MARIUS BARTH

FORSCHUNGSMETHODEN IM BIOLOGIEUNTERRICHT

Inhalt

Vorwort 5

1 *Messen* 7

1.1 *Skalen und Skalenniveaus* 9

1.2 *Messbarkeit und Messfehler* 10

Aufgaben 10

2 *Beschreiben und Zusammenfassen* 11

2.1 *Häufigkeitsverteilungen* 11

2.2 *Statistische Kennwerte* 12

Aufgaben 14

3 *Wahrscheinlichkeitstheorie und inferenzstatistische Grundlagen* 17

3.1 *Wahrscheinlichkeitsbegriffe* 18

3.2 *Wahrscheinlichkeits- und Wahrscheinlichkeitsdichteverteilungen* 19

3.3 *Stichprobenkennwerteverteilung* 21

Aufgaben 21

4 *Hypothesen und Hypothesentest* 23

4.1 *Statistisches Hypothesenpaar* 23

4.2 *Entscheidungsfehler* 24

Aufgaben 25

5 Zusammenhänge zwischen zwei Variablen 27

5.1 Andere Formen von Zusammenhängen 29

5.2 Regression 30

6 Der Einfluss von Drittvariablen 33

6.1 Das Simpson-Paradox 33

6.2 Kontrolle von Drittvariablen 36

Lösungen 37

Kapitel 1: Messen 37

Kapitel 2: Beschreiben und Zusammenfassen 38

Vorwort

Diese Materialien sollen die im Rahmen der Lehrveranstaltung *Didaktische Forschungsprojekte* vermittelten Grundkenntnisse zur Statistik in leicht verständlicher Form zusammenfassen.

1
Messen

Bei einer Datenerhebung geht es darum, bestimmte *Merkmale* – Eigenschaften der untersuchten Merkmalsträger (Personen, Gruppen, Objekte, ...) – zu erfassen. Die individuellen *Ausprägungen* der Merkmale (die Werte der Variablen) werden dann zu einem Datensatz zusammengeführt und gespeichert. Weil diese Merkmale zwischen den Merkmalsträgern variieren können, spricht man auch von *Variablen*. (Unveränderliche Größen nennt man *Konstanten*.) Damit man die Merkmalsausprägungen in einem Datensatz speichern und weiter verarbeiten kann, muss man sie zunächst in symbolische Zeichen (z.B. Zahlen) "übersetzen". Diesen Vorgang nennen wir *Messen*.

Aus dem Alltag ist uns das Messen von vielen Merkmalen vollkommen vertraut: Wenn wir die Körpergröße einer Person wissen wollen, nehmen wir z.B. ein Maßband zur Hand, messen sie kurz und schreiben uns die Körpergröße in Zentimetern oder Metern auf. Wenn wir unser Körpergewicht kennen wollen, wiegen wir uns auf einer Personenwaage und notieren das Gewicht in Kilogramm. Bei diesen Merkmalen ist uns also schon eine Übersetzung von Merkmalsausprägungen in Zahlen geläufig.

Bei anderen Merkmalen ist es aber gar nicht so einfach, einer bestimmten Ausprägung auch einen ganz bestimmten Zahlenwert zuzuordnen. Würden wir z.B. eine Schulklasse nach ihren Lieblingsgerichten fragen, bekämen wir Antworten wie "Pizza", "Mangoldquiche" oder "Reis mit Ketchup", aber keine Zahlen, die wir uns ohne weiteres notieren können. Um aber dennoch auch solche Merkmale messen zu können, benötigen wir also eine Übersetzung von diesen Begriffen in Zahlen. Es gibt eine Reihe von Regeln, die einem dabei helfen eine möglichst sinnvolle Übersetzung anzufertigen. Das Ziel dieser Regeln ist es, dass sich die *Beziehungen*, die zwischen den Ausprägungen des Merkmals herrschen, auch in den Beziehungen zwischen den Zahlen wiederfinden (man spricht von den *erhaltenen Relationen*).

1. Gleichheit und Verschiedenheit: Wenn sich die Ausprägungen eines Merkmals unterscheiden, sollen sich auch die Zahlen, in die man sie übersetzt, unterscheiden (z.B. bekommen alle Schüler, die "Pizza" gesagt haben, eine 1, all jene, die "Mangoldquiche" gesagt haben, aber eine 2). Wenn die Ausprägungen aber gleich sind, dann sollen auch die zugeordneten Zahlen gleich sein (z.B. bekommen wirklich alle Schüler, die "Pizza" gesagt haben, auch eine 1).
2. Ordnung: Manchmal lassen sich die Ausprägungen eines Merkmals in eine sinnvolle Reihenfolge bringen, indem man die Ausprägungen nach ihrer Größe, Stärke oder Intensität ordnet. Würde man z.B. einen einzelnen Schüler zu mehreren Gerichten befragen, wie gerne er diese mag, und ihn seine Antworten "auf einer Skala von 1 bis 10" geben lassen (mit einer 10 als bestem Wert), könnte man sich ziemlich sicher sein, dass eine 10 eine größere Vorliebe wiederspiegelt als eine 9. Ob aber der Unterschied zwischen einer 9 und einer 10 *genauso groß* ist wie der Unterschied zwischen einer 8 und einer 9, könnte man nicht sicher sagen. Ein anderes Beispiel für ordinalskalierte Variablen sind Rangfolgen, also z.B. die Gold-, Silber- und Bronzemedaillen bei den olympischen Spielen. Wir wüssten, dass eine Goldmedaille eine bessere Leistung als eine Silbermedaille anzeigen soll; vielleicht war aber der Leistungsunterschied zwischen Gold und Silber viel kleiner (oder viel größer) als der zwischen Silber und Bronze.
3. Größe der Verschiedenheit: Zusätzlich zu der Ordnung der Merkmalsausprägungen ist es manchmal möglich, auch Asussagen darüber zu treffen, *wie groß* die Unterschiede zwischen zwei Merkmalsausprägungen ist. Dies ist z.B. Bei der Temperatur in Grad Celsius möglich: Der Temperatur*unterschied* zwischen 10°C und 20°C ist genauso groß wie der Temperaturunterschied zwischen 20°C und 30°C. Man kann allerdings nicht sagen, dass 30°C warmes Wasser dreimal so warm ist wie 10°C warmes Wasser.
4. Verhältnis der Merkmalsausprägung: Zusätzlich kann es möglich sein, auch Aussagen über das Verhältnis mehrerer Merkmalsausprägungen zueinander treffen. Dies ist bei vielen physikalischen Größen wie z.B. der Masse eines Körpers möglich: 100 kg sind 100-mal so viel wie 1 kg. Solche Variablen haben einen natürlichen Nullpunkt, z.B. besagt eine Masse von 0 kg, dass tatsächlich *keine* Masse vorhanden ist.
5. Absolute Werte: Schließlich kann es noch möglich sein, dass das Merkmal in einer natürlichen Einheit vorliegt, d.h. auch die Abstände zwischen zwei Merkmalsausprägungen sind natürlich gegeben. Dies ist der Fall für absolute Häufigkeiten, also z.B. die Häufigkeit, wie oft sich jemand in einer Unterrichtsstunde gemel-

det hat.

1.1 Skalen und Skalenniveaus

Die Zuordnung von bestimmten Zahlen zu Ausprägungen eines Merkmals nennen wir *Skala*. Je nachdem, welche der Regeln beim Übersetzen in Zahlen berücksichtigt wurden, unterscheidet man unterschiedliche *Skalenniveaus*.

1. Eine Skala, die nur die Regel von Gleichheit/Verschiedenheit befolgt, nennen wir *Nominalskala*.
2. Eine Skala, die zusätzlich zu Gleichhheit/Verschiedenheit auch die Rangreihenfolge der Merkmalsausprägungen in Zahlen abbildet, nennen wir *Ordinalskala*.
3. Eine Skala, die zusätzlich auch die Größe der Verschiedenheit abbildet, nennen wir *Intervallskala*.
4. Eine Skala, die zusätzlich auch das Verhältnis der Merkmalsausprägungen abbildet, nennen wir *Verhältnisskala*.
5. Eine Skala, die zusätzlich auch eine natürliche Maßeinheit abbildet, nennen wir *Absolutskala*. Mit diesem Skalenniveau haben wir regelmäig dann zu tun, wenn wir betrachten, wie *häufig* ein Merkmal aufgetreten ist bzw. wie *viele* Beobachtungen einer bestimmten Art wir machen. Ein Beispiel wäre die Anzahl der Versuche, die jemand braucht, um ein Rad zu schlagen.

Tabelle 1.1 gibt eine Übersicht über die erhaltenen Beziehungen (= Relationen) für die fünf vorgestellten Skalenniveaus. Das Skalenniveau entscheidet häufig darüber, welche Aussagen über eine Variable sinnvoll zu treffen sind und (später) wie wir sie beschreiben, zusammenfassen und (später) auswerten können. Je *höher* das Skalenniveau ist, d.h. je mehr der Relationen erhalten sind, desto mehr Aussagen sind sinnvoll zu treffen. Deshalb ist es wünschenswert, ein Merkmal möglichst immer auf einem möglichst hohen Skalenniveau zu messen.

Tabelle 1.1: Übersicht über die vorgestellten *Skalenniveaus* und die jeweils erhaltenen *Relationen*.

Skala	Gleichheit/ Verschiedenheit	Ordnung	Größe der Verschiedenheit	Verhältnisse	absolute Werte
Nominal-	ja	nein	nein	nein	nein
Ordinal-	ja	ja	nein	nein	nein
Intervall-	ja	ja	ja	nein	nein
Verhältnis-	ja	ja	ja	ja	nein
Absolut-	ja	ja	ja	ja	ja

1.2 *Messbarkeit und Messfehler*

Jede Messung ist mit einem mehr oder weniger großen *Messfehler* behaftet. Damit ist gemeint, dass der Messwert niemals exakt der Ausprägung des Merkmals entspricht. Hierfür gibt es mindestens zwei Gründe:

Der erste Grund liegt darin, dass jedes reale Messinstrument hat nur eine begrenzte Genauigkeit besitzt. Wenn wir z.B. ein Lineal mit einer Millimetereinteilung verwenden, können wir nicht auf den Zehntel- oder Hundertstelmillimeter genau ablesen. Auch wenn wir uns mehrmals nacheinander mit einer Personenwaage wiegen, wird sie nicht immer exakt den gleichen Wert anzeigen, obwohl es unwahrscheinlich ist, dass sich unser Gewicht innerhalb eines Augenblicks geändert hat – die Abweichungen entstehen durch Ungenauigkeiten des Messinstruments.

Der zweite Grund liegt darin, dass viele Merkmale, für die wir uns interessieren, gar nicht *direkt messbar* sind. Beispielsweise soll es ja vielleicht in einer Klassenarbeit darum gehen, die Fähigkeit oder das Wissen der Schüler in einem bestimmten Bereich zu messen. Die Note der Klassenarbeit (die in diesem Fall unser Messinstrument für die Fähigkeit sein soll), wird aber auch durch viele andere Dinge beeinflusst, z.B. ob man am Tag der Klassenarbeit einen guten Tag hatte, ob man mit der Art, wie der Lehrer die Aufgaben stellt, zurechtkommt, man während der Arbeit auf Toilette musste und deshalb Zeit verloren hat, und viele andere mögliche Störeinflüsse.

Aufgaben

Überlegt in 2er- oder 3er-Gruppen, welches Skalenniveau die folgenden Variablen aufweisen. Hierzu is es sinnvoll, zunächst zu überlegen, welches Merkmal die Variable wohl eigentlich erfassen soll und welche Relationen (= Beziehungen zwischen den Merkmalsausprägungen) erhalten sind.

a) Das bei uns geläufige metrische System zur Messung von Distanzen in Millimetern, Zentimetern, Metern, oder Kilometern;
b) die Nummern der Straßenbahnen der KVB (Kölner Verkehrsbetriebe);
c) Schulnoten;
d) Postleitzahlen.

2
Beschreiben und Zusammenfassen

Die *deskriptive Statistik* beschäftigt sicht mit der Frage, wie Merkmale bzw. Variablen sinnvoll beschrieben und zusammengefasst werden können. Wichtige Methoden hierzu sind *Häufigkeitsverteilungen* und *statistische Kennwerte*, die wir in diesem Kapitel vorstellen werden.

2.1 Häufigkeitsverteilungen

Um ein Merkmal zu beschreiben, kann man sich dessen *Häufigkeitsverteilung* anschauen. Die Häufigkeitsverteilung eines Merkmals ist charakterisiert durch (1) die Gesamtheit der unterschiedlichen Merkmalsausprägungen und (2) die Häufigkeit, mit der diese Ausprägungen vorkommen. Sie lässt sich in Form von Häufigkeitstabellen oder in einem Säulendiagramm wie Abbildung 2.1 darstellen.

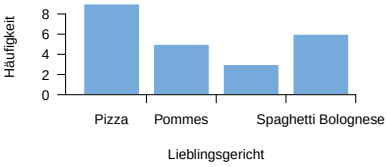


Abbildung 2.1: Die Häufigkeitsverteilung der nominalskalierten Variable *Lieblingsgericht*. Die x-Achse bezeichnet die Ausprägung des Merkmals, die Höhe der Verteilung (y-Wert) gibt Aufschluss über die Häufigkeit der zugehörigen Merkmalsausprägung.

In diesem ersten Beispiel gibt es nur vier unterschiedliche Merkmalsausprägungen (Lieblingsgerichte), deshalb lässt sich die Häufigkeitsverteilung noch leicht in einem Diagramm darstellen. Wenn es aber sehr viele unterschiedliche Merkmalsausprägungen gibt, könnte man aber auch *Kategorien von Merkmalsausprägungen* bilden, um diese dann als sog. *sekundäre Häufigkeitsverteilung* darzustellen. Z.B. wäre es denkbar, dass die Schüler viele unterschiedliche Nudelgerichte angegeben hätten, die man dann zu der Kategorie "Nudelgerichte" zusammengefasst hätte.

Betrachtet man die Häufigkeitsverteilung eines Merkmals wie der Körpergröße von Schülern in einer Klasse, merkt man schnell, dass man nicht für jede Merkmalsausprägung eine eigene Säule zeichnen möchte: Hat man die Körpergröße z.B. auf den Zentimeter genau gemessen, wird es in einer Klasse kaum zwei Schüler geben, die genau gleich groß sind und jede Säule hätte die Höhe 1. Deshalb ist es bei intervall- und verhältnisskalierten Variablen oft sinnvoll, gleich eine sekundäre Häufigkeitsverteilung zu zeichnen. Die häufigste Form der Darstellung ist bei solchen Variablen eine besondere Form des Säulendiagramms, das *Histogramm*: Bei ihm ist auch die Breite

der Säulen und somit die Fläche der Säulen interpretierbar; man sieht so anhand der Flächen relativ schnell, wie sich das Merkmal auf dessen Ausprägungsbereiche verteilt.

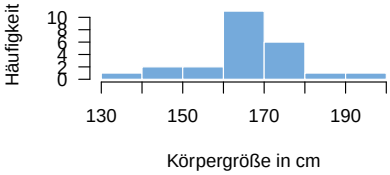


Abbildung 2.2: Histogramm der gemessenen Körpergröße in einer 10. Klasse.

2.2 *Statistische Kennwerte*

Um die Häufigkeitsverteilung eines Merkmals zusammenfassend zu beschreiben, verwendet man zusätzlich *statistische Kennwerte*, die bestimmte Eigenschaften der Häufigkeitsverteilung in Zahlen zusammenfassen sollen. Hierzu unterscheidet man (1) *Lagemaße* bzw. *Maße der zentralen Tendenz* und (2) *Streuungs-* bzw. *Dispersionsmaße*. Lagemaße beantworten die Frage "Wo auf der Skala liegen die Messwerte?", Streuungsmaße beantworten die Frage "Wie stark unterscheiden sich die Messwerte voneinander?".

Sowohl für die Lage als auch die Streuung gibt es viele unterschiedliche Kennwerte – welchen Kennwert man am besten verwendet, hängt nämlich davon ab, welches Skalenniveau die Variable, die wir beschreiben möchten, aufweist.

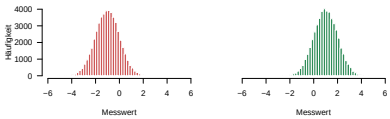


Abbildung 2.3: Zwei Häufigkeitsverteilungen mit unterschiedlicher Lage, aber gleicher Streuung.

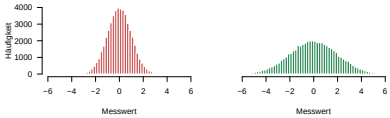


Abbildung 2.4: Zwei Häufigkeitsverteilungen mit gleicher Lage, aber unterschiedlicher Streuung.

2.2.1 *Lagemaße*

Modus bzw. *Modelwert*. Der Modus *Mo* ist derjenige Wert einer Variablen, der *am häufigsten* vorkommt. Man "berechnet" ihn also, indem man zählt, wie häufig jede einzelne Merkmalsausprägung vorkommt und die häufigste Ausprägung aufschreibt. Der Modus ist schon ab Nominalskalenniveau interpretierbar. In unseren Beispiel aus Kapitel 1, in dem wir eine Schulklasse nach ihren Lieblingsgerichten gefragt haben, wäre also der Modus der Variable Lieblingsgericht der Wert "Pizza" – genau jene Antwort, die am häufigsten genannt wurde. Hätten wir den einzelnen Gerichten Zahlen zugeordnet, z.B. "Pizza" = 1, "Pommes" = 2, usw., wäre der Modus $Mo = 1$. Das ist genau der Zahlenwert, den wir zuvor der Antwort "Pizza" zugeordnet hatten.

Median. Der Median *Md* ist derjenige Wert einer Variablen, der alle Werte, die man voher der Größe nach geordnet hat, genau in der Mitte halbiert. Um ihn zu berechnen, ordnet man also zunächst alle Werte der Größe nach und schaut dann, welcher Wert genau in der Mitte dieser geordneten Zahlenreihe steht. Der Median ist dann sinnvoll interpretierbar, wenn die Variable mindestens Ordinalskalenniveau aufweist, sich die Werte also sinnvoll der Größe, Intensität oder Stärke nach ordnen lassen. Ein gutes Beispiel hierfür sind Schulnoten.

Mittelwert. Der Mittelwert *M* ist der Durchschnitt der Werte einer Variablen. Ihn kann man am leichtesten bestimmen, indem man (1) alle beobachteten Werte aufsummiert und dann (2) durch die Anzahl der beobachteten Werte teilt. Damit der Mittelwert sinnvoll inter-

pretierbar ist, muss die Variable mindestens Intervallskalennievau aufweisen. Ein gutes Beispiel hierfür ist die Temperatur in Grad Celsius: Haben wir einen Eimer mit 10°C warmen Wasser und einem Eimer mit 20°C warmen Wasser, können wir sagen, dass die Temperatur des wassers in den beiden Eimern im Durchschnitt oder im Mittel 15°C beträgt. Das geht deshalb, da der Unterschied von 15°C zu 10°C genauso groß ist wieder der Unterschied von 15°C zu 20°C (nämlich jeweils 5°C).

2.2.2 *Streuungsmaße*

Range bzw. *Variationsbreite*. Der Range v (engl. für Reichweite/Bandbreite) bezeichnet dem Umfang der Werte einer Variable. Im Falle einer nominalskalierten Variable berechnet er sich als die Anzahl unterschiedlicher Merkmalsausprägungen, die beobachtet wurden (wenn z.B. nur drei unterschiedliche Lieblinsgerichte genannt, ist der Range $v = 3$) Im Falle von mindestens ordinalskalierten Variablen bezeichnet er die Differenz zwischen dem höchsten und niedrigsten beobachteten Wert der Variablen. (Besser ist es im Falle von nominal- oder ordinalskalierten Variablen, den sog. *relativen Informationsgehalt* H zu berechnen.)

Interquartilsbereich. Der Interquartilsbereich *IQB* bezeichnet den Bereich der Werte einer Variable, der die Hälfte aller Beobachtungen umfasst, die, wenn man die Werte der Reihe nach ordnet, in der Mitte liegen. Man kann also auch sagen, dass man die Werte der Variable der Reihe nach ordnet, das unterste Viertel (mit den kleinsten Werten) und das oberste Viertel (mit den größten Werten) abschneidet, und dann schaut, welche Werte die Variable an diesen Schnittpunkten hat. Da man die Werte hierfür in eine Ordnung bringen muss, kann man den Interquartilsbereich erst ab Ordinalskalenniveau berechnen.

Standardabweichung. Die Standardabweichung *SD* ist das wohl wichtigste und am häufigsten verwendete Maß der Dispersion. Leider lässt sich es sich nicht ganz so leicht berechnen, denn man muss (1) für jeden Wert der Variablen den Abstand zum Mittelwert der Variablen berechnen, (2) diese Abstände dann jeweils quadrieren, (3) die quadrierten Abstände aufsummieren, (4) diese Summe durch die Anzahl der Beobachtungen teilen und (5) aus dieser Summe die Wurzel ziehen. Das klingt ziemlich kompliziert, muss man aber zum Glück praktisch nie von Hand machen. Um zu verstehen, was die Standardabweichung ausdrückt, kann man sich vorstellen, dass sie ungefähr dem Wert entspricht, den die Messwerte (betragsmäßig) durchschnittlich vom Mittelwert abweichen (wobei große Abweichungen aber etwas stärker gewichtet werden). Es lässt sich erst ab

Intervallskalenniveau berechnen, da erst hier die Größen von Abständen bzw. Unterschieden sinnvoll interpretierbar sind.

Tabelle 1.2 zeigt einen Überblick, ab welchem Skalenniveau man welchen der vorgestellen Kennwerte berechnen darf.

Tabelle 2.1: Übersicht über die vorgestellten Lage- und Streuungsmaße und auf welchem Skalenniveau man sie sinnvoll berechnen kann.

Skala	Modus	Median	Mittelwert	Range	*IQB*	*SD*
Nominal-	ja	nein	nein	ja	nein	nein
Ordinal-	ja	ja	nein	ja	ja	nein
Intervall-	ja	ja	ja	ja	ja	ja
Verhältnis-	ja	ja	ja	ja	ja	ja
Absolut-	ja	ja	ja	ja	ja	ja

Aufgaben

1. Zeichnet ein Histogramm der Variable *Alter* in der folgenden Tabelle.

Tabelle 2.2: Die Häufigkeitsverteilung der Variablen *Geschlecht* und *Alter* in einer Erstsemesterveranstaltung der Uni Köln.

Nr.	Geschlecht	Alter	Nr.	Geschlecht	Alter
1	weiblich	33	21	männlich	19
2	weiblich	28	22	weiblich	21
3	männlich	21	23	männlich	27
4	weiblich	26	24	weiblich	44
5	weiblich	25	25	weiblich	24
6	weiblich	22	26	weiblich	29
7	weiblich	22	27	weiblich	27
8	männlich	37	28	weiblich	19
9	weiblich	27	29	weiblich	21
10	weiblich	25	30	weiblich	25
11	weiblich	21	31	weiblich	25
12	weiblich	19	32	weiblich	24
13	weiblich	26	33	weiblich	21
14	weiblich	19	34	männlich	28
15	weiblich	19	35	weiblich	37
16	weiblich	22	36	weiblich	23
17	weiblich	31	37	weiblich	20
18	weiblich	24	38	weiblich	21

Nr.	Geschlecht	Alter	Nr.	Geschlecht	Alter
19	weiblich	27	39	weiblich	19
20	weiblich	24	40	weiblich	22

2. Zeichnet ein Histogramm der Variable *Alter* aus der vorigen Aufgabe. Zeichnet dieses Mal jedoch eine sekundäre Häufigkeitsverteilung, beginnend mit der Kategorie "18-21 Jahre".
3. Berechnet den Modus der Variable *Geschlecht* aus Aufgabe 1.
4. Berechnet den Mittelwert der Variable *Alter* aus Aufgabe 1.
5. Erläutert, warum es nicht sinnvoll ist, den Median der Variable *Geschlecht* zu berechnen.

3
Wahrscheinlichkeitstheorie und inferenzstatistische Grundlagen

In den vorausgehenden Kapitel haben wir uns damit beschäftigt, Merkmale zu messen, zu beschreiben und zusammenzufassen. Ziel jeder Wissenschaft ist es jedoch, ausgehend von spezischen Beobachtungen auf *allgemeingültige* Aussagen schließen zu können.

Hierzu ist es hilfreich, zwischen der Population bzw. Grundgesamtheit und der Stichprobe zu unterscheiden: Die *Population* ist ein Begriff für alle potentiell untersuchbaren Merkmalsträger (und damit ein Begriff für das, was "im Allgemeinen" gilt), die *Stichprobe* ist die Teilmenge der Population, deren Merkmalsausprägung wir gemessen haben. Will man also allgemeingültige Aussagen treffen können, muss man von der Stichprobe auf die Eigenschaften der Population schließen.

Wir müssen uns aber klarmachen, dass wir immer dann, wenn wir von spezifischen Beobachtungen zu allgemeingültigen Aussagen gelangen wollen, das Risiko eingehen, dass unsere Beobachtungen immer nur einen Teil der Wirklichkeit wiederspiegeln und wir zufällig genau jenen Teil der Wirklichkeit nicht beobachten, der unserer allgemeingültigen Aussage widerspricht. Diese Quelle von Fehlentscheidungen – den sog. *Stichprobenfehler* – können wir niemals mit absoluter Sicherheit ausschließen, wenn wir nicht die gesamte Population beobachten können; Ziel der *Inferenzstatistik* ist es deshalb abzuschätzen, mit welcher Sicherheit (oder Unsicherheit) dennoch von der Stichprobe auf die Population geschlossen werden kann.

Um dies tun zu können, benötigen wir immer ein *Modell* der Merkmalsausprägungen in der Population; haben wir ein solches Modell, können wir das Ziehen unserer Stichprobe aus der Population als ein sog. *Zufallsexperiment* auffassen. Ein Zufallsexperiment ist definiert als die Durchführung eines *Zufallsvorgangs*, d.h. eines Vorgangs, der zu unvorhersehbaren und sich gegenseitig ausschließenden Ergebnissen führt, unter kontrollierten Bedingungen. Die

Ergebnisse eines Zufallsexperiments nennen wir *Zufallsvariable*.

Ein gutes Beispiel für ein Zufallsexperiment ist der einfache Münzwurf: Der Zufallsvorgang – das Werfen einer Münze – kann beliebig oft und unter kontrollierten Bedingungen durchgeführt werden. Die resultierende Zufallsvariable sind die Häufigkeiten, mit denen die Münze Kopf oder Zahl gezeigt hat. Nehmen wir nun ein Modell unseres Merkmals in der Population an, das besagt, dass beide Seiten gleich häufig vorkommen, lässt sich mit Methoden der Inferenzstatistik berechnen, wie wahrscheinlich ein Ergebnis unter der Annahme dieses Modells waren. An dieser Stelle kann man auch schon sehen, dass das zugrundeliegende Populationsmodell oft "nur" ein *Modell* und keine allumfassende Beschreibung der Wirklichkeit ist, denn es gibt ja vmtl. auch noch den sehr seltenen Fall, dass eine Münze auf ihrer Kante stehen bleibt, dieser Fall ist aber im Modell nicht berücksichtigt.

Die Annahme, dass das Ziehen unserer Stichprobe aus der Population den Regeln eines Zufallsexperiments folgt, erlaubt es uns, die Wahrscheinlichkeit bestimmter Ergebnisse, die wir in unserer Stichprobe beobachten, für ein bestimmtes Modell der Population zu bestimmen. Um dies zu erläutern betrachten wir zunächst, welche Vorstellungen vor Wahrscheinlichkeit es gibt und wie man Wahrscheinlichkeiten von Ergebnissen eines Zufallsexperiments berechnen kann.

3.1 Wahrscheinlichkeitsbegriffe

Wahrscheinlichkeit ist uns aus dem Alltag ein vertrauter Begriff. Beispielsweise würden die meisten Leute darin übereinstimmen, dass es "unwahrscheinlich" ist, fünfmal nacheinander eine Sechs zu würfeln. Wir verwenden aber den Begriff Wahrscheinlichkeit mit mindestens zwei unterschiedlichen Bedeutungen; es ist z.B. etwas anderes zu sagen "Ein fairer Würfel zeigt mit einer Wahrscheinlichkeit von 1/6 eine Sechs." als zu zu sagen "Morgen wird es wahrscheinlich regnen.". Diese Intuition findet sich auch in unterschiedlichen Wahrscheinlichkeitsbegriffen in der Wahrscheinlichkeitstheorie wieder.

Klassicher Wahrscheinlichkeitsbegriff. Der klassische Wahrscheinlichkeitsbegriff entspricht unserem obigen Beispiel des Würfelwurfs. Hierbei geht man davon aus, dass man ein und denselben Zufallsvorgang beliebig häufig wiederholen kann.

Die klassische Definition einer Wahrscheinlichkeit (nach Pierre-Simon Laplace) lautet: Wahrscheinlichkeit ist der Anteil der günstigen Fälle an der Gesamtzahl der Fälle. Ist das Ereignis A der "günstige" Fall, berechnet sich dessen Wahrscheinlichkeit $p(A)$ als

$$p(A) = \frac{n_A}{N_{\text{gesamt}}}$$

wobei n_A die Anzahl der günstigen Fälle und N_{gesamt} die Gesamtzahl der Fälle bezeichnet.

Die frequentistische Definition einer Wahrscheinlichkeit (nach Richard Edler von Mises) erweitert diesen Gedanken, indem sie annimmt, dass es eine *wahre* Wahrscheinlichkeit gibt, die wir mit einer wachsenden Anzahl an Versuchen bzw. Beobachtungen immer genauer schätzen können. Die Definition besagt, dass bei prinzipiell unendlich häufiger Durchführung des Versuchs der *Grenzwert* des Anteils der günstigen Fälle dieser wahren Wahrscheinlichkeit $\pi(A)$ ("pi von A") enspricht.

$$\pi(A) = \lim_{N \to \infty} \frac{n_A}{N}$$

Man kann also auch sagen, dass sich der Anteil der günstigen Fälle $\frac{n_A}{N}$ immer weiter der Wahrscheinlichkeit $\pi(A)$ annähert, je größer N, also die Anzahl an Versuchen oder Wiederholungen, wird. Diese Überlegungen bilden die Grundlage der klassischen – frequentistischen – Statistik.

Subjektiver Wahrscheinlichkeitsbegriff. Im Gegensatz dazu geht es beim subjektiven Wahrscheinlichkeitsbegriff darum, die subjektive Einschätzung der Sicherheits des Eintretens eines Ereignisses auszudrücken, z.B. "Morgen wird es mit einer Wahrscheinlichkeit von 80% regnen". Es geht hierbei darum, das Ausmaß des Vertrauens, das ein vernünftiger Akteur in eine Aussage setzen würde, auszudrücken. Dieser Wahrscheinlichkeitsbegriff erlaubt Aussagen über einzelne Ereignisse, nicht nur um "prinzipiell beliebig häufig wiederholbare" Ereignisse (wie bei dem klassichen Wahrscheinlichkeitsbegriff). Darüber hinaus erlaubt er die Kombination vielfältiger Informationen zu einem Urteil (z.B. Vorwissen, Plausibilitätsüberlegungen, etc.). Dieser Wahrscheinlichkeitsbegriff bildet die Grundlage der sog. Bayes-Statistik (die wir im Rahmen dieser Veranstaltung nicht vertiefen werden).

3.2 *Wahrscheinlichkeits- und Wahrscheinlichkeitsdichteverteilungen*

Wahrscheinlichkeits- und Wahrscheinlichkeits*dichte*verteilungen dienen der Beschreibung der möglichen Ergebnisse eines Zufallsexperiments. Wahrscheinlichkeitsverteilungen kommen dabei bei *diskreten*, Wahrscheinlichkeitsdichteverteilungen bei *stetigen* Zufallsvariablen zur Anwendung. Eine diskrete Variable liegt dann vor, wenn es endlich viele Ausprägungen eines Merkmals gibt (z.B. die wählbaren

Parteien bei der Bundestagswahl), *oder* wenn es theoretisch unendlich viele, aber abzählbare Ausprägungen eines Merkmals gibt (z.B. die Anzahl der Versuche, bis jemand eine Aufgabe gelöst hat – theoretisch gibt es unendlich viele mögliche Ausprägungen des Variable *Anzahl Versuche*, man kann aber trotzdem die Anzahl der Versuche abzählen.) Gibt es jedoch überabzählbar unendlich viele Merkmalsausprägungen, sprechen wir von einer *stetigen* oder *kontinuierlichen* Variable. "Überabzählbar" bedeutet, dass innerhalb eines Werteintervalls beliebig viele Zwischenwerte liegen können, man also die Gesamtheit möglicher Werte nicht abzählen kann (typische Beispiele sind die Körpergröße in Meter oder das Gewicht in Kilogramm).

Wahrscheinlichkeitsverteilungen. Eine Wahrscheinlichkeitsverteilung gibt für jeden Wert einer *diskreten* Zufallsvariable die Auftretenswahrscheinlichkeit an. Eine Wahrscheinlichkeitsverteilung lässt sich als Säulendiagramm wie Abbildung 3.1 darstellen.

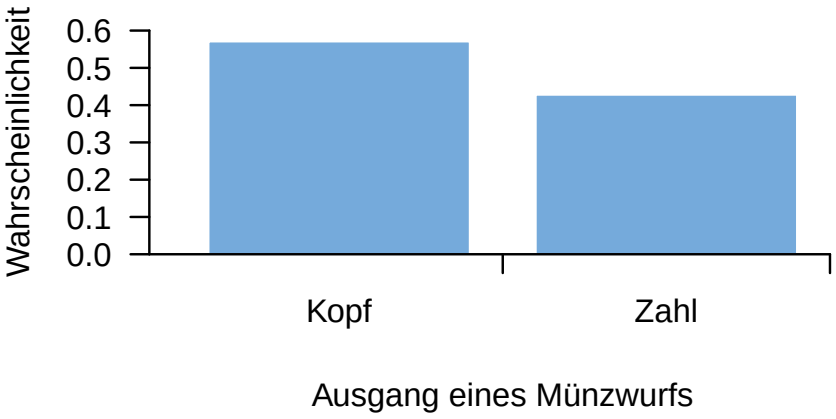


Abbildung 3.1: Die Wahrscheinlichkeitsverteilung der Variable *Ausgang eines Münzwurfs*. Die x-Achse bezeichnet die Ausprägung des Merkmals, die y-Achse die zugehörige Wahrscheinlichkeit.

Wahrscheinlichkeitsdichteverteilungen. Eine Wahrscheinlichkeitsdichteverteilung (wie in Abbildung 3.2) beschreibt ein Zufallsexperiment mit überabzählbar unendlich vielen möglichen Ereignissen. (Man kann man sie sich auch als diskrete Wahrscheinlichkeitsverteilung mit unendlich kleinen Kategoriebreiten vorstellen.) Das hat zur Folge dass, je kleiner man das betrachtete Werteintervall wählt, auch die Wahrscheinlichkeit immer kleiner wird. Geht die Breite des Intervalls gegen 0, wird die zugehörige Wahrscheinlichkeit unendlich klein; für einen einzelnen fixen Wert ist die Wahrscheinlichkeit gleich 0. Deshalb können Wahrscheinlichkeiten immer nur für Bereiche (Intervalle) angegeben werden. Die Kurve selbst beschreibt die Wahrscheinlichkeitsdichte. Die Fläche unter der Kurve in einem Bereich gibt die Wahrscheinlichkeit dafür an, dass ein Wert in diesen Bereich fällt.

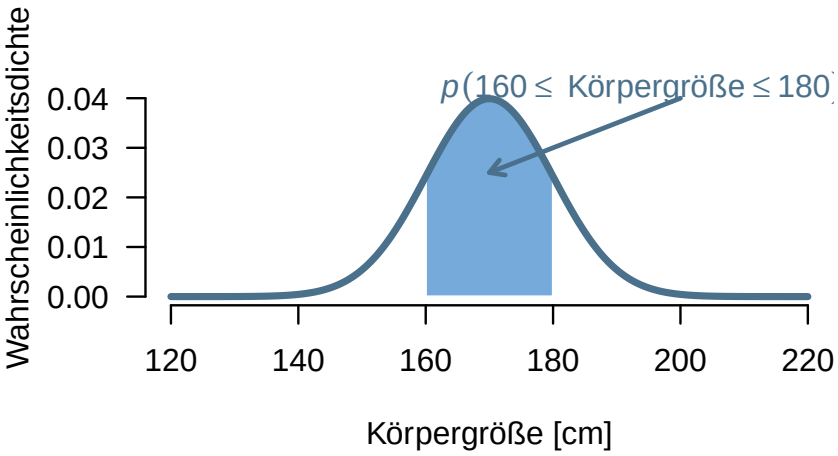


Abbildung 3.2: Die Wahrscheinlichkeitsdichteverteilung der Variable *Körpergröße*. Die x-Achse bezeichnet die Ausprägung des Merkmals, die y-Achse die zugehörige Wahrscheinlichkeitsdichte. Wahrscheinlichkeiten können für Intervalle von Merkmalsausprägungen angegeben werden, indem die Fläche unterhalb der Kurve in diesem Intervall bestimmt wird: In diesem Beispiel beträgt die Fläche im Intervall [160, 180] ungefähr .68, also ist $p(160 \leq \text{Körpergröße} \leq 180) = .68$.

3.3 *Stichprobenkennwerteverteilung*

Wir haben in Kapitel 2.2 schon verschiedene Stichprobenkennwerte, also Kennwerte zur zusammenfassenden Beschreibung eines Merkmals einer Stichprobe, kennengelernt (z.B. Mittelwert und Standardabweichung). Möchten wir allgemeingültige Aussagen treffen, interessieren wir uns aber für nicht für die Kennwerte der konkreten Stichprobe, sondern für jene der zugrundeliegenden Population. Nehmen wir nun ein bestimmtes Modell des Merkmals in der Population an, können wir mithilfe der Stichprobenkennwerte *Schätzer* der entsprechenden Populationskennwerte bestimmen. Hierbei gilt, dass mit wachsender Stichprobengröße die Genauigkeit der Schätzung zunimmt.

In der klassichen (= frequentistischen) Inferenzstatistik verwendet man die oben beschriebene frequentistische Definition einer Wahrscheinlichkeit zusätzlich, um Verteilungen von Stichprobenkennwerten über viele Stichproben hinweg vorherzusagen, diese sind Wahrscheinlichkeits- bzw. Wahrscheinlichkeitsdichteverteilungen. Hiermit ist es auch möglich, die Wahrscheinlichkeit zu bestimmen, dass ein bestimmter Stichprobenkennwert unter der Annahme eines bestimmten Modells des Merkmals in der Population auftreten kann.

Eine besonders wichtige Stichprobenkennwerteverteilung ist Verteilung des Stichprobenmittelwerts, die in folgender App dargestellt wird.

Öffne App in neuem Fenster

Aufgaben

Verwendet die App, um die folgenden Fragen zu beantworten:

1. Wie verändert sich die Verteilung der Messwerte in der Stichprobe, wenn die Stichprobengröße N größer oder kleiner wird?
2. Wie verändert sich die Verteilung der Stichprobenmittelwerte, wenn die Stichprobengröße N größer oder kleiner wird?
3. Wie verändert sich die Verteilung der Stichprobenmittelwerte, wenn Körpergröße in Wirklichkeit eine kleinere oder größerere Variabilität (entspricht hier der Standardabweichung) aufweist?
4. Welche drei Maßnahmen sollte ich (in diesem Beispiel) ergreifen, um möglichst viel über die "Wirklichkeit" zu erfahren?

4
Hypothesen und Hypothesentest

Wissenschaftliche Untersuchungen werden durchgefürt, um eine bestimmte Forschungsfrage zu beantworten. Diese Forschungsfrage wird als sog. *inhaltliche Hypothese* formuliert. Hypothesen sind Erwartungen über Unterschiede zwischen oder Zusammenhänge von Variablen, die vor einer Untersuchung formuliert werden.

Beispiele für inhaltliche Hypothesen sind:

- Niderländer sind größer als Deutsche.
- Deutsche sind im Mittel größer als 170 cm.
- Nichtraucher treiben mehr Sport als Raucher.
- Menschen mit hohen Cholesterinspiegeln entwickeln im Laufe ihres Lebens eher eine Herzkrankheit.
- Menschen, die sich in ihrer Jugend wenig bewegen, haben später auch mehr Zivilisationskrankheiten.
- Es gibt einen Zusammenhang zwischen dem Einkommen der Eltern und dem eigenen Einkommen.

Davon unterscheidet man statistische Hypothesen. Sie beziehen sich auf statistische Kennwerte und deren Verteilungen.

Beispiele für statistische Hypotehesen:

- Der Mittelwert der Körpergröße der Niederländer ist größer als der der Deutschen.
- Der Mittelwert der Körpergröße der Deutschen ist größer als 170 cm.
- ...

4.1 Statistisches Hypothesenpaar

Es werden immer zwei komplementäre statistische Hypothesen formuliert: die Nullhypothese und die Alternativhypothese. Die Nullhypothese H_0 besagt in der Regel, dass es keinen Unterschied zwischen zwei Populationen in der Ausprägung eines Merkmals (bzw. keinen

Zusammenhang zwischen zwei Merkmalen in einer Population) gibt. Die Alternativhypothese H_1 besagt, dass es einen Unterschied (bzw. einen Zusammenhang) gibt. H_0 und H_1 sind komplementär: Das bedeutet entweder die eine oder die andere Hypothese trifft zu, es gibt keine dritte Möglichkeit.

Das Ziel des Signifikanz- bzw. Hypothesentests ist es nun häufig, zwischen diesen beiden Hypothesen *entscheiden* zu können. Hierzu wird der Stichprobenkennwert (z.B. der Stichprobenmittelwert) mit der Stichprobenkennwerteverteilung verglichen, die unter Annahme der *Nullhypothese* entstehen würde. Ist der Stichprobenmittelwert unter Annahme der Nullhypothese wenig wahrscheinlich, *verwirft* man die Nullhypothese und entscheidet sich für die Alternativhypothese (sie ist ja die einzige Alternative zur als "falsch" befundenen Nullhypothese). Ist der Stichprobenmittelwert aber unter Annahme der Nullhypothese recht wahrscheinlich, bleibt man bei der Nullhypothese (denn sie hat ja den Stichprobenmittelwert gut vorhersagen können).

Für eine echte Entscheidung *zwischen* diesen beiden Hypothesen reicht dieses Vorgehen eigentlich nicht aus: Man hat ja noch nicht überprüft, wie wahrscheinlich oder unwahrscheinlich der Stichprobenmittelwert unter Annahme der Alternativhypothese gewesen wäre. Eigentlich können wir nur sagen, wie wahrscheinlich der Stichprobenkennwert unter Annahme der Nullhypothese war.

4.2 Entscheidungsfehler

Entscheiden wir zwischen zwei Hypothesen, besteht auch immer eine gewisse Wahrscheinlichkeit, dass wir einen Entscheidungsfehler begehen und uns für die falsche der beiden Hypothesen entscheiden. Das Verwerfen der Nullhypothese, obwohl sie stimmt, nennt man α-Fehler (oder Fehler 1. Art), das Beibehalten der Nullhypothese, obwohl sie nicht stimmt, nennt man β-Fehler (oder Fehler 2. Art).

Wann genau entscheidet man sich aber nun für oder gegen die Nullhypothese? Es hat sich eingebürgert, sich gegen die Nullhypothese zu entscheiden, wenn die Wahrscheinlichkeit des Stichprobenkennwerts unter Annahme der Nullhypothese kleiner oder gleich 5% ist. (Diesen Wert nennt man auch "α-Fehlerniveau von 5%".) Das bedeutet aber auch, dass der Stichprobenkennwert unter Annahme der Nullhypothese nicht unmöglich war, sondern eben "nur" eine 5%-ige Wahrscheinlichkeit hatte.

Es ist an dieser Stelle wichtig festzustellen, dass wir anhand einer einzelnen Stichprobe und ihres einen Kennwerts nichts über die Wahrscheinlichkeit wissen, ob wir gerade einen α-Fehler begehen oder nicht: Es kann ja gut sein, dass die Nullhypothese "wahr" ist

und wir nur zufällig eine jener 5% von Stichproben aus der Population gezogen haben, die eben unter der Nullhypothese etwas weniger wahrscheinlich waren.

Was ist dann aber der Vorteil davon, überhaupt einen Signifikanztest zu rechnen? Der Vorteil liegt darin, dass wir, wenn wir viele Untersuchungen durchführen, uns "nur" in 5% der Fälle, in denen die Nullhypothese "wahr" ist, gegen sie entscheiden. Wenn wir die gleiche Untersuchung mehrfach durchführen – oder jemand anderes unsere Untersuchung wiederholt – fällt es eher auf, wenn wir uns einmal falsch entschieden haben und der Fehler kann korrigiert werden.

Aufgaben

Öffne App in neuem Fenster

Verwendet die App, um die folgenden Fragen zu beantworten:

1. Wenn Ihr 20 Stichproben zieht und anhand deren Mittelwert zwischen H_0 und H_1 entscheiden müsst, wie häufig begeht Ihr einen α-Fehler?
2. Nehmt nun an, dass der wahre Mittelwert nicht 170 cm, sondern 180 cm beträgt. Wie häufig begeht Ihr einen β-Fehler?
3. Nehmt nun an, dass der wahre Mittelwert nicht 170 cm, sondern 190 cm beträgt. Wie häufig begeht Ihr einen β-Fehler?
4. Mit welchen Größen lässt sich das Verhältnis von α- und β-Fehler beeinflussen?

5
Zusammenhänge zwischen zwei Variablen

Häufig möchte man nicht nur wissen, wie eine einzelne Variable verteilt ist, sondern vielmehr, wie zwei oder mehr Variablen miteinander *zusammenhängen*. Der Zusammenhang zweier Variablen meint deren gemeinsames Variieren (ihre *Kovariation*) – also ein gemeinsames Auftreten von hohen oder niedrigen Werten: Ein *gleichsinniger* oder *positiver* Zusammenhang bedeutet, dass hohe Werte in der einen Variable mit hohen Werten in der anderen Variable einhergehen; ein *gegenläufiger* oder *negativer* Zusammenhang liegt dann vor, wenn hohe Werte in der einen Variable mit niedrigen Werte in der anderen Variable einhergen.

Um Zusammenhänge bildlich darzustellen, verwendet man häufig ein *Streudiagramm* wie in Abbildung 5.1.

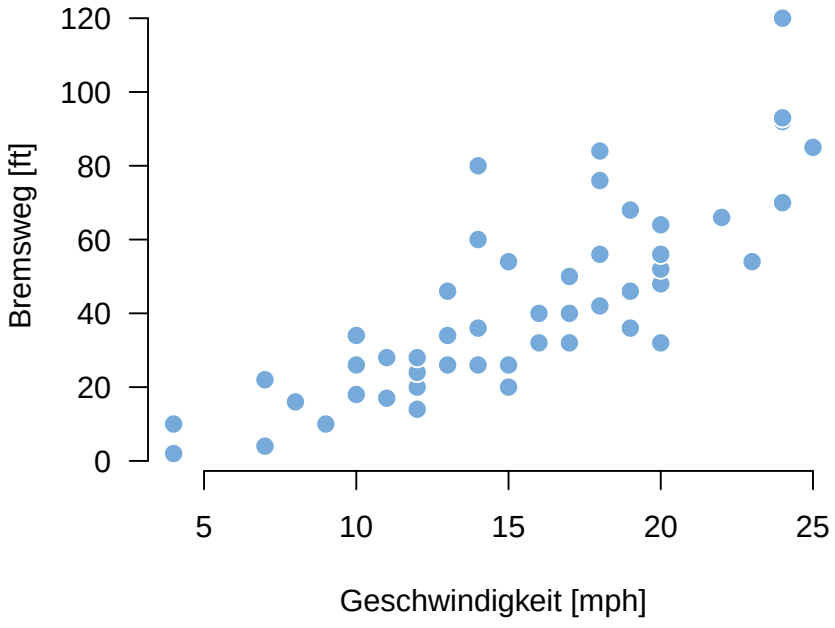


Abbildung 5.1: Der Zusammenhang zwischen der Geschwindigkeit eines Fahrzeugs und dem Bremsweg, um das Fahrzeug aus dieser Geschwindigkeit zum Stillstand zu bringen. Die Daten stammen aus den 1920er Jahren.

Neben der *Richtung* (positiv–negativ) eines Zusammenhangs ist

auch die *Form* eines Zusammenhangs relevant. Abbildung 5.2 zeigt einerseits, wie ein positiver im Vergleich zu einem negativen Zusammenhang aussieht (Tafel A vs. B), aber auch wie ein *Nullzusammenhang* oder ein quadratischer Zusammenhang aussehen könnten. Darüber hinaus sind natürlich viele andere Formen, z.B. kubische oder umgekehrt-U-förmige Zusammenhänge möglich.

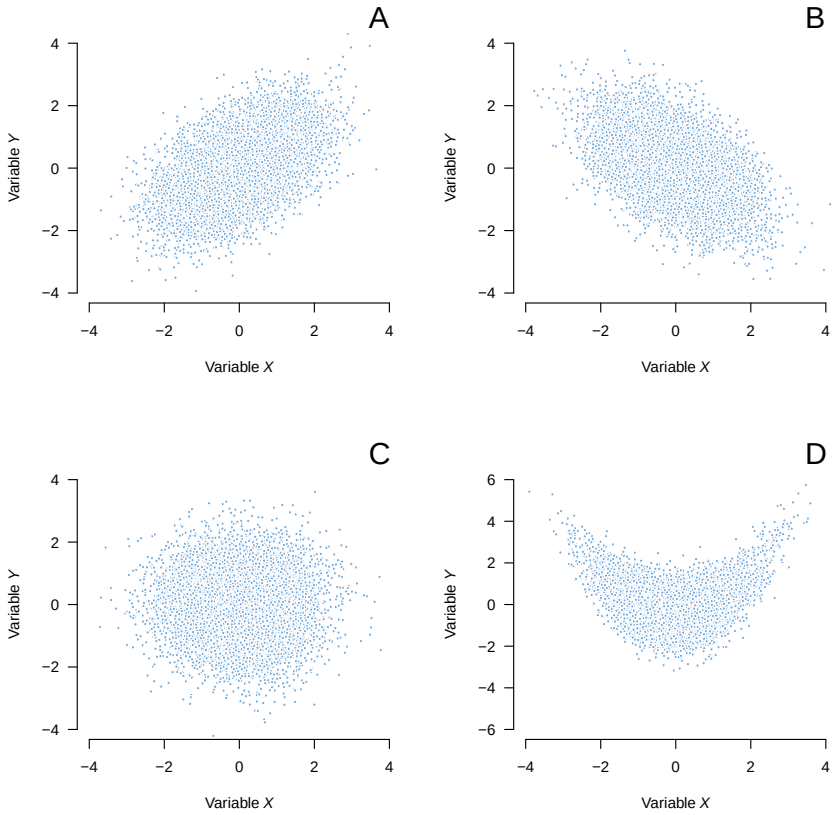


Abbildung 5.2: Vier unterschiedliche Formen von Zusammenhängen. Tafel A zeigt einen positiv-linearen, Tafel B einen negativ-linearen Zusammenhang. Tafel C zeigt einen Nullzusammenhang – es besteht kein Zusammenhang zwischen X und Y. Tafel D zeigt einen quadratischen Zusammenhang.

Die dritte wichtige Eigenschaft eines Zusammenhangs ist dessen *Stärke*: Hiermit beschreibt man, wie eng das Verhältnis der beiden Variablen zueinander ist. Die Stärke eines Zusammenhangs lässt sich mithilfe eines geeigneten statistischen Kennwerts quantifizieren. Welcher statistische Kennwert in einer bestimmten Anwendung geeignet ist, hängt von der Form des Zusammenhangs, dem Skalenniveau der beteiligten Variablen und weiteren Voraussetzungen einzelner statistischer Kennwerte bzw. Verfahren ab. Die Frage, welcher Kennwert geeignet ist, stellt dabei eine der Fragen dar, mit denen sich ein großer Teil der statistischen Literatur beschäftigt, den wir aber im Rahmen dieser Veranstaltung nicht vertiefen werden.

Ein besonders häufig verwendetes Maß für den Zusammenhang zweier Variablen ist der Korrelationskoeffizient r nach Pearson (er

eignet sich zur Quantifizierung des linearen Zusammenhangs zweier intervallskalierter Variablen). Er ist ein standardisiertes Maß für die *Richtung* und die *Stärke* eines linearen Zusammenhangs und hat einen Wertebereich von -1 bis +1: Negative Werte zeigen einen negativen, positive Werte zeigen einen positiven Zusammenhang an. Ein Wert von 0 bedeutet, dass kein (linearer) Zusammenhang besteht.

Abbildung 5.3 zeigt unterschiedlich starke (linear-positive) Zusammenhänge.

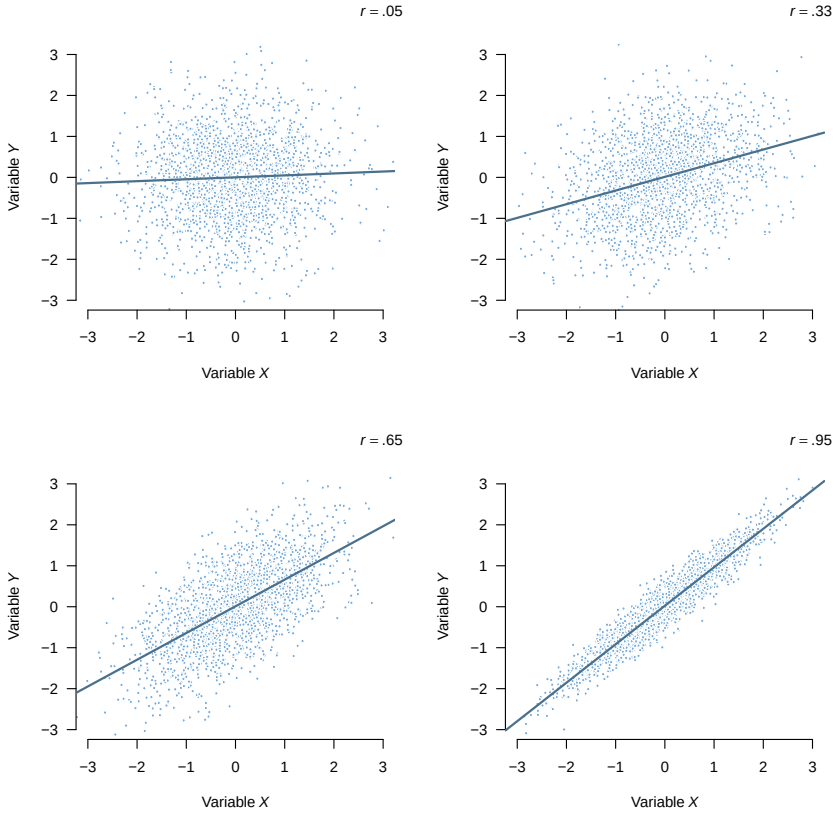


Abbildung 5.3: Vier unterschiedlich *starke* Zusammenhänge.

5.1 *Andere Formen von Zusammenhängen*

Die obigen Beispiele zeigen ausschließlich Zusammenhänge zwischen zwei jeweils mindestens intervallskalierten (man spricht auch von *metrischen* oder *kardinalskalierten*) Variablen. Zusammenhänge lassen sich aber auch sinnvoll zwischen Variablen unterschiedlicher Skalenniveaus bestimmen. So zeigt zum Beispiel Abbildung 5.4 den Zusammenhang zwischen einer intervallskalierten Variable (hier der IQ-Wert) und einer nominalskalierten Variable (hier die Zulassung zum Studium). Genauer ausgedrückt wird hier auf die x-Achse die

intervallskalierte Variable gezeichnet, auf die yAchse zeichnen wir aber die *Wahrscheinlichkeit* dafür, dass eine bestimmte Ausprägung der nominalskalierten Variable vorliegt. In diesem (fiktiven) Beispiel ist es so, dass es einen Zusammenhang zwischen der Intelligenz und der Wahrscheinlichkeit zugelasse zu werden gibt.

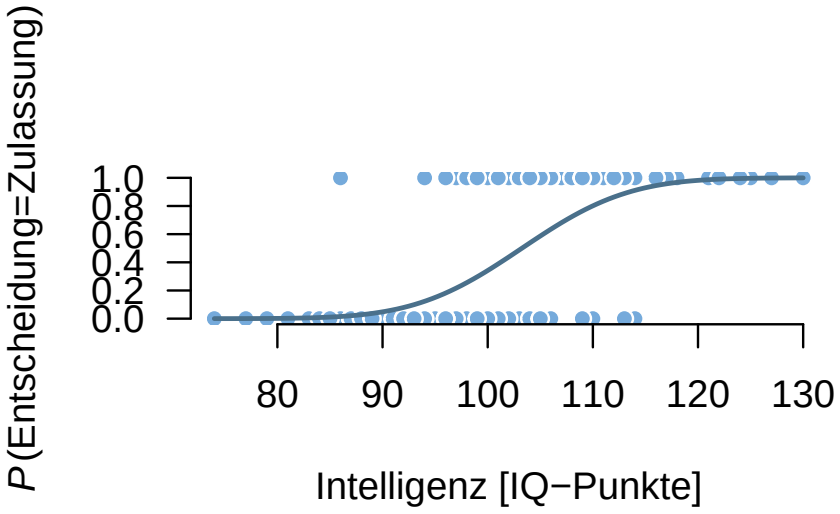


Abbildung 5.4: Der Zusammenhang zwischen dem Abschneiden in einem allgemeinen Intelligenztest (x-Achse) und der Wahrscheinlichkeit, für den Studiengang *Verteidigung gegen die dunklen Künste* (y-Achse) zugelassen zu werden.

Abbildung 5.5 zeigt die gleichen Daten bzw. auch den gleichen Zusammenhang: Wie man sehen kann, kann man einen solchen Unterschied zwischen zwei Gruppen (Bewerber, die zugelassen wurden, sind auch im Mittel intelligenter) auch so verstehen, dass es einen *Zusammenhang zwischen Gruppenzugehörigkeit und Intelligenz gibt*.

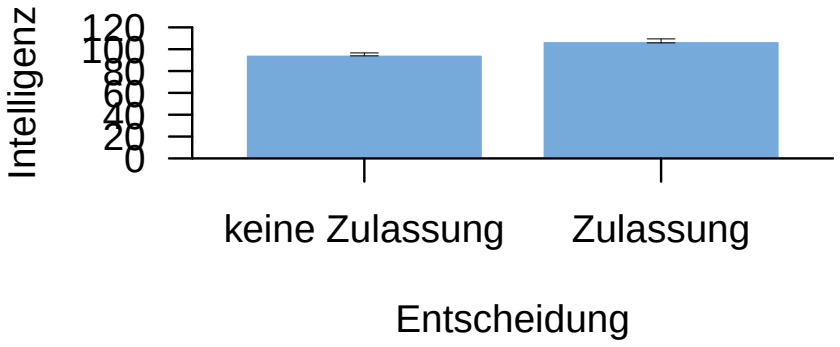


Abbildung 5.5: Der Zusammenhang zwischen der Zulassung zum Studiengang *Verteidigung gegen die dunklen Künste* (x-Achse) und dem Abschneiden in einem allgemeinen Intelligenztest (y-Achse). Für jeden

5.2 *Regression*

Hierdurch ist es nun also auch möglich, von den Werten der einen Variablen auf die Werte der anderen Variablen schließen. Mit welcher

Genauigkeit man dies kann, hängt von der *Stärke* des Zusammenhangs ab.

6
Der Einfluss von Drittvariablen

Bisher haben wir uns damit beschäftigt, den Zusammenhang zwischen zwei Variablen zu beschreiben. Im folgenden Kapitel soll es weiterhin darum gehen, den Zusammenhang zwischen zwei Variablen zu beschreiben; darüber hinaus soll es aber darum gehen, diesen Zusammenhang von dem Einfluss einer weiteren Variable (einer Drittvariable) zu bereinigen.

Warum dies wichtig ist, lässt sich anhand des *Simpson-Paradoxes* verdeutlichen, das im folgenden Abschnitt vorgestellt werden soll.

6.1 Das Simpson-Paradox

Das Simpson-Paradox stellt eine besondere Form von Zusammenhang dar, bei dem es zunächst so aussieht, dass es einen positiven (oder negativen) Zusammnhang zwischen einer Prädiktor- und einer Kriteriumsvariable gibt. Berücksichtigt man jedoch den Einfluss einer weiteren (Prädiktor-)Variable (der *Drittvariable*) auf die Kriteriumsvariable, dreht sich der Zusammenhang zwischen der ursprünglichen Prädiktor- und der Kriteriumsvariable um.

Für solch eine Situation gibt es einen prominenten Fall: Im Jahr 1973 wurde die University of California-Berkeley verklagt, Frauen bei der Zulassung zur Graduate School zu benachteiligen. Die Daten sprachen zunächst deutlich dafür, dass hier aufgrund des Geschlechts diskriminiert wurde, denn von den Männern wurden 44% der Bewerber zugelassen, von den Frauen jedoch nur 35%. Anders ausgedrückt hatten Männer also eine ungefähr 20% höhere Wahrscheinlichkeit zugelassen zu werden als Frauen. Abbildung 6.1 zeigt diesen zunächst bebobachteten Zusammenhang.

Bickel, Hammel und O'Connell (**?**) untersuchten dieses Phänomen genauer: Hierzu betrachteten sie die Zulassungszahlen nicht nur getrennt nach Geschlecht, sondern auch zusätzlich danach, welches Department der Universität die Bewerber zugelassen hat. Abbildung 6.2 zeigt den jeweiligen Anteil zugelassener Bewerber getrennt nach

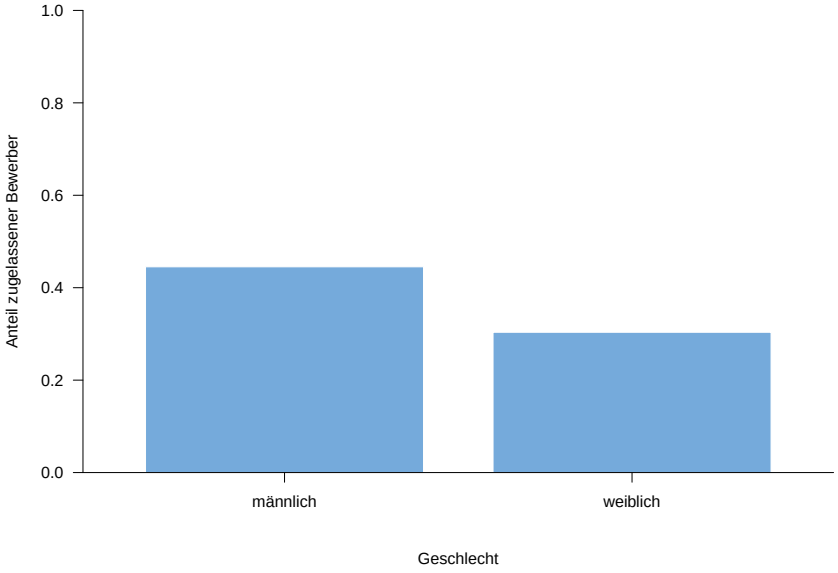


Abbildung 6.1: Der Anteil zugelassener Studierender an der Gesamtzahl der Studienbewerberinnen und -bewerber, getrennt nach *Geschlecht*.

Geschlecht und Department (für die sechs größten Departments).

Man kann erkennen, dass es offenbar innerhalb jedes Departments kaum Unterschiede bzgl. des Anteils zugelassener Bewerber gibt, Frauen und Männer also ungefähr die gleiche Wahrscheinlichkeit hatten, zugelassen zu werden. Die Autoren fanden sogar einen sehr kleinen Effekt, dass Frauen geringfügig "bevorzugt" wurden – die Richtung des Zusammenhangs dreht sich um, es handelt sich also um einen Fall von Simpsons Paradox.

Aber wie erklären die Autoren, dass sich der ursprünglich beobachtete Effekt umkehrt, wenn man die Daten der unterschidlichen Departments getrennt betrachtet? Betrachten wir noch einmal Abbildung 6.2: Auch wenn sich die Balken zwischen Männern und Frauen kaum unterscheiden, ist es doch offensichtlich, dass die Departments (unabhängig vom Geschlecht) sehr unterschiedlich große Anteile an Bewerbern zulassen. So ist es z.B. viel leichter, in Department A zugelassen zu werden als in Department F. Zusätzlich hierzu muss man berücksichtigen, dass sich Männer und Frauen unterschiedlich häufig in den unterschiedlichen Departments beworben haben, nämlich Frauen häufiger in den "schweren" und Männer häufiger in den "leichten" Departments. Diesen Zusammenhang zeigt Abbildung 6.3.

Zusammenfassend war es also nicht so, dass die Universität im Rahmen des Zulassungsverfahrens Frauen benachteiligte. Die unterschiedlichen Fächerpräferenzen von Männern und Frauen und die unterschiedlichen Zulassungsquoten der Fächer (Departments) führte jedoch dazu, dass Frauen eine geringere Wahrscheinlihckeit hatten, zugelassen zu werden. Bickel, Hammel und O'Connell (**?**) machen

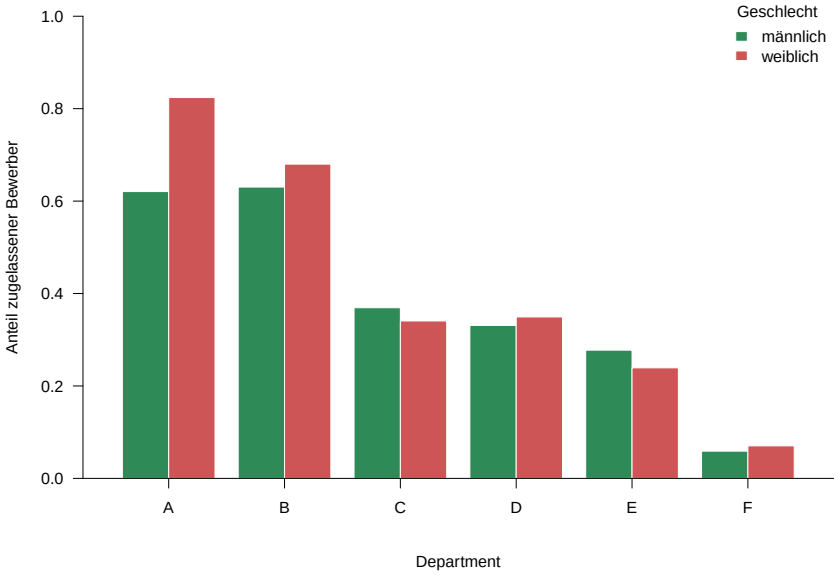


Abbildung 6.2: Der Anteil zugelassener Studierender an der Gesamtzahl der Studienbewerberinnen und -bewerber, getrennt nach *Geschlecht* und zulassendem *Department*.

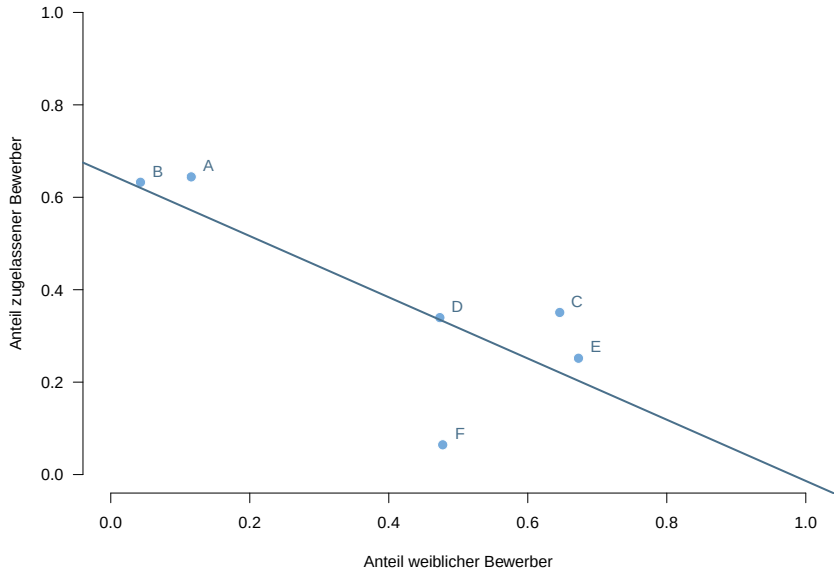


Abbildung 6.3: Streudiagramm des Anteils zugelassener Bewerberinnen und Bewerber (y-Achse) und des Anteils weiblicher Bewerberinnen (x-Achse) für die sechs größten Departments der University of California-Berkeley.

aber auch deutlich, dass dies nicht bedeutet, dass die Welt diskriminierungsfrei ist: Vielmehr lässt sich einerseits die Frage stellen, (1) warum überhaupt Männer und Frauen unterschiedliche Fächerpräferenzen haben und (2) warum in den von Frauen bevorzugten Fächern offenbar pro Bewerber weniger Studienplätze zur Verfügung stehen:

> Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

6.2 Kontrolle von Drittvariablen

In der Realität ist es meist so, dass es nicht nur eine, sondern viele Ursachen für die beobachtete Ausprägung eines Merkmals gibt (man spricht von *Multi-Determiniertheit*).

Lösungen

Kapitel 1: Messen

Überlegt in 2er- oder 3er-Gruppen, welches Skalenniveau die folgenden Variablen aufweisen:

a) Das bei uns geläufige metrische System zur Messung von Distanzen in Millimetern, Zentimetern, Metern, oder Kilometern;
b) die Nummern der Straßenbahnen der KVB (Kölner Verkehrsbetriebe);
c) Schulnoten;
d) Postleitzahlen.

Das metrische System weist Verhältnisskalenniveau auf: Die Relation *Gleichheit/Verschiedenheit* ist erhalten, denn 1 *m* ist etwas anderes als 2 *m*, aber 1 *m* ist immer das gleiche. Die Relation *Ordnung* ist erhalten, denn 1 *m* ist kürzer als 2 *m* ist kürzer als 3 *m*. Die Relation *Größe der Verschiedenheit* ist erhalten, denn der Unterschied zwischen 1 *m* und 2 *m* ist genauso groß wie der Unterschied zwischen 2 *m* und 3 *m*. Die Relation *Verhältnis der Merkmalsausprägung* ist erhalten, 2 *m* sind doppelt so lang wie 1 *m* und 4 *m* sind doppelt so lang wie 2 *m*. Die Relation *absoluter Werte* ist nicht gegeben, denn die Länge eines Meters ist willkürlich gewählt, man hätte genauso gut eine kürzere oder längere Länge festlegen können. (*Übrigens*: Ein Meter entspricht genau der Strecke, die Licht in $\frac{1}{299{,}792{,}458} s$ im Vakuum zurücklegt.) Da alle Relationen außer der Relation absoluter Werte gegeben sind, ist Verhältnisskalenniveau gegeben.

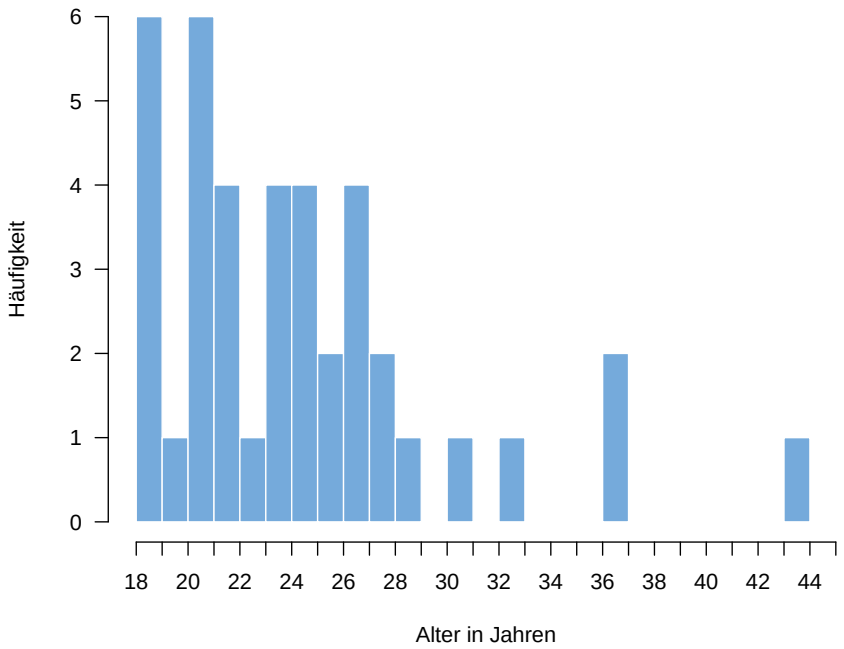
Die Nummern der Straßenbahnen der KVB weisen *bestenfalls* Nominalskalenniveau auf: Es ist vermutlich so gedacht, dass die Nummer die gefahrene Strecke anzeigen soll, das ist auch häufig der Fall. Unterschiedlich Nummern bezeichnen unterschiedliche Strecken, gleiche Nummern bezeichnen oft, aber nicht immer die gleiche Strecke: Gelegentlich fährt eine Linie 1 z.B. nicht zur Endhaltestelle Bensberg, sondern zur Endhaltestelle Zündorf. Man kann also nicht davon ausgehen, dass zwei Straßenbahnen mit der gleichen Nummer auch die gleiche Strecke fahren.

Über Schulnoten lässt sich immer wieder gut streiten. Ist die Benotung fair, ist die Relation von *Gleichheit/Verschiedenheit* erhalten, denn zwei unterschiedliche Schulleistungen sollten sich auch in unterschiedlichen Noten niederschlagen. Ungefähr gleiche Leistungen werden auch gleich benotet. Die Relation der *Ordnung* ist ebenfalls erhalten, denn eine 1 ist besser als eine 2 ist besser als eine 3. Die Relation der *Größe der Verschiedenheit* ist aber vermutlich meist selbst bei den besten und bemühtesten Lehrern nicht gegeben: Oft ist es der Leistungsunterschied zwischen einer 1 und einer 2 eben doch nicht genauso groß wie der Unterschied zwischen einer 2 und einer 3. Deshalb kann man sagen, dass Schulnoten *Ordninalskalenniveau* aufweisen. (Wir werden später sehen, warum es dann auch eigentlich nicht geschickt ist, eine "Durchschnittsnote" zu berechnen.)

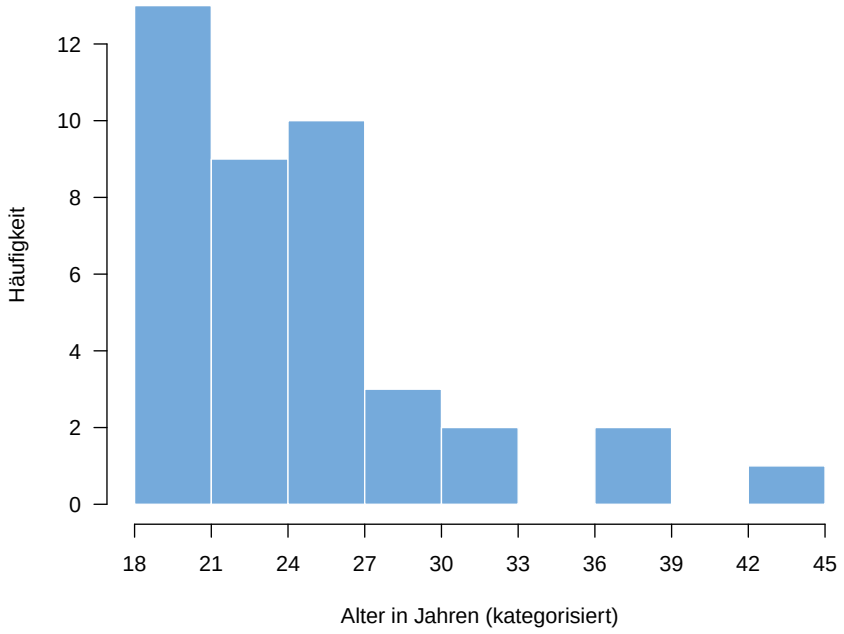
Postleitzahlen weisen *Nominalskalenniveau* auf: Nur die Relation von *Gleichheit/Verschiedenheit* ist erhalten: Unterschiedliche Zahlen bedeuten unterschiedliche Postleitzahlenbereiche (z.B. Stadtteile), gleiche Postleitzahlen bedeuten, dass es auch der gleiche Postleitzahlenbereicht ist. Die Relation der *Ordnung* ist schon nicht mehr wirklich gegeben: Zwar kann man eine grobe Ordnung nach Regionen feststellen, diese ist aber sehr unsystematisch.

Kapitel 2: Beschreiben und Zusammenfassen

1. Zeichnet ein Histogramm der Variable *Alter* in der folgenden Tabelle.



2. Zeichnet ein Histogramm der Variable *Alter* aus der vorigen Aufgabe. Zeichnet dieses Mal jedoch eine sekundäre Häufigkeitsverteilung, beginnend mit der Kategorie "18-21 Jahre".



3. Berechnet den Modus der Variable *Geschlecht* aus Aufgabe 1.

   $Mo =$ "weiblich"

4. Berechnet den Mittelwert der Variable *Alter* aus Aufgabe 1.

   $M = 24.85$

5. Erläutert, warum es nicht sinnvoll ist, den Median der Variable *Geschlecht* zu berechnen.

Es ist nicht sinnvoll, den Median von Geschlecht zu berechnen, da sich die Merkmalsausprägungen (weiblich vs. männlich) in keine sinnvolle Ordnung bringen lassen. Entsprechend weist die Variable Geschlecht nur Nominalskalenniveau auf.