# Statistik leicht gemacht

Guido Strunk

# Statistik leicht gemacht

[Arbeitstitel]

Complexity-Research, Forschung & Lehre, Verlag

**Statistik leicht gemacht**
[Arbeitstitel]

ISBN 978-3-903291-[\ \]



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 (1879–1955) 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 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 sinnlos. Da man den Zahlen selbst nicht ansieht, wofür sie stehen und was daher eine sinnvolle mathematische Behandlung ausmacht, ergeben sich hier einige gravierende Schwierigkeiten.

Zusammenfassend kann gesagt werden, dass bei einer Messung empirische Gegebenheiten durch Zahlen repräsentiert werden. Ziel ist es, die Unterschiede, 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. Gemeint ist, dass Messdaten nach der Messung durchaus noch verändert wer-

den 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<br>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<br>Beispiel: Alter |
| Verhältnis | Verhältnisse zwischen den Zahlen (erlaubt ist die Multiplikation/Division mit Konstanten) | Verhältnisse zwischen den Zahlen<br>Beispiel: Gehalt |
| Absolut | Nichts darf verändert werden. | Verhältnisse zwischen den Zahlen, Kardinalzahl<br>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 Sor-

tierung 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 möglich. Wenn ich frage, wie viele Buchstaben dieser Satz hat, dann ist die Antwort eine ganz bestimmte Zahl, und diese Zahl kann man nicht mehr sinnvoll in andere Einheiten umrechnen.

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. 19) stellt die zulässigen und üblichen Methoden dar, um diese Daten deskriptiv auszuwerten.

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 empirischen Sachverhalte zueinander interpretiert werden. Beispiel: höchster Bildungsabschluss: Pflichtschule = 1, Matura/Abitur = 2, Studium = 3, ... Höhere Bildungsabschlüsse erhalten höhere Zahlen. Da aber die genaue Höhe der Zahlen keine Bedeutung hat, 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 Eigenschaft der Zahlen, größer oder kleiner sein zu können. 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 der Daten (vgl. Tabelle 2, S. 19).

Das nächst höhere Skalenniveau ist das der *Intervallskala*: Die Abstände zwischen den Zahlen können bei einer Intervallskala sinnvoll interpretiert werden. Dies gilt jedoch nicht für Zahlenverhältnisse. Beispiel: Alter gemessen in Jahren. Wenn eine Person 2 Jahre älter ist als eine andere, dann bleibt das so, auch wenn Zeit vergeht. Das Intervall, der Abstand, beträgt 2 Jahre. Wenn aber eine Person in einem Moment genau doppelt so alt ist wie eine andere, ist sie es am nächsten Tag oder in der nächsten Stunde oder Minute schon nicht mehr. Das Zahlenverhältnis ist also nicht sinnvoll interpretierbar. Es ändert sich mit jeder Sekunde, die vergeht. Nachträglich kann eine Intervallskala durchaus noch in andere Maßzahlen umgerechnet werden, wenn die Abstände dabei nicht zerstört werden und interpretierbar bleiben. So kann das 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. 19).

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). Die nächsthöhere ist die *Verhältnisskala*: Zahlenverhältnisse können hier erstmals sinnvoll interpretiert werden. Beispiel: 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 verstrichenen 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 ver-

schoben 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. 19).

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.

## 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 aber auch wenig Sinn, Kennwerte zu berechnen, die dann 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 erkennbar, ob es sich um eine Nominalskala (die Zahlen sind Be-

zeichnungen 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. 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.

| Beispiele | Art der Erhebung (Beispiele) | Niveau |
|---|---|---|
| Alter | Alter:_____<br>Geburtsjahr:_____<br>❍ 1 10-15 Jahre<br>❍ 2 16-25 Jahre<br>❍ 3 26-35 Jahre<br>❍ 4 älter als 35 Jahre | Intervall<br>Ordinal |
| Geschlecht | ❍ 0 männlich<br>❍ 1 weiblich<br>❍ 2 divers | Nominal |
| Beruf | ❍ 1 Arbeitslosigkeit<br>❍ 2 Arbeitsunfähigkeit<br>❍ 4 Unselbstständig erwerbstätig<br>❍ 5 Selbstständig erwerbstätig<br>❍ 6 Studium<br>❍ 7 Ausbildung<br>❍ 8 Hausfrau / Hausmann<br>❍ 9 Rente / Pension<br>❍ 10 Wehr- / Zivildienst<br>❍ –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 sogar empirisch nachgewiesen werden.

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

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 in Englisch 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.

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 Fragestellungen 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 Ra-

tingskalen als das zentrale Forschungsinstrument in den Sozialwissenschaften, und es ist durchaus üblich, dass mehrere Ratings durch Mittelwerte zusammengefasst werden (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 tatsächlich kein Intervallskalenniveau vorliegt. 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 daher häufig auf Methoden angewiesen, die eigentlich Intervallskalenniveau voraussetzen.

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 nicht die Abstände zwischen den Zahlen. Gehälter haben aber durchaus interpretierbare Abstände. Auch die Zahlenverhältnisse sind hier interpretierbar. Man könnte also sehr genaue Mittelwerte bestimmen. 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 wenigen, 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 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 ihnen, 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, aber keine Zwischenstufe (vielleicht) vorgesehen ist. Wenn es nur zwei diskrete Möglichkeiten gibt, werden diese Messungen auch als dichotome (griechisch) oder binäre (lateinisch) Messungen 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.]

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*<br>*1: weiblich*<br>*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 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, den seltensten Wert ebenfalls 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. 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)<br><br>*Bei 1/0-Kodierung der dichotomen Variable ist die Produkt-Moment-Korrelation identisch mit der Punktbiserialen Korrelation.* | Rangkorrelation (Spearman)<br><br>*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).<br><br>*Bei 1/0-Kodierung der dichotomen Variablen ist die Produkt-Moment-Korrelation identisch mit Phi.* | Biseriale Rangkorrelation (Alternativ: U-Test) |
| Ordinal | | | Rangkorrelation<br><br>*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 tatsächlich korreliert. Das würde jedoch nicht bedeuten, dass die Störche die Kinder bringen. 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.

**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 interpretiert 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 Ver-

hältnisskala 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 berück-

sichtigt 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 Regressionsana-

lyse 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 da-

mit 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 Personalauswahl 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.

# 4 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 $\overline{x}$ | 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.

# 5 Literaturverzeichnis


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

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.

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

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

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

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

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