gesamten Forschungsprozess eine zentrale Rolle. Sie sind allerdings nicht so klar definiert und mit Indikatoren nur schwer abbildbar wie beispielsweise die Reliabilität. Daher ist es kaum möglich, konkrete Empfehlungen abzugeben oder Kennwerte für das Vorliegen von Objektivität zu berechnen. In Studien wird häufig darauf hingewiesen, dass Objektivität durch standardisierte Erhebungsinstrumente, ein standardisiertes Vorgehen, eine Erhebung mittels Computer, automatisierte Kodierung, eine standardisierte und automatisch ablaufende Auswertung sowie vorher festgelegte Interpretationsregeln erreicht wird. Nicht jede Studie umfasst all diese Maßnahmen.

In der Regel wird also auf Maßnahmen verwiesen, die in einer Studie ergriffen wurden, um die Objektivität zu erhöhen. Daran schließt sich der Hinweis an, dass aufgrund dieser Maßnahmen davon ausgegangen wird, dass Objektivität erreicht wurde. Im Fall des Gütekriteriums der Objektivität ist das jedoch nicht einheitlich prüfbar. Hinzu kommt, dass nicht jedes Forschungsprojekt für ein standardisiertes Vorgehen mit geschlossenen Fragen im Rahmen einer Onlinebefragung geeignet ist. Mitunter geht es um weitaus offenere Inhalte, wie sie beispielsweise bei offenen, qualitativen Interviews eine Rolle spielen. Auch hierfür gibt es Möglichkeiten und Methoden, um bei aller Offenheit möglichst objektiv zu arbeiten. So können die Interviews beispielsweise aufgezeichnet, Transkriptionsregeln berücksichtigt und eine festgelegte Auswertungsmethode befolgt werden.

Die Objektivität wurde bereits als grundlegende Eigenschaft wissenschaftlichen Arbeitens benannt. Andere Gütekriterien bauen darauf auf. Versäumnisse in diesem Bereich wirken sich daher direkt auf die Reliabilität aus. Wenn in Intelligenztests beispielsweise Wortergänzungen erfragt werden und die Auswertenden selbst entscheiden können, ob ein ergänztes Wort wirklich passt, kann es passieren, dass sie einmal so und ein anderes Mal anders entscheiden. Die Reliabilität der Ergebnisse, die durch die Übereinstimmung der Ergebnisse bei einer Messwiederholung oder mit anderen Überprüfungsmethoden definiert ist, kann nicht größer sein, als es die Objektivität zulässt. Fehlende Objektivität verringert also immer auch die Reliabilität. Und was nicht reliabel ist, kann in weiterer Folge auch nicht valide sein.

Die drei Hauptgütekriterien Objektivität, Reliabilität und Validität bauen also wie eine Stufenpyramide aufeinander auf. Die Objektivität ist die Voraussetzung dafür, dass etwas reliabel erfasst werden kann. Die Reliabilität kann dabei nicht besser sein als die Objektivität. Das bedeutet, dass die Stufe der Reliabilität in der Regel kleiner ausfällt als die der Objektivität.

Reliabilität wurde oben bereits als Messfehlerfreiheit definiert. Sie liegt dann vor, wenn die Varianz der wahren Werte exakt der Varianz der gemessenen Werte entspricht. Der Quotient aus beiden Varianzen kann maximal 1 erreichen, was einer perfekten Reliabilität entspricht. Praktisch kann dies, wie ebenfalls schon dargestellt, durch verschiedene Methoden geprüft werden (Retest-Reliabilität, Odd-Even-Vergleich, Cronbachs Alpha etc.).

Eine hohe Reliabilität ist in der Regel eine gute Voraussetzung für eine hohe Validität. In der Praxis fällt diese jedoch in der Regel geringer aus als die Reliabilität, auf der sie aufbaut. Eine ausschließlich an den Reliabilitätskennwerten orientierte Itemauswahl kann problematisch sein. So besteht die Gefahr, dass eine überoptimierte Skala aus inhaltlich eingeschränkten Items besteht, die nur einen leicht zugänglichen Aspekt des Konstrukts erfassen. Dies entspricht jedoch häufig nicht dem, wofür die Skala eigentlich gedacht war. Genau das ist das Thema der Validität. Sie beschreibt die inhaltliche Passung der Skala zum Untersuchungszweck. Es geht also darum, dass das inhaltliche Ziel der Studie tatsächlich erreicht wird. Wenn beispielsweise der IQ gemessen werden soll, muss sich zeigen lassen, dass tatsächlich der IQ und nicht das Ausmaß an Prüfungsangst gemessen wird. Wenn das Ziel der Messung darin besteht, als Studieneingangstest darüber zu entscheiden, wer in ein Studium aufgenommen wird und wer nicht, sollte das Messinstrument nachweisen können, dass es die Eignung für das Studium tatsächlich feststellen kann. Das gilt auch für Einstellungstests, wie sie beispielsweise von Personalabteilungen durchgeführt werden.

Die Validität sollte auch geprüft werden, wenn ein neues hypothetisches Konstrukt etabliert werden soll. Dabei steht die Frage im Vordergrund, ob sich das neue Konstrukt klar genug von anderen Konstrukten, mit denen es verwechselt werden könnte, unterscheidet.

Die Ziele von Messungen bzw. Studien sind sehr vielfältig. Die Validität fragt danach, ob das zentrale Thema, das Ziel bzw. das Konstrukt tatsächlich inhaltlich passend behandelt wurde. Somit ist die Validität die zentrale Größe, auf die die Pyramide der Hauptgütekriterien abzielt. Was nützt eine

objektive und reliable Messung, wenn nicht das gemessen wird, was gemessen werden sollte? Verglichen mit ihrer großen Bedeutung für die Brauchbarkeit einer Studie wird die Validität erstaunlich selten überprüft. Das liegt daran, dass es dabei so viel zu berücksichtigen gibt. So existieren für verschiedene Studienziele auch verschiedene Validitäten. Hinzu kommt, dass die Prüfung der Validität idealerweise auch Daten anderer Quellen zu Vergleichszwecken berücksichtigen würde. Das stellt einen hohen zusätzlichen Aufwand dar.

Ein mögliches Ziel einer Befragung könnte beispielsweise darin bestehen, die Einstellung zur Mülltrennung zu erfragen. Hier ist es nachvollziehbar, dass die Befragten sozial erwünscht antworten und angeben, sehr für die Mülltrennung zu sein. Die Validität dieser Befragungsmethode ließe sich testen, indem die Mülltonnen in den befragten Wohnanlagen kontrolliert würden. Dabei sollte die erfragte Einstellung zur Mülltrennung mit der tatsächlichen Mülltrennung korrelieren. Die Validität wäre also durch eine solche Korrelation mit Zahlen zwischen 0 und 1 konkret bezifferbar. Eine weitere Möglichkeit wäre der Vergleich von IQ-Tests mit Urteilen von Fachleuten oder mit Schulnoten, da die Korrelation in diesem Fall gering ist. Die Überprüfung eines Befragungsinstruments anhand realer empirischer Gegebenheiten ist die wohl beste Form der Validitätsprüfung, aber wegen des hohen Aufwands besonders schwer durchzuführen.

Stattdessen wird seit einigen Jahren zunehmend auf deutlich einfachere Methoden zurückgegriffen, die jedoch nur sehr eingeschränkte Aspekte der Validität betreffen.

5 Glossar für einige wichtige statistische Begriffe

Abhängige bzw. unabhängige Variablen. Je nach Theorie die einer Untersuchung zugrunde liegt, gibt es Vermutungen über die Kausalrichtung in der Variablen zueinander stehen. Eine Variable kann dann als Ursache und eine andere als Wirkung aufgefasst werden. Die Wirkung wird als *abhängige Variable* bezeichnet, da sie von der Ursache abhängig ist. Die Ursache wird hingegen als *unabhängige Variable* bezeichnet, wenn im Rahmen der Theorie und/ oder der Untersuchung keine zusätzlichen Variablen berücksichtigt werden, die die Ursache beeinflussen. Bei vielen Studien steht eine einzige zentrale Größe als abhängige Variable im Vordergrund. Darauf können mehrere unabhängige Variablen einen Einfluss haben. So wird die körperliche Gesundheit (abhängige Variable) bestimmt von verschiedenen unabhängigen Variablen (Sport, Ernährung, Gesundheitsverhalten, Alter, genetischen Prädispositionen etc.). Als *Kontrollvariablen* bezeichnet man solche unabhängigen Variablen, die einen Einfluss haben könnten und daher statistisch berücksichtigt werden müssen, aber nicht zentrales Forschungsinteresse der Studie sind.

Abhängige Daten. Bei Interventionsstudien sind die Ergebnisse nach der Intervention abhängig von den Merkmalsausprägungen vor der Intervention. Damit eine Veränderung sichtbar wird müssen verschiedene Messzeitpunkte miteinander verglichen werden. Die Messwerte stehen dadurch in einer zeitlichen Abhängigkeit. Für abhängige Daten sind besondere statistische Testverfahren vorgeschlagen worden. Für einige Auswertungen kann es zudem sinnvoll sein den Unterschied zwischen den Zeitpunkten zu berechnen. Dieses sog. *Delta* lässt sich dann statistisch als eine Größe für die Wirkung der Intervention verwenden.

Abweichungsmaße, Streuungsmaße. Mittelwerte und andere ↗ *Maße der zentralen Tendenz* geben zwar einen Eindruck über die Daten insgesamt, die einzelnen Messwerte weichen jedoch in der Regel auch von der zentralen Tendenz ab. Diese Abweichungen werden durch Abweichungsmaße er-

fasst. ↗ *Standardabweichung, Streuung oder Varianz* zeigen wie intervallskalierte Messwerte (↗ Messung) vom arithmetischen Mittel abweichen. Abweichungen vom Median werden als Interquartilsabstand angegeben. Dabei wird die Messwerteverteilung nach der Größe sortiert, die ersten 25% der Daten sind das erste Quartil, die 50%-Grenze ist der Median und die ersten 75% der Daten sind das dritte Quartil. Der Interquartilsabstand ist der Bereich vom ersten zum dritten Quartil. 50% aller Daten liegen innerhalb dieser Grenzen. Bei nominalen Daten kann der Modalwert den häufigsten Wert angeben. Eine Abweichung ist hier durch Angabe der Häufigkeit und der Prozentzahl ersichtlich. Auch kann es hier sinnvoll sein, zusätzlich den seltensten Wert zu bestimmen.

Alpha-Fehler, Beta-Fehler, *Test Power*. Ein Signifikanztest (↗ statistische Signifikanz) befindet den Unterschied zwischen zwei Kennwerten (z. B. Mittelwerten) dann als signifikant, wenn der Unterschied so groß ist, dass es nach den Gesetzen der Wahrscheinlichkeitsrechnung nur eine geringe Wahrscheinlichkeit (↗ P-Wert) dafür gibt, dass *kein Unterschied* besteht. Wie gering sollte dafür die Wahrscheinlichkeit sein? Es handelt sich um eine Übereinkunft, dass üblicherweise bei einer Wahrscheinlichkeit von 5% (und darunter) von ↗ Signifikanz gesprochen wird. Das bedeutet dann, dass ein Signifikanztest, der zwei Kennwerte mit einer Wahrscheinlichkeit von 5% für ähnlich hält, zu dem Schluss kommt, dass eine Signifikanz vorliegt. Die Annahme, dass die Kennwerte ähnlich sind, wird daher verworfen und die dazu passende Alternativhypothese (↗ Hypothesenarten) wird akzeptiert. Die Signifikanzgrenze von 5% gibt dabei gleichzeitig den Fehler der Entscheidung über eine Signifikanz an. Denn mit einer 5%igen Wahrscheinlichkeit ist die Nullhypothese ja korrekt. Wenn sie dennoch verworfen wird – und damit Signifikanz angenommen wird – ist dies mit einer Wahrscheinlichkeit von 5% falsch. Wenn man aufgrund eines Signifikanztests davon ausgeht, dass eine Signifikanz vorliegt, macht man mit eben jener Wahrscheinlichkeit die als Signifikanzgrenze festgelegt wird einen Fehler. Dieser Fehler wird *Alpha-Fehler* genannt (Fehler 1. Art). Mitunter wird auch nur von Alpha gesprochen. Dieses Alpha ist nicht zu verwechseln mit *Cronbachs Alpha* einem Maß für die internen Konsistenz einer Fragebogenskala. Der Alpha-Fehler ist der Fehler fälschlicherweise von Signifikanz auszugehen obwohl keine vorliegt. Demgegenüber steht ein *Beta-Fehler* (Fehler 2. Art) dieser gibt an wie hoch die Wahrscheinlichkeit dafür ist fälschlicherweise davon auszugehen, dass keine Signifikanz vorliegt.

Hier wird also eine vorliegende Signifikanz übersehen. Den Alpha-Fehler kann man für eine Untersuchung frei wählen. So ist Alpha = 5% eine übliche Signifikanzgrenze. Diese kann auch strenger gewählt werden, z. B. mit 1% oder 0,1%. Der Beta-Fehler hängt aber von Umständen ab, die nicht beeinflusst werden können. Der Beta-Fehler kann nicht direkt gewählt und festgelegt werden. Der Beta-Fehler ist z. B. auch gegeben durch die Empfindlichkeit des eingesetzten statistischen Testverfahrens. Testverfahren mit einem hohen Beta-Fehler übersehen Signifikanzen. Eins minus Beta wird als *Test-Power* bezeichnet. Das ist die Wahrscheinlichkeit mit der ein Test eine vorliegende Signifikanz auch findet.

Alpha-Fehler-Adjustierung. In der Regel sind Signifikanztests in der Lage, nur zwei Kennwerte (z. B. Mittelwerte) miteinander zu vergleichen. Einige Fragestellungen bzw. Hypothesen machen mehrere Vergleiche zwischen jeweils zwei Kennwerten nötig, um die Hypothese insgesamt prüfen zu können. Beantworten z. B. drei Personengruppen einen Fragebogen (Gruppe A, B, C), so kommt man auf insgesamt drei paarweise Vergleiche (A mit B; A mit C und B mit C). Allgemein gilt: Anzahl der Vergleiche = [Anzahl der Gruppen mal [Anzahl der Gruppen minus Eins]] geteilt durch Zwei. So ergeben sich für vier Gruppen bereits: (4 x 3)/2 = 6 Vergleiche. Wenn die Hypothese relativ offen formuliert ist und generell nach Unterschieden zwischen den Gruppen gefragt wird, so wächst die Wahrscheinlichkeit, einen Unterschied zu finden, je mehr Vergleiche möglich werden. Da man ja bei jedem Paarvergleich einen ↗ Alpha-Fehler von 5% begeht, summieren sich die Fehler von Paarvergleich zu Paarvergleich. Bei drei Vergleichen macht man also einen viel höheren Fehler als bei nur einem. Höhere Fehler als 5% sind jedoch nach der oben angesprochenen Vereinbarung nicht signifikant. Um insgesamt nur auf einen Fehler von 5% zu kommen, müssen für jeden Einzelvergleich strengere Alpha-Fehler-Grenzwerte festgelegt werden. Für 3 Vergleiche ergibt sich z.B. ein Wert von 1,7%, bei vier Vergleichen sind es 1,3%, bei 10 Vergleichen 0,5%, usw. Eine solche Anpassung der Signifikanzgrenze für mehrfaches Testen heißt Alpha-Fehler-Adjustierung. Eine Alternative für die Berechnung vieler Signifikanztests, die nur jeweils zwei Kennwerte vergleichen, ist die Varianzanalyse (↗ Varianzanalyse, ANOVA) oder die multiple Regression.

Chi-Quadrat-Test. Der Chi-Quadrat-Test ermöglicht den Vergleich von erwarteten Häufigkeiten mit tatsächlich beobachteten Häufigkeiten. Erwartet man aufgrund von Vorerfahrungen oder aus der Literatur zum Beispiel, dass jeder vierte männliche Österreicher Raucher ist, so würde man bei 100 befragten Personen 25 Raucher erwarten. Der Chi-Quadrat-Test vergleicht die erwarteten 25 Raucher dann mit den tatsächlich im Rahmen einer Befragung vorgefundenen Rauchern. Im Rahmen eines Chi-Quadrat-Tests können beliebig viele verschiedene Häufigkeiten miteinander verglichen werden. So ergibt sich beim Chi-Quadrat-Test auf eine Gleichverteilung die erwartete Häufigkeit als Mittelwert der beobachteten Häufigkeiten. Aufgrund geringer Voraussetzungen kann der Chi-Quadrat-Test immer berechnet werden, wenn es um Häufigkeiten geht und eine bestimmte oder mehrere bestimmte Häufigkeiten erwartet werden können. Der Chi-Quadrat-Test ermittelt einen Chi-Quadrat-Wert, für den zusammen mit den sog. Freiheitsgraden (in der Regel Zahl der Messwerte minus eins) die Wahrscheinlichkeit bekannt ist. Die Wahrscheinlichkeit ist das Ergebnis des Tests. Man spricht von einer ↗ statistischen Signifikanz, wenn diese Wahrscheinlichkeit kleiner als der vorher festgelegte ↗ Alpha-Fehler ist. Da der Chi-Quadrat-Test für ganz verschiedene Hypothesen über die Verteilung von Häufigkeiten angewendet werden kann gibt es zahlreiche sehr unterschiedliche Chi-Quadrat-Tests. Hier ist darauf zu achten den konkreten Zweck des Tests genau zu prüfen und ihn im Rahmen einer wissenschaftlichen Arbeit nicht einfach nur als Chi-Quadrat-Test zu bezeichnen sondern z. B. als Chi-Quadrat-Test für Kreuztabellen (4-Felder-Tabellen) oder auf Gleichverteilung oder auf Normalverteilung. Der Test arbeitet mit einer Näherungsgleichung, die bei keinen Stichproben auch mal daneben liegen kann. Der Chi-Quadrat-Test wird daher heute viel seltener verwendet als noch vor einigen Jahren.

Fishers exakter Test. Ein besonders *sicherer* Test ist Fishers exakter Test, da er kaum an Voraussetzungen gebunden ist und immer berechnet werden kann, wenn es um den Vergleich zweier Prozentzahlen (bzw. Häufigkeiten) geht. Er ist damit eine exakte und bessere Alternative für den ↗ Chi-Quadrat-Test für Kreuztabellen (4-Felder-Tabellen). Eine Berechnung durch einen Computer ist aber rechenintensiv und kommt an Grenzen, wenn die Stichprobengröße ca. den Wert 1000 erreicht. Neben der exakten Variante dieses Tests gibt es für große Stichproben daher auch Näherungsformeln über den ↗ T-Test, die jedoch mit Vorsicht zu genießen sind. Fishers exakter

Test liefert die Wahrscheinlichkeit für die Übereinstimmung zweier Prozentzahlen (bzw. Häufigkeiten). Die Wahrscheinlichkeit ist das Ergebnis des Tests. Man spricht von einer ↗ statistischen Signifikanz, wenn diese Wahrscheinlichkeit kleiner als der vorher festgelegte ↗ Alpha-Fehler ist.

Hypothesenarten. Hypothesen werden aus Theorien bzw. konkreten theoretischen Annahmen und bereits publizierten wissenschaftlichen Studien logisch abgeleitet. Sie formulieren auf Grundlage dieser theoretischen Annahmen eine Vorhersage für den konkreten Fall der geplanten Untersuchung. Die Hypothese, die aus den theoretischen Grundlagen folgt heißt „*Alternativhypothese*". Statistisch geprüft wird die Verneinung der Alternativhypothese. Diese wird als „*Nullhypothese*" bezeichnet. Ist die Nullhypothese unwahrscheinlich, wird von statistischer ↗ Signifikanz gesprochen. Eine unwahrscheinliche Nullhypothese wird verworfen und die Alternativhypothese wird akzeptiert. Hypothesen die statistisch geprüft werden sollen, sollten idealerweise direkt so formuliert werden, dass der statistische Zugang in der Hypothese bereits deutlich wird. Die Statistik kennt *Unterschiedshypothesen* – bei denen Unterschiede zwischen zweien oder mehreren Untersuchungsgruppen (↗ Kategorien) – vermutet werden, und sie kennt *Zusammenhangshypothesen* – bei denen Zusammenhänge zwischen zweien oder mehreren Variablen vermutet werden. Hypothesen über *keinen* Unterschied und *keinen* Zusammenhang können nicht auf Signifikanz geprüft werden und sollten daher vermieden werden. Hypothesen können *einseitig* oder *zweiseitig* formuliert werden (↗ P-Wert). Im einseitigen Fall wird die Richtung des Unterschieds (z. B. Hypothese: Das Gehalt der Männer ist höher als das der Frauen.) oder des Zusammenhangs (z. B. Hypothese: Es besteht ein positiver Zusammenhang zwischen Alkohol und Krebsrisiko.) in der Hypothese konkret genannt. Ob diese Form der Hypothese gewählt wird, hängt von der Theorie bzw. der Studienlage ab. Ist diese uneindeutig, wird zweiseitig formuliert und offen gelassen, welche Kategorie höhere Werte zeigt (z. B. Hypothese: Die Schulklassen unterscheiden sich in ihren Leistungen.) oder welcher Zusammenhang vorliegt (z. B. Hypothese: Es besteht ein Zusammenhang zwischen Musikgenuss und Nervosität.).

Kategoriale Daten. Bei Unterschiedshypothesen werden Unterschiede in einer ↗ abhängigen Variablen für verschiedene Untersuchungsgruppen –

oder allgemeiner: Kategorien – angenommen. So werden z. B. bei der Hypothese, dass es einen Lohnunterschied zwischen Frauen und Männern gibt, die beiden Kategorien Frauen und Männer als unabhängige Variable verwendet und der Lohn innerhalb der beiden Kategorien ermittelt. Eine Kategorie umfasst in der Regel mehrere Untersuchungseinheiten, also nicht nur einen Mann. Kategorien sind entweder direkt durch die Erhebung gegeben (z. B. Geschlecht wurde direkt erhoben) oder müssen nach der Erhebung gebildet werden (z. B. Alter wurde erhoben und Altersgruppen werden später gebildet). Nominale Daten bilden automatisch Kategorien. Ordinale Daten oder Intervalldaten und andere höhere Datenniveaus können nachträglich in Kategorien unterteilt werden (↗ Messung, Messniveau, Skalenniveau).

x / y	Intervall	Dichotom (2-stufig: z.B. ja/nein)	Ordinal
Intervall	Produkt-Moment-Korrelation (Pearson)	Punktbiseriale Korrelation (Alternativ: T-Test) *Bei 1/0-Kodierung der dichotomen Variable ist die Produkt-Moment-Korrelation identisch mit der Punktbiserialen Korrelation.*	Rangkorrelation (Spearman) *Bei Kodierung der Ordinalskala mit 1, 2, 3, … ist der Wert mit der Produkt-Moment-Korrelation identisch.*
Dichotom (2-stufig: z. B. ja/nein)		Phi-Koeffizient (über Chi-Quadrat). *Bei 1/0-Kodierung der dichotomen Variablen ist die Produkt-Moment-Korrelation identisch mit Phi.*	Biseriale Rangkorrelation (Alternativ: U-Test)
Ordinal			Rangkorrelation *Bei Kodierung der Ordinalskalen mit 1, 2, 3, … ist der Wert mit der Produkt-Moment-Korrelation identisch.*

Tabelle 3: Korrelationsverfahren in Abhängigkeit vom Skalenniveau bzw. der Stetigkeit

Die Tabelle zeigt, dass bei geeigneter Kodierung von dichotomen und ordinalen Daten immer die Produkt-Moment-Korrelation (Pearson) berechnet werden kann. Das liegt daran, dass die Gleichungen für die anderen Korrelationsverfahren aus der Produkt-Moment-Korrelation abgeleitet wurden und mit diesen identisch sind, sobald die Kodierung den angegebenen Regeln folgt.

Korrelationen. Eine Korrelation beschreibt den statistischen Zusammenhang zwischen zwei Merkmalen. Beide Merkmale müssen in unterschiedlichen Ausprägungen vorkommen können. Ist das nicht der Fall, so kann keine Korrelation berechnet werden. Wird z. B. die Frage danach gestellt, ob die Zahl der Geburten und die Zahl der Störche einen Zusammenhang (also eine Korrelation) aufweist, so muss sowohl die Zahl der Störche, als auch die Zahl der Geburten variieren können. Es bietet sich hier an, die Zahl der Geburten und die Zahl der Störche pro Monat zu erheben. Dadurch erhält man Zahlenpaare aus Geburtenzahl und Storchenpopulation für jeden Monat. Es stellt sich nun die Frage, ob sich die Zahl der Störche und die Zahl der Geburten über das Jahr hinweg in die gleiche Richtung entwickelt, also ob mit ansteigender Zahl der Geburten auch die Zahl der Störche wächst und ob mit sinkender Zahl der Geburten auch die Zahl der Störche abnimmt. Ist es so, dass die Zahl der Störche und die Zahl der Geburten sich jeweils in die gleiche Richtung entwickeln, so spricht man von einer positiven Korrelation. Steigt jedoch die Zahl der Geburten, immer wenn die Zahl der Störche abnimmt (und umgekehrt: die Zahl der Geburten sinkt und gleichzeitig nimmt die Zahl der Störche zu), so spricht man von einer negativen Korrelation. Korrelationen können Zahlenwerte zwischen –1 und +1 annehmen. Dabei zeigt das Vorzeichen an, ob es sich um einen positive oder um eine negative Korrelation handelt. Je näher die Zahlenwerte bei 1 (bzw. –1) liegen, desto „perfekter" ist der Zusammenhang. Ist der Zahlenwert jedoch 0, dann liegt gar keine Korrelation – also auch kein Zusammenhang – vor. Viele Zusammenhänge, die z. B. in der Psychologie beschrieben werden, haben relativ kleine Werte um 0,3 (bzw. –0,3), wohingegen z. B. in der Physik nicht selten Korrelationen um 0,9 (bzw. –0,9) gefunden werden können. Die Höhe einer Korrelation zu interpretieren ist daher nicht leicht. Es gibt zwar allgemein akzeptierte Einteilungen aber im konkreten Anwendungsfall können auch andere Grenzen für eine ausreichende oder nicht ausreichende Korrelation sinnvoll sein. Allgemein gilt eine Korrelation ohne Berücksichtigung des Vorzeichens ab 0,1 als klein, ab 0,3 als mittel und ab 0,5 als groß (Cohen, 1992). Ob eine Korrelation nicht eventuell doch auf das Fehlen einer Korrelation (Null-Korrelation) hinweist, kann nur durch einen Signifikanztest (↗ Statistische Signifikanz) entschieden werden. Erst, wenn eine Korrelation sich als signifikant herausstellt, kann sie interpretiert werden. Ist sie nicht signifikant, so kann man nicht davon ausgehen, dass ein Zusammenhang beobachtet wurde. Ist sie jedoch signifikant, so bedeutet das noch nicht, dass der beobachtete Zusammenhang kausal zu interpretieren ist. Es gibt vielleicht Stu-

dien, die zeigen, dass die Zahl der Störche mit der Zahl der Geburten in einigen Gegenden im Verlauf des Jahres signifikant korreliert ist. Das würde jedoch nicht bedeuten, dass die Störche die Kinder bringen. Korrelationen sind immer nur Beobachtungen gleich- oder gegengerichteter Entwicklungen in Variablen. Die Ursachen für eine gemeinsam in die gleiche Richtung gehende Entwicklung können allein aus der Korrelation nicht ersehen werden. Es gibt verschiedene statistische Verfahren die die Korrelation in Abhängigkeit von der ↗ Messung, dem Messniveau, dem Skalenniveau berechnen. In der Regel wird dabei von linearen Zusammenhängen ausgegangen. Liegen tatsächlich nichtlineare Zusammenhänge vor, kann die Korrelationsberechnung fälschlicherweise auf eine Null-Korrelation verweisen. Die neuere Komplexitäts- und Chaosforschung kennt auch Korrelationen für beliebige nichtlineare Zusammenhänge, wie z. B. die Mutual Information (vgl. Strunk, 2019).

Mann-Whitney-U-Test. Besteht der Verdacht, dass die Voraussetzungen für einen ↗ T-Test verletzt sein könnten, kann am besten der U-Test von Mann und Withney berechnet werden.

Messung, Messniveau, Skalenniveau. Bei einer Messung werden empirische Gegebenheiten mit Zahlen abgebildet. Das Ziel ist dabei die Unterschiede, Ähnlichkeiten oder Relationen, in denen die empirischen Gegebenheiten zueinander stehen, mit den Zahlen bestmöglich wiederzugeben. Nach der Messung liegen nur mehr die Zahlen vor und es muss mitgeteilt werden und bekannt sein, wie die Messzuordnung erfolgte und was man aus den Zahlen ablesen darf und was aufgrund der Messung nicht interpretiert werden kann. *Nominalskala*: Die Zahlen werden ein-eindeutig den Objekten zugeordnet. Die Zahlen können die Objekte identifizieren. Die Höhe der Zahlen hat keinerlei Bedeutung. Beispiel: Zahlencode für Berufe, Bäcker:in = 234, Professor:in = 43, … *Ordinalskala*: Die Anordnung der Zahlen gemäß ihrer Größe entsprich einer Ordnung der empirischen Gegebenheiten. Diese wird aber nur grob wiedergegeben oder ist tatsächlich nur grob vorhanden. So kann der Abstand der Zahlen zueinander nicht als Abstand der empirischen Gegebenheiten zueinander interpretiert werden. Beispiel: höchster Bildungsabschluss: Pflichtschule = 1, Abitur = 2, Studium = 3, ... *Intervallskala*: Die Abstände zwischen den Zahlen können sinnvoll interpretiert werden. Zahlenverhältnisse können nicht sinnvoll interpre-

tiert werden. Beispiel: Alter gemessen in Jahren. Wenn eine Person 2 Jahre älter ist als eine andere, wird das so bleiben, auch wenn Zeit vergeht. Wenn eine Person *exakt doppelt so alt* ist wie eine andere, ist das am nächsten Tag oder in der nächsten Stunde oder Minute schon nicht mehr korrekt. Das Zahlenverhältnis ist also nicht vernünftig interpretierbar. *Verhältnisskala*: Zahlenverhältnisse sind sinnvoll interpretierbar. Beispiel: Gehalt. Eine Verhältnisskala erfordert einen inhaltlich klaren und unveränderbaren Nullpunkt. Das ist beim Gehalt gegeben. Die exakte Zahlengröße ist bei dieser Skala nach der Messung immer noch veränderbar (z. B. kann das Gehalt in verschiedenen Währungen angegeben werden). *Absolutskala*: Bei einer Absolutskala ist eine nachträgliche Umrechnung der Zahlen z. B. in andere Maßeinheiten unsinnig. Z. B. ist die Anzahl der Personen in einem Raum eine Zahl, die exakt diese Anzahl angibt und sinnvoll nicht mehr verändert werden sollte. Je nach Skalenniveau sind also verschiedene Eigenschaften interpretierbar und daher passende statistische Verfahren zu wählen.

Maße der zentralen Tendenz, (Mittelwert, Median, Modalwert). Messwerte einer Stichprobe unterscheiden sich in der Regel. *Maße der zentralen Tendenz* werden eingesetzt, um mit einer einzigen Zahl die Messwerte einer größeren Gruppe, z. B. einer gesamten Stichprobe zusammenzufassen. Je nach Skalenniveau (↗ Messung) kann ein arithmetischer Mittelwert (Intervallskalenniveau), der Median (Ordinalskala) oder der Modalwert (Nominalskalenniveau) benutzt werden. Der Mittelwert, als Summe aller Messwerte, geteilt durch die Anzahl der Messwerte liegt exakt in der Mitte der Messwerteverteilung. Er berücksichtigt dabei die Abstände zwischen den Messwerten. So ist der Mittelwert empfindlich gegenüber extremen Zahlenwerten, auch dann, wenn diese nur selten in der Stichprobe vorkommen. Der Median weist eine solche Empfindlichkeit nicht auf. Der Median ist die Mitte der nach der Größe sortierten Messwerte. Er teilt die Daten in zwei Hälften, so dass 50% der Messwerte kleiner als der Median sind und 50% darüber liegen. Da der Median die Abstände zwischen den Messwerten nicht berücksichtigt, kann er auch für ordinale Daten benutzt werden. Demgegenüber kann bei nominalen Daten nur der Modalwert herangezogen werden. Das ist der Messwert, der insgesamt am häufigsten vorkommt. Bei einigen Fragestellungen ergibt es sich, dass Mittelwert, Median und Modalwert exakt den gleichen Wert aufweisen. Dies ist jedoch nicht immer der Fall. Aus der Anordnung der drei Werte kann man Informationen über die Verteilung der Messwerte in der Stichprobe gewinnen. Bei Merkmalen,

die durch extreme Antworten verzerrt sein könnten, ist der Median eventuell eine gute Wahl, auch dann, wenn die Daten Intervallskalenniveau aufweisen. Maße der zentralen Tendenz geben einen Eindruck über die Daten, die allerdings von dieser Tendenz in der Regel auch abweichen. Diese Abweichungen werden durch ↗ Abweichungsmaße erfasst.

P-Wert. Das Ergebnis eines Signifikanztests (↗ statistische Signifikanz) ist die Wahrscheinlichkeit dafür, dass die Nullhypothese – also das Gegenteil der eigentlich in der Hypothese formulierten Aussage – zutrifft (↗ Hypothesenarten). Es wird also die Wahrscheinlichkeit dafür bestimmt, dass der vermutete Unterschied nicht besteht bzw. der vermutete Zusammenhang nicht vorliegt. Da Wahrscheinlichkeit auf Englisch *Probability* heißt, wird sie mit dem Buchstaben „p" abgekürzt. p kann jedoch grundsätzlich auf zwei verschiedene Arten berechnet werden. p kann 1-*seitig* oder auch 2-*seitig* bestimmt werden. Welche der beiden Berechnungen im Einzelfall anzugeben ist, entscheidet sich durch die Hypothese, die mit dem Signifikanztest beantwortet werden soll. Eine zweiseitige ↗ Hypothese prüft, ob ein Unterschied besteht, ohne genauere Vermutungen darüber anzustellen, in welche Richtung der Unterschied weisen könnte. Eine einseitige Fragestellung geht darüber hinaus. Sie prüft nicht nur, ob allgemein ein Unterschied besteht, sondern zudem, ob er in die erwartete Richtung geht. Der 2-seitige Wert wird also bei ungerichteten Signifikanztests angegeben. Er ist immer exakt doppelt so hoch wie der entsprechende 1-seitige Wert. Der 1-seitige Wert hat es damit „leichter" signifikant zu werden, erfordert aber die genauere Hypothese. Der P-Wert gibt also die Wahrscheinlichkeit für die Nullhypothese an. Daraus lässt sich nicht errechnen, wie hoch die Wahrscheinlichkeit für die Alternativhypothese ist. Ist der P-Wert klein, wird die Nullhypothese verworfen. Die Alternativhypothese wird als Alternative zur Nullhypothese in diesem Fall akzeptiert. Sie ist dadurch weder bewiesen noch kann man davon ausgehen, dass Eins minus dem P-Wert die Wahrscheinlichkeit für die Alternativhypothese darstellt. Es kann nämlich sehr viele Alternativhypothesen geben. Die Wahrscheinlichkeit für die Alternativhypothese kann aus logischen Gründen niemals bestimmt werden. Der P-Wert für die Nullhypothese wird in einigen Statistikprogrammen als „Signifikanz" bezeichnet. Das ist genau genommen nicht korrekt, weil die Signifikanz eben erst vorliegt, wenn der P-Wert besonders klein ist.

Regressionsanalyse. Eine Regressionsanalyse untersucht den Einfluss einer oder mehrerer unabhängiger Variablen auf eine einzige abhängige Variable. Obwohl das Messniveau der unabhängigen Variablen und die Art der Kodierung nicht beliebig sind, kann die Regressionsanalyse leicht mit unterschiedlichen Skalenniveaus für die unabhängigen Variablen umgehen. Auf diese Weise können verschiedene Variablen gleichzeitig berücksichtigt werden. Nominale oder ordinale Variablen werden z. B. als dichotome Dummy-Variablen mit den Werten 1 und 0 kodiert. Intervallskalierte unabhängige Variablen können direkt verwendet werden. Wechselwirkungen zwischen Variablen können ebenfalls berücksichtigt werden, indem die interagierenden Variablen zuvor miteinander multipliziert werden und dieses Produkt zusätzlich als neue unabhängige Variable berücksichtigt wird. Insgesamt ist die Regressionsanalyse damit sehr flexibel und vielseitig einsetzbar. Sie hat inzwischen viele klassische Tests wie den einfachen T-Test abgelöst. Eine Regressionsanalyse liefert zwei unterschiedliche Informationen. (1) Ihr Ergebnis ist eine Gleichung, mit der bei gegebenen unabhängigen Variablen der Wert der abhängigen Variable geschätzt werden kann. In diesem Sinne „lernt" die Regressionsanalyse aus den gegebenen Daten, wie der Zusammenhang zwischen den unabhängigen Variablen und der abhängigen Variable ist. Die ermittelte Regressionsgleichung kann dann verwendet werden, um aus den unabhängigen Variablen eine noch unbekannte abhängige Variable zu berechnen. Beispielsweise könnte eine Gleichung für das Krebsrisiko aus Studien abgeleitet werden, in die leicht erfassbare Variablen wie Rauchen, Alkoholkonsum, Größe, Gewicht, Alter, Geschlecht usw. eingehen. Die Verwendung von Regressionsgleichungen zur Vorhersage abhängiger Variablen erfordert eine besondere Überprüfung der Regressionsmodelle. Wie gut ein Modell insgesamt funktioniert, kann z. B. durch die Gesamtkorrelation R angegeben werden. Sie gibt an, wie stark die Vorhersage mit den tatsächlichen Werten der abhängigen Variablen korreliert. Wenn R hoch ist und nahe bei eins liegt, ist das Modell perfekt. Diese Perfektion ist jedoch nicht immer das Ziel einer Regressionsanalyse. (2) Die Regressionsanalyse kann auch verwendet werden, um den Zusammenhang zwischen den unabhängigen Variablen und der abhängigen Variablen statistisch zu testen. Hier steht nicht die Vorhersage, sondern der statistische Test im Vordergrund. Der Vorteil der Regressionsanalyse liegt darin, dass alle unabhängigen Variablen gleichzeitig untersucht und in ihrer Wechselwirkung berücksichtigt werden. Wenn z. B. in einer Stichprobe die befragten Unternehmerinnen im Durchschnitt älter sind als die befragten abhängig Beschäftigten, kann es sein, dass ein mit dem T-Test

ermittelter Lohnunterschied nicht auf die Tätigkeit, sondern auf das Alter und die Berufserfahrung zurückzuführen ist. Der T-Test berücksichtigt diese gleichzeitig wirkenden Zusammenhänge nicht. In der Regressionsanalyse mit den beiden Variablen Alter und Tätigkeit (selbständig vs. unselbständig) kann dann das Alter möglicherweise eine Signifikanz aufweisen und die Art der Tätigkeit keine Signifikanz. Da bei der Regressionsanalyse mehrere unabhängige Variablen gleichzeitig berücksichtigt werden, sind die Ergebnisse nicht immer einfach zu interpretieren. Die Signifikanz einer Variable kann nur interpretiert werden, wenn berücksichtigt wird, dass gleichzeitig andere Variablen im Regressionsmodell verwendet werden. Wird z. B. das Alter nicht in die Analyse einbezogen, ergeben sich völlig andere Ergebnisse. Die Regressionsanalyse ist immer nur so gut wie ihre Modellannahmen, d. h. die Annahmen, welche Variablen überhaupt einbezogen werden. Dies mag gegen diese Form der Analyse sprechen. Aber das wäre voreilig. Denn die Welt ist nun einmal ein Geflecht aus vielen Variablen und die Regressionsanalyse geht in die richtige Richtung, indem sie diese zusammen betrachtet. Sie ist daher einer einfachen Statistik, die nur zwei Untersuchungsgruppen vergleicht, vorzuziehen. Dass die Regressionsanalyse schwieriger zu interpretieren ist, liegt also eher an der Kompliziertheit der „realen Welt“ als an der Methode. Scheinbar paradoxe Effekte treten z. B. auf, wenn Variablen signifikant werden, die bei genauer Betrachtung gar nicht direkt mit der abhängigen Variable korrelieren. So hat z. B. der Kaffeekonsum keinen Einfluss auf den IQ, wird aber signifikant, wenn der Kaffee in die Regressionsgleichung aufgenommen wird. Kaffee als Ursache für übermäßige Nervosität beim IQ-Test kann die Ergebnisse verzerren und wird daher signifikant. Eine solche Variable wird als Suppressorvariable bezeichnet. Es gibt verschiedene Arten von Regressionsmodellen. Die korrekte Variante wird in der Regel von der Art der abhängigen Variable bestimmt. Ist die abhängige Variable intervallskaliert, wird die klassische lineare multiple Regression durchgeführt, für dichotome abhängige Variablen (ja/nein) wird die binär-logistische Regression verwendet, für die Zeit bis zum Eintreten eines Ereignisses die Cox-Regression. Nichtlineare Zusammenhänge können durch vorherige mathematische Transformationen der unabhängigen Variablen überprüft werden. Die theoretische Planung einer Regressionsanalyse kann recht aufwendig sein. Zum Beispiel müssen mehrere Variablen gleichzeitig berücksichtigt werden. Es werden daher auch Verfahren vorgeschlagen, die aus einem Pool von Variablen schrittweise und automatisiert die signifikanten Variablen auswählen bzw. die nicht signifikanten Variablen schrittweise ausschließen.

Standardabweichung, Streuung, Varianz. Die Standardabweichung, auch Streuung genannt, ist – vereinfacht gesprochen – ein Wert für die mittlere Abweichung der Messwerte vom Mittelwert (ohne Berücksichtigung der Abweichungsrichtung). Die Standardabweichung bzw. Streuung gibt damit einen Eindruck von der Variationsbreite der Antworten und damit zum Teil auch über die Messgenauigkeit. Bei ideal normalverteilen Messwerten liegt der ↗ Mittelwert zusammen mit dem ↗ Median und dem ↗ Modalwert exakt in der Mitte der Messwerteverteilung. Insgesamt 68% aller Antworten befinden sich dann in dem Messwertebereich zwischen dem Mittelwert minus der Streuung und dem Mittelwert plus der Streuung. Beispiel: Ein IQ-Test weist in der Regel einen Mittelwert von 100 und eine Streuung von 10 auf. Damit liegen 68% aller Menschen mit ihrem IQ zwischen einem IQ von 90 und 110. Die Varianz ist das Quadrat der Streuung bzw. Standardabweichung.

Statistische Signifikanz. (Statistische Bedeutsamkeit) Jeder im Rahmen einer Messung gewonnene Messwert ist mit einem gewissen Fehler behaftet. Die Ergebnisse einer Befragung sind daher nie exakt. Die Genauigkeit einer Messung kann in vielen Fällen mit Hilfe der Wahrscheinlichkeitsrechnung angegeben werden. In diesem Sinne bezeichnet z. B. die ↗ Streuung die Schwankungsbreite der Messwerte um den Mittelwert. Wenn nun zwei Kennwerte verglichen werden sollen, z. B. der Mittelwert des Gehalts für selbstständig tätige Menschen mit dem Mittelwert des Gehalts für Unselbstständige, so muss immer auch mitbedacht werden, dass beide Messwerte ungenau sind. Ein per Augenschein sichtbarer Unterschied in den Mittelwerten bedeutet nicht automatisch, dass sich die beiden Untersuchungsgruppen tatsächlich unterscheiden. Denn dieser Unterschied könnte auf Messfehler und natürliche Schwankungen innerhalb der Untersuchungsgruppen zurückzuführen sein. Ein statistischer Signifikanztest beantwortet die Frage, ob ein per Augenschein sichtbarer Unterschied zwischen zwei Kennwerten (z. B. Mittelwerten) durch Messungenauigkeiten, Fehlerschwankungen etc. erklärt werden kann. Erst wenn die Wahrscheinlichkeit dafür, dass ***kein Unterschied*** vorliegt, gering ist und unter der vorher festgelegten Signifikanzgrenze (in der Regel 5%, ↗ Alpha-Fehler) liegt, sagt man, dass die Unterschiede statistisch signifikant sind. D.h., dass ein statistischer Signifikanztest niemals behaupten würde, dass ein Unterschied tatsächlich besteht. Statistisch signifikant heißt nur, dass es ***unwahrscheinlich (aber nicht unmöglich) ist, dass kein Unterschied besteht.*** Je

nach erhobenen Daten müssen verschiedene Verfahren für die Signifikanzprüfung angewandt werden. Wichtige Testverfahren sind z. B.: ↗ T-Test, ↗ Fishers exakter Test, ↗ Chi-Quadrat-Test, ↗ Mann-Withney-U-Test, ↗ Varianzanalyse, Signifikanzprüfung einer ↗ Korrelation. Das wichtigste Ergebnis eines Testes ist die Wahrscheinlichkeit (↗ P-Wert) dafür, dass sich die Kennwerte nicht unterscheiden. Diese Wahrscheinlichkeit wird mit einem vorher festgelegten Grenzwert, der Signifikanzgrenze (↗ Alpha-Fehler) verglichen.

T-Test. Ein meistverwendete Signifikanztest für den Vergleich von zwei Mittelwerten ist der T-Test (↗ statistische Signifikanz). Der T-Test besitzt jedoch einige Voraussetzungen, die erfüllt sein müssen, damit er berechnet werden kann. Diese Voraussetzungen sind nicht immer erfüllt. Zu den Grundvoraussetzungen gehört u.a., dass mit gutem Gewissen ein ↗ Mittelwert und die dazu gehörige ↗ Streuung berechnet werden können. Die Verteilung der Mittelwerte muss einer Normal- bzw. T-Verteilung folgen, was bei kleinen Stichproben Probleme machen kann. Bei Stichproben mit einer Gruppengröße von mindestens 25 bis 50 Personen pro Untersuchungsgruppe, liegt automatisch eine Normalverteilung der Mittelwerte vor (↗ Zentraler Grenzwertsatz), so dass dann kein Problem bei der Anwendung des T-Tests besteht. Der T-Test berechnet einen t-Wert, für den zusammen mit den sog. Freiheitsgraden (in der Regel: Zahl der Messwerte minus eins) die Wahrscheinlichkeit bekannt ist. Die Wahrscheinlichkeit ist das Ergebnis des Tests. Man spricht von einer ↗ statistischen Signifikanz, wenn diese Wahrscheinlichkeit kleiner als der vorher festgelegte ↗ Alpha-Fehler ist.

Validität, prognostische. Wenn Personalauswahlverfahren eingesetzt werden, erhofft man sich von ihnen Hinweise, die es tatsächlich ermöglichen, unter den Bewerberinnen und Bewerbern die am besten geeigneten Kandidatinnen und Kandidaten zu finden. Die Verfahren sollen also im weitesten Sinne „Eignung" feststellen. Ob ein Verfahren tatsächlich das misst, was es zu messen vorgibt, hier die „Eignung", wird als *Validität* des Verfahrens bezeichnet. Zur Feststellung der Validität wird in der Regel eine ↗ Korrelation zwischen den Ergebnissen des eingesetzten Verfahrens und passender Außenkriterien (z. B. Leistungsbeurteilung durch Vorgesetzte) berechnet. Damit ist die Validität quantifizierbar mit Werten zwischen Null und Eins, wobei hohe Werte einer hohen Validität entsprechen. Da es bei der Perso-

nalauswahl darum geht, die Eignung zu prognostizieren und als passende Außenkriterien Merkmale in Frage kommen, die in der Zukunft liegen, spricht man von einer prognostischen Validität, also von der Fähigkeit des eingesetzten Verfahrens, Vorhersagen über die Verwendbarkeit einer Bewerberin eines Bewerbers zu erstellen. Wie hoch die Validität im Idealfall sein soll, hängt vom Einsatzziel (z. B. von der Anzahl der wahrscheinlich ohnehin geeigneten Bewerberinnen und Bewerbern: sind wahrscheinlich ohnehin alle für die Stelle geeignet, kann die Auswahl einfach gehalten werden) und vom Aufwand (Kosten vs. Nutzen) ab. Eine hohe Validität wird Verfahren mit einem Wert über 0,3 zugesprochen. Hierzu gehört z. B. das Assessment Center, wohingegen Bewerbungsunterlagen, Schulnoten und graphologische Gutachten deutlich darunter liegen.

Varianzanalyse (englische Abkürzung: ANOVA). In der Regel sind Signifikanztests in der Lage nur zwei Kennwerte (z. B. Mittelwerte) miteinander zu vergleichen. Einige Fragestellungen machen daher mehrere Vergleiche zwischen jeweils zwei Messwerten nötig, um die Frage insgesamt beantworten zu können. Beantworten drei Personengruppen einen Fragebogen (Gruppe A, B, C), so kommt man auf insgesamt drei paarweise Vergleiche (A mit B; A mit C und B mit C). Obwohl es hier möglich ist, jede Kombination der Gruppen einzeln zu vergleichen und eine Alpha-Fehler-Adjustierung vorzunehmen (↗ Alpha-Fehler-Adjustierung), ist eine Varianzanalyse eleganter und weniger aufwändig zu rechnen. Die Varianzanalyse löst das Problem durch einen Trick: Es werden im Wesentlichen zwei Varianzen (↗ Standardabweichung, Streuung, Varianz) ermittelt und diese mit einem F-Test verglichen. Es werden also auch hier nur zwei Kennwerte (hier Varianzen) durch den Test verglichen. Die eine Varianz ist die innerhalb der Gruppen, die andere ist die zwischen den Gruppen. Sind die Unterschiede (ermittelt durch die Varianz) zwischen den Gruppen größer als die Unterschiede innerhalb der Gruppen, so unterscheiden sich die Gruppen. Allerdings ist dann noch nicht bekannt, welche der Gruppen sich voneinander unterscheiden. Um dies herauszufinden werden anschließend paarweise Vergleiche durchgeführt. Für eine Varianzanalyse werden klar abgegrenzte Untersuchungsgruppen (↗ Kategorien) benötigt. Die Regressionsanalyse ist eine flexiblere Alternative, wenn solche Kategorien nicht vorliegen.

Zentraler Grenzwertsatz. Ist eine Untersuchungsstichprobe groß, so ergibt sich unabhängig von der Verteilung der Rohdaten für den Mittelwert eine Normalverteilung. Diesen Zusammenhang kann man sich wie folgt vorstellen: Es wird aus einer größeren Stichprobe eine begrenzte Zufallsauswahl getroffen und für diese Zufallsauswahl ein Mittelwert berechnet. Dies wird mehrfach wiederholt. Jeder berechnete Mittelwert beruht dann nur auf einer Zufallsauswahl und stimmt damit mit dem echten Mittewert nur mehr oder weniger gut überein. Es zeigt sich, dass die Mittelwerte der Zufallsauswahlen um den echten Mittelwert normalverteilt streuen und zwar unabhängig von der Verteilung der eigentlichen Rohwerte. Testverfahren wie der ↗ T-Test oder die ↗ Varianzanalyse benötigen solche normalverteilten Mittelwerte. Diese sind nach dem zentralen Grenzwertsatz für große Stichproben immer gegeben. Was groß ist und was nicht hängt vom jeweiligen Lehrbuch ab. Einige sagen, dass 25 Personen pro Untersuchungs-Gruppe genügen, andere fordern 30 und ganz strenge sogar 50 Messwerte. Da für kleine Stichproben der zentrale Grenzwertsatz nicht gilt, kann eine Normalverteilung der Mittelwerte nur dann erwartet werden, wenn auch die Rohwerte normalverteilt sind. Dies muss für kleine Stichproben geprüft werden. Bei großen Stichproben ist eine solche Prüfung irreführend und sollte unterbleiben.

6 Darstellung und Abkürzungen

Die folgenden Tabellen zeigen übliche Abkürzungen und ihre Bedeutung. Dabei werden auch Beispiele für die Darstellung von Ergebnissen im Text und Tabellen angeführt.

AM oder M oder $\bar{x}$	Mittelwert (arithmetisches Mittel; Mean)
SD oder s oder Std.	Standardabweichung (Standard Deviation; Streuung)
MD	Median
IQR	Interquartilsabstand
df oder FG	Freiheitsgrade (degrees of freedom)
N	Anzahl bzw. Größe einer Grundgesamtheit oder Stichprobe.
n	Anzahl bzw. Häufigkeit von Untersuchungsobjekten mit einer bestimmten Eigenschaft.
P	Wahrscheinlichkeit (kann Werte zwischen 0 und 1 annehmen. 0,6 bedeutet also eine Wahrscheinlichkeit von 60%).
p-2-seitig	Wahrscheinlichkeit dafür, dass etwas nicht signifikant ist (2-seitig getestet).
p-1-seitig	Wahrscheinlichkeit dafür, dass etwas nicht signifikant ist (1-seitig getestet).
*	Der Unterschied ist signifikant bei einem Alphafehler von 5% (p <= 0,05).
**	Der Unterschied ist hoch signifikant bei einem Alphafehler von 1% (p <= 0,01).

Tabelle 4: Allgemein gebräuchliche Abkürzungen

Es kann nie schaden, Abkürzungen im Text einzuführen oder unter Tabellen zu erklären. Allerdings werden einige Abkürzungen auch in guten Zeitschriften nicht mehr erklärt, weil sie als üblich vorausgesetzt werden.

Im Text werden Besonderheiten hervorgehoben. Die Tabelle sollte nicht nacherzählt werden, aber einige auffällige Aspekte sollten besprochen werden. In der Regel wird auf Aspekte eingegangen, die die Stichprobe anschaulich beschreiben: „... Die meisten befragten sind Frauen (62 %, n = 235) und mit rund 66 % (n = 250) ist der größte Teil der Stichprobe berufstätig. Das Durchschnittsalter liegt bei 47,3 Jahren (SD = 10) ...“

Tabelle 1: Deskriptive Ergebnisse

	AM bzw. %	SD bzw. n	N
Frauen [1 vs. Männer]	62,01%	235	379
Männer [0 vs. Frauen]	37,99%	144	379
Alter [Jahre]	47,32	10,02	377
Berufstätig [1/0]	65,96%	250	379
Berufserfahrung [Jahre]	15,32	12,28	378
Skala 1: Neurotizismus [1–6]	5,25	1,32	372
Skala 2: Gewissenhaftigkeit [1–6]	4,98	1,04	372

Tabelle 1 (Alternative): Deskriptive Ergebnisse

	AM bzw. %	SD bzw. n	N	1.	2.	3.	4.	5.
1. Frauen [1 vs. Männer 0]	62,01%	235	379					
2. Alter [Jahre]	47,32	10,02	377	,062				
3. Berufstätig [1/0]	65,96%	250	379	-,223**	,527**			
4. Berufserfahrung [Jahre]	15,32	12,28	378	-,152**	,875**	,254**		
5. Skala 1: Neurotizismus [1–6]	5,25	1,32	372	,004	,002	-,354**	-,214**	
6. Skala 2: Gewissenhaftigkeit [1–6]	4,98	1,04	372	,006	,004	,257**	,325**	,078

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Tabelle 5: Darstellung deskriptiver Ergebnisse

Es ist üblich als erste Ergebnistabelle einen Überblick über alle erhobenen Daten mittels deskriptiver Statistiken zu präsentieren. Gerade gute Zeitschriften bevorzugen dabei Tabellen in denen möglichst viele Informationen untergebracht werden. So kombinieren beide Tabellen deskriptive Ergebnisse für Intervallskalen (AM und SD) mit Ergebnissen für Nominalskalen (% und n) in den gleichen Spalten. Das ist schwerer zu lesen, aber es ermöglicht die Gesamtdarstellung aller relevanten Variablen. In eckigen Klammern wird zudem der Messbereich bzw. die Maxeinheit angegeben. Denn sonst wäre nicht klar, was die Zahlen eigentlich inhaltlich bedeuten. In vielen Sozialwissenschaften (z. B. Psychologie und BWL) wird auch eine vollständige Korrelationstabelle verlangt, die gerne auch in diese erste Ergebnisstabelle integriert wird. Mitunter wird hier ein Querformat nötig. Auf auffällige Ergebnisse wird im Text eingegangen. Die Korrelationen werden nur bei Bedarf diskutiert. Alle Zahlen wurden für die Beispieltabellen frei erfunden.

r	Korrelation.
p-2-seitig	Wahrscheinlichkeit dafür, dass eine Korrelation Null ist.
*	Die Korrelation ist mit einem Alphafehler (einer Wahrscheinlichkeit) von 5% (p <= 0,05) Null. Die Korrelation ist signifikant.
**	Die Korrelation ist mit einem Alphafehler (einer Wahrscheinlichkeit) von 1% (p <= 0,01) Null. Die Korrelation ist hoch signifikant.

Im Text: „... Es besteht also ein hoch signifikanter Zusammenhang zwischen den Sonnenstunden im August und der Menge der verkauften Sonnenkreme (r = 0,52; p-2-seitig = 0,001) ...“

Tabelle 1: Interkorrelationsmatrix der Skalen

N = 506	**Skala 1**	**Skala 2**	**Skala 3**	**Skala 4**	**Skala 5**
Skala 2	0,758 **				
Skala 3	0,924 **	0,723 **			
Skala 4	0,815 **	0,589 **	0,292 *		
Skala 5	0,810 **	0,491 **	0,587 **	0,517 **	
Skala 6	0,849 **	0,599 **	0,062	0,706 **	0,562 **

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.
Skala 1: Neurotizismus
Skala 2: Offenheit für neue Erfahrungen
Skala 3: Extraversion
Skala 4: Selbstdarstellung
Skala 5: Führungsmotivation

Tabelle 6: Darstellung von Korrelationen

Werden Korrelationen im Text angesprochen, wird an geeigneter Stelle eine Klammer gesetzt, die die üblichen, relevanten Kennwerte enthält. Die Abkürzungen r und p-2-seitig etc. werden nicht gesondert erklärt. Sie werden üblicherweise als bekannt vorausgesetzt. In vielen wissenschaftlichen Zeitschriften werden Tabellen mit Überschriften über der Tabelle verwendet. Typische Korrelationstabellen haben ungefähr den Aufbau, wie er hier gezeigt wird.

t	Prüfgröße für den T-Test.
df	Freiheitsgrade (degrees of freedom).
p-2-seitig	Wahrscheinlichkeit dafür, dass zwei Mittelwerte sich nicht signifikant unterscheiden (2-seitig getestet).
p-1-seitig	Wahrscheinlichkeit dafür, dass zwei Mittelwerte sich nicht signifikant unterscheiden (1-seitig getestet).
*	Der Unterschied ist signifikant bei einem Alphafehler von 5% (p <= 0,05)
**	Der Unterschied ist hoch signifikant bei einem Alphafehler von 1% (p <= 0,01)

Im Text: „... Es besteht ein hoch signifikanter Unterschied zwischen dem Verhalten der beobachteten Fußballfans der beiden Vereine (t = 3,52; df = 255; p-2-seitig = 0,003) ...“

Tabelle 2: T-Test für die Unterschiede zwischen den Fans

	Fanclub A			Fanclub B					
	AM	SD	n	AM	SD	n	t	df	p
Skala 1	5,25	1,32	500	6,00	1,12	420	2,57	918	0,004**
Skala 2	4,98	1,04	499	5,98	1,20	420	2,62	917	0,003**

** Die Unterschiede sind auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Unterschiede sind auf dem Niveau von 0,05 (2-seitig) signifikant.
Skala 1: Neurotizismus [1–6]
Skala 2: Aggressivität [1–6]

Tabelle 2 (Alternative): T-Test für die Unterschiede zwischen den Fans

	Fanclub A		Fanclub B				
	AM (SD)	n	AM (SD)	n	t	df	p
Skala 1	5,25 (1,32)	500	6,00 (1,12)	420	2,57	918	0,004**
Skala 2	4,98 (1,04)	499	5,98 (1,20)	420	2,62	917	0,003**

** Die Unterschiede sind auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Unterschiede sind auf dem Niveau von 0,05 (2-seitig) signifikant.
Skala 1: Neurotizismus [1–6]
Skala 2: Aggressivität [1–6]

Tabelle 7: Darstellung von T-Tests

Werden T-Tests im Text angesprochen, wird an geeigneter Stelle eine Klammer gesetzt, die die üblichen, relevanten Kennwerte enthält. Die Abkürzungen für diese Kennwerte werden in der Regel nicht gesondert erklärt. Sie werden als bekannt vorausgesetzt. Typische Tabellen für Vergleiche von Mittelwerten stellen diese nebeneinander dar und zeigen dann ob es Unterschiede gibt. Die Spalten für t und df werden inzwischen bei einigen Zeitschriften eingespart. Beide Kennwerte sind für den T-Test zwar relevant, aber wenn es letztlich um die Wahrscheinlichkeit (ganz rechts) geht, können sie auch weggelassen werden.

7 Literaturverzeichnis

Aristoteles (2019/4. Jhdt. v. Chr.) *Methaphysik. Ins Deutsche übertragen und eingeleitet von Adolf Lasson*. Grafrath: Boer Verlag

Bergkvist, L. & Rossiter, J. R. (2007) The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of marketing research*, 44 (2), 175-184

Binet, A., Simon, T. & Vaney, X. (2003/1905) Recherches de pédagogie scientifique. *Monitor on Psychology*, 34 (2), 233-274

Bortz, J. & Döring, N. (2002) *Forschungsmethoden und Evaluation*. Berlin, Heidelberg: Springer

Bortz, J., Lienert, G., A. & Boehnke, K. (2000) *Verteilungsfreie Methoden in der Biostatistik*. Berlin: Springer

Cohen, J. (1992) A Power Primer. *Psychological Bulletin*, 112 (1), 155-159

Einstein, A. (2002/1918-1921) *Geometrie und Erfahrung*. Princeton: Princeton University Press

Galilei, G. (1953/1623) *Il Saggiatore*. Wikisource, https://it.m.wikisource.org/wiki/Il_Saggiatore – Abgefragt am: 08.12.2022.

Goodman, S. (2008) A dirty dozen: twelve p-value misconceptions. *Seminars in hematology*, 45 (3), 135-140

Herrmann, D. (2014) *Die antike Mathematik. Geschichte der Mathematik in Alt-Griechenland und im Hellenismus*. Berlin: Springer Spektrum

Klein, I. (2004) *Skalentypen und Statistik: ein Kommentar zu Velleman & Wilkinson (1993)*. Diskussionspapier,

Krey, O. (2012) *Zur Rolle der Mathematik in der Physik. Wissenschaftstheoretische Aspekte und Vorstellungen Physiklernender*. Berlin: Logos-Verlag

Likert, R. (1932) A Technique for the Measurements of Attitudes. *Archives of Psychology*, 140 (22), 5-55

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F. & Turkheimer, E. (2012) Intelligence: new findings and theoretical developments. *American Psychologist*, 67 (2), 130-160

Norman, G. (2010) Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15, 625-632

Rossiter, J. R. (2002) The C-OAR-SE procedure for scale development in marketing. *International journal of research in marketing*, 19 (4), 305-335

Stevens, S. S. (1946) On the theory of scales of measurement. *Science*, 103 (2684), 677-680

Strunk, G. (2019) *Leben wir in einer immer komplexer werdenden Welt? Methoden der Komplexitätsmessung für die Wirtschaftswissenschaft.* Wien: Complexity-Research

Strunk, G. (2022) *Das Verfassen einer wissenschaftlichen Abschlussarbeit. Hinweise zu Themenfindung und Form. Allgemeine Version.* https://www.complexity-research.com/pdf/Seminare/Wiss_Arb_Allgemein_MK.pdf – Abgefragt am: 04.10.2023.

Suppes, P. (1957) *Introduction to Logic. Princeton: D. van Norstrand Co*: Inc

Thomas, M. (2019) Mathematization, not measurement: A critique of Stevens' scales of measurement. *Journal of Methods and Measurement in the Social Sciences*, 10 (2), 76-94

Thomson, W. (1889) Electrical units of measurement. In: Thomson, W. (Hrsg.) *Popular lectures and addresses. Nature Series. Vol. I. Constitution of matter*. London: Macmillan, S. 73-136

Thurstone, L. L. (1927) A Law of Comparative Judgment. *Psychological Reviews*, 34, 273-286

Westermann, R. (1985) Empirical tests of scale type for individual ratings. *Applied Psychological Measurement*, 9 (3), 265-274