Mathe 2: Statistik
TIT17/TIM17

Dr. M. Oettinger 2019

Inhaltsverzeichnis

1 Einleitung **4**

1.1 Was ist Statistik? 4

1.2 Historisches 5

1.3 Häufige Probleme der Statistik 6

2 Element Zufall **9**

2.1 Ereignisse 10

2.2 der Begriff der Wahrscheinlichkeit 11

2.2.1 Wahrscheinlichkeit und relative Häufigkeit 12

2.2.2 Axiome der Wahrscheinlichkeitsrechnung 13

2.3 Laplace-Experimente 14

2.4 Rechnen mit Wahrscheinlichkeiten 16

3 Grundlagen der deskriptiven Statistik **17**

3.1 Begriffe 17

3.2 Klassifizierung statistischer Merkmale 17

3.3 Darstellung statistischer Information 18

3.3.1 Altersverteilung 19

3.3.2 Häufigkeitsverteilung 20

3.3.3 Klassierte Daten: Verteilung der Körpergröße 23

3.4 Kumulierte Häufigkeitsverteilungen 25

4 Statistische Analyse **29**

4.1 Lagemaße 29

4.1.1 das arithmetische Mittel 29

4.1.2 Alternative Berechnung des arithmetischen Mittels . . . 34

4.1.3 arithmetisches Mittel bei klassierten Daten 34

4.1.4 das geometrische Mittel 36

4.1.5 Harmonisches Mittel 39

4.1.6 Median 41

4.1.7 Ermittlung des Medians bei klassierten Daten 44

4.1.8 der Modus 47

4.1.9 Quantile 48

4.1.10 Zusammenfassung: Lageparameter 52

4.1.11 Übungsaufgaben zu den Lageparametern 53

4.2 Streuung 54

4.2.1 Spannweite 54

4.2.2 mittlere absolute Abweichung 56

4.2.3 empirische Varianz und Standardabweichung 58

4.2.4 Variationskoeffizient 62

4.3 Schiefe 63

4.3.1 Statistische Momente 65

4.3.2 Streuungs- und Schiefemaße 66

4.4 Übungsaufgaben zu Streuungs- und Schiefemaßen 68

4.5 Konzentration und Disparität 69

4.5.1 Lorenzkurve 70

4.5.2 GINI-Koeffizient 72

4.5.3 Maximalwert des GINI-Koeffizienten 75

4.5.4 normierter GINI-Koeffizient 76

4.6 absolute Konzentration 77

4.6.1 Übungsaufgaben 80

5 Bivariate Verteilungen **82**

5.1 Kreuztabellen 82

5.2 Lineare Regression 83

5.2.1 Die Kovarianz 85

5.2.2 Lineare Regression 87

5.3 Der Korrelationskoeffizient 90

5.3.1 empirischer Korrelationskoeffizient 91

5.3.2 Interpretation des Korrelationskoeffizienten 91

5.3.3 Das Bestimmtheitsmaß R^2 93

A Lösungen zu den Übungsaufgaben im Skript **100**

A.1 Lageparameter 100

A.2 Streuungsmaße 100

A.3 Konzentration 101

1 Einleitung

1.1 Was ist Statistik?

Diese einfach klingende kurze Frage ist alles andere als einfach zu beantworten. Während die meisten Wissenschaften zumindest formal eine klare Definition besitzen und sich deshalb eindeutig von anderen Wissenschaften abgrenzen können, gelingt dies bei der Statistik nicht so einfach. Ein Anhaltspunkt dafür sind schon die vielen sehr unterschiedlichen Definitionen von 'Statistik', die in der Literatur zu finden sind.

In der deutschen Sprache hat das Wort 'Statistik' unterschiedliche Bedeutungen:

- Statistik im Sinne einer Sammlung von Daten (Synonym für Tabelle)
- Statistik im Sinne einer Kennzahl (aus dem englischen *statistic*)
- Statistik als Aktivität der Datensammlung oder -erhebung
- Statistik als wissentschaftliche Disziplin, die Lehre von Methoden zum Umgang mit quantitativen Informationen (Daten)

Wir werden hier die Definition des Duden (Das große Wörterbuch der deutschen Sprache) benutzen:

Statistik , die, -, -n: 1. Wissenschaft von der zahlenmäßigen Erfassung, Untersuchung u. Auswertung von Massenerscheinungen. 2. schriftlich fixierte Zusammenstellung, Aufstellung der Ergebnisse von Massenuntersuchungen, meist in Form von Tabellen od. grafischen Darstellungen.

Die Bedeutung der Statistik liegt in ihrer Fähigkeit, komplexe Datenmengen durch Reduktion verständlich darzustellen und verallgemeinerte Schlüsse von vorhandenen Daten auf zukünftige Daten oder verallgemeinerte Populationen zu liefern (ein bekanntes Beispiel hierfür sind Umfragen in der Politik). Die häufigsten Probleme sind dabei beschränkte Datenmengen (Stichproben). Die Beschränkung der Datenmenge ist meist aus praktischen Gründen notwendig (bei einer Umfrage über einzu erwartendes Wahlergebnis ist es beispielsweise schlicht nicht möglich, alle Wähler in Deutschland zu befragen).

Grundkenntnisse der Statistik ermöglichen es ...
kleine statistische Anwendungsprobleme mit den eigenen Daten selbst zu lösen;
bei größeren Problemen sinnvoll mit einem beratenden Statistiker zusammen zu arbeiten;
die Statistik in Veröffentlichungen (wenigstens in den Grundzügen) zu verstehen;
die vielen missbräuchlichen Anwendungen und Fehler leichter zu durchschauen und selbstständig zu beurteilen.

1.2 Historisches

Die Statistik hat zwei vollkommen unterschiedliche Wurzeln. Eine wichtige Grundlage wurde bereits im 17.Jahrhundert mit der *Wahrscheinlichkeitsrechnung* gelegt, als sich bedeutende Mathematiker wie Pascal oder Laplace und Glücksspieler wie Girolamo Cardano für die Mechanismen bzw. den Determinismus von Glücksspielen zu interessieren begannen. Determinismus bedeutet hier die Möglichkeit, auch über zufällige Ereignisse sichere Aussagen machen zu können, wenn man diese Ereignisse nur oft genug wiederholt. Diese Möglichkeit war zuvor einfach nicht denkbar. Erst als die Vereinbarkeit von Determinismus und Wahrscheinlichkeiten erkannt wurde, konnte die Wahrscheinlichkeitstheorie wissenschaftlich behandelt und entwickelt werden.

Der zweite wichtige Ausgangspunkt lag in der 'Zustandsbeschreibung des Staates' (lat. Status: Zustand). Bereits im 16. Jahrhundert wurden in vielen Pfarrgemeinden Geburten und Sterbefälle aufgezeichnet. Die Erhebung dieser Daten war auch für die Regierungen vieler Staaten von Interesse, wurde aber in verschiedenen Regionen mit sehr unterschiedlicher Konsequenz und Genauigkeit vorangetrieben. Ab dem 19. Jahrhundert wurde die Wissenschaft Statistik mit der Gründung von statistischen Gesellschaften (v.a. in England) erstmals institutionalisiert. Gleichzeitig war man sich einig, eine ganz bestimmte Richtung vertreten zu wollen. Die Statistik sollte zu damaligen Zeitpunkt möglichst objektiv neutrales Wissen ansammeln und Aufzeichnungen zur Verfügung stellen, keinesfalls aber über Ursachen und Wirkungen nachdenken. Die Herausforderung bestand damals in der Verwaltung, Handhabung und v.a. *Beschreibung* großer Datenmengen. Selbstverständlich wurden aber auch bereits zu dieser Zeit statistische Erkenntnisse als Grundlage für wichtige Entscheidungen, etwa in der Ökonomie oder der Gesetzgebung verwendet.

1.3 Häufige Probleme der Statistik

Gibt es schlechte Statistik? Ja, leider nur allzuviel davon! Die Statistik erlaubt es, große Mengen an erhobenen Daten einfach und verständlich - oft in einer einzigen Kennzahl ausgedrückt - darzustellen. Das genaue Vorgehen bei der Erhebung von Daten, aber auch bei deren Weiterverarbeitung mittels statistischer Modelle, bleibt dabei oft im Dunklen. Teils aus Unkenntnis, teils aber auch beabsichtigt (mit Zahlen, die man durch Anfertigen einer eigenen Statistik in die gewünschte Richtung verändern kann, lässt sich vortrefflich Werbung oder Politik machen) werden ständig sachlich falsche oder zumindest schlechte, wenig aussagekräftige Statistiken in Umlauf gebracht. Einige Beispiele für schlechte Statistik:

Relevanz der Stichprobe: Jahr für Jahr besagt die Statistik dass Ausländer, gemessen an ihrem Anteil an der Bevölkerung, einen überproportional hohen Prozentsatz der verurteilten Straftäter stellen.
Die Zahl stimmt - die Interpretation der Zahl durch (hauptsächlich) die Boulevardpresse ist jedoch meist falsch. Denn mitgerechnet werden bei den Ausländern auch Touristen, Durchreisende, illegal Eingewanderte, Nato-Soldaten und Personen, die nur eingereist sind, um Straftaten zu begehen. Ein weiterer sehr heikler Faktor der Berechnung: es werden auch Straftaten gezählt, die überhaupt nur von Ausländern begangen werden können: Verstöße gegen das Ausländergesetz und Asylverfahrensgesetz.
Ein weiteres Beispiel ist die Statistik zur Arbeitslosenzahl. Die von der Arbeitsagentur veröffentlichten Zahlen decken sich nie mit denen des statistischen Bundesamtes.

Kausalität und Koinzidenz: Oft wird (beispielsweise in der Werbung) vom gleichzeitigen auftretenzweier Tatbestände (Koinzidenz) auf eine Kausalität zwischen beiden (ursächlicher Zusammenhang oder Beeinflussung) geschlossen. Ein etwas konstruiertes Beispiel dafür ist die Geschichte der Klapperstörche und der Geburtenrate. Nehmen wir mal an, in Schweden sei die Geburtenrate besonders hoch, ebenso die Zahl der Störche. In einem Vergleichsort wie Berlin ist die Geburtenrate sehr niedrig, und es gibt wenig Störche. Daraus könnte man schließen, dass die Störche die Kinder bringen, und Tatsächlich gibt es hier und da eine gleichzeitige Zunahme von Storchen- und Kinderzahl - aber beide Tatbestände hängen nicht ursächlich miteinander zusammen, sondern sind

jeder für sich die Folge einer dritten Größe: In Schweden ist es besonders ländlich, Störche haben auf dem Land größere Überlebenschancen, und Menschen kommen hier ebenfalls auf eine höhere Geburtenrate als in Großstädten.

Männer mit wenig Kopfhaar verdienen mehr Geld. Natürlich tun sie das - Männer mit Haarschwund sind tendenziell älter und verdienen *deshalb* meist mehr.
Es sterben mehr Menschen in Krankenhäusern als zu Hause. Natürlich ist das so - in Krankenhäusern befinden sich viele Erkrankte, deren Sterberisiko höher ist.

Umfragen und ihre Teilnehmer: Kritisch sollte man auch Statistiken begegnen, die auf Umfragen beruhen. Wer im Yachthafen fragt: 'Wie viel verdienen Sie im Monat?', darf die Antworten nicht als repräsentativ für die ganze Bevölkerung ansehen, weil überdurchschnittlich viele Gutverdienende sich die Zeit am Wochenende beim segeln vertreiben.
Die amerikanische Militärregierung ließ nach dem Krieg in Deutschland den Ernährungszustand der Deutschen ermitteln und stellte dafür Waagen an Bahnhöfen und öffentlichen Plätzen auf. Ausschließlich gesunde Menschen gerieten in die Stichprobe, hungernde Bettlägerige nicht.
Auch die Fragetechnik kann eine Statistik deutlich verändern. Fragte man Firmenchefs, ob sie etwas dagegen hätten, wenn ihre Angestellten beim Arbeiten essen, würden sie wohl mit Ja antworten. Fragte man dieselben Firmenchefs, ob sie etwas dagegen hätten, wenn ihre Angestellten beim Essen arbeiten, würden sie (vermutlich) eher mit Nein antworten. Eine Umfrage über 'Abtreibung' fällt anders aus als eine zum Thema 'Schutz des ungeborenen Lebens'.
Es gibt Umfragen, bei denen man von vornherein nicht mit einer ehrlichen Antwort rechnen kann: 'Schlagen Sie Ihre Kinder?'

Unklare Begriffe: Unsinnig wird eine Statistik, wenn sie mit schwammigen Begriffen hantiert, wie etwa: 'Ist Fliegen sicher?' - der Begriff 'sicher' kann nicht eindeutig definiert werden. Man liest häufig, dass Fliegen statistisch gesehen sicherer als Autofahren ist: auf eine Milliarde Passagierkilometer kommen im Flugverkehr 0,3 Tote, beim Autofahren sind es vier. Legt man der Statistik aber nicht die zurückgelegten Kilometer zu Grunde, sondern die Anzahl der Reisen, sieht das Bild ganz anders aus: Auf eine Milliarde Flüge kommen 55 Tote, auf eine Milliarde Autofahrten 45. Eine Fluglinie würde aus vorhandenen Daten andere Schlüsse ziehen als ein Autoverleih!

Bei der Erstellung, aber auch beim Lesen von Statistiken ist es also durchaus angebracht, einige wichtige Fragen kritisch zu beleuchten. Ist die Stichprobe so angelegt, dass ein repräsentativer Wert zu erwarten ist? Ist sie ausreichend groß? Wie hängen die betrachteten Größen miteinander zusammen? Wie wurden die Daten in einer Umfrage erhoben? Mit etwas gesundem Menschenverstand können Fehler bei der Interpretation oder Durchführung meist relativ leicht vermieden oder erkannt werden.

2 Element Zufall

Nicht nur in den Naturwissenschaften ist es wichtig, die studierten Phänomene möglichst genau beschreiben zu können. Wir alle wüssten manchmal ganz gern über Dinge, die uns oder unsere Umgebung beeinflussen genau genug Bescheid, um Voraussagen über das zukünftige Geschehen machen zu können. Die Natur setzt dieser Bestrebung allerdings durch das Element Zufall recht enge Grenzen.

Von manchen Ereignissen sagen wir, dass sie zufällig geschehen. Damit drücken wir aus, dass wir diese Ereignisse nicht mit Sicherheit vorhersehen können. Der Grund für die Unvorhersehbarkeit eines solches Experiments kann eine grundsätzliche *Unbestimmtheit* sein, Beispiele hierfür sind die Freiheit menschlicher Erkenntnisse oder das physikalische Verhalten sehr kleiner Teilchen. Genauso kann aber aber auch schlichte *Unkenntnis* bzw. die Unmöglichkeit, die relevanten Einflüsse während des Ereignisses zu erfassen, dazu führen, dass ein Ereignis nicht vorhersehbar ist. Beispiele sind hier der Würfel oder auch das Wetter an einem bestimmten Ort - bei beiden sind die von außen einwirkenden Einflüsse bekannt und die Auswirkungen berechenbar, es ist aber nicht möglich, den Ausgangszustand genau genug zu erfassen. Für die hier verwendete mathematische Wahrscheinlichkeitsrechnung spielen die Gründe für die Unvorhersehbarkeit der Geschehnisse keine Rolle.

Wenn sich die Mathematik mit dem Zufall beschäftigt, so benötigt sie Modelle von Situationen, deren Ausgang unsicher ist, die sich aber mit mathematischen Mitteln beschreiben lassen. Derartige Modelle nennen wir (ideale) Zufallsexperimente (oder Zufallsversuche). Die anschaulichsten Zufallsexperimente stammen aus einem Bereich, der einerseits sehr strenge und wohldefinierte Regeln besitzt, bei dem aber andererseits die Unsicherheit ausdrücklich gewünscht ist: dem Glücksspiel, das auch der Ausgangspunkt für die Entwicklung der mathematische Behandlung von Wahrscheinlichkeiten war.

Beispiel: eines Zufallsexperiments: Werfen eines (idealen) Würfels.

Ideal bedeutet, dass der Würfel jeder Augenzahl exakt die gleiche Chance gibt - an diese Voraussetzung kann man sich in der Realität zwar recht gut annähern, sie ist letzten Endes aber nicht erreichbar. Die möglichen Versuchsausgänge sind hier natürlich die erreichbaren Augenzahlen $(1, 2, 3, 4, 5, 6)$.

Beispiel: es werden zwei unterscheidbare (ideale) Würfel geworfen.

Dabei sollen die beiden Würfel unabhängig voneinander fallen, d.h. das Verhalten des einen soll das Verhalten des anderen nicht beeinflussen. Notiert man das Ergebnis in Klammern in der Reihenfolge (Erg. Würfel 1, Erg. Würfel 2), so sind die Ergebnisse die 36 möglichen Paare von Augenzahlen: $(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),...(6,4),(6,5),(6,6)$

Wie diese Beispiele zeigen, ist ein Zufallsexperiment immer eine gedankliche Konstruktion. Es muss, wie alle mathematischen Konstruktionen, *wohldefiniert* sein. Und genau wie in anderen Gebieten der Mathematik können gedankliche Konstruktionen meist nur näherungsweise auf die Wirklichkeit angewandt werden (z.B. auf einen realistischen Würfel). Jedes (ideale) Zufallsexperiment besitzt eine festgelegte Menge möglicher Versuchsausgänge. Jeder Versuchsausgang wird auch **Elementarereignis** genannt. Die Menge aller Elementarereignisse nennen wir den **Ereignisraum**. Für die obigen Beispiele sind die Ereignisräume die Mengen der möglichen Versuchsausgänge. Für den einzelnen Würfel ist also der Ereignisraum die Menge der Augenzahlen $\{1,2,3,4,5,6\}$

2.1 Ereignisse

Der Begriff **Ereignis** beschreibt eine Zusammenfassung von Versuchsausgängen (also Elementarereignissen). Präziser ausgedrückt ist ein Ereignis eine Teilmenge des Ereignisraumes. Jedes einzelne Elementarereignis ist auch ein Ereignis, aber es gibt im Allgemeinen mehr davon. Für den einzelnen Würfel ist beispielsweise 'die Augenzahl ist zwei' ein Ereignis (2 ist eine Teilmenge des Ereignisraumes), 'die Augenzahl ist gerade' ebenfalls (dies entspricht der Teilmenge $\{2,4,6\}$ des Ereignisraumes). Wie diese Beispiele zeigen, können Ereignisse oft verbal als 'Aussagen' formuliert werden, die eine Beschreibung ihrer Elemente darstellen. Wichtig ist dabei, dass jede solche Aussage eine Teilmenge des Ereignisraumes eindeutig festlegt (es kann manchmal schwierig sein, alle ihre Elemente aufzulisten).

Wird ein Zufallsexperiment ausgeführt, so sagen wir, dass ein Ereignis A eintritt, wenn der Ausgang des Versuchs in der Menge A enthalten ist. Wurde im Beispiel etwa 'Augenzahl 4' gewürfelt (das ist der Versuchsausgang), so ist damit das Ereignis 'Die Augenzahl ist gerade' eingetreten. Die Ereignisse 'Die Augenzahl ist 2' und 'Die Augenzahl ist ungerade' sind nicht eingetreten. Wichtig: *Versuchsausgang* und *Ereignis* sind im Allgemeinen unterschiedlich - mit einem Versuchsausgang treten meist viele unterschiedliche Ereignisse ein!

2.2 der Begriff der Wahrscheinlichkeit

Bei zufallsbehafteten Ereignissen oder Experimenten kann die Mathematik keine Aussagen über das Eintreffen oder Ausbleiben eines bestimmten Ausgangs treffen. Dennoch kann auch das Element des Zufalls unter gewissen Bedingungen mathematisch erfaßt werden. Es ist nämlich möglich, ein Maß für die Sicherheit (oder Unsicherheit) anzugeben, die mit einer Aussage verbunden ist. Ein solches Maß ist die Wahrscheinlichkeit.

Die Wahrscheinlichkeitsrechnung ordnet jedem Ereignis eines Zufallsexperiments eine Wahrscheinlichkeit für sein Eintreten zu. Nennen wir ein Ereignis A, so wird die ihm zugeschriebene Wahrscheinlichkeit mit $p(A)$ oder p_A bezeichnet. (Der Buchstabe p stammt vom englischen *probability*). Andere Bezeichnungen, die Sie in der Literatur finden, sind $P(A), P_A$ und Prob(A). Die Wahrscheinlichkeit für das Eintreten eines Ereignisses A ist immer eine reelle Zahl, für die

$$0 \leq p(A) \leq 1 \tag{1}$$

gilt. Die beiden Extremfälle geben absolute Sicherheit an

- Ist $p(A) = 1$, so tritt A mit Sicherheit ein
- Ist $p(A) = 0$, so tritt A mit Sicherheit nicht ein

Die Werte dazwischen drücken *Grade* an Sicherheit aus. Je größer die Wahrscheinlichkeit $p(A)$, umso eher ist anzunehmen, dass das Ereignis A eintritt. Was aber bedeutet das genau? Wie sind die Grade an Sicherheit, die durch Wahrscheinlichkeiten ausgedrückt werden, definiert?

2.2.1 Wahrscheinlichkeit und relative Häufigkeit

Bevor wir zur Berechnung von Wahrscheinlichkeiten kommen, müssen wir wissen, was sie bedeuten. Gehen wir von einem der einfachsten Zufallsexperimente aus: dem Würfel (Beispiel 2). Das Maß für die Sicherheit, die höchste Augenzahl 6 zu würfeln, könnte so formuliert werden: 'Ungefähr bei jedem sechsten Würfel-Versuch wird die Augenzahl 6 auftreten' oder auch 'Unter 6 Würfel-Versuchen wird ungefähr 1 mal die Augenzahl 6 auftreten'. Bei lediglich 6 Versuchen besteht keine Sicherheit, dass die gewünschte Augenzahl genau einmal eintritt, also würfeln wir öfter: 'Unter 6000 Würfel-Versuchen wird ungefähr 1000 mal die Augenzahl 6 auftreten'. Das klingt schon plausibler. Geht man noch einen Schritt weiter, so erhält man

'Unter einer sehr großen Zahl n von Würfel-Versuchen wird ungefähr $n/6$ mal die Augenzahl 6 auftreten'

Allgemein lässt sich formulieren: Wenn ein Zufallsexperiment in identischer Weise n mal durchgeführt wird und dabei genau m mal das Ereignis A eintritt, so heißt der Quotient

$$h(A) = \frac{m}{n} \tag{2}$$

die **relative Häufigkeit**, mit der das Ereignis A eingetreten ist. Die relative Häufigkeit wird nicht bei jeder Reihe von n Durchführungen des Versuchs gleich sein. Wenn aber n sehr groß ist, so wird sich jedes Mal ungefähr die gleiche relative Häufigkeit ergeben. Lässt man nun n gedanklich in einem Grenzprozess über jede Schranke wachsen, so nimmt die relative Häufigkeit einen festen, nur vom Zufallsexperiment und dem betrachteten Ereignis A abhängigen Wert annehmen. Diesen Wert nennen wir die Wahrscheinlichkeit des Ereignisses.

Die **Wahrscheinlichkeit** eines Ereignisses ist die vorausgesagte relative Häufigkeit seines Eintretens für eine gegen unendlich strebende Anzahl n von Durchführungen des betreffenden Zufallsexperiments (3)

Bemerkung: da man n in der Wirklichkeit nicht unendlich groß machen kann, handelt es sich hier, wie beim Begriff des Zufallsexperiments (siehe oben), um eine mathematische Idealisierung.

2.2.2 Axiome der Wahrscheinlichkeitsrechnung

Die Wahrscheinlichkeitsrechnung ist ein Teilgebiet der Mathematik. Es ist üblich, an den Anfang einer mathematischen Theorie einige Axiome zu setzen, aus denen sich dann alle weiteren Sätze dieser Theorie ableiten lassen. Die Axiome selbst werden sind nicht beweisbar, sie werden als gegeben angenommen. In der Regelbesitzen sie jedoch einen verständlichen Bezug zur Anschauung. Wir werden auch in der Wahrscheinlichkeitsrechnung auf diese Weise vorgehen und beginnen daher mit dem Axiomensystem, das 1935 von Kolmogorov[1] eingeführt wurde. Dieses Axiomensystem stellt die Grundlage der modernen Wahrscheinlichkeitsrechnung dar.

Axiom 1 (Nichtnegativität):

$$P(A) \leq 0 \tag{4}$$

Wahrscheinlichkeiten sind nichtnegative, relle Zahlen, die den Ereignissen zugeordnet sind.

Axiom 2 (Normierung):

$$P(\Omega) = 1 \tag{5}$$

Die Wahrscheinlichkeit des sicheren Ereignisses Ω ist 1, womit eine Normierung der Wahrscheinlichkeit erfolgt. Aus den beiden ersten Axiomen ergibt sich, dass Wahrscheinlichkeiten reelle Zahlen sind, die im Intervall $[0; 1]$ liegen.

Anschaulich gilt für den oben beschriebenen Zusammenhang zwischen relativer Häufigkeit und Wahrscheinlichkeit

- die relative Häufigkeit jedes Ereignisses A erfüllt stets $0 \leq h(A) \leq 1$, und daher gilt dies auch für jede Wahrscheinlichkeit. (Beweis: Tritt das Ereignis bei n-maliger Durchführung des Zufallsexperiments m mal ein, so gilt $0 \leq m \leq n$, woraus die Behauptung folgt).
- Tritt ein Ereignis A mit Sicherheit ein, so tritt es bei n-maliger Durchführung des Zufallsexperiments immer, d.h. n mal ein. Seine relative Häufigkeit ist gleich $n/n = 1$, und daher ist $p(A) = 1$.
- Tritt ein Ereignis A mit Sicherheit nicht ein, so tritt es bei n-maliger Durchführung des Zufallsexperiments nie, d.h. 0 mal ein. Seine relative Häufigkeit ist gleich $0/n = 0$, und daher ist $p(A) = 0$.

[1] Andrei Nikolajewitsch Kolmogorow, russischer Mathematiker, 1903-1987

2.3 Laplace-Experimente

Die einfachsten Zufallsexperimente sind dadurch gekennzeichnet, dass jeder Versuchsausgang gleich wahrscheinlich ist. Wir nennen sie Laplace-Experimente. Ein typisches Beispiel ist der (ideale) Würfel. Selbst wenn wir die Wahrscheinlichkeiten für das Eintreten der einzelnen Augenzahlen nicht kennen, sorgt seine perfekte (ideale) Form dafür, dass sie alle gleich groß sind. Diese Information reicht bereits aus, um sie konkret zu berechnen: Wird n mal gewürfelt, so sagen wir für große n und wegen der Gleichberechtigung der Augenzahlen voraus, dass *jede* gegebene Augenzahl $n/6$ mal eintreten wird. Die entsprechende Wahrscheinlichkeit ist mit (3) dann $(n/6)/n = 1/6$.

Axiom 3 (Additivität):

$$P(A \cup B) = P(A) + P(B), \text{ falls } A \cap B = \emptyset \tag{6}$$

Die Wahrscheinlichkeit einer Vereinigung disjunkter Ereignisse ist gleich der Summe der Einzelwahrscheinlichkeiten.

Beispiel:
$P(A)$ ist die Wahrscheinlichkeit, mit dem idealen Würfel eine 2 (Ereignis A) zu werfen, und $P(B)$ die Wahrscheinlichkeit des Ereignisses $B = 5, 6$. Da beide Ereignisse disjunkt sind, also keine Schnittmenge aufweisen, berechnet sich die Wahrscheinlichkeit dafür, dass A oder B eintritt nach Axiom 3 als Summe der Einzelwahrscheinlichkeiten:

$$P(A \cup B) = P(2; 5; 6) = \frac{3}{6} = \frac{1}{2}$$
$$P(A) + P(B) = P(2) + P(5; 6) = \frac{1}{2} + \frac{2}{6} = \frac{1}{2}$$

Ereignisse können also auch komplexer sein: sie sind Zusammenfassungen von Versuchsausgängen. So ist für den (idealen) Würfel auch 'Die Augenzahl ist gerade' ein Ereignis. Wie groß ist die Wahrscheinlichkeit für sein Eintreten? Dazu überlegen wir: Unter den 6 möglichen Augenzahlen (die möglichen Fälle) sind 3 geradzahlig (nämlich 2, 4 und 6). Jeder einzelne dieser günstige Fälle (und auch jeder einzelne ungünstige Fall) tritt bei n-maligem Würfeln für großes n gleich oft ein, nämlich $n/6$ mal, d.h. sein relativer Anteil ist 1/6. Jetzt

muss lediglich gezählt werden: der relative Anteil der günstigen Fälle (gerade Augenzahl) ist dreimal so groß wie der relative Anteil jeder einzelnen Augenzahl, also $3/6 = 1/2$. Daher ist die Wahrscheinlichkeit, eine gerade Augenzahl zu würfeln, genau $1/2$.

Hinter diesem Argument steckt eine Regel, die für beliebige Laplace-Experimente anwendbar ist und die Berechnung von Wahrscheinlichkeiten auf das Abzählen von Fällen reduziert. Die Anzahl aller möglichen Versuchsausgänge eines Laplace-Experiments (d.h. die Zahl der Elemente seines Ereignisraums) wird als 'Zahl der möglichen Fälle' bezeichnet. Alle diese Fälle sind für ein Laplace-Experiment gleich wahrscheinlich. Sei nun A ein betrachtetes Ereignis. Es besteht aus einer Anzahl bestimmter Versuchsausgänge ('Zahl der günstigen Fälle'), der Zahl der Elemente, die das Ereignis A - als Teilmenge des Ereignisraums - besitzt, oder, wiederum anders ausgedrückt, die Zahl der möglichen Versuchsausgänge, aus deren Eintreten das Eintreten von A folgt. Dann ist die Wahrscheinlichkeit für das Eintreten des Ereignisses A durch den Quotienten

$$p(A) = \frac{\text{Zahl der günstigen Fälle}}{\text{Zahl der möglichen Fälle}} \tag{7}$$

gegeben

Beispiel:
Um beim Werfen zweier Würfel die Wahrscheinlichkeit des Ereignisses 'die Summe der Augenzahlen ist gerade' zu berechnen, benötigt man

- die Zahl der möglichen Fälle. Sie beträgt 36 (s. Beispiel oben).
- die Zahl der Fälle, in denen die Summe der Augenzahlen gerade ist. Jeder Würfel hat 3 gerade und 3 ungerade Augenzahlen, also gibt es insgesamt 9 Versuchsausgänge der Form (gerade, gerade) und 9 Versuchsausgänge der Form (ungerade, ungerade) und damit 18 Ergebnisse mit gerader Summe.

Damit wird die Berechnung mit (7) ganz einfach:

$$p(\text{Die Summe der Augenzahlen ist gerade}) = 18/36 = 1/2.$$

2.4 Rechnen mit Wahrscheinlichkeiten

Wir gehen von einem Zufallsexperiment und dessen Ereignisraum aus. Zur Erinnerung:

- Der Ereignisraum - im Folgenden mit E bezeichnet - ist die Menge aller Versuchsausgänge (Elementarereignisse).
- Ein Ereignis ist eine Zusammenfassung von Versuchsausgängen und als Teilmenge in E enthalten.

Ereignisse können in verschiedener Weise in Beziehung zueinander stehen, und ein Ereignis kann aus anderen Ereignissen konstruiert werden. Da Ereignisse Teilmengen des Ereignisraums sind, können ihre Beziehungen in Begriffen der Mengenlehre ausgedrückt werden, sie können wie Mengen miteinander verknüpft werden.

Die Wahrscheinlichkeit des Komplementärereignisses $\bar{A}$ (d.h. es tritt *nicht* A ein) berechnet sich durch

$$P(\bar{A}) = 1 - P(A), \tag{8}$$

da entweder A oder $\bar{A}$ eintritt (die Summe der beiden Wahrscheinlichkeiten muss eins ergeben).

Beispiel:
A sei das Ereignis, eine 2 beim einmaligen Würfelwurf zu erzielen. Wie groß ist die Wahrscheinlichkeit für das Komplementärereignis $\bar{A}$?

$$P(\bar{A}) = P(1;3;4;5;6) = \frac{5}{6}$$
$$P(\bar{A}) = 1 - P(A) = 1 - \frac{1}{6} = \frac{5}{6}$$

3 Grundlagen der deskriptiven Statistik

3.1 Begriffe

Statistische Einheit, Merkmalsträger: Personen, Gegenstände aber auch Ereignisse wie Geburten oder Todesfälle, die (üblicherweise) in einer Stichprobe untersucht werden.

Merkmale: die bei einer statistischen Einheit interessierenden Eigenschaften, z.B. die Haar- oder Augenfarbe bei Personen, werden Merkmale genannt.

Merkmalsausprägungen: Alternativen, die von einer bei einer statistischen Einheit interseeierenden Eigenschaft angenommen werden können. Beispiele für Merkmalsausprägungen sind 'blond', 'rothaarig' oder 'schwarz' für die Eigenschaft 'Haarfarbe'.

Grundgesamtheit / statistische Masse: ist die Menge aller relevanten statistischen Einheiten mit übereinstimmenden sachlichen, räumlichen und zeitlichen Identifikationskriterien.

Bestandsmasse: statistische Einheiten mit einer von Null verschiedenen Lebensdauer. Beispielsweise stellt die Masse der Einwohner der Stadt Ravensburg eine Bestandsmasse dar, ebenso die Menge der Touristen, die den Bodensee besuchen. Das wesentliche Kriterium für eine Bestandsmasse ist: die Erfassung der Zahl der Einheiten, die zur Bestandsmasse gehören, erfolgt zu einem festgelegten *Zeitpunkt*, nicht über einen längeren Zeitraum hinweg.

Bewegungsmasse oder Ereignismasse: statistische Einheiten einer solchen Bewegungsmasse treten nur punktuell auf, sie haben keine von Null verschiedene Lebensdauer. Beispiele sind die Zahl der Geburten innerhalb eines Jahres, aber auch die Zuzüge zur Stadt Ravensburg. Wesentliches Charakteristikum einer solchen statistischen Masse ist: da die statistischen Einheiten keine Lebensdauer haben, erfolgt ihre Erfassung über einen längeren Zeitpunkt hinweg, nicht zu einem bestimmten Zeitpunkt.

3.2 Klassifizierung statistischer Merkmale

Nominale Merkmale: sind Merkmale, deren Merkmalsausprägungen keine natürliche Rangfolge aufweisen. Einzelne Merkmalsausprägungen kön-

nen deshalb nur danach beurteilt werden, ob sie entweder gleich oder aber verschieden sind. Beispiele für nominale Merkmale:

- Familienstand mit den Ausprägungen ledig, verheiratet, geschieden und verwitwet
- Geschlecht mit den Ausprägungen männlich und weiblich (Schnecken sind ausgenommen)
- Staatsangehörigkeit, Bundesland

Ordinale Merkmale: die Merkmalsausprägungen eines solchen Merkmals weisen eine natürliche Rangfolge auf. Beispiele für ordinale Merkmale sind:

- Klausurnoten mit den Ausprägungen sehr gut, gut, befriedigend, ausreichend, ...
- Hotelgüteklassen
- die Qualität von Statistikvorlesungen mit den Ausprägungen unter aller Sau, miserabel und erträglich

Kardinale bzw. metrische oder quantitative Merkmale: die Merkmalsausprägungen lassen sich in reellen Zahlen erfassen und weisen damit natürlich auch die Ordnungseigenschaften reeller Zahlen auf. Kardinale Merkmale können weiter in diskrete oder stetige Merkmale unterteilt werden:

Diskrete Merkmale: hier ist die Zahl der Merkmalsausprägungen entweder endlich oder abzählbar unendlich (die Merkmalsausprägungen besitzen keine obere oder untere Grenze, können aber mit natürlichen Zahlen durchnummeriert werden). Beispiele sind Semesterzahlen, Einwohnerzahlen etc.

Stetige Merkmale: die Zahl der Merkmalsausprägungen ist überabzählbar unendlich (Körpergewicht, Körpergröße, Alter).

3.3 Darstellung statistischer Information

Zur Darstellung von Information verwendet die Statistik üblicherweise Individualwerte (Einzeldaten) oder aber klassierte Daten (in denen die Information aus einer Stichprobe in Datenklassen, also Intervallen festgelegter Breite, erhoben oder angegeben wird). Für die Beispiele im folgenden Abschnitt werden der Anschaulichkeit halber anonymisierte Daten der Teilnehmer eines Statistik-Kurses aus Ravensburg benutzt.

3.3.1 Altersverteilung

In der nachfolgenden Tabelle sind als einfaches Beispiel das Geschlecht, die Körpergröße und das Alter der Teilnehmer eines Statistik-Kurses in Ravensburg aufgeführt.

w	180	18
w	168	28
w	167	23
w	176	20
w	168	20
w	162	19
w	166	20
m	183	21
m	175	29
w	168	21
w	172	19
w	164	21
w	165	20
w	177	21
m	181	19
m	176	21
w	176	19
w	178	20
m	199	22
w	160	20
w	168	21
m	181	24
w	170	19
m	184	22
w	171	19
m	185	25
m	184	40

Tabelle 1: die zu Beginn erhobenen Daten.

Die relevante Information des Merkmals Alter in der ersten Spalte der Tabelle lässt sich kürzer durch eine sog. *Urliste*, einen Vektor der einzelnen Daten, darstellen:

$$\{x_i\} = (18, 28, 23, 20, 20, 19, 20, 21, 29, 21, 19, 21, 20, 21, 19,$$

$$21, 19, 20, 22, 20, 21, 24, 19, 22, 19, 25, 40)$$

Dabei wird noch nicht auf Information verzichtet, sofern die Reihenfolge der Personen der Reihenfolge der Daten in der Urliste entspricht. Ebensogut könnte jedoch ein Vektor von $n = 32$ Zahlen benutzt werden, der die Alterswerte bereits in geordneter Form enthält:

$$(18, 19, 19, 19, 19, 19, 19, 20, 20, 20, 20, 20, 20, 21, 21,$$

$$21, 21, 21, 21, 22, 22, 23, 24, 25, 28, 29, 40)$$

Daraus geht allerdings nicht mehr hervor, welche Person mit welchem Alter verknüpft ist. Dieselbe Information liefert eine Tabelle, die die *absoluten* Häufigkeiten h_i bzw. *relativen* Häufigkeiten $f_i = h_i/n$ für den i-ten in der Stichprobe auftretenden Wert enthält:

Alter x_i	Häufigkeit h_i	$h_i \cdot x_i$	rel. Häufigkeit f_i
18	1	18	0.037
19	6	114	0.222
20	6	120	0.222
21	6	126	0.222
22	2	44	0.074
23	1	23	0.037
24	1	24	0.037
25	1	25	0.037
28	1	28	0.037
29	1	29	0.037
40	1	40	0.037
Summe	591		1,00

Tabelle 2: absolute Häufigkeiten h_i und relative Häufigkeiten f_i zum Alter.

3.3.2 Häufigkeitsverteilung

In einer Stichprobe vom Umfang n vorhandene Information eines kardinalen Merkmals X kann, sofern der Umfang der Stichprobe nicht allzugroß ist, allgemein natürlich auch in Form einer Urliste oder eines Vektors $(x_1, x_2, x_3, ...x_n)$ der bereits geordneten Werte angegeben werden. Sehr oft werden darüberhinaus einzelne Merkmalswerte mehrfach in einer Stichprobe beobachtet, so dass tatsächlich nur $m < n$ der enthaltenen Merkmalswerte verschieden sind.

Diese können wieder in einem neuen Vektor $(x_1, x_2, x_3, ...x_m)$ zusammengefaßt werden, der dann lediglich die verschiedenen auftretenden Merkmalswerte, üblicherweise in geordneter Reihenfolge, enthält. Um die erhobene Information sinnvoll wiederzugeben, muss zusätzlich ein Vektor der auftretenden absoluten oder relativen Häufigkeiten $(h_1, h_2, h_3, ...h_m)$ bzw. $(f_1, f_2, f_3, ...f_m)$ angegeben werden, der als (absolute bzw. relative) Häufigkeitsverteilung bezeichnet wird.

X	x_1	x_2	x_3	...	x_m	Summe
$h_i = h(X = x_i)$	h_1	h_2	h_3	...	h_m	n
$f_i = f(X = x_i)$	f_1	f_2	f_3	...	f_m	1

Häufig wird allerdings zur Veranschaulichung der in einer Stichprobe enthaltenen Information auf grafische Illustrationen zurückgegriffen. Bei unklassierten Daten kann dies Beispielsweise in Form von Stab-, Balken- oder Kreisdiagrammen geschehen, bei klassierten Daten werden normalerweise Histogramme verwendet.

Beispiel: Stab- und Balkendiagramm der Altersverteilung im Kurs.

Die absoluten Häufigkeiten, mit denen die unterschiedlichen Lebensalter im Kurs auftreten, sehen in einem Stabdiagramm folgendermaßen aus:

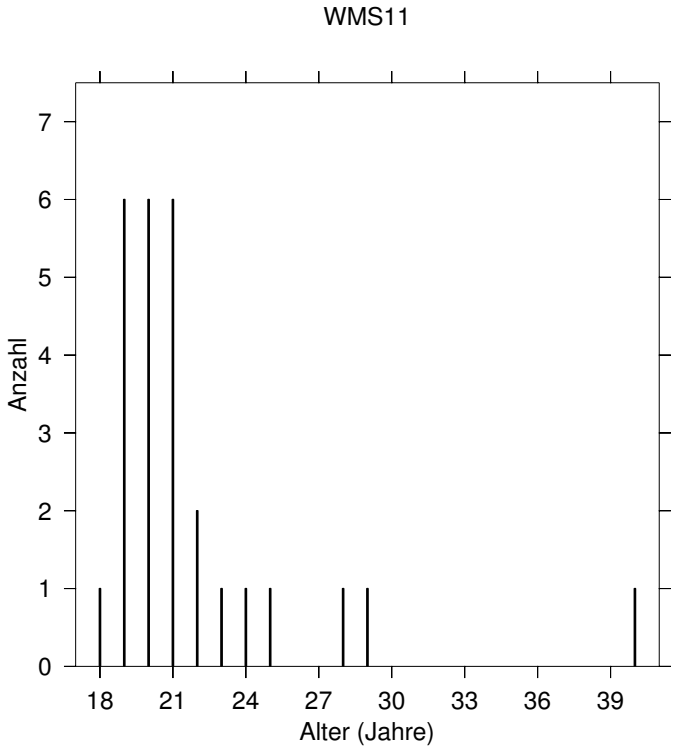

Abbildung 1: Altersverteilung im Stabdiagramm: absolute Häufigkeiten

Als Balkendiagramm ergibt sich folgendes Bild:

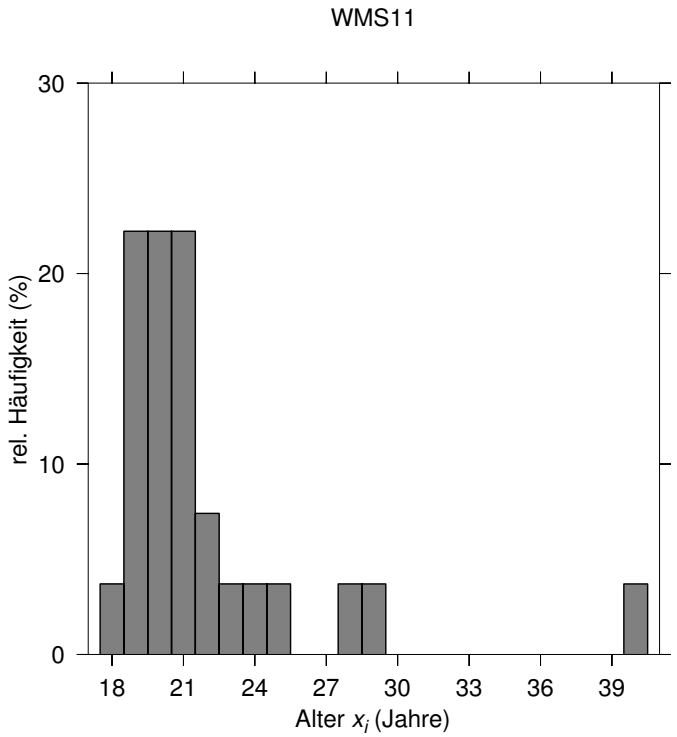

Abbildung 2: Altersverteilung im Balkendiagramm: relative Häufigkeiten, angegeben in %.

3.3.3 Klassierte Daten: Verteilung der Körpergröße

Für diskrete Merkmale sind erhobene Daten nur an den Stellen $x = x_i$ empirisch gehaltvoll. Liegt jedoch eine stetiges Merkmal vor, so ist für x jeder einzelne Wert möglich. In diesem Fall ist es sinnvoll, schon bei der Erhebung der Daten benachbarte Beobachtungswerte vordefinierten Intervallen zuzuordnen, den sog. Klassen. Die Zahl und die Größe dieser Klassen wird vom Untersuchungsziel und den Möglichkeiten der Datenerhebung bestimmt.

Die folgende Verteilung gibt Auskunft über die Verteilung der Körpergröße der $n = 27$ erfassten Teilnehmer des Kurses. Die Individuellen Werte des Merkmals *Größe* werden in $k = 6$ Größenklassen[2] eingeteilt.

[2] $]a; b]$: die Klasse erstreckt sich von a bis b, wobei a nicht enthalten ist.

Größenklasse	absolute Häufigkeit h_k	relative Häufigkeit f_k	Breite Δx_k	Dichte f_k^*
[1,50 ; 1,65]	4	0.148	0.15	0.988
]1,65 ; 1,70]	7	0.259	0.05	5.185
]1,70 ; 1,75]	3	0.111	0.05	2.222
]1,75 ; 1,80]	6	0.222	0.05	4.444
]1,80 ; 1,85]	6	0.222	0.05	4.444
]1,85 ; 2,05]	1	0.037	0.2	0.185

Tabelle 3: die Größenverteilung im Kurs in klassierter Form. Die Dichte f_k^* ist der Quotient f_k/Δ_k.

Im Gegensatz zur vorherigen Tabelle der Verteilung des Lebensalters ist hier bereits Information in Form einzelner Körpergrößen vernichtet worden. Während bei einem geringen Stichprobenumfang (hier $n = 27$ Werte) diese Reduktion der Daten nicht nötig gewesen wäre, ist sie bei bei größer angelegten Stichproben unumgänglich: man stelle sich alleine die Verteilung der Einkommen deutscher Haushalte ohne die Reduktion durch klassierte Angaben vor!

Zur grafischen Darstellung klassierter Daten werden sinnvollerweise Histogramme herangezogen. Histogramme sind eine Form der Auftragung, die an ein Balkendiagramm erinnert, sie zeichnen sich aber dadurch aus, dass die wiedergegebene Information (die relative Häufigkeit f_i eines Merkmalswerts) in der Fläche des Balkens enthalten ist.

Beispiel: Histogramm der Verteilung der Körpergröße im Kurs.

Liegen Stichprobeninformationen in Form klassierter Daten vor, so sollten zur Illustration Histogramme verwendet werden. Die *Flächen* der aufgetragenen Rechtecke oder Balken entsprechen dabei per Konstruktion den relativen Häufigkeiten der darzustellenden Klasse. Um dies zu erreichen, wird auf der Abszisse (x-Achse) die Klassenbreite Δx_i aufgetragen, auf der Ordinate (der y-Achse) die Dichte

$$f_i^* = \frac{f_i}{\Delta x_i}$$

Damit ergibt sich für die Fläche des i-ten Rechtecks (das die relative Häufigkeit der Klasse i symbolisiert) das Produkt der Breite und der Höhe

$$\text{Fläche} \;=\; \text{Breite} \;\times\; \text{Höhe} \;= \Delta x_i \cdot f_i^* = \Delta x_i \frac{f_i}{\Delta x_i} = f_i \qquad (9)$$

Falls die Klassenbreiten eines klassierten Datensatzes alle gleich breit sind, kann auf die Angabe von Dichten verzichtet werden - dies ist allerdings in den allermeisten Fällen nicht so. Insbesondere sind die Breiten der untersten und der obersten Klasse oft verschieden von den übrigen Breiten. Die Verteilung der Körpergröße im Kurs zusammen mit den nach Formel (9) berechneten Dichten ist in Tabelle 3 aufgeführt und hier im Histogramm dargestellt.

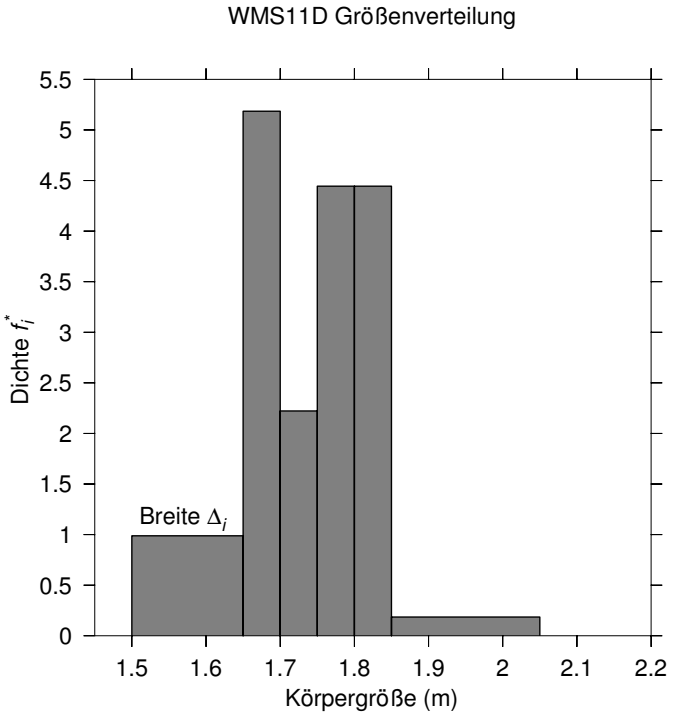

Abbildung 3: Größenverteilung im Histogramm: die relativen Häufigkeiten ergeben sich durch Multiplikation der aufgetragenen Dichte mit der Klassenbreite Δ_i

3.4 Kumulierte Häufigkeitsverteilungen

Wir gehen von einer Stichprobe des Umfangs n aus, die $m \leq n$ unterschiedliche, geordnete Ausprägungen eines ordinalen Merkmals X enthält. Die relativen Häufigkeiten, mit denen die einzelnen Merkmalsausprägungen auftreten, sind durch $(f_1, f_2, ...F_m)$ gegeben. Unter der **kumulierten absoluten bzw.**

relativen Häufigkeit H_i bzw. F_i versteht man die Summe der absoluten oder relativen Häufigkeiten für alle Merkmalsausprägungen bis zum Niveau i.

In einer Stichprobe des Umfangs n eines kardinalen Merkmals X mit $m \leq n$ verschiedenen, geordneten Merkmalswerten $(x_1, x_2, ...x_m)$ treten diese mit den relativen Häufigkeiten $(f_1, f_2, ...f_m)$ auf. Die kumulierte absolute (relative) Häufigkeit H_i (F_i) ist die Summe dieser Häufigkeiten für all diejenigen Merkmalswerte, die kleiner oder gleich dem jeweiligen Wert x_i sind.

$$H_i = \sum_{x_j \leq x_i} h_j \quad \text{bzw.} \quad F_i = \sum_{x_j \leq x_i} f_j \tag{10}$$

Die dadurch gebildeten Vektoren $(H_1, H_2, ...H_n)$ bzw $(F_1, F_2, ...F_n)$ geben die kumulierte absolute bzw. relative Häufigkeitsverteilung für den Vektor $(x_1, x_2, ...x_n)$ der einzelnen Merkmalswerte an.

Beispiel: kumulierte Häufigkeiten für die Altersverteilung.
Die kumulierten absoluten und relativen Häufigkeiten der Lebensalter der Kursteilnehmer lauten

Alter	h_i	H_i	f_i	F_i
18	1	1	1/27	1/27
19	6	7	6/27	7/27
20	6	13	6/27	13/27
21	6	19	6/27	19/27
22	2	21	2/27	21/27
23	1	22	1/27	22/27
24	1	23	1/27	23/27
25	1	24	1/27	24/27
28	1	25	1/27	25/27
29	1	26	1/27	26/27
40	1	27	1/27	27/27
Summe	27		27/27	

Tabelle 4: kumulierte Häufigkeiten H_i und F_i für die Altersverteilung.

Liegt die Stichprobeninformation für ein kardinales Merkmal X in Form von klassierten Daten vor mit l Klassen vor, werden die kumulierten relativen Häufigkeiten F_k gebildet aus der Summe der relativen Häufigkeiten für die Klassen

1 bis l. Die kumulierte relative Häufigkeit F_k wird der oberen Grenze x_k^o der k-ten Klasse zugeordnet. Bei l Klassen muss die Summe der relativen Häufigkeiten für die Klassen 1 bis l ergo Eins ergeben: $F_l = 1$.

Die Punkte $(x_1^o, F_1); (x_2^o, F_2); ...; (x_l^o, F_l)$ stellen die Eckpunkte des sogenannten Verteilungspolygons dar. Zur Skizzierung des Verteilungspolygons werden diese Eckpunkte jeweils durch eine Gerade verbunden, wobei der zusätzliche Punkt $(x_1^u, 0)$ mit der unteren Grenze x_1^u der 1. Klasse den Startpunkt bildet. Das Verteilungspolygon durch eine Parallele zur Abszisse (x-Achse) in Höhe von 1, die beim letzten Eckpunkt $(x_l^o, F_l = 1)$ beginnt, vervollständigt werden.

Beispiel: Verteilungspolygon bei klassierten kardinalen Daten.
Die folgende Tabelle stellt die klassierten Daten der Verteilung der Körpergröße der Kursteilnehmer dar:

Größenklasse	h_k	f_k	F_k (kumuliert)	Eckpunkte (x_k^o, F_k)
bis 1.50	0	0	0	$(1.50; 0)$
$]1,50; 1,65]$	4	$0,148$	0.148	$(1.65; 0.148)$
$]1,65; 1,70]$	7	$0,259$	0.407	$(1.70; 0.407)$
$]1,70; 1,75]$	3	0.111	0.519	$(1.75; 0.519)$
$]1,75; 1,80]$	6	0.222	0.741	$(1.80; 0.741)$
$]1,80; 1,85]$	6	0.222	0.963	$(1.85; 0.963)$
$]1,85; 2,05]$	1	0.037	1.0	$(2.05; 1.0)$

Tabelle 5: Klassengrenzen und kumulierte Häufigkeiten F_k zur Konstruktion des Verteilungspolygons.

Das zugehörige Verteilungspolygon ist in Abb. 4 dargestellt. Die durch die Punkte gekennzeichneten Eckpunkte des Polygons können der letzten Spalte der Tabelle entnommen werden. Sie werden aus der jeweiligen oberen Klassengrenze x_k^o und der zugehörigen kumulierten relativen Häufigkeit F_k gebildet. Das Verteilungspolygon (die Funktion $F(x)$) ergibt sich durch die Verbindung der Eckpunkte mit Geraden.

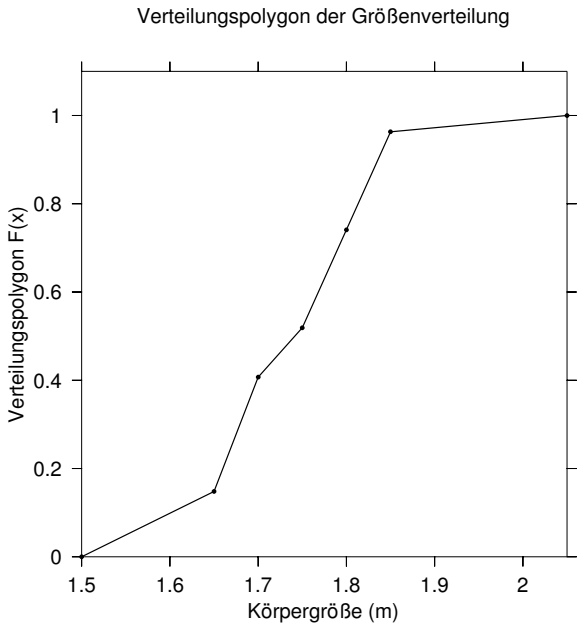

Abbildung 4: Größenverteilung im Verteilungspolygon

4 Statistische Analyse

Statistische Parameter (auch als statistische Maßzahlen bezeichnet) sind charakteristische (Zahlen-)Werte, die eine Menge von Beobachtungen *einfach* beschreiben. Der Zweck ist die Verdichtung von Daten einer Stichprobe in einzelne, möglichst einfache Parameter. Dabei wird stets Information vernichtet, dieser Informationsverlust muss jedoch für eine bessere Übersicht in Kauf genommen werden. Für eine Menge von Beobachtungen lassen sich viele solcher Maßzahlen angeben, wir werden davon einige der am häufigsten benutzten kennenlernen.

Lagemaße: geben für eine Stichprobe repräsentative, typische Werte an (beispielsweise einen Durchschnittswert)

Streuungsmaße: geben an wie dicht (oder wie weit entfernt) einzelne Merkmalswerte bei einem Mittelwert liegen

Schiefemaße: liefern Information über die Symmetrie oder Asymmetrie einer Verteilung von Daten

4.1 Lagemaße

Lagemaße sind Werte, die für eine gegebene Stichprobe einen einzelnen, für die vorliegenden Daten repräsentativen Wert angeben, beispielsweise einen Mittelwert. Sie müssen dabei nicht selbst Werte aus dem Bestand des vorliegenden Datenmaterials sein. So spricht beispielsweise bei einer Erhebung von Lebensaltern in ganzen Jahren nichts gegen einen Mittelwert, der als Wert zwischen zwei vollen Jahren angegeben wird.

4.1.1 das arithmetische Mittel

Das arithmetische Mittel ist der am weitesten verbreitete Mittelwert, es wird häufig in der Werbung oder in politischen Umfragen verwendet. Strenggenommen kann ein arithmetisches Mittel nur für kardinale Merkmale berechnet werden, oft wird es aber auch für ordinale Merkmale verwendet (teilweise unsinnig: eine Hotelbewertung von $3,4$ Sternen hat keine Bedeutung!). Das arithmetische Mittel einer Datenmenge von n kardinalen Merkmalen kann über

die folgende Überlegung einfach hergeleitet werden: wir gehen von einer einfachen Stichprobe von Merkmalswerten x_i eines kardinalen Merkmals X aus. Die Summe der Merkmalswerte ist also

$$S = x_1 + x_2 + x_3 + \cdots + x_n = \sum_{i=1}^{n} x_i$$

Als typischen Wert x für das vorliegende Datenmaterial wählen wir denjenigen Wert, der n-mal summiert denselben Wert S ergibt:

$$\underbrace{x + x + x + \cdots + x}_{n \cdot x} = \sum_{i=1}^{n} x = S = \sum_{i=1}^{n} x_i$$
$$\Leftrightarrow n \cdot x = \sum_{i=1}^{n} x_i \Leftrightarrow x = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Dieser Wert ist das *arithmetische Mittel*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ mit } x_i = (x_1, x_2, x_3, \ldots x_n) \tag{11}$$

Beispiel: Altersverteilung im Kurs.

die Einzelwerte des Alters der einzelnen Teilnehmer in Jahren sind in der folgenden Tabelle aufgeführt:

i	Alter (x_i)
1	19
2	21
3	19
4	18
5	20
6	19
7	20
8	21
9	21
10	25
11	20
12	20
13	20
14	22
15	21
16	21
17	20
18	19
19	20
20	20
21	21
22	20
23	26
24	21
25	20
26	23
27	20
28	40
29	20
30	20
31	19
32	29

Tabelle 6: Altersverteilung im Kurs.

Zur Berechnung des arithmetischen Mittels $\bar{x}$ wird die Summe der einzelnen Lebensalter durch den Umfang n der Stichprobe - also die Zahl der erfassten Personen - geteilt (Angabe in Jahren):

$$\bar{x} = \frac{685}{32} = 21,42$$

Beim Vergleich mit der grafischen Darstellung der erhobenen Daten in Abb. 1 wird deutlich, dass der Mittelwert hier nur eingeschränkt sinnvoll eingesetzt werden kann. Der Mittelwert erscheint im Vergleich mit der Grafik als zu groß - er wird durch die Ausreißer auf der rechten Seite zu höherem Alter hin verschoben.

Eigenschaften des arithmetischen Mittels

1. Schwerpunkteigenschaft:

$$n\bar{x} = x_1 + x_2 + x_3 + ... + x_n \Longleftrightarrow x_1 + x_2 + x_3 + ... + x_n - n\bar{x} = 0 \tag{12}$$

umsortieren der Summanden liefert

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + ... + (x_n - \bar{x}) + = 0$$

Abweichungen der Einzelwerte vom arithmetischen Mittel heben sich in der Summe auf.

2. Für eine Stichprobe $(x_1, x_2, x_3, ...x_n)$ ist das atithmetische Mittel die Lösung des Minimalisierungsproblems

$$\min_y \sum_{i=1}^{n} (x_i - y)^2$$

Bei gegebenen (x_i) ist die Summe $\sum_i (...)$ eine Funktion $f(y)$. Das arithmetische Mittel ist der Wert y, der die Summe der quadrierten Abweichungen minimiert (Erinnerung: eine Funktion $f(y)$ besitzt ein Minimum an der Stelle, an der die erste Ableitung verschwindet und die zweite Ableitung positiv ist).

$$\frac{df(y)}{dy} = 0$$

$$0 = \frac{df(y)}{dy} \sum_i (x_i - y)^2 = (-2) \sum_i (x_i - y)$$

$$\sum_i (x_i - y) \Longleftrightarrow ny = \sum_i x_i$$

$$y = \frac{1}{n} \sum_i x_i$$

Für ein Minimum muss ferner die zweite Ableitung positiv sein:

$$\frac{d^2}{(dy)^2}\sum_i (x_i - y)^2 = \frac{d}{dy}(-2)\sum_i (x_i - y) = \sum_i (-2)(-1) = 2n > 0$$

3. Lineare Transformation des arithmetischen Mittels:
Geht ein kardinales Merkmal Y durch eine allgemeine lineare Transformation

$$Y = a + bX$$

aus einem kardinalen Merkmal X hervor, so ergibt sich das arithmetische Mittel $\bar{y}$ des Merkmals Y aus derselben linearen Transformation aus dem arithmetischen Mittel $\bar{x}$ des Merkmals X:

$$\bar{y} = a + b\bar{x}$$

Jeder Wert x_i ergibt durch eine lineare Transformation einen Wert $y_i = a + bx_i$. Das arithmetische Mittel der Merkmalswerte y_i ergibt sich nach (11) durch

$$\begin{aligned}\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_i (a + bx_i) = \frac{1}{n}\sum_i a + \frac{1}{n}\sum_i bx_i \\ = \frac{n}{n}a + b\frac{1}{n}\sum_i x_i = a + b\bar{x}\end{aligned} \tag{13}$$

Beispiel: Umrechnung zwischen Fahrenheit und Celsius

Die Temperatur T_F in Grad Fahrenheit ergibt sich aus der Temperatur T_C in Grad Celsius nach der Vorschrift

$$T_\mathsf{F} = \frac{9}{5}T_\mathsf{C} + 32$$

Die Temperaturen $x_1 = 10, x_2 = 20, x_3 = 30$ Grad Celsius können damit in die Werte $y_1 = 50, y_2 = 68, y_3 = 86$ Grad Fahrenheit umgerechnet werden. Für die arithmetischen Mittel ergeben sich die Werte

$$\bar{x} = 20°\text{Celsius und } \bar{y} = \frac{50 + 68 + 86}{3} = 68°\text{Fahrenheit}$$

Genausogut kann der Mittelwert $\bar{y}$ aber über die lineare Transformation bestimmt werden:

$$\bar{y} = \frac{9}{5}\bar{x} + 32 = \frac{9}{5}\cdot 20 + 32 = 68$$

4.1.2 Alternative Berechnung des arithmetischen Mittels

Kommen einzelne Merkmalswerte mehrfach vor und gibt es in Wirklichkeit nur $m < n$ verschiedene Merkmalswerte $(x_1, x_2, ...x_m)$, die mit den absoluten Häufigkeiten $(h_1, h_2, ...h_m)$ auftreten, so lässt sich das arithmetische Mittel folgendermaßen berechnen:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} h_i x_i = \frac{h_1 x_1 + h_2 x_2 + h_3 x_3 + ...h_m x_m}{n} \tag{14}$$

Manchmal ist aber auch die folgende Form, die die relativen Häufigkeiten f_i benutzt, praktischer:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} h_i x_i = \frac{h_1}{n}x_1 + \frac{h_2}{n}x_2 + \frac{h_3}{n}x_3 + ...\frac{h_m}{n}x_m = \sum_{i=1}^{m}\frac{h_i}{n}x_i = \sum_{i=1}^{m} f_i x_i \tag{15}$$

Die einzelnen Faktoren f_i, mit denen die Merkmalswerte x_i multipliziert werden, könne als Faktoren aufgefaßt werden, mit denen die jeweiligen Merkmalswerte gewichtet werden. Im allgemeinen sind diese Gewichtungsfaktoren natürlich nicht identisch. Die letzte Formel (15) macht plausibel, weshalb hier vom *gewogenen* arithmetischen Mittel gesprochen wird. Das arithmetische Mittel, das für n verschiedene Einzelwerte $(x_1, x_2, ...x_n)$ gebildet wird, kann ebenfalls als ein gewogenes Mittel betrachtet werden, allerdings mit n völlig identischen Gewichten $1/n$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \frac{1}{n}x_3 + ...\frac{1}{n}x_n$$

4.1.3 arithmetisches Mittel bei klassierten Daten

Während die allgemeine Formel (11) zur Berechnung des arithmetischen Mittels bei klassiertem Datenmaterial wegen der Unkenntnis der einzelnen Werte nicht benutzt werden kann, findet die Berechnung des gewogenen Mittels nach Formel (15) auch bei klassierten Daten Anwendung. Anstelle der jeweiligen Merkmalswerte können dabei die Klassenmitten zur Berechnung herangezogen werden. Man geht gedanklich also davon aus, dass sich die Merkmalswerte gleichmäßig in jeder einzelnen Klasse verteilen (dies wird nur selten der Fall sein) und erhält so eine Näherung für das arithmetische Mittel des Datenmaterials.

Beispiel: Berechnung des arithmetischen Mittels bei klassierten Daten.

Für die klassierten Daten (Verteilung der Körpergröße) des Beispiels in Abschnitt **??** ergibt sich bei Verwendung der jeweiligen Klassenmitten eine mittlere Körpergröße der 20 Personen im Kurs von

$$\bar{x} = \frac{4 \cdot 1,425 + 8 \cdot 1,77 \cdot 1,8 + 1 \cdot 2,025}{20} = 1,696$$

(alle Angaben in m). Berechnet man das arithmetische Mittel direkt d.h. ohne Einteilung der Daten in Klassen, so ergibt sich ein Wert von $1,718$ m. Das arithmetische Mittel der klassierten Daten kann lediglich als Näherung interpretiert werden.

Beispiel: Gewogenes arithmetisches Mittel.

Der Primärenergieverbrauch (PEV) pro Kopf (in t Steinkohle-Einheiten (SKE)) im Jahr 2000 ist für verschiedene Kontinente und Regionen der Welt in der folgenden Tabelle aufgelistet, ebenso wie die Anteile an der gesamten Weltbevölkerung von 6.057 Millionen Menschen und der Zahl der Einwohner in diesen Regionen (Quellen: Weltbank, UN):

Region	PEV/Kopf	Anteil	Einwohner in Mio.
Europa	4,5	9,51 %	576
Ehemalige UdSSR	4,5	4,81 %	291
Nordamerika	11,4	5,18 %	314
Mittel- und Südamerika	1,4	8,57 %	519
Afrika	0,5	13,11 %	794
Asien, Ozeanien	1,1	8,84 %	3.564

Tabelle 7: Primärenergieverbrauch pro Kopf im Jahr 2000.

Der Pro-Kopf-Verbrauch ist in Nordamerika im Durchschnitt zweieinhalb mal so hoch wie in Europa und den Regionen der ehemaligen UdSSR. Diese Zahl ist ein simpler Indikator dafür, in welcher Region der Welt die Potenziale zur Energieeinsparung bzw. zur Verbesserung der Energieeffizienz am größten sind. Für den weltweiten durchschnittlichen Primärenergieverbrauch pro Kopf des Jahres 2000 ergibt sich nach der Formel

(15) für das gewogene arithmetische Mittel:

$$4,5 \cdot 0,0951 + 4,5 \cdot 0,0481 + 11,4 \cdot 0,0518 + 1,4 \cdot 0,0857$$

$$+0,5 \cdot 0,1311 + 1,1 \cdot 0,5884 = 2,15.$$

Ebensogut hätte der gesuchte Wert mit Hilfe der absoluten Einwohnerzahlen und den Durchschnittswerten des Pro-Kopf-Verbrauchs für die verschiedenen Regionen berechnet werden können:

$$\frac{4,5 \cdot 576 + 4,5 \cdot 291 + 11,4 \cdot 314 + 1,4 \cdot 519 + 0,5 \cdot 794 + 1,1 \cdot 3,564}{6,058}$$

Der folgende Abschnitt zeigt, dass es bei kardinalen Merkmalen keinesfalls immer sinnvoll ist - bei ordinalen und nominalen ist es ohnehin nicht zulässig - das arithmetische Mittel zur Berechnung eines mittleren Wertes anzuwenden. Als Faustregel gilt: Während das arithmetische Mittel bei additiven Zusammenhängen zur Durchschnittsbildung angewandt wird, findet das im folgenden Abschnitt diskutierte geometrische Mittel bei multiplikativen Zusammenhängen Anwendung. Das ebenfalls noch zu besprechende harmonische Mittel wird bei der Mittelwertbildung von Quotienten angewandt - nicht immer allerdings, wie eines der folgende Beispiele zeigen wird.

4.1.4 das geometrische Mittel

Beispiel:

Ein Sparbrief der Spaßkasse Nirgendwo verspricht bei Anlage einer Summe von $K_0 = 10.000$ im 1. Jahr einen Zins von $q_1 = 6\%$, im 2. Jahr von $q_2 = 7\%$ und im 3. Jahr von $q_3 = 8\%$. Nach drei Jahren erfolgt die Rückzahlung. Der hypothetische Kapitalbetrag nach Ende des ersten Jahres lautet

$$K_1 = K_0 + q_1 \cdot K_0 = (1 + q_1)K_0 = 10.600$$

Der durch das Ausklammern von K_0 entstehende Ausdruck $(1+q_1)$ wird auch als *Kapitalwachstumsfaktor* bezeichnet.

Die geometrische Folge der Kapitalbeträge K_1, K_2 und K_3 errechnet sich wie folgt:

$$\begin{aligned} K_1 &= (1+q_1)K_0 = 10.600 \\ K_2 &= (1+q_2)K_1 = 1,07 \cdot 10.600 = 11.342 \\ K_3 &= (1+q_3)K_2 = (1+q3)K_2 = 1,08 \cdot K_2 \\ &= (1+q_3)(1+q_2)K_1 = 1,08 \cdot 1,07 \cdot K_1 \\ &= (1+q_3)(1+q_2)(1+q_1)K_0 = 1,08 \cdot 1,07 \cdot 1,06 \cdot K_0 \end{aligned}$$

Wie lässt sich hier ein mittlerer Zinssatz ermitteln? Wir fragen nach dem Zinssatz, der beim selben Kapitaleinsatz K_0 nach drei Jahren denselben Endbetrag K_3 ergibt:

$$K_0(1+q)(1+q)(1+q) = (1+q)^3 K_0 = K_3$$

Nach Einsetzen von $K_3 = (1+q_3)(1+q_2)(1+q_1)K_0$ und Kürzen von K_0 erhält man

$$\begin{aligned} (1+q)^3 &= (1+q_3)(1+q_2)(1+q_1) \\ \Longleftrightarrow q &= \sqrt[3]{(1+q_3)(1+q_2)(1+q_1)} - 1 \end{aligned} \tag{16}$$

Mit Zahlenwerten ergibt sich für unser Beispiel das arithmetische Mittel

$$\bar{q} = \frac{8\% + 7\% + 6\%}{3} = 7\%.$$

Berechnet man dagegen das geometrische Mittel, so folgt

$$q = \sqrt[3]{1,08 \cdot 1,07 \cdot 1,06} - 1 = 0,06997 = 6,997\%.$$

Der Unterschied von $0,003\%$ ist in diesem Fall zwar sehr gering, bei höheren Zinssätzen oder starken Wertveränderungen wie bei einer Aktie kann es aber deutliche Konsequenzen haben, wenn fälschlicherweise das arithmetische Mittel zur Berechnung durchschnittlicher Renditen benutzt wird.

Der durchschnittliche Kapitalwachstumsfaktor $\bar{x}_G := 1+q$ resultiert nach Gleichung (16) aus einer speziellen Mittelung der einzelnen Kapitalwachstumsfaktoren $x_1 := 1+q_1$, $x_2 := 1+q_2$, $x_3 := 1+q_3$ für die einzelnen Jahre:

$$\bar{x}_G := 1+q = \sqrt[3]{(1+q_3)\cdot(1+q_2)\cdot(1+q_1)} = \sqrt[3]{x_3 \cdot x_2 \cdot x_1}. \tag{17}$$

Der aus einer solchen Mittelwertbildung resultierende Wert wird als *geometrisches* Mittel bezeichnet.

Defnition des geometrischen Mittels

Für n einzelne Werte $(x_1, x_2, ...x_n)$ erhält man durch Verallgemeinerung der Formel (17) eine allgemeingültige Formel für das geometrische Mittel x_G:

$$\bar{x}_G := \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n}. \tag{18}$$

Treten dabei einzelne Merkmalswerte mehrfach auf und sind in Wirklichkeit nur $m < n$ Merkmalswerte $(x_1, x_2, ...x_m)$ voneinander verschieden, die mit den absoluten Häufigkeiten $(h_1, h_2, ..., h_m)$ bzw. den relativen Häufigkeiten $(f_1, f_2, ..., f_m)$ auftreten, so berechnet sich das geometrische Mittel durch

$$\bar{x}_G = \sqrt[n]{(x_1)^{h_1} \cdot (x_2)^{h_2} \cdot ...(x_m)^{h_m}} = (x_1)^{h_1/n} \cdot (x_2)^{h_2/n} \cdot ...(x_m)^{h_m/n} \tag{19}$$

bzw.

$$\bar{x}_G = (x_1)^{f_1} \cdot (x_2)^{f_2} \cdot ...(x_m)^{f_m}. \tag{20}$$

Beispiel: mittlere Rendite einer Aktieninvestition.

Herr Andy Theke kauft eine Aktie zum Kurs von $K_0 = 100$. Genau ein Jahr später ist dieselbe Aktie nur noch die Hälfte wert, exakt $K_1 = 50$. Die - nicht besonders gute - Rendite der Aktie betrug also im ersten Jahr genau $q_1 = -50\%$. Im zweiten Jahr allerdings steigt der Kurs der Aktie wieder um 100%, so dass er am Ende des zweiten Jahres wieder auf dem ursprünglichen Niveau von $K_0 = K_2 = 100$ angekommen ist. Andys Finanzberater B. Trüger berechnet die durchschnittliche Rendite mit dem arithmetischen Mittel - das ergibt hier einen Wert von

$$\bar{q} = \frac{q_1 + q_2}{2} = \frac{-50\% + 100\%}{2} = +25\%$$

Es ist leicht einzusehen, dass dies nicht das richtige Mittel für die Rendite sein kann, da die Aktie nach den zwei Jahren ihren Wert überhaupt nicht gesteigert hat. Die Gesamtrendite beträgt 0%, das ist in diesem Fall auch die tatsächliche durchschnittliche Rendite. Die Ursache für

den Trugschluss auf Basis des arithmetischen Mittels ist hier der unterschiedliche Ausgangskurs, auf den sich die prozentualen (also relativen) Werte q_1 und q_2 beziehen.

Mit Hilfe des geometrischen Mittels der Kapitalwachstumsfaktoren ergäbe sich dagegen der korrekte Wert:

$$\bar{q}_G = \sqrt[2]{(1+q_1)\cdot(1+q_2)} - 1 = \sqrt{0,5\cdot 2} - 1 = 0$$

Dieses Beispiel zeigt deutlich, weshalb durchschnittliche Renditen korrekterweise nicht mittels des arithmetischen Mittels berechnet werden sollten.

4.1.5 Harmonisches Mittel

Bei kardinalen Merkmalen besteht zur Mittelwertbildung nicht nur die Auswahl zwischen arithmetischem oder geometrischem Mittel. In manchen Fällen ist es sogar keinesfalls sinnvoll, einen dieser beiden Mittelwerte anzuwenden. Das folgende Beispiel soll diesen Sachverhalt verdeutlichen:

Beispiel: Berechnung der mittleren Geschwindigkeit.

Herr Arno Nym legt mit seinem PKW eine Strecke s von 300km zurück - die ersten 100km erreicht er auf der Autobahn eine Durchschnittsgeschwindigkeit von $v_1 = 120$km/h, auf dem zweiten Teilstück von ebenfall 100km Länge erreicht er wegen zunehmendem Verkehr noch $v_2 = 100$km/h Durchschnittsgeschwindigkeit, die letzten 100km auf einer Bundesstraße legt er mit einem Schnitt von lediglich $v_3 = 80$km/h zurück.
Wie hoch war die durchschnittliche Geschwindigkeit auf der gesamten Strecke? Man könnte auch hier auf die Idee kommen, den gesuchten Schnitt mit Hilfe des arithmetischen Mittels $\bar{v}$ zu berechnen:

$$\bar{v} = \frac{v_1 + v_2 + v_3}{3} = \frac{120 + 100 + 80}{3} \text{ km/h } = 100 \text{ km/h}$$

Rechnet man jedoch nach der intuitiven Formel

Durchschnittliche Geschwindigkeit = Gesamtweg geteilt durch Gesamtzeit

die gesamte Wegstrecke $s = 300$km durch die gesamte Fahrzeit

$$T = \frac{100\text{km}}{120\text{km/h}} + \frac{100\text{km}}{100\text{km/h}} + \frac{100\text{km}}{80\text{km/h}} = 5/6\text{h} + 1\text{h} + 5/4\text{h} = 37/12\text{h}$$

so ergibt sich die korrekte Durchschnittsgeschwindigkeit von

$$\bar{v}_H = \frac{s}{T} = \frac{300\text{km}}{37/12\text{h}} = 97,30\text{km/h}$$

Natürlich ist auch hier der Unterschied zwischen arithmetischem Mittel $\bar{v}$ und dem korrekt gerechneten Wert kaum der Rede wert, das Beispiel illustriert aber, welche Art von Mittelwert hier logisch richtig ist.

Mit Hilfe der allgemeinen Bezeichnungen v_1, v_2, v_3 für die Geschwindigkeiten lautet die analoge Formel zur Berechnung der Durchschnittsgeschwindigkeit

$$\bar{v}_H = \frac{1}{\frac{1}{3}\left(\frac{1}{v_1} + \frac{1}{v_2} + \frac{1}{v_3}\right)} \tag{21}$$

Durch Verallgemeinerung der Formel (21) erhält man die Definition des *harmonischen Mittels*.

Definition des harmonischen Mittels

Für n Einzelwerte $(x_1, x_2, x_3, ...x_n)$ ergibt sich das harmonische Mittel nach

$$\bar{x}_H = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}\right)} \tag{22}$$

Sind aus den n Einzelwerten lediglich $m < n$ Merkmalswerte $(x_1, x_2, ...x_n)$ verschieden, die mit den absoluten Häufigkeiten $(h_1, h_2, ...h_m)$ bzw. den relativen Häufigkeiten $(f_1, f_2, ...f_m)$ auftreten, so lässt sich das harmonische Mittel nach

$$\bar{x}_H = \frac{1}{\frac{1}{n}\left(\frac{h_1}{x_1} + \frac{h_2}{x_2} + ... + \frac{h_m}{x_m}\right)} \tag{23}$$

bzw.

$$\bar{x}_H = \frac{1}{\frac{1}{n}\left(\frac{f_1}{x_1} + \frac{f_2}{x_2} + ... + \frac{f_m}{x_m}\right)} \tag{24}$$

4.1.6 Median

Statt des arithmetischen, geometrischen oder harmonischen Mittels kann bei Merkmalswerten oder Merkmalsausprägungen, die eine Rangfolge besitzen, das Konzept des *Medians* (auch als Zentralwert bezeichnet) benutzt werden. Dies ist bei ordinalen oder kardinalen Merkmalen der Fall, bei nominalen Merkmalen kann der Median wegen der fehlenden Rangfolge nicht benutzt werden.

Der Median ermittelt für gegebenes Datenmaterial bei ordinalen Merkmalen diejenige Merkmalsausprägung $\bar{x}_Z$ (bzw. denjenigen Merkmalswert bei kardinalen Merkmalen), die es gestattet, das vorhandene Datenmaterial in zwei möglichst gleichgroße Hälften aufzuteilen. Es ist also auch beim Median die Ermittlung eines mittleren Wertes (eines typischen Wertes für das vorliegende Datenmaterial) beabsichtigt.

Beschreibende Definition des Medians bei kardinalen Merkmalen

Der Median $\bar{x}_Z$ ist derjenige Merkmalswert eines kardinalen Merkmals X, den mindestens 50% aller Merkmalswerte einer Stichprobe vom Umfang n unterschreiten oder höchstenfalls erreichen **und** den mindestens 50% aller Merkmalswerte überschreiten oder zumindest erreichen.

Beispiel: Ermittlung des Medians über Stamm-Blatt-Darstellung.

Im Beispiel der Altersverteilung im Kurs in 3.3.1 lässt sich der Median mithilfe einer sogenannten Stamm-Blatt-Darstellung ermitteln, die die Merkmalswerte in eine geordnete Reihenfolge bringt.

1	8
1	9 9 9 9 9
2	0 0 0 0 0 0 0 0 0 0 0 0 0
2	1 1 1 1 1 1 1
2	2
2	3
2	5
2	6
2	9
4	0

Tabelle 8: Stamm-Blatt-Darstellung zur Ermittlung des Medians.

Auf der linken Seite des Trennstriches sind die Zehnerstellen des Lebensalters aufgetragen. Rechts des Striches repräsentiert jede einzelne Ziffer eine Einerstelle des jeweiligen Lebensalters. Man erkennt, dass das Alter von 20 Jahren die Stichprobe in zwei Teile trennt - leider nicht mit dem gleichen Umfang, so dass man nicht von Hälften sprechen sollte: In der einen Teilmenge sind 19 von 32 Personen enthalten (also 59% aller Personen), deren Lebensalter kleiner oder gleich dem Wert von 20 Jahren ist. In der zweiten Teilmenge sind alle diejenigen Personen zusammengefasst, deren Alter größer oder gleich dem Wert 20 ist, genau 26 von 32 Personen (also 82%). Der Wert von 20 stellt in diesem Fall den Median (oder das 50%-Quantil) dar.

Beispiel: Berechnung des Medians.

Bei Vorliegen eines Vektors $(x_1, x_2, ...x_11)$ von 11 *geordneten* Einzelwerten ist unabhängig vom tatsächlichen Aussehen der Einzelwerte der 6. Wert des Vektors der Median, denn beim 6. Wert einer geordneten Reihe von 11 Werten sind stets 6 von 11 Werten kleiner oder gleich dem Wert x_6 und ebenso 6 von 11 Werten größer oder gleich x_6:

$$\bar{x}_Z = x_6.$$

Diese beschreibende Definition des Medians legt diesen leider nicht immer eindeutig fest. Wären statt der 11 Werte im Vektori $(x_1, x_2, ...x_12)$ 12 Werte vorhanden, so kämen nach der obigen Definition zwei Werte in Frage - der 6. und der 7. Wert der geordneten Reihe erfüllen beide das genannte Kriterium. Um die Zweideutigkeit zu beseitigen, wird der Median in einem solchen Fall in eindeutiger Weise als das arithmetische Mittel der beiden in Frage kommenden Werte festgelegt:

$$\bar{x}_Z = \frac{x_6 + x_7}{2}$$

Als Beispiel könnte der folgende Vektor von 12 Zahlen vorgelegen haben:

$$(0, 1, 1, 2, 3, 3, 4, 4, 9, 9, 10, 32)$$

Der Median lautet in diesem Fall

$$\bar{x}_Z = \frac{x_6 + x_7}{2} = \frac{3+4}{2} = 3,5$$

Dieser Wert teilt die Menge von 12 Zahlen in zwei gleichgroße Hälften, Teilmengen gleichen Umfangs, auf:

$$(0, 1, 1, 2, 3, 3) \text{ und } (4, 4, 9, 9, 10, 32)$$

Es ist sofort Einsichtig, dass der Median, wie die Werte anderer Mittelwertkonzepte auch, nicht im vorliegenden Datenmaterial enthalten sein muss. Man erkennt ebenfalls, dass - im Gegensatz zum arithmetischen Mittel - beim Median nicht alle Werte einer vorliegenden Stichprobe in dessen Berechnung einfließen. Besonders der kleinste und der größte Datenwert spielen für die Berechnung des Medians i.A. keine Rolle. Diese Eigenschaft macht der Median robust gegenüber positiven und negativen Ausreißern bzw. dem Auftreten von extremen Werten. Aus diesem Grund wird zur Darstellung des durchschnittlichen Einkommens - bei dem praktisch immer Ausreißer bei Spitzengehältern vorliegen - inzwischen häufig der Median herangezogen.

Definition des Medians bei n Einzelwerten eines kardinalen Merkmals

Bezeichnet $(x_1, x_2, x_3, ...x_n)$ einen Vektor von geordneten Merkmalswerten eines kardinalen Merkmals, so ist der Median $\bar{x}_Z$ in eindeutiger Weise definiert durch

$$\bar{x}_Z = \begin{cases} x_i & \text{mit } i = (n+1)/2 \text{ für ungerade } n, \\ \dfrac{x_i + x_{i+1}}{2} & \text{wobei } i = n/2 \text{ für gerade } n. \end{cases} \tag{25}$$

Einige Eigenschaften des Medians sind:

Minimumseigenschaft des Medians: für eine gegebene Stichprobe von n Einzelwerten $(x_1, x_2, x_3, ...x_n)$ eines kardinalen Merkmals ist der Median $\bar{x}_Z$ die Lösung des Minimierungsproblems

$$\min_y \sum_{i=1}^{n} |x_i - y| \,,$$

wobei die Summe $\sum |x_i - y|$ bei gegebenen $(x_1, x_2, x_3, ...x_n)$ eine Funktion $f(y)$ allein der Variablen y ist. Der Median ist dre Wert, der die Summe der *absoluten* Abweichungen (die Beträge der Abweichungen) der Werte x_i von y minimiert. Diese Eigenschaft lässt sich nicht einfach mithilfe der Differentialrechnung beweisen, da die Betragsfunktion nicht differenzierbar ist!

Robustheit: Da nicht alle Werte einer Stichprobe in die Berechnung des Medians einfließen, ist er (im Gegensatz zu dem des arithmetischen Mittels) gegenüber Ausreißern robust: er wird durch das Auftreten einzelner, extremer Werte *nicht* beeinflusst.

Anwendbarkeit: das Konzept des Medians ist bei kardinalen und ordinalen Merkmalen anwendbar. Bei nominalen Merkmalen ist er aber nicht zu verwenden, da das Konzept eine Rangfolge oder Ordnung unter den Merkmalswerten oder Merkmalsausprägungen voraussetzt.

4.1.7 Ermittlung des Medians bei klassierten Daten

Für klassierte Daten kann der Median grafisch mithilfe des Verteilungspolygons (vgl. Abschnitt 3.4) $F(x)$ ermittelt werden. Der Median $\bar{x}_Z$ (das 50%-Quantil) ist der zum Funktionswert $F = 0,5 = 50\%$ gehörige x-Wert:

$$F(\bar{x}_Z) = 0,5.$$

Beispiel: Grafische Bestimmung des Medians bei klassierten Daten.

Mit Hilfe des bereits konstruierten Verteilungspolygons $F(x)$ der Körpergrößen im Kurs (vgl. Abb. 4) lässt sich der Median $\bar{x}_Z$ über die Bestimmung des zum Funktionswert $F(x) = 0,5$ gehörenden x-Werts ermitteln:

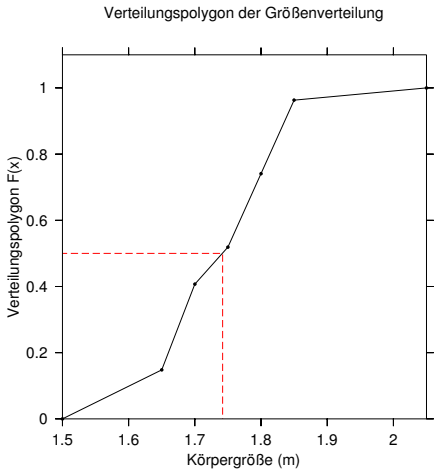

Abbildung 5: Bestimmung des Medians im Verteilungspolygon

In diesem Beispiel liegt der Median offensichtlich innerhalb der dritten Größenklasse $[x_3^u; x_3^o) = [1,70; 1,75)$m. Die Eckpunkte dieser Klasse $(1,70; 0,38)$ und $(1,75; 0,52)$ werden bei der Darstellung durch eine Gerade verbunden. Die Verhältnisse zwischen zwei beliebigen Funktionswerten $F(x)$ und den dazugehörigen x-Werten geben die Steigung der interpolierten Geraden innerhalb der betreffenden Klasse an. Sie sind also innerhalb einer Größenklasse alle gleich. So gilt beispielsweise:

$$\frac{F(x_3^o) - F(x_3^u)}{x_3^o - x_3^u} = \frac{F(\bar{x}_Z) - F(x_3^u)}{\bar{x}_Z - x_3^u},$$

woraus sich mit der Definition des Medians $F(\bar{x}_Z) = 0,5$ die Lage des Medians (in m) der Größenverteilung ergibt:

$$\begin{aligned} \bar{x}_Z &= x_3^u + (x_3^o - x_3^u) \cdot \frac{F(\bar{x}_Z) - F(x_3^u)}{F(x_3^o) - F(x_3^u)} \\ &= 1,70 + (1,70 - 1,65) \cdot \frac{0,5 - 0,38}{0,52 - 0,38} = 1,74 \end{aligned} \tag{26}$$

Der Wert von $\bar{x}_Z = 1,74$m für den Median weicht in diesem Beispiel kaum vom arithmetischen Mittel $\bar{x} = 1,73$m ab, da die Verteilung keine nennenswerten Ausreißer nach oben oder unten aufweist.

Durch Verallgemeinerung der im letzten Beispiel zur Berechnung des Medians benutzten Gleichung ergibt sich eine allgemeine Formel zur Berechnung des

Medians bei klassierten Daten:

$$F(\bar{x}_Z) = x_j^u + (x_j^o - x_j^u)\frac{F(\bar{x}_Z) - F(x_j^u)}{F(x_j^o) - F(x_j^u)} = x_j^u + (x_j^o - x_j^u)\frac{F(0,5) - F(x_j^u)}{F(x_j^o) - F(x_j^u)} \tag{27}$$

Diese Formel resultiert aus der Bedingung $F(\bar{x}_Z) = 0,5$ und der Konstruktion des Verteilungspolygons, bei der die Eckpunkte $(x_j^u; F(x_j^u))$ und $(x_j^o; F(x_j^o))$ durch eine Gerade verbunden werden. Um den Median mit dieser Formel berechnen zu können, muss allerdings zunächst festgestellt werden (beispielsweise grafisch oder mit Hilfe einer Tabelle), in welche Klasse j der Median fällt.

Definition des Medians bei ordinalen Merkmalen

Der Median $\bar{x}_Z$ ist diejenige Merkmalsausprägung eines ordinalen Merkmals, die mindestens 50% aller Merkmalsausprägungen einer Stichprobe vom Umfang n unterschreiten oder höchstenfalls erreichen **und** die mindestens 50% aller Merkmalsausprägungen überschreiten oder zumindest erreichen.

Beispiel: Berechnung des Medians bei ordinalen Daten.

Mr. Stu Dent, der an einer dualen Hochschule im tiefen Süden Deutschlands einen Statistik-Kurs besucht, soll angeben, wie regelmäßig er die 12 für ihn relevanten Veranstaltungen des Sommersemesters 2010 besucht hat. Seine Antworten bilden das Merkmal X, die Häufigkeit der Teilnahme an den Veranstaltungen, wobei die Ausprägungen folgendermaßen kodiert sind: 0 (nie), 1 (sehr selten), 2 (selten), 3 (oft), 4 (meistens), 5 (immer). Die absoluten, relativen und kumulierten relativen Häufigkeiten für seine Teilnahmehäufigkeit an den 12 Veranstaltungen sind in der folgenden Tabelle zusammengefasst:

Ausprägung	absolute Häufigkeit h_i	relative Häufigkeit f_i	kum. relative Häufigkeit F_i
5 (immer)	1	1/12	12/12
4 (meistens)	3	3/12	11/12
3 (oft)	2	2/12	8/12
2 (selten)	3	3/12	6/12
1 (sehr selten)	1	1/12	3/12
0 (nie)	2	2/12	2/12

Tabelle 9: Teilnahmehäufigkeiten des Herrn Dent.

Mr. Dent besucht also beispielsweise 3 von 12 Veranstaltungen nur selten. Die kumulierte relative Häufigkeit von 6/12 besagt: 6 von 12 Veranstaltungen hat er höchstenfalls selten, wenn nicht gar seltener (bei einer Veranstaltung) oder nie (zwei Veranstaltungen) besucht. Nach der Definition des Medians für ordinale Merkmale kommen zwei Merkmalsausprägungen für den Median in Frage: Die Merkmalsausprägung 'selten', aber auch die Merkmalsausprägung 'häufig'. Im Gegensatz zum oben besprochenen Fall kardinaler Merkmale kann diese Zweideutigkeit im vorliegenden Fall nicht beseitigt werden: Die Bildung des arithmetischen Mittels von 'selten' und 'oft' ergibt keinen Sinn! Beide Merkmalsausprägungen könnten als Median festgelegt werden.

4.1.8 der Modus

Bei Vorliegen eines nominalen Merkmals kann keines der bisher diskutierten Mittelwertskonzepte angewendet werden. Für diesen speziellen Fall existiert das Konzept des Modus oder der modalen Klasse.

Definition des Modus und der modalen Klasse

Bei *nominalen* bzw. *ordinalen* Merkmalen ist der Modus die am häufigsten auftretende Merkmalsausprägung. Bei Vorliegen von Einzelwerten eines *kardinalen* Merkmals ist der Modus oder Modalwert $\bar{x}_M$ der am häufigsten auftretende Merkmalswert. Liegen statt Einzelwerten klassierte Daten eines kardinalen Merkmals vor, wird diejenige Klasse, welche die größte absolute und damit natürlich auch die größte relative Häufigkeit aufweist, modale Klasse genannt.

Beispiel: Modus und modale Klasse.

Im Beispiel der Verteilung der Körpergröße (s. Abschnitt 3.3.3) ist die Größenklasse $[1,65; 1,75[$ die modale Klasse, denn 8 von 20 - und damit die meisten - Teilnehmer weisen eine Körpergröße auf, welche in diese Klasse fällt.

Hinsichtlich der Noten der Statistik-Klausur dieses Kurses stellt die Note 'sehr gut' den Modus dar, beim Geschlecht die Ausprägung 'weiblich'.

Es gibt durchaus Beispiele, bei denen es keinen eindeutigen Modus gibt - in diesem Fall treten gleiche Werte der Häufigkeiten für unterschiedliche Merkmalsausprägungen auf. Auch im obigen Beispiel des ordinalen Merkmals 'Häufigkeit der Teilnahme an den Veranstaltungen' gibt es keinen eindeutigen Modus: 'meistens' bzw. 'selten' sind die beiden gleichermaßen häufig auftretenden Merkmalsausprägungen, die für Mr. Dent und seine Besuchshäufigkeit der Veranstaltungen charakteristisch sind.

4.1.9 Quantile

Neben den Mittelwerten sind noch weitere Lageparameter in der Statistik von wesentlicher Bedeutung, die sogenannten Quantile. Das Konzept des Quantils ist eine Verallgemeinerung des Konzeptes des Medians: beim Median handelt es sich um nichts anderes als ein spezielles Quantil nämlich das 50%-Quantil. Von Interesse könnten daneben beispielsweise auch das 25%-Quantil oder das 75%-Quantil sein, oft auch unteres und oberes Quartil genannt. Beim Vergleich von Einkommensverteilungen verschiedener Länder besitzen zum Beispiel auch Dezentile - das sind die 10% -, 20% - usw. Quantile - eine gewisse Relevanz.

Beschreibende Definition des p-Quantils x_p bei kardinalen Merkmalen: das P%-Quantil bzw. p-Quantil x_p ist derjenige Merkmalswert eines kardinalen Merkmals X, den mindestens $P\%$ aller Merkmalswerte einer Stichprobe vom Umfang n unterschreiten oder höchstenfalls erreichen und den mindestens $(100-P)$% aller Merkmalswerte überschreiten oder zumindest erreichen. Dabei ist $0 < P < 100, p = P/100$ und $0 < p < 1$.

Man beachte: Ebenso wie die entsprechende Definition des Medians ist diese Definition nicht eindeutig.

Wie beim Median können natürlich auch Quantile aus einer geordneten Reihe von Einzelwerten ermittelt werden. Die konkreten Werte müssen dazu prinzipiell nicht bekannt sein.

Beispiel: Intuitive Berechnung des unteren Quartils.

Hätte man beispielsweise $n = 11$ geordnete Einzelwerte im Vektor $(x_1, x_2, ..., x_1 1)$ vorliegen, so ist - ungeachtet des tatsächlichen Aussehens der Merkmalswerte - der 3. Wert das 25% -Quantil (auch als Quartil bezeichnet). Der Wert x_3 erfüllt die beiden Bedingungen, die an ein $0,25$-Quantil gestellt werden: es ist der erste Wert, der die Stichprobe in zwei Teilmengen aufteilt, wobei eine Teilmenge einen Umfang von mindestens einem Viertel aller Werte besitzt, was bei x_1, x_2, x_3 der Fall wäre, während die andere - hier die Menge $x_3, x_4, ..., x_1 1$ mindestens einen Umfang von $3/4$ aller Werte haben soll. Im konkreten Beispiel der bereits geordneten Stichprobe

$$(0, 1, 1, 2, 2, 4, 4, 6, 9, 13, 17)$$

lautet das untere Quartil $x_{0,25} = 1$. Der Wert 1 ergibt eine nach der Definition des 25%-Quantils geforderte Aufteilung der Stichprobe in $0, 1, 1$ und $1, 1, 2, 2, 4, 4, 6, 9, 13, 17$. Durch Division der Gesamtzahl von 11 Elementen durch 4 ergibt sich ein Hinweis auf die Aufteilung der zwei Teilmengen. Leider liefert die Division mit $2,75$ hier aber keine ganze Zahl. Dennoch ist damit klar, dass die kleinere Teilmenge mindestens 3 Elemente - die nächstgrößerere ganze Zahl nach $2,75$ - der Stichprobe enthalten muss, während die andere Teilmenge aus mindestens 9 Elemente bestehen muss - die nächstgrößere ganze Zahl bezogen auf $8,25 = 3/4 \cdot 11$.

12 (anstatt 11) geordnete Einzelwerte wie beispielsweise $(0, 1, 1, 2, 2, 3, 4, 4, 6, 9, 13, 17)$, lassen sich mathematisch gesehen eingängiger in zwei Teilmengen mit ungefähr $1/4$ bzw. $3/4$ aller Merkmalswerte aufteilen: $1/4 \cdot 12 = 3$ sollte die Zahl der Elemente der einen Teilmenge sein und $3/4 \cdot 12 = 9$ die Zahl der anderen. Allerdings kämen nach der beschreibenden Definition für das untere Quartil zwei Werte in Frage: Der 3. und der 4. Wert in der geordneten Reihenfolge - beide Werte, x_3 und x_4 , würden die Bedingungen für ein unteres Quartil erfüllen. Um diese Zweideutigkeit zu beseitigen, kann wie bei der Definition des Medians das arithmetische Mittel der beiden in Frage kommenden Werte als unteres Quartil festgelegt werden:

$$x_{0,25} = \frac{x_3 + x_4}{2} = \frac{1 + 2}{2} = 1,5$$

Der Wert $1,5$, nach Definition das untere Quartil, teilt die Menge von 12 Zahlen in zwei Teilmengen des Umfangs $1/4$ bzw. $3/4$ aller Merkmalswerte auf: $0, 1, 1$ und $2, 2, 3, 4, 4, 6, 9, 13, 17$.

Nach dieser ausführlichen Darstellung der intuitiven Ermittlung von Quantilen am Beispiel des unteren Quartils dient die folgende formale Definition mehr der Vollständigkeit und Vergleichbarkeit mit der statistischen Literatur denn als praktisch handhabbare Möglichkeit zur Ermittlung von Quantilen bei Individualdaten eines kardinalen Merkmals. Während die für die Definition notwendige Notation bereits eine gewisse Gedächtnisleistung erfordert, ist es unwahrscheinlich, dass jemand die Definition reproduzieren kann, ohne das dahinter liegende Prinzip verstanden zu haben. Ist das Prinzip aber erst verstanden, ist die Definition zur Bestimmung von Quantilen eigentlich überflüssig.

Definition des p-Quantils xp bei n Einzelwerten eines kardinalen Merkmals Bezeichnet $(x_1, x_2, ..., x_n)$ einen Vektor geordneter, individueller Merkmalswerte eines kardinalen Merkmals X, so wird das p-Quantil x_p in eindeutiger Weise definiert durch

$$x_p := \begin{cases} x_i \text{ ,wobei } i = [n \cdot p] + 1, \text{ falls } n \cdot p \text{ nicht ganzzahlig ist,} \\ \dfrac{x_i + x_{i-1}}{2} \text{ ,wobei } i = [n \cdot p], \text{ falls } n \cdot p \text{ ganzzahlig ist.} \end{cases} \tag{28}$$

Dabei stellen die eckigen Klammern die sogenannten GAUSS-Klammern dar. $[n \cdot p]$ bezeichnet die größte ganze Zahl, die kleiner oder gleich dem Ausdruck $n \cdot p$ innerhalb der Klammer ist. Für den Median x_Z, das 50%-Quantil $x_{0,5}$, ergibt sich aus der allgemeinen Definition des p-Quantils (28) mit $p = 0,5 = 1/2$ natürlich wieder die Definition (27) des Medians bei Individualdaten eines kardinalen Merkmals.

Beispiel: Berechnung des unteren Quartils nach Definition (28).

Zur Berechnung des unteren Quartils ergibt sich für den Vektor

$$(0, 1, 1, 2, 2, 4, 4, 6, 9, 13, 17)$$

von $n = 11$ geordneten Zahlen über die allgemeine Definition ein Wert von $x_{0,25} = x_3 = 1$, denn $n \cdot p = 11 \cdot 0,25 = 2,75$. Es ist der erste Teil der Definition anzuwenden, wobei $[n \cdot p] = [2,75] = 2$ und daher der Index des Kandidaten für das gesuchte untere Quartil $i = 2 + 1 = 3$ lautet.

Für den Vektor

$$(0, 1, 1, 3, 2, 2, 4, 4, 6, 9, 13, 17)$$

von $n = 12$ geordneten Zahlen ergäbe sich wie oben auch $x_{0,25} = (x3 + x4)/2 = (1 + 2)/2 = 1,5$. Wegen $n \cdot p = 12 \cdot 0,25 = 3$ wird in diesem Fall der zweite Teil der Definition benutzt, wobei $[n \cdot p] = [3] = 3$.

p-Quantil bei klassierten Daten

Liegen klassierte Daten vor, so kann das p-Quantil x_p grafisch mit Hilfe des Verteilungspolygons $F(x)$ bestimmt werden. Das p-Quantil ist in diesem Fall als der zum Funktionswert $F(x) = p$ gehörige Variablenwert definiert:

$$F(x_p) = p \tag{29}$$

Es kann aber ebenfalls, analog zum Fall des Medians, rechnerisch durch eine allgemeine Formel berechnet werden:

$$x_p = x_j^u + (x_j^o - x_j^u) \cdot \frac{F(x_p) - F(x_j^u)}{F(x_j^o) - F(x_j^u)} = x_j^u + (x_j^o - x_j^u) \cdot \frac{p - F(x_j^u)}{F(x_j^o) - F(x_j^u)} \tag{30}$$

Sie ergibt sich aus der obigen Bedingung (29) und der Konstruktion des Verteilungspolygons, bei der die Eckpunkte $(x_j^u; F(x_j^u))$ und $(x_j^o; F(x_j^o))$ durch eine Gerade verbunden werden. Vor der Bestimmung eines Quantils über die Formel (30) muss wieder - grafisch oder mit Hilfe einer Tabelle - ermittelt werden, in welche Klasse j das p-Quantil entfällt. Setzt man für $p = 0,5$, so ergibt sich aus (30) natürlich wieder die Formel (27) zur Berechnung des Medians für klassierte Daten.

Das p-Quantil bei ordinalen Merkmalen

das P%-Quantil bzw. p-Quantil x_p ist diejenige Merkmalsausprägung eines ordinalen Merkmals X, den mindestens $P\%$ aller Merkmalsausprägungen einer Stichprobe vom Umfang n unterschreiten oder höchstenfalls erreichen und den mindestens $(100 - P)$% aller Merkmalswerte überschreiten oder zumindest erreichen. Dabei gilt wieder $0 < P < 100, p = P/100$ und $0 < p < 1$.

Beispiel: Berechnung von Quantilen bei ordinalen Daten.

Anhand der Tabelle der absoluten, kumulierten und relativen Häufigkeiten für die Teilnahme von Stu Dent (bekannt aus Abschnitt 4.1.7) an den relevanten Veranstaltungen seines Studienganges lassen sich das obere und das untere Quartil bestimmen.

Ausprägung	absolute Häufigkeit h_i	relative Häufigkeit f_i	kum. relative Häufigkeit F_i
5 (immer)	1	1/12	12/12
4 (meistens)	3	3/12	11/12
3 (oft)	2	2/12	8/12
2 (selten)	3	3/12	6/12
1 (sehr selten)	1	1/12	3/12
0 (nie)	2	2/12	2/12

Tabelle 10: kumulierte Teilnahmehäufigkeiten des Herrn Dent.

Bei einer Gesamtzahl von 12 Veranstaltungen ist das untere Quartil (das 25%-Quantil) die Ausprägung 'sehr selten': damit lassen sich die Veranstaltungen in zwei Teilmengen aufteilen, von denen die eine die Ausprägungen 'nie' und 'sehr selten' umfasst (3 von 12 Veranstaltungen oder $1/4$ aller Veranstaltungen). Die zweite Teilmenge umfasst alle Ausprägungen ab einer Teilnahmehäufigkeit von 'sehr selten' (10 von 12 Veranstaltungen oder $5/6 > 3/4$ der Veranstaltungen).

Das obere Quartil oder 75%-Quartil ist die Ausprägung 'meistens': 11 von 12 Veranstaltungen, also mehr als 75% weisen eine Teilnahmehäufigkeit von höchstens 'meistens' auf, während 4 von 12, also mehr als 25% mindestens 'meistens' besucht wurden.

4.1.10 Zusammenfassung: Lageparameter

Lageparameter geben für eine Stichprobe repräsentative, für die erhobene Verteilung typische Werte an. Die Lageparameter selbst müssen dabei nicht Bestand des erhobenen Datenmaterials sein.

Bei nominalen (qualitativen) Merkmalen ist als einziger der diskutierten Lageparameter der Modus bestimmbar. Er ist die am häufigsten auftretende Merkmalsausprägung und nicht notwendigerweise eindeutig bestimmt.

Bei ordinalen Merkmalen (mit Rangfolge) können neben dem Modus oft auch Quantile sowie als Sonderfall des 50%-Quantils der Median berechnet und angegeben werden.

Bei kardinalen Merkmalen sind darüber hinaus die Konzepte des arithmetischen, geometrischen und des harmonischen Mittels anwendbar. Für die Mittelwertbildung bei kardinalen Merkmalen gilt die folgende Faustregel: bei additiven Zusammenhängen findet das arithmetische Mittel $\bar{x}$ Anwendung, bei multiplikativen Zusammenhängen das geometrische Mittel $\bar{x}_G$. Das harmonische Mittel $\bar{x}_H$ kann bei Mittelwertbildung von Quotienten verwendet werden. Strenggenommen findet keines dieser drei Mittelwertkonzepte bei ordinalen oder gar bei nominalen Merkmalen Anwendung.

4.1.11 Übungsaufgaben zu den Lageparametern

- Herr B. ist selbstständiger Statistiker und erwirtschaftet im Jahr 2000 während der allgemeinen Börsen-Euphorie mit Aktienanalysen ein Ergebnisplus von 35% im Vergleich zum Vorjahr. Im Jahr 2000 ändern sich die Voraussetzungen, ab dem 2.Halbjahr kann sich kaum noch jemand für Aktien begeistern. Dadurch ergeben sich natürlich auch weniger Aufträge für Herrn B., er verzeichnet im Jahr 2001 ein Ergebnisminus von ebenfalls 35%. Ist das Ergebnis seiner Statistik-Tätigkeit gleich hoch wie während des Aktien-Booms? Wie hoch ist die durchschnittliche Ergebnisentwicklung innerhalb dieser zwei Jahre?

- Der Gummistiefel- und Teebeutelverkäufer Manfred M. betreibt einen mobilen Verkaufsstand in der Ostfriesenstraße, in der seine gesamte Stammkundschaft wohnt. Bei genauer Kenntnis seiner Kundschaft und deren Wohnsitze will er seinen Standort möglichst so wählen, dass er den Umsatz maximieren kann. Seine 15 Hauptkunden wohnen allesamt in der Ostfriesenstraße, im k-ten Haus, das x_k m vom Anfang der Straße entfernt ist, wohnen n_k seiner Stammkunden:

x_k	0	10	20	30	35	50
n_k	3	4	1	2	3	2

  In welcher Entfernung vom Anfang der Ostfriesenstraße wird Manfred M. seinen Standort wählen?

4.2 Streuung

Die beiden Verteilungen in Abb. 6 besitzen dasselbe arithmetische Mittel. Man erkennt sofort, dass der Lageparameter Mittelwert nicht ausreicht, um die Verteilungen sinnvoll zu beschreiben, bei der Bestimmung geht zuviel der relevanten Information verloren. Die beiden Verteilungen unterscheiden sich aber deutlich in der typischen Abweichung einzelner Werte vom gemeinsamen arithmetischen Mittel.

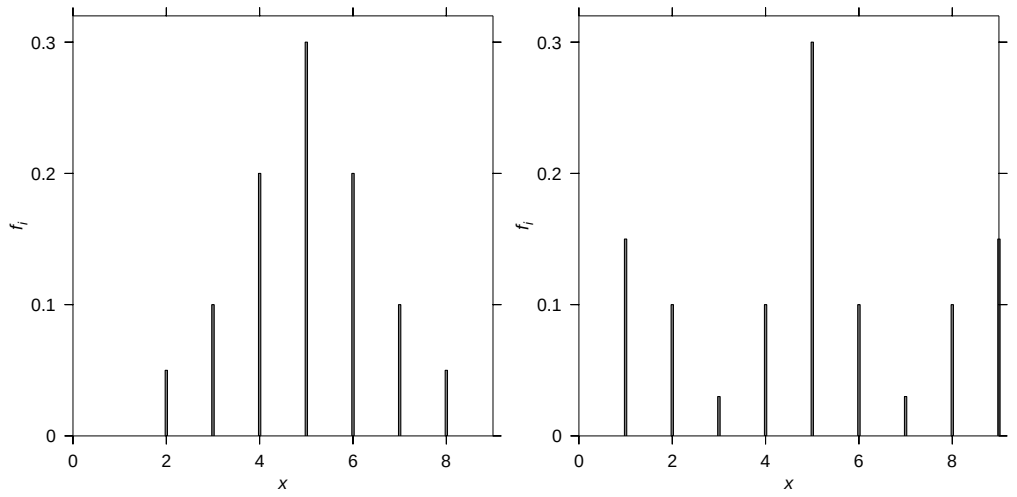

Abbildung 6: Vergleich zweier Verteilungen mit dem gleichen arithmetischen Mittel

Während Lagemaße wie Mittelwerte typische Werte einer Stichprobe repräsentieren, treffen die Streuungsmaße eine Aussage darüber, ob die verschiedenen Merkmalswerte dicht bei einem Mittelwert liegen oder sich in mehr oder weniger großen Abständen davon befinden. Da Abstände aber nur für kardinale Merkmale sinnvoll zu definieren sind, können auch die Streuungsmaße strenggenommen nur für kardinale Merkmale sinnvoll definiert werden.

4.2.1 Spannweite

Die Spannweite ist das primitivste Streuungsmaß, sie gibt die Differenz zwischen dem größten und dem kleinsten Merkmalswert einer Stichprobe $(x_1, x_2, ...x_n)$ eine kardinalen Merkmals X an:

$$s_W = \mathsf{max}\{x_1, x_2, ...x_n\} - \mathsf{min}\{x_1, x_2, ...x_n\}. \tag{31}$$

Bei klassierten Daten ist das Maximum bzw. Minimum jedoch nicht bekannt, hier wird die Spannweite als Differenz zwischen der oberen Klassengrenze der obersten von k Klassen x_k^o und der untersten Klassengrenze der ersten Klasse x_1^u definiert:

$$s_W := x_k^o - x_1^u \text{ für klassierte Daten in } k \text{ Klassen} \tag{32}$$

- Die Aussagekraft der Spannweite für eine Stichprobe $(x_1, x_2, ..., x_n)$ ist sehr eingeschränkt, da s_W nur aus zwei Werten dieser Stichprobe berechnet wird: die einfache Berechnung wird mit einem hohen Informationsverlust erkauft. Die Spannweite ist lediglich eine Angabe zur Größe des Bereichs, aus dem die Stichprobenwerte stammen. Handelt es sich bei einem der beiden Werte oder gar bei beiden um Ausreißer, ist s_W wenig aussagekräftig für das zu analysierende Datenmaterial.

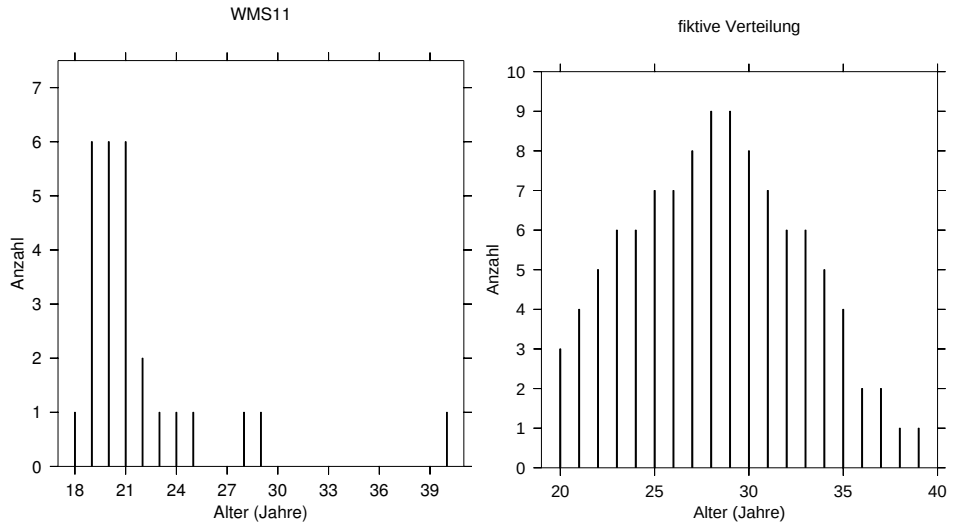

Abbildung 7: Stabdiagramme zweier Verteilungen ähnlicher Spannweite

Das linke Stabdiagramm zeigt wieder die relativen Häufigkeiten f_i der Lebensalter im Kurs, während das rechte Diagramm eine fiktive Altersverteilung mit der Spannweite von $s_W = 39 - 20 = 19$ Jahren zeigt. Die Spannweite der Lebensalter beträgt für die Altersverteilung in WMS11 $40 - 18 = 22$ Jahre. Ohne den Ausreißer von 40 Jahren bei der linken Verteilung betrüge die Spannweite nur 11 Jahre, was die geringere Streuung der Lebensalter deutlich besser wiedergeben würde.

- Spannweiten verschiedener Stichproben unterschiedlichen Umfangs lassen sich nicht miteinander vergleichen, da bei der Berechnung der Spannweite die Größe des Stichprobenumfangs nicht berücksichtigt wird.

4.2.2 mittlere absolute Abweichung

Es liegt nahe, die Streuung unter den Merkmalswerten $(x_1, x_2, ..., x_n)$ eines kardinalen Merkmals X mittels der Summe der Abweichungen der Einzelwerte x_i von einem Mittelwert, beispielsweise dem arithmetischen Mittel $\bar{x}$, messen zu wollen. Diese Summe der Einzelabweichungen

$$\sum_{i=1}^{n}(x_i - \bar{x})$$

ist jedoch nach Gleichung (12), begründet in der Schwerpunkteigenschaft des arithmetischen Mittels, stets Null - positive und negative Abweichungen heben sich in der Summe auf.

Um ein Maß für die Summe der Abweichungen vom Mittelwert zu erhalten, können die Absolutbeträge dieser Abweichungen benutzt werden. Dies geschieht bei der Berechnung der *mittleren absoluten Abweichung* der Merkmalswerte $(x_1, x_2, ..., x_n)$ eines kardinalen Merkmals X vom arithmetischen Mittel $\bar{x}$

$$d_{\bar{x}} := \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|. \tag{33}$$

bzw. bei der mittleren absoluten Abweichung vom Median $\bar{x}_Z$:

$$d_{\bar{x}_Z} := \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_Z|. \tag{34}$$

Oft wird das zweite Maß bevorzugt, weil der Median $\bar{x}_Z$ diese Summe der absoluten Abweichungen minimiert (siehe Eigenschaften des Medians).

Wegen der Beträge werden die Werte der beiden Maße nicht negativ. Sie können nur dann den Wert Null annehmen, wenn alle Einzelwerte identisch sind - damit ist automatisch auch der jeweilige Mittelwert festgelegt. In diesem Fall mit

$$x_1 = x_2 = ... = x_n = \bar{x} \text{ bzw. } \bar{x}_Z$$

verschwinden alle Differenzen in den Definitionen (33) und (34). Es gibt natürlich keine Streuung unter den Merkmalswerten, was sich folgerichtig im Wert Null der beiden Streuungsmaße niederschlägt.

Beispiel: Berechnung der mittleren absoluten Abweichungen vom Median und arithmetischen Mittel.

Der Median der Altersverteilung im Kurs hat einen Wert von 21 Jahren (vgl. Abschnitt 4.1.6). Die mittlere absolute Abweichung der Lebensalter vom Median in Jahren beträgt

$$\begin{aligned} d_{\bar{x}_Z} = \frac{1}{32}(&\cdot|18-20| + 5\cdot|19-20| + 13\cdot|20-20| + 7\cdot|21-20| \\ &+ |22-20| + |23-20| + |25-20| + |26-20| + |29-20| \\ &+ |40-20|) = \frac{1}{32}\cdot \approx 1,84. \end{aligned}$$

Die mittlere absolute Abweichung vom arithmetischen Mittel $\bar{x} = 21,4$ Jahre ist wie in den meisten Fällen ähnlich, aber etwas größer:

$$\begin{aligned} d_{\bar{x}} = \frac{1}{32}(&\cdot|18-21,4| + 5\cdot|19-21,4| + 13\cdot|20-21,4| + 7\cdot|21-21,4| \\ &+ |22-21,4| + |23-21,4| + |25-21,4| + |26-21,4| + |29-21,4| \\ &+ |40-21,4|) = \frac{1}{32}\cdot \approx 2,28. \end{aligned}$$

Sind für $m < n$ verschiedene Einzelwerte $(x_1, x_2, ..., x_m)$ die absoluten bzw. relativen Häufigkeiten $(h_1, h_2, ..., h_m)$ bzw. $(f_1, f_2, ..., f_m)$ gegebe, so können die mittleren absoluten Abweichungen folgendermaßen berechnet werden:

$$d_{\bar{x}} := \frac{1}{n}\sum_{i=1}^{m}|x_i - \bar{x}|\cdot h_i \text{ bzw. } d_{\bar{x}} := \sum_{i=1}^{m}|x_i - \bar{x}|\cdot f_i \tag{35}$$

und

$$d_{\bar{x}_Z} := \frac{1}{n}\sum_{i=1}^{m}|x_i - \bar{x}_Z|\cdot h_i \text{ bzw. } d_{\bar{x}_Z} := \sum_{i=1}^{m}|x_i - \bar{x}_Z|\cdot f_i \tag{36}$$

Zur Berechnung sowohl der mittleren absoluten Abweichungen vom arithmetischen Mittel als auch vom Median werden alle Werte der Stichprobe $(x_1, x_2, ..., x_n)$ verwendet. Bei der Berechnung dieser beiden Maße wird also im Gegensatz zur Spannweite keine Information mißachtet. Außerdem sollte ein zum Vergleich der Streuungen verschiedener Stichproben konzipiertes Streuungsmaß sinnvollerweise den Umfang der jeweiligen Stichprobe berücksichtigen. Das ist bei den beiden Maßen (33) und (34) der Fall. Das am

meisten verwendete Streuungsmaß ist allerdings die empirische Standardabweichung, die ganz wesentlich auf der empirischen Varianz basiert.

4.2.3 empirische Varianz und Standardabweichung

Die *durchschnittliche quadratische Abweichung* der Einzelwerte einer Stichprobe $(x_1, x_2, ..., x_n)$ von Merkmalswerten eines kardinalen Merkmals X vom arithmetischen Mittel $\bar{x}$ wird als die empirische Varianz s^2 bezeichnet:

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{37}$$

Mit Hilfe der 2. binomischen Formel kann die empirische Varianz umgeschrieben werden:

$$\begin{aligned} s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - 2\bar{x}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) + \frac{1}{n}\cdot n\bar{x}^2, \end{aligned}$$

wobei der mittlere Term wegen $1/n\sum x_i = \bar{x}$ genau $-2\bar{x}^2$ ergibt und der letzte Term sich auf $\bar{x}^2$ reduziert. Damit ergibt sich eine analoge Formel

$$s^2 := \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2, \tag{38}$$

die sich meist einfacher berechnen lässt.

Die empirische Varianz s_2 nimmt genau dann den Wert Null an, wenn jede einzelne der quadratischen Differenzen verschwindet. Das ist dann und nur dann der Fall, wenn alle Merkmalswerte identisch sind. In diesem Fall existiert keine Streuung unter den Merkmalswerten, es ist in nur ein Merkmalswert vorhanden, der dann auch das arithmetische Mittel darstellt. Wie bei den mittleren absoluten Abweichungen vom Median bzw. vom arithmetischen Mittel ist der Wert Null bei der empirischen Varianz der einzige ausgezeichnete Wert: er charakterisiert eine Situation ohne Streuung. Dies stellt aber eine höchst seltene und unter statistischen Gesichtspunkten eher langweilige Situation dar.

Sind für die $m < n$ tatsächlich verschiedenen Einzelwerte $(x_1, x_2, ..., x_n)$ die absoluten bzw. relativen Häufigkeiten $(h_1, h_2, ..., h_m)$ bzw. $(f_1, f_2, ..., f_m)$ gegeben, kann die empirische Varianz wie folgt bestimmt werden:

$$s^2 := \frac{1}{n} \sum_{i=1}^{m} (x_i - \bar{x})^2 \cdot h_i, \tag{39}$$

bzw.

$$s^2 := \sum_{i=1}^{m} (x_i - \bar{x})^2 \cdot f_i. \tag{40}$$

Würde mit dem kardinalen Merkmal X zum Beispiel die Körpergröße in Meter bezeichnet, wäre die empirische Varianz s^2 von der Dimension 'Körpergröße im Quadrat' und hätte die Einheit m^2, also die Einheit einer Fläche! Das kann nicht sinnvoll sein - die empirische Varianz ist daher ein nicht zu interpretierendes Maß. Stattdessen stellt die Wurzel aus der empirischen Varianz ein interpretierbares Maß dar, das als empirische Standardabweichung bezeichnet wird.

Beispiel: Berechnung der empirischen Varianz und Standardabweichung.

Das bereits bekannte arithmetische Mittel der Altersverteilung im Kurs beträgt $21,4$ Jahre. Damit kann die empirische Varianz s^2 nach Formel (37) berechnet werden:

$$\begin{aligned} s^2 = \frac{1}{32} \big[&\cdot (18 - 21,4)^2 + 5 \cdot (19 - 21,4)^2 + 13 \cdot (20 - 21,4)^2 \\ &+ 7 \cdot (21 - 21,4)^2 + (22 - 21,4)^2 + (23 - 21,4)^2 \\ &+ (25 - 21,4)^2 + (26 - 21,4)^2 + (29 - 21,4)^2 + (40 - 21,4)^2 \big] \\ &\approx 15,87. \end{aligned}$$

Zur einfacheren Berechnung kann natürlich genausogut die zweite Form (38) herangezogen werden:

$$\begin{aligned} s^2 = \frac{1}{32} \big(&18^2 + 5 \cdot 19^2 + 13 \cdot 20^2 + 7 \cdot 21^2 + 24^2 + 25^2 \\ &+ 29^2 + 31^2 + 39^2 \big) - 21,4^2 \approx 15,87. \end{aligned}$$

Das Ergebnis für die Varianz beträgt $20,64$ Jahre im Quadrat. Interpretierbar ist allerdings die Standardabweichung als Wurzel der Varianz. Die Standardabweichung (die Wurzel der durchschnittlichen quadratischen Abweichung der einzelnen Werte vom arithmetischen Mittel) beträgt in diesem Beispiel $s = 3,98$ Jahre.

Definition der empirischen Standardabweichung

Die empirische Standardabweichung s ist definiert als die *positive* Quadratwurzel der empirischen Varianz:

$$s := \sqrt{s^2} > 0 \tag{41}$$

- Nicht-Negativität: die empirische Varianz als Summe lauter quadratischer Terme sowie die Wurzel daraus, die empirische Standardabweichung, sind stets größer oder gleich Null. Beide Maße weisen nur dann einen Wert von Null auf, wenn alle Merkmalswerte eines kardinalen Merkmals identisch sind. Dies zeichnet den Wert von Null als einzigen aller Werte aus, den beide Maße annehmen können. Üblicherweise sind jedoch die Werte der empirischen Varianz und Standardabweichung von Null verschieden, weil immer eine gewisse Streuung unter den Merkmalswerten einer Stichprobe vorhanden ist (andernfalls benötigt man eigentlich keine Statistik).

- Transformationseigenschaften von empirischer Varianz und Standardabweichung: geht ein kardinales Merkmal Y durch eine allgemeine lineare (und reelle) Transformation aus einem kardinalen Merkmal X hervor, dessen arithmetisches Mittel $\bar{x}$ und empirische Varianz s_X^2 bekannt sind, so kann die empirische Varianz s_Y^2 des Merkmals Y aus der empirischen Varianz s_X^2 des Merkmals X berechnet werden, ohne dass die einzelnen Werte x_i einer Stichprobe $({}_x1, x_2, ..., x_n)$ durch die lineare Transformation $y_i = a + b \cdot x_i$ in jeweils einen Merkmalswert des Merkmals Y umgerechnet werden:

$$s_Y^2 = b^2 \cdot s_X^2. \tag{42}$$

Für die Stichprobe $({}_y1, y_2, ..., y_n)$ ergibt sich nach einer linearen Transformation der Einzelwerte die folgende empirische Varianz s_Y^2,

$$\begin{aligned} s_Y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_1 - \bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n}[(a + b \cdot x_i) - (a + b \cdot \bar{x})]^2 = \frac{1}{n}\sum_{i=1}^{n} b^2 \cdot (x_i - \bar{x})^2 \\ &= b^2 \cdot s_X^2, \end{aligned}$$

wobei die Transformationseigenschaft für das arithmetische Mittel, $\bar{y} = a + b \cdot \bar{x}$ (Formel (13)) und die Definition (37) der empirischen Varianz verwendet wurde. Durch Bildung der Quadratwurzel[3] ergibt sich die Standardabweichung

$$s_Y = \sqrt{s_Y^2} = \sqrt{b^2 \cdot s_X^2} = |b| \cdot s_X.$$

Beispiel: Lineare Tranformation der Standardabweichung

Gemessene Werte in Zoll (X) lassen sich durch die genäherte Transformation

$$Y = 2,5 \cdot X$$

in Werte in cm (Y) umrechnen. Es handelt sich hierbei wieder um eine lineare Transformation, allgemein durch $y = a + b \cdot x$, wobei hier das konstante Glied a verschwindet, es gilt $b = 2,5$.

Einige der Felgen auf dem Hof des örtlichen Autohändlers haben die folgenden Durchmesser:

Y [in cm]	35	$37,5$	40	$37,5$	$42,5$
X [in Zoll]	14	15	16	15	17

Tabelle 11: Felgengrößen.

[3] Die empirische Standardabweichung ist als die *positive* Quadratwurzel der empirischen Varianz definiert. Daher muss für die Berechnung mit beliebigem b der Betrag benutzt werden.

Das arithmetische Mittel der Felgengrößen beträgt $\bar{x} = 15.4$Zoll, die Standardabweichung $s_x = 1.0198$ Zoll. Das arithmetische Mittel in cm erhält man nach Definition als

$$\bar{y} = \frac{1}{5}(35 + 37,5 + 40 + 37,5 + 42,5) = 38,5,$$

die Standardabweichung als

$$s_y = \sqrt{\frac{1}{5}(35^2 + 37,5^2 + 40^2 + 37,5^2 + 42,5^2) - 38,5^2} = 2,55$$

Einfacher ist es aber, das arithmetische Mittel und die Standardabweichung in cm durch die lineare Transformation aus den bereits bekannten Werten für das arithmetische Mittel $\bar{x}$ und die Standardabweichung s_x in Zoll zu berechnen:

$\bar{y} = a + b\bar{x} = 2,5 \cdot 15,4 = 38,5$ und $s_y = |b| \cdot s_x = 2,5 \cdot 1,0198 = 2,55$

4.2.4 Variationskoeffizient

Wie die Spannweite und die mittleren absoluten Abweichungen vom arithmetischen Mittel oder vom Median ist auch die empirische Standardabweichung ein Maß für die absolute Streuung. Diese Maße sind im Allgemeinen dimensionsbehaftet sie hängen von der Einheit, in der ein Merkmal gemessen wird, ab. Relative Streuungsmaße sind dagegen dimensionslos. Ein Beispiel eines solchen relativen Maßes ist der *Variationskoeffizient*.

Definition des Variationskoeffizienten

Für ein kardinales Merkmal X mit arithmetischem Mittel $\bar{x}$ und empirischer Standardabweichung s_X ist der Variationskoeffizient v_X definiert durch

$$v_X := \frac{s_X}{\bar{x}}, \tag{43}$$

das absolute Streuungsmaß s_X wird ins Verhältnis zum durchschnittlichen Niveau - ausgedrückt durch das arithmetische Mittel - des Merkmals X gesetzt. Der Variationskoeffizient v_X ist als Quotient zweier Größen gleicher Dimension und Einheiten dimensions - und einheitenlos.

Beispiel: Variationskoeffizient

Die Streuung der Daten zur Felgengröße im letzten Beispiel ist natürlich unabhängig von der Wahl der Messung in cm oder Zoll. Die beiden unterschiedlichen Werte $s_x = 1,0198$ Zoll bzw. $s_y = 2,55$ cm für die Standardabweichung scheinen aber etwas anderes nahezulegen (faktisch bedeutet die lineare Transformierbarkeit der Standardabweichung eben gerade, dass die Streuung dieselbe ist).

Der Grund dafür ist, dass die Standardabweichung ein Maß für die *absolute* Streuung ist, dessen Wert von der gewählten Einheit abhängt, in der das untersuchte Merkmal (die Felgengröße) gemessen wird. Der in cm gemessene Wert $s_y = 2,55$ ist größer als der in Zoll gemessene Wert $s_x = 1,0198$, weil die Felgengrößen in cm größere Zahlenwerte aufweisen als die in Zoll gemessenen Werte.

Dieser Skaleneffekt kann dadurch vermieden werden, dass *relative* anstatt absoluter Streuungsmaße verwendet werden. Ganz allgemein setzen relative Streuungsmaße absolute Streuungsmaße ins Verhältnis zum durchschnittlichen Niveau, das ein untersuchtes Merkmal aufweist. Ein Beispiel für ein solches relatives Streuungsmaß ist der Variationskoeffizient: er ist für die (an der gleichen statistischen Gesamtheit erhobenen) Merkmale X und Y identisch, da es sich bei den beiden um denselben Sachverhalt handelt:

$$v_X = \frac{s_x}{\bar{x}} = \frac{1,0198 \text{ Zoll}}{15,4 \text{ Zoll}} = 0,066 = 6,6\%,$$

$$v_Y = \frac{s_y}{\bar{y}} = \frac{2,55 \text{ cm}}{38,5 \text{ cm}} = 0,066 = 6,6\%$$

Diese Zahl lässt sich folgendermaßen interpretieren: die empirische Standardabweichung der Felgengrößen der untersuchten Felgen beträgt 6,6% des Mittelwertes. Es spielt dabei keine Rolle, in welchen Einheiten die betrachteten Merkmale gemessen werden.

4.3 Schiefe

Ob eine Verteilung als symmetrisch oder unsymmetrisch bzw. schief zu bezeichnen ist, kann anhand von Stab- und Balken-Diagrammen bzw. Histogrammen sehr leicht ermittelt werden. Beispielsweise sind die beiden in Abb.

6 dargestellten Verteilungen symmetrisch. Die in Abb. 8 gezeigte Altersverteilung wird als schief - insbesondere als linkssteil oder synonym als rechtsschief - bezeichnet.

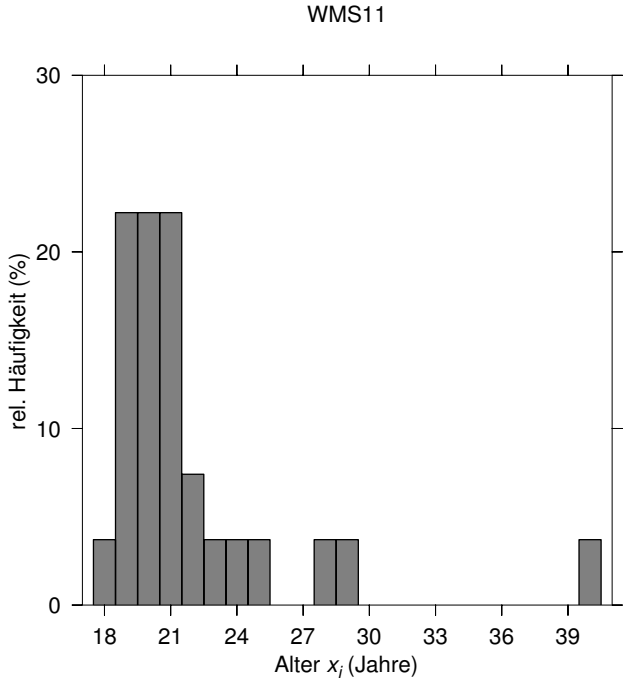

Abbildung 8: die Altersverteilung im Kurs ist linkssteil bzw. rechtsschief. Mit den bereits bestimmten Werten $\bar{x} = 21.4, \bar{x}_Z = 20$ und $\bar{x}_M = 20$ ergibt sich die Ungleichung $\bar{x} \geq \bar{x}_Z \geq \bar{x}_M$.

Eine einfache Möglichkeit zur Einschätzung der Schiefe eingipfliger Verteilungen, welche durch die Eindeutigkeit des Modus charakterisiert sind, bietet die folgende Faustregel, die als Kriterium Beziehungen zwischen dem arithmetischen Mittel $\bar{x}$, dem Median $\bar{x}_Z$ und dem Modus $\bar{x}_M$ benutzt:

FECHNERsche Lageregel: ist eine eingipflige Verteilung

linkssteil,	so gilt in der Regel	$\bar{x} \geq \bar{x}_Z \geq \bar{x}_M$
rechtssteil,	so gilt in der Regel	$\bar{x} \leq \bar{x}_Z \leq \bar{x}_M$
symmetrisch,	so gilt immer	$\bar{x} = \bar{x}_Z = \bar{x}_M$

Aufgrund der allgemein gültigen Regel $\bar{x} = \bar{x}_Z = \bar{x}_M$ bei symmetrischen, eingipfligen Verteilungen sind bei Kenntnis des arithmetischen Mittels einer symmetrischen Verteilung auch Modus und Median bekannt.
Linkssteile bzw. rechtsschiefe Verteilungen sind von großer empirischer Bedeutung. Beispielweise sind Verteilungen, welche Einkommens- und Vermögensverhältnisse unterschiedlicher Personengruppen wiedergeben, typischerweise linkssteil. Charakteristisch für diese Verteilungen ist insbesondere, dass

der Median $\bar{x}_Z$ kleiner als das arithmetische Mittel $\bar{x}$ ist: wenige Bezieher hoher Einkommen verleihen dem arithmetischen Mittel einen hohen Wert $\bar{x}$, während sich die große Masse der Einkommensbezieher am unteren Rand einer typischen Einkommensverteilung konzentriert. Rechtssteile bzw. linksschiefe Verteilungen sind hingegen tendenziell dadurch charakterisiert, dass der Median $\bar{x}_Z$ größer als das arithmetische Mittel $\bar{x}$ ist.

Maßzahlen zur Quantifizierung der Schiefe einer Verteilung stützen sich auf das dritte zentrale Moment[4].

4.3.1 Statistische Momente

Geht man von einer Stichprobe von n verschiedenen Einzelwerten $(x_1, x_2, ..., x_n)$ aus, so ist das statistische Moment r-ter Ordnung um einen festen Bezugspunkt a definiert durch

$$m_r(a) := \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r. \tag{44}$$

Is der Bezugspunkt $a = 0$, so spricht man vom *gewöhnlichen Moment r-ter Ordnung*

$$m_r(0) := \frac{1}{n}\sum_{i=1}^{n} x_i^r, \tag{45}$$

das arithmetische Mittel der Merkmalswerte $(x_1, x_2, ..., x_n)$ ergibt sich damit als ein spezielles statistisches Moment - es ist das gewöhnliche Moment erster Ordnung (mit $a = 0$ und $r = 1$):

$$m_1(0) := \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \tag{46}$$

und damit das gewöhnliche Moment mit der größten Bedeutung. Stellt das arithmetische Mittel $\bar{x}$ den Bezugspunkt a dar, so wird das Moment als *zentrales Moment r-ter Ordnung* bezeichnet:

$$m_r(\bar{x}) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r. \tag{47}$$

[4]...selbst Statistiker haben ihre Momente!

Die empirische Varianz stellt also das zentrale Moment 2-ter Ordnung (auch das zweite zentrale Moment genannt) dar

$$m_2(\bar{x}) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2. \tag{48}$$

Von den zentralen Momenten hat die empirische Varianz als zentrales Moment 2-ter Ordnung die größte Bedeutung.

4.3.2 Streuungs- und Schiefemaße

Streuungsmaße erlauben es, Aussagen über die Variabilität von Merkmalswerten innerhalb einer Stichprobe zu treffen. Die Frage, welches Streuungsmaß im Einzelnen heranzuziehen ist, kann dabei pauschal nicht beantwortet werden. Prinzipiell unterscheidet man zwischen absoluten Streuungsmaßen, beispielsweise der Spannweite, den mittleren absoluten Abweichungen vom Median bzw. dem arithmetischen Mittel oder der empirischen Varianz bzw. Standardab weichung, und relativen Streuungsmaßen, zum Beispiel dem Variationskoeffizienten. Während die absoluten Streuungsmaße im Allgemeinen dimensions- und einheitenbehaftet sind, besitzen relative Streuungsmaße weder Dimension noch Einheit, da bei diesen ein absolutes Streuungsmaß auf einen Lageparameter bezogen wird (die beiden Größen besitzen dieselbe Einheit). Beim Variationskoeffizienten wird beispielsweise das Verhältnis von empirischer Standardabweichung und arithmetischem Mittel gebildet. Denkbar wäre zum Beispiel auch das Verhältnis von empirischer Standardabweichung zum Median. Die größte Beliebtheit genießt in der statistischen Literatur die empirische Standardabweichung, während die einfach zu berechnende Spannweite ebenfalls oft Anwendung findet, allerdings bei der Existenz von Ausreißern einen völlig falschen Eindruck vermittelt. Allgemein gilt die folgende Kette von Ungleichungen zwischen mittlerer absoluter Abweichung vom Median, der empirischen Standardabweichung und der Spannweite:

$$d_{\bar{x}_Z} \leq s \leq s_W \tag{49}$$

Diese Ungleichung kann als einfacher Hinweis benutzt werden, ob die Standardabweichung korrekt berechnet wurde: Der Wert s muss kleiner sein als der Wert s_W, den die Spannweite aufweist. Sinnvollerweise sollten alle Maße den Wert Null annehmen, wenn keine Streuung unter den Merkmalswerten existiert, also alle Beobachtungen ein und denselben Wert haben. Dieser Fall kommt jedoch in der Praxis normalerweise nicht vor. Üblicherweise weisen

die Beobachtungen in einer Stichprobe mehr oder weniger stark voneinander abweichende Werte auf[5]. Diese Streuung der Werte wird je nach Maß quantifiziert durch die absoluten oder quadratischen Abstände der Einzelwerte vom gewählten Bezugspunkt wie dem arithmetischen Mittel oder dem Median. Einfache Differenzen der Merkmalswerte speziell vom arithmetischen Mittel sind wegen der Schwerpunkteigenschaft des arithmetischen Mittels nicht geeignet, sie heben sich in der Summe immer auf.

Streng genommen setzt die Berechnung von Streuungsmaßen ein Vorliegen kardinaler Merkmale voraus. Bei ordinalen Merkmalen kann man sich aber mit Hilfsgrößen wie beispielsweise dem Quartilsabstand - dem Abstand zwischen oberen und unterem Quartil - zu behelfen. Bei kardinalen Merkmalen hat der Quartilsabstand gegenüber der Spannweite den Vorteil, nicht anfällig gegenüber Ausreißern zu sein.

[5]Wenn die Beobachtungen einer Stichprobe keine Streuung aufweisen, ist strenggenommen keine Statistik vonnöten.

4.4 Übungsaufgaben zu Streuungs- und Schiefemaßen

- Landtagswahlen in 7 fiktiven Bundesländern einer föderalen Republik brachten den Parteien A und B die folgenden Ergebnisse (in Prozent):

Bundesland	1	2	3	4	5	6	7
Partei A	5,6	6,3	6,6	6,9	7,1	7,6	6,1
Partei B	40,4	41,9	47,9	40,4	48,9	41,4	42,9

  Die arithmetischen Mittel der Ergebnisse der beiden Parteien lauten $\bar{x} = 6,6\%$ für die Partei A, $\bar{y} = 43,4\%$ für Partei B. Gerno Osterwelle, Vorsitzender der Partei A behauptet während eines Interviews trotzig: "Unser Ergebnis ist in allen Ländern ziemlich gleich, während die Ergebnisse der Partei B wesentlich weniger stabil sind." Ist diese Behauptung korrekt?

- Die Beschäftigten eines Unternehmens erhalten im Mittel einen Monatslohn von $\bar{x} = 2200$ Euro bei einer Standardabweichung von $s_X = 800$ Euro. Dank eines Wirtschaftsaufschwungs war die letzte Lohnverhandlung für die Mitarbeiter erfolgreich: Das Monatsgehalt jedes Beschäftigten wird um 10% angehoben und es werden in Zukunft 960 Euro Urlaubsgeld gewährt. Wie ändern sich Mittelwert, Varianz und Standardabweichung der Gehälter der Mitarbeiter sowie der Variationskoeffizient?

- Das Merkmal X besitze das arithmetische Mittel $\bar{x}$ und die Standardabweichung s_X. Das Merkmal Y entstehe aus dem Merkmal X und der linearen Transformation $Y = b \cdot X$. Zeigen Sie, dass der Variationskoeffizient v_Y für das Merkmal Y mit dem Variationskoeffizienten v_X für das Merkmal X übereinstimmt. Träfe das auch zu, wenn die lineare Transformation $Y = a + b \cdot X$ lauten würde?

4.5 Konzentration und Disparität

Konzentration im ökonomischen Sinne kann zweierlei bedeuten:

- Die Konzentration von beispielsweise Marktanteilen, also von ökonomischer Macht, auf genau eine Wirtschaftseinheit (Monopol) oder auf lediglich einige wenige Wirtschaftseinheiten (Oligopol).
- Die Existenz erheblicher Unterschiede zwischen den Anteilen von Wirtschaftseinheiten am Gesamtbetrag eines relevanten Merkmals wie beispielsweise dem Umsatz.

Im ersten Fall ist das relevante Kriterium die geringe Anzahl an Wirtschaftseinheiten, also der Aspekt der absoluten Anzahl an Merkmalsträgern (absolute Konzentration oder Konzentration im engeren Sinne). Im zweiten Fall hingegen ist der Aspekt der Ungleichheit (Disparität) unter den Wirtschaftseinheiten bezüglich eines Merkmals, nicht aber deren absolute Anzahl (relative Konzentration oder Konzentration im weiteren Sinne).

> Beispiel: zur absoluten und relativen Konzentration.
> Eine Aussage im Sinne einer relativen Konzentration ist beispielsweise: 2% der Bevölkerung lateinamerikanischer Staaten besitzen mehr als 90% des Geldvermögens dieser Staaten. In der Aussage tauchen ausschließlich relative Werte (angegeben in Prozenten) auf: Diese relativen Werte geben den Anteil am Gesamtwert des untersuchten Merkmals (das Geldvermögen) an, den ein bestimmter Anteil von Merkmalsträgern aufweist.
>
> Eine Aussage im Sinne der absoluten Konzentration wäre dagegen: Auf dem deutschen Energiemarkt haben nur zwei Konzerne zusammen einen Marktanteil von etwa 80%. Die Merkmalsträger sind in absoluter Anzahl angegeben, die Zahl ist zudem sehr gering.

Der Unterschied zwischen absoluter und relativer Konzentration wird besonders deutlich bei einer Gleichverteilung, bei welcher der Gesamtwert eines Merkmals, beispielsweise das Geldvermögen völlig gleichmäßig auf alle einzelnen Merkmalsträger verteilt ist. Unabhängig von der Zahl der Merkmalsträger existiert bei einer Gleichverteilung per Definition keine relative Konzentration, sie ist selbst bei einer Verteilung des Gesamwerts auf lediglich zwei

Merkmalsträger genau Null. Je kleiner aber die Zahl der Merkmalsträger ist, desto größer ist die absolute Konzentration.

Der folgende Abschnitt beginnt mit der Herleitung des wohl bekanntesten Ungleichheitsmaßes, dem GINI-Koeffizienten. Seine Interpretation basiert auf der Lorenzkurve, mit deren Hilfe Ungleichheit-Situationen illustriert werden können. In der Folge wird der HERFINDAHL-Index diskutiert, welcher wohl das populärste Maß zur Erfassung absoluter Konzentrationen darstellt.

Statistische Maße zur Messung relativer Konzentration berücksichtigen nur den Aspekt der Ungleichheit (Disparität), wohingegen Maße zur Messung der absoluten Konzentration neben der Disparität auch den Aspekt der absoluten Anzahl erfassen.

4.5.1 Lorenzkurve

Bei der Bestimmung einer relativen Konzentration oder der Ungleichheit geht es um die Frage, ob ein großer Anteil am Gesamtwert eines Merkmals wie beispielsweise dem Energieverbrauch, um den es im folgenden Beispiel geht, auf einen geringen Anteil aller Merkmalsträger entfällt.

Beispiel: fiktive Werte zum Energieverbrauch auf Gliese 581c.
Die Bevölkerung des Exoplaneten Gliese 581c (GB für *G*liese-*B*evölkerung) wird häufig in abwertender Weise aufgeteilt in die sogenannte 'erste', 'zweite' und 'dritte' Welt, womit die Bevölkerung der Industrieländer, der Schwellenländer respektive der Entwicklungsländer gemeint ist. Der jährliche Gliese-Energieverbrauch (GEV) teilt sich auf diese drei 'Welten' in etwa wie folgendermaßen auf:

	Anteil der GB f_i	kum. Ant. der GB F_i	Anteil am GEV q_i	kum. Ant. am GEV Q_i	$(F_i; Q_i)$
3. Welt $(i = 1)$	60%	60%	10%	10%	$(0,6; 0,1)$
2. Welt $(i = 2)$	30%	90%	30%	40%	$(0,9; 0,4)$
1. Welt $(i = 3)$	10%	100%	60%	100%	$(1; 1)$

Ein geringer Anteil der Gliese-Bevölkerung von etwa 10% der in den Industrieländern lebenden Menschen beanspruchen demnach einen großen Anteil, ca. 60% der jährlich verbrauchten Energie, während ca. 60% der Bevölkerung mit der geringen Menge von ca. 10% auskommt.

Die Lorenzkurve L wird konstruiert mit Hilfe der Eckpunkte $(F_i; Q_i)$, allgemein gebildet aus den kumulierten relativen Anteilen einer untersuchten Gruppe an der Gesamtheit (im Beispiel die Bevölkerung)

$$F_i = \sum_{k=1}^{i} f_k = f_1 + f_2 + ... + f_i \tag{50}$$

und deren kumuliertem Anteil

$$Q_i = \sum_{k=1}^{i} q_k = q_1 + q_2 + ... + q_i \tag{51}$$

am Gesamtwert des betrachteten Merkmals. Ergänzt werden diese Punkte noch um den Ursprung $(0; 0) = (F_0; Q_0)$, der den Ausgangspunkt der Lorenzkurve bildet. Die Lorenzkurve selbst besteht aus dem Polygonzug, der die Punkte $(F_0; Q_0), ..., (F_i; Q_i), ..., (F_n; Q_n)$ durch Geraden miteinander verbindet.

Beispiel: Lorenzkurve des Gliese-Energieverbrauchs.
Für das obige Beispiel des Gesamtenergieverbrauchs auf Gliese 581c lauten die Eckpunkte der Lorenzkurve: $(F_0; Q_0) = (0; 0), (F_1; Q_1) = (0,6; 0,1), (F_2; Q_2) = (0,9; 0,4)$ und $(F_3; Q_3) = (1; 1)$. Die Lorenzkurve L verbindet nun diese Eckpunkte durch Geraden.

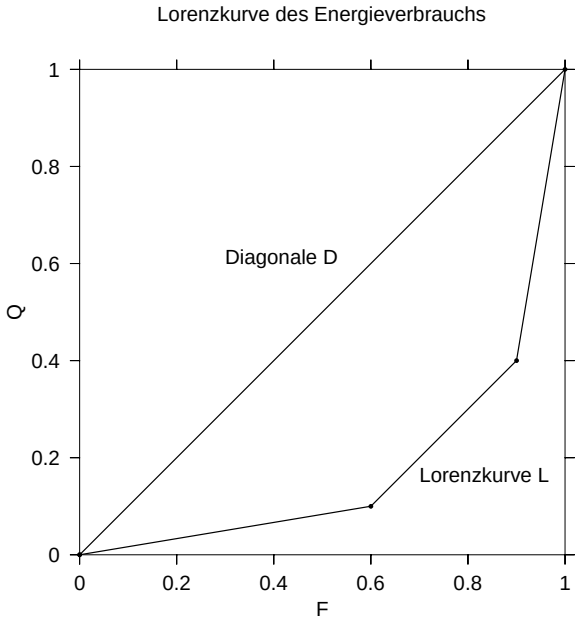

Abbildung 9: Die Lorenzkurve zur Illustration des fiktiven Energieverbrauchs.

Die Diagonale D im Schaubild ist die Kurve, die den Zustand der Gleichverteilung darstellt. Je stärker eine Lorenzkurve L von dieser Diagonalen abweicht, desto größer ist die Ungleichheit innerhalb der Verteilung der Merkmale auf einzelne Merkmalsträger. Mit anderen Worten: je größer die Abweichung der Lorenzkurve von der Diagonalen, desto stärker ist die relative Konzentration (bezogen auf ein bestimmtes Merkmal) innerhalb der betrachteten Grundgesamtheit.

4.5.2 GINI-Koeffizient

Ein Maß für die Abweichung der Lorenzkurve L von der Diagonalen D (gewissermaßen den 'Bauch' der Lorenzkurve) ist der GINI-Koeffizient. Im extremen Grenzfall, der in der Realität allerdings nicht auftreten kann, entspricht dieser Bauch gerade der gesamten Fläche unter der Diagonalen und damit der Fläche eines Dreiecks.

Der GINI-Koeffizient G mißt die Fläche zwischen der Diagonalen D und der Lorenzkurve L und setzt sie ins Verhältnis zur Fläche des Dreiecks unter der Diagonalen, die wegen der Konstruktion über kumulierte relative Werte den Betrag$1/2$ aufweist:

$$G := \frac{\text{Fläche zwischen } D \text{ und } L}{\text{Dreiecksfläche unter } D} = \frac{\text{Fläche zwischen } D \text{ und } L}{1/2} \\ = 2 \cdot \text{Fläche zwischen } D \text{ und } L \tag{52}$$

Im Falle völliger Gleichverteilung - bei der natürlich keine Konzentration vorliegt - weicht die Lorenzkurve nicht von der Diagonalen ab. Die Fläche zwischen der Diagonalen und der Lorenzkurve ist in diesem Fall Null, der Wert des GINI-Koeffizienten ist damit ebenfalls 0. Im Falle einer extremen Ungleichverteilung kommt die Fläche zwischen der Diagonalen D und der Lorenzkurve L der Dreiecksfläche unter D sehr nahe, allerdings ohne sie jemals zu erreichen. Durch die Division durch die Zahl $1/2$ (d.h. die Multiplikation mit 2) besitzt der GINI-Koeffizient die folgende Bandbreite:

$$0 \leq G < 1. \tag{53}$$

Die Fläche zwischen D und L, und damit den GINI-Koeffizienten, gewinnt man aus der Fläche des Dreiecks unter der Diagonalen D durch Subtraktion der Flächen aller Trapeze, die unterhalb der Lorenzkurve liegen. In Abb. 10 ist eine Lorenzkurve mit der Konstruktion der Trapeze unter der Kurve exemplarisch dargestellt. Die Fläche der einzelnen dargestellten Trapeze errechnet

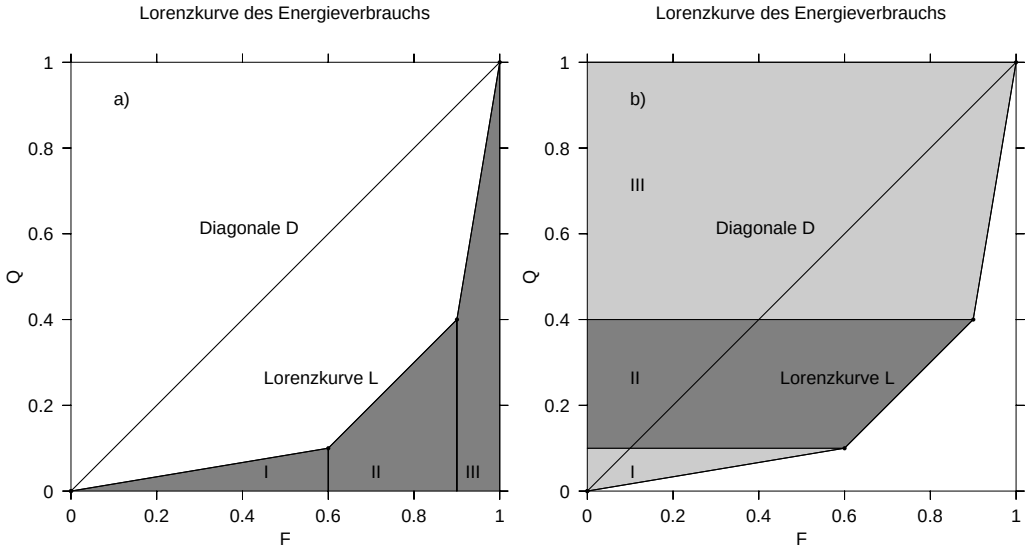

Abbildung 10: a) Herleitung einer Formel für den GINI-Koeffizienten. b) Alternative Berechnung.

sich aus der Länge der Grundseite $F_i - F_{i-1} = f_i$, multipliziert mit der durchschnittlichen Höhe $(Q_i + Q_{i-1})/2$.

$$\begin{aligned} G &= 2 \cdot \left(\frac{1}{2} - \text{ Summe der Flächen unter } L \text{ (Trapeze)} \right) \\ &= \frac{2}{2} - 2 \cdot \sum_{i=1}^{n} f_i \frac{Q_{i-1} + Q_i}{2} \\ &= 1 - \sum_{i=1}^{n} f_i (Q_{i-1} + Q_i) \end{aligned} \tag{54}$$

Dabei gilt für die Eckpunkte der Diagonalen D immer $Q_0 = 0, F_0 = 0$ und $Q_n = 1, F_n = 1$.

Beispiel: Ungleichheit beim fiktiven Energieverbrauch auf Gliese 581c.
Der GINI-Koeffizient für unser Beispiel des fiktiven Energieverbrauchs der Bevölkerung des Exoplaneten ergibt sich nach der hergeleiteten Formel (54) zu

$$G = 1 - \sum_{i=1}^{n} f_i(Q_{i-1} + Q_i) = 1 - (0.6 \cdot 0.1 + 0.3 \cdot (0.1 + 0.4) + 0.1 \cdot (0.4 + 1))$$

$$= 1 - 0.6 \cdot 0.1 - 0.3 \cdot (0.1 + 0.4) - 0.1 \cdot (0.4 + 1) = 0.65$$

Als Alternative lässt sich der GINI-Koeffizient auch berechnen, indem von der Summe der Flächen der Trapeze *oberhalb* der Lorenzkurve die Fläche des Dreiecks oberhalb der Diagonalen abgezogen wird. So errechnet sich die Fläche des in Abb. 10 b) skizzierten Trapezes II beispielsweise, indem die Länge der Grundseite (auf der y-Achse) $Q_i - Q_{i-1} = q_i$ mit der mittleren Höhe $(F_{i-1} + F_i)/2$ multipliziert wird

$$G = 2 \cdot \left(\text{ Summe der Flächen der Trapeze } - \frac{1}{2} \right) = 2 \cdot \sum_{i=1}^{n} q_i \frac{F_{i-1} + F_i}{2} - 1.$$

$$= \sum_{i=1}^{n} q_i (F_{i-1} + F_i) - 1 \qquad (55)$$

Beispiel: Marktmacht innerhalb einer Branche
Die fünf Hersteller von Kolbenrückholfedern in den USA erzielten im vorigen Jahr die folgenden Umsätze:

	U_1	U_2	U_3	U_4	U_5	Summe
Umsatz	600	1500	900	1800	1200	6000

Ordnet man die Unternehmen nach der Größe des Umsatzes und ermittelt die Anteile an der Gesamtheit sowie die Marktanteile (MA) in absoluter und kumulierter Form, so ergibt sich das folgende Bild:

	Umsatz	$f_i = \frac{1}{n}$	$F_i = i \cdot \frac{1}{n}$	MA	kum. MA	$(F_i; Q_i)$
$U_1(i=1)$	600	20%	20%	10%	10%	$(0.2; 0.10)$
$U_2(i=2)$	900	20%	40%	15%	25%	$(0.4; 0.25)$
$U_3(i=3)$	1200	20%	60%	20%	45%	$(0.6; 0.45)$
$U_4(i=4)$	1500	20%	80%	25%	70%	$(0.8; 0.70)$
$U_5(i=5)$	1800	20%	100%	30%	100%	$(1.0; 1.00)$

Aus diesen Daten kann nun mithilfe der Formel (54) der gesuchte Wert des GINI-Koeffizienten berechnet werden. Benötigt werden lediglich die Anteile an der Gesamtheit f_i (im Beispiel sind alle $f_i = 20\%$) sowie die kumulierten Marktanteile Q_i:

$$G = 1 - \sum_{i=1}^{n} f_i(Q_{i-1} + Q_i)$$

weil alle f_i gleich

$$\begin{aligned} &= 1 - f_i \cdot \sum_{i=1}^{n}(Q_{i-1} + Q_i) \\ &= 1 - 0.2 \cdot (0.1 + (0.1 + 0.25) + (0.25 + 0.45) \\ &+ (0.45 + 0.7) + (0.7 + 1)) \\ &= 1 - 0.2 \cdot 2 \cdot (0.1 + 0.25 + 0.45 + 0.7 + 1/2) = 0.2 \end{aligned}$$

Der Wert $G = 0.2$ des GINI-Koeffizienten deutet auf eine relativ geringe Konzentration des Umsatzes in der Branche hin. Dabei muss aber berücksichtigt werden, dass der maximale Wert G_{max} des GINI-Koeffizienten von der Zahl der untersuchten statistischen Einheiten n abhängt, für unseren Fall von 5 Unternehmen ergibt sich ein maximaler GINI-Koeffizient von $G_{\text{max}}(5) = 0.8$. Diesen Wert würde man erhalten, wenn sich der gesamte Umsatz der Branche auf ein einziges Unternehmen konzentrierte.

4.5.3 Maximalwert des GINI-Koeffizienten

Abb. 11 zeigt eine Lorenzkurve für den (wenig realistischen) Fall, in dem sich der Gesamtwert eines Merkmals auf eine einzelne aus n untersuchten statistischen Einheiten konzentriert. In diesem speziellen Fall lässt sich der GINI-Koeffizient relativ leicht berechnen, indem von der Fläche $1/2$ des Dreiecks unterhalb der Diagonalen D die Fläche des Dreiecks unter der Lorenzkurve L abgezogen wird:

$$G_{\text{max}}(n) = 2 \cdot \left(\frac{1}{2} - \frac{1}{2} \cdot \frac{1}{n} \cdot 1\right) = 1 - \frac{1}{n} = \frac{n-1}{n}. \tag{56}$$

Damit bewegt sich der Wert des GINI-Koeffizienten G im Intervall

$$0 \leq G \leq \frac{n-1}{n}.$$

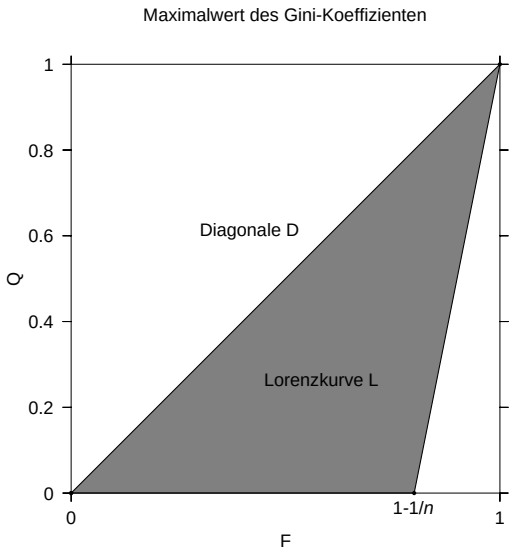

Abbildung 11: zur Bestimmung des Maximalwerts des GINI-Koeffizienten

Es ist sofort ersichtlich, dass der Wert des GINI-Koeffizienten den Wert 1 nicht überschreitet. Er erreicht ihn bei einer endlichen Zahl untersuchter Einheiten aber auch nie, selbst wenn sich das untersuchte Merkmal auf eine dieser Einheiten konzentriert. Im Beispiel der $n = 5$ Unternehmen kann der GINI-Koeffizient maximal $(5 - 1)/5 = 0.8$ erreichen, sofern sich der gesamte Umsatz auf ein einzelnes Unternehmen konzentriert - intuitiv erwartet man hier bei vollständiger Konzentration einen Wert von 1. Der berechnete Wert suggeriert also eine etwas geringe Konzentration als in der Realität vorhanden.

4.5.4 normierter GINI-Koeffizient

In Relation zum Maximalwert G_{max} des GINI-Koeffizienten beschreibt der ermittelte Wert G eine größere Konzentration als der reine Zahlenwert erwarten lässt. Im Beispiel steht der ermittelte Wert von $G = 0.2$ für eine Konzentration von

$$\frac{0.2}{4/5} = \frac{5 \cdot 0.2}{4} = 0.25.$$

Diese Überlegung legt die Bildung eines normierten GINI-Koeffizienten G_{norm} nahe, der durch Division des GINI-Koeffizienten G durch seinen Maximalwert G_{max} gebildet wird:

$$G_{\text{norm}} := \frac{G}{G_{\text{max}}(n)} = \frac{n-1}{n} \cdot G \tag{57}$$

Damit gilt für den Wertebereich des normierten GINI-Koeffizienten

$$0 \leq G_{\text{norm}} \leq 1$$

$$\text{wobei } G_{\text{norm}} = \begin{cases} 0 & \text{bei bei gleichmäßiger Verteilung der Merkmalswerte} \\ 1 & \text{bei vollständiger Konzentration} \end{cases}$$

Mit Hilfe des normierten GINI-Koeffizienten lässt sich der Grad der relativen Konzentration bzw. Ungleichheit zwischen zwei Stichproben unterschiedlichen Umfangs n miteinander vergleichen. Bei Stichproben mit großem Umfang ist die normierung des GINI-Koeffizienten jedoch oft nicht nötig, denn es gilt für den Normierungsfaktor

$$\frac{n}{n-1} \longrightarrow 1 \text{ für } \qquad n \longrightarrow \infty,$$

so dass der normierte GINI-Koeffizient für große n gegen den Wert des GINI-Koeffizienten strebt. Eine Normierung kann außerdem nur dann vorgenommen werden, wenn der Umfang der Stichprobe n bekannt ist - beispielsweise auf Basis einer relativen Häufigkeitsverteilung kann zwar der GINI-Koeffizient G, nicht aber der normierte GINI-Koeffizient G_{max} berechnet werden.

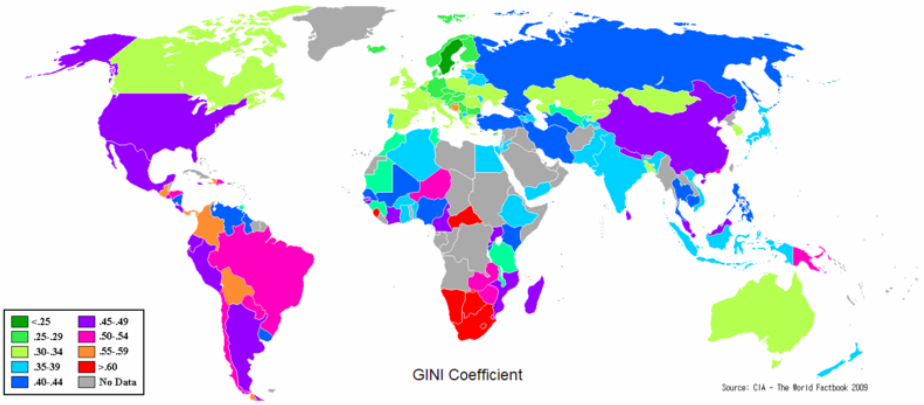

Abbildung 12: Beispiel der Anwendung des Gini-Koeffizienten: die Disparität der Einkommensverteilung pro Familie [2](CIA Factbook, 2009 [3])

4.6 absolute Konzentration

Der HERFINDAHL-Index ist das bekannteste Maß zur Messung absoluter Konzentrationen. Für eine Stichprobe bestehend aus n verschiedenen statistischen Einheiten, die auf die Merkmalswerte $x_1, x_2, ..., x_n$ eines Merkmals X

untersucht werden, ist der HERFINDAHL-Index definiert als

$$H := q_1^2 + q_2^2 + ... + q_n^2 = \sum_{i=1}^{n} q_i^2, \tag{58}$$

wobei q_i der Anteil am Gesamtwert des Merkmals X ist, der auf die statistische Einheit i entfällt:

$$q_i := \frac{x_i}{x_1 + x_2 + ... + x_n} = \frac{x_i}{\sum\limits_{i=1}^{n} x_i}.$$

Beispiel: Konzentrationsmessung.
In einem Markt seien fünf Unternehmen mit den Marktanteilen $60\%, 10\%, 5\%, 20\%$ und 5% tätig. Der GINI-Koeffizient berechnet sich nach Gleichung (54):

$$G = 1 - \sum_{i=1}^{n} f_i(Q_{i-1} + Q_i) \text{ mit} \qquad Q_i = \sum_{k=1}^{i} q_k.$$

Da die einzelnen $f_i = f = 0.2$ hier alle gleich sind, ergibt sich für den GINI-Koeffizienten

$$\begin{aligned} G &= 1 - f \cdot \sum_{i=1}^{n} (Q_{i-1} + Q_i) \\ &= 1 - 0.2 \cdot [(0 + 0.05) + (0.05 + 0.1) + (0.1 + 0.2) + (0.2 + 0.4) + (0.4 + 1)] \\ &= 0.2 \cdot 2.5 = 0.5. \end{aligned}$$

Der HERFINDAHL-Index zur Messung der Konzentration in diesem Markt beträgt für unser Beispiel

$$H = (0.6)^2 + (0.1)^2 + (0.05)^2 + (0.2)^2 + (0.05)^2 = 0.415$$

e beiden Unternehmen mit dem geringsten Marktanteil beschließen eine Fusion, um sich besser im Markt zu positionieren. Nach der Fusion wären dann nur noch 4 Unternehmen am Markt, der HERFINDAHL-Index betrüge dann

$$H = (0.6)^2 + 2 \cdot (0.1)^2 + (0.2)^2 = 0.42,$$

t sich gegenüber dem Wert vor der Fusion leicht erhöht (die absolute Konzentration ist gestiegen, weil weniger Unternehmen am Markt sind).

Der GINI-Koeffizient weist dagegen nach der Fusion einen geringfügig kleineren Wert auf:

$$G = 1 - 0.25 \cdot [(0 + 0.1) + (0.1 + 0.2) + (0.2 + 0.4) + (0.4 + 1)] = 0.4,$$

da durch die Fusion zwei Unternehmen mit einem Marktanteil von jeweils 10% entstanden sind, während die beiden Unternehmen mit einem Anteil von je 5% verschwunden sind. Obwohl das Unternehmen mit 60% Marktanteil weiterhin ein deutliches Übergewicht besitzt, hat sich durch die Fusion die Ungleichheit leicht reduziert.

Sind die Anteile von n statistischen Einheiten am Gesamtwert eines untersuchten Merkmals X allesamt gleich groß ($x_1 = x_2 = ... = x_n$), so ergibt sich für die einzelnen q_i ein Wert von

$$q_i = \frac{x_1}{x_1 + x_2 + ... + x_n} = \frac{x_1}{n \cdot x_1} = \frac{1}{n},$$

für den HERFINDAHL-Index ergibt sich dann ein Wert von

$$H = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = n \cdot \left(\frac{1}{n}\right)^2 = \frac{1}{n}.$$

Je kleiner die Zahl der betrachteten statistischen Einheiten, desto größer ist der HERFINDAHL-Index H. Im Fall extremer Konzentration vereinigt eine einzige statistische Einheit den gesamten Wert des Merkmals auf sich - in diesem Fall ist der Anteil dieser statistischen Einheit am Merkmalswert $q_k = 1$, die Anteile der anderen betrachteten Einheiten verschwinden $q_{i \neq k} = 0$. Für den Herfingahl-Index bedeutet dies

$$H = q_k^2 + \sum_{i=1; i \neq k}^{n} q_i^2 = 1 + \sum_{i=1; i \neq k}^{n} 0 = 1.$$

Der Wertebereich des HERFINDAHL-Index erstreckt sich natürlich vom Wert für den Fall völliger Gleichverteilung bis zu seinem Wert bei völliger Konzentration:

$$1/n \leq H \leq 1, \tag{59}$$

je größer der Wert des HERFINDAHL-Index, desto größer ist die Konzentration.

Beispiel: HERFINDAHL-Index und Fusionen.
In einem Markt mit n unterschiedlichen Unternehmen, die die Marktanteile $(q_1, q_2, ...q_n)$ besitzen, soll das Unternehmen k mit dem Unternehmen $k+1$ fusionieren. Bei dieser Fusion nimmt die absolute Konzentration zu (das letzte Beispiel hat gezeigt, dass dies für die relative Konzentration nicht unbedingt der Fall sein muss). Die Zunahme der absoluten Konzentration spiegelt sich in der Zunahme des HERFINDAHL-Index wieder:

$$H = q_1^2+...+(q_k+q_{k+1})^2+...+q_n^2 = q_1^2+...+q_k^2+ \underbrace{2q_kq_{k+1}}_{\text{zusätzlicher Term } >0} +q_{k+1}^2+...+q_n^2$$

Der zusätzliche Term $2q_kq_{k+1}$ ist immer positiv, wenn die Marktanteile q_k und q_{k+1} der beiden Unternehmen positiv sind.

Wenn insbesondere die Marktanteile aller Unternehmen vor der Fusion gleich $(q_1, q_2, ..., q_n) = (1/n, 1/n, ..., 1/n)$ und damit $H = 1/n$, so ergibt sich nach der Fusion von Unternehmen k mit Unternehmen $k+1$ der HERFINDAHL-Index

$$H = \left(\frac{2}{n}\right)^2 + \sum_{i=1}^{n-1}\left(\frac{1}{n}\right)^2 = \frac{4}{n^2} + (n-1)\cdot\left(\frac{1}{n}\right)^2 = \frac{3}{n^2} + \frac{1}{n}.$$

4.6.1 Übungsaufgaben

- Zwei verschiedene Märkte sollen jeweils von 10 unterschiedlichen Firmen beliefert werden. Die Marktanteile verteilen sich wie folgt:

Markt M_1	9 Firmen mit je 50/9% Anteil	1 Firma mit 50% Marktanteil
Markt M_2	5 Firmen mit je 2% Anteil	5 Firmen mit je 18% Marktanteil

  1. Zeichnen Sie für die beiden Märkte die Lorenzkurven.
  2. Welcher der Märkte kann im ökonomischen Sinne als konzentrierter bezeichnet werden?
  3. Berechnen Sie den GINI-Koeffizienten für die beiden Märkte. Welcher Schluss kann aus dem Ergebnis gezogen werden?

- Der HERFINDAHL-Index als Maß für absolute und relative Konzentration. In einem Markt mit insgesamt 1000 Unternehmen soll auf 999 Unternehmen ein jeweils gleicher, aber verschwindend geringer Anteil entfallen - diese Unternehmen sollen gemeinsam einen Marktanteil von insgesamt 1% haben. Das letzte Unternehmen soll sich den größten Anteil von 99% gesichert haben. Zeigen Sie, dass der HERFINDAHL-Index, obwohl bei einer Anzahl von insgesamt 1000 Unternehmen nicht von absoluter Konzentration gesprochen werden kann, einen Wert nahe 1 anzeigt - und damit trotzdem eine große (relative) Konzentration anzeigt.

5 Bivariate Verteilungen

Bisher wurden lediglich Verteilungen eines einzelnen Merkmals (z.B. die Körpergröße bestimmter Personen) betrachtet. Liegen Stichproben vor, die mehrere Merkmalswerte erfassen (die prinzipiell voneinander abhängig sein können), so spricht man von multivariaten Verteilungen. Der Einfachheit halber werden wir uns hier auf Verteilungen zweier Merkmale, sog. bivariate Verteilungen, beschränken. Die interessante Frage in der deskriptiven Statistik ist hier natürlich die nach der Beziehung der einzelnen Merkmale untereinander.

5.1 Kreuztabellen

Eine Darstellungsform bivariater Beziehungen für alle Arten von Merkmalen sind zweidimensionale Häufigkeitstabellen, die Häufigkeitsvergleiche für Paare von Merkmalsausprägungen erlauben, sogenannte Kontingenz- oder Kreuztabellen. Die erhobenen Daten werden zweidimensional für eines der Merkmale in Zeilen, für das zweite in Spalten dargestellt.

Merkmal	Merkmal X					
Y	x_1	...	x_j	...	x_k	$\sum$
y_1	h_{11}	...	h_{1j}	...	h_{1k}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_1	h_{i1}	...	h_{ij}	...	h_{ik}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_1	h_{m1}	...	h_{mj}	...	h_{mk}	$n_{m.}$
$\sum$	$n_{.1}$	...	$n_{.j}$	...	$n_{.k}$	n

In der Tabelle bedeuten x_i die i-te Merkmalsausprägung (bzw. Merkmalswert) des Merkmals X, h_{ij} die absoluten Häufigkeiten, mit denen die Ausprägung x_j und y_i auftreten.

Durch diese Art der Darstellung lassen sich Häufungen von Paaren feststellen. Falls die erhobenen Merkmale unabhängig voneinander sind, erwartet man bei einer solchen Auftragung von *relativen* Häufigkeiten dieselben Werte in den jeweiligen Spalten bzw. Zeilen (eben gerade, weil die *relativen* Häufigkeiten unabhängig vom jeweiligen zweiten Merkmal dieselben sind).

Beispiel: Abhängigkeit der Mathematiknote vom Vertiefungsfach.
Untersucht werden soll die Abhängigkeit der Mathematiknote vom vom jeweiligen Vertiefungsfach. Dazu wurden 25 Studenten zu ihrem Vertiefungsfach und zur im letzten Semester erzielten Mathematiknote befragt und die Zahl der Studenten mit Vertiefungsfach Y und Mathematiknote X notiert:

Mathematiknote	Vertiefungsfach			
Y	Hallen-Halma	Häkeln	Zitronenfalten	
2	1	6	3	10
3	4	3	3	10
4	0	1	4	5
$\sum$	5	10	10	25

Nach Berechnung der relativen Häufigkeiten ergibt sich das folgende Bild:

Mathematiknote	Vertiefungsfach			
Y	Hallen-Halma	Häkeln	Zitronenfalten	
2	0,2	0,6	0,3	0,4
3	0,8	0,3	0,3	0,4
4	0	0,1	0,4	0,2
$\sum$	1	1	1	1

Tabelle 12: Mathenoten und Vertiefungsfach

Innerhalb der Zeilen bzw. Spalten der Verteilung relativer Häufigkeiten ergeben sich unterschiedliche Werte. Dies bedeutet, dass die Mathematiknote, die ein Student im letzten Semester erzielte, nicht von seinem Vertiefungsfach unabhängig war.

5.2 Lineare Regression

Bei sehr vielen Beobachtungswerten stetiger Merkmale ist eine graphische Darstellung als Streudiagramm übersichtlicher. Dabei werden die Merkmalswerte des Merkmals Y über denen des Merkmals X aufgetragen.

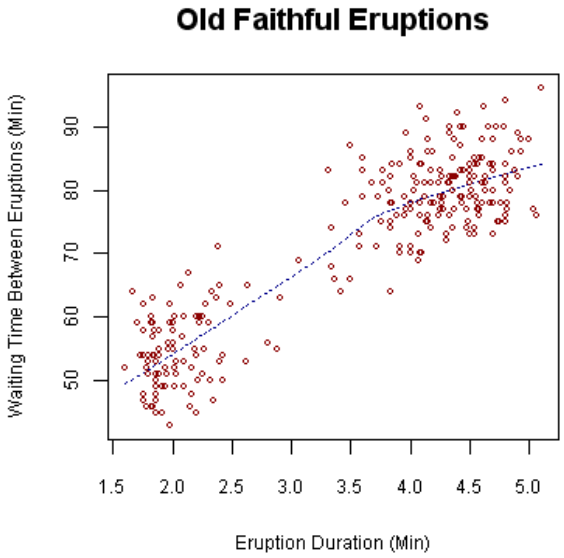

Abbildung 13: Messwerte des Geysirs 'Old faithful' zur Pause zwischen den Ausbrüchen und der Dauer der Ausbrüche [1]. Die Verteilung der Daten legt einen Zusammenhang zwischen Dauer und Pause nahe.

In einer solchen Auftragung zeigt sich eine Abhängigkeit der Merkmale voneinander in einer Häufung von Datenpunkten entlang von Geraden (bei linearer Abhängigkeit) bzw. allgemeiner Funktionen (im Falle nicht-linearer Abhängigkeiten). Wir wollen natürlich versuchen, diese Funktionen aus dem vorliegenden Datenmaterial zu bestimmen. Dazu betrachten wir als einfaches Beispiel die folgenden Daten:

Beispiel: lineare Regression.
Wir nehmen an, dass ein Hersteller von Kolbenrückholfedern mehrere Modelle (nummeriert mit dem Index i) zum Preis von jeweils x_i im Angebot hat. Um den Absatz zu optimieren, soll eine Preis-Absatz-Funktion ermittelt werden.

	Preis	Menge			
i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	20	0	5	-5	-25
2	16	3	1	-2	-2
3	15	7	0	2	0
4	16	4	1	-1	-1
5	13	6	-2	1	-2
6	10	10	-5	5	-25
Summe	90	30	0	0	-55
Mittelwert	15	5			

Tabelle 13: Preise verschiedener Modelle von Kolbenrückholfedern

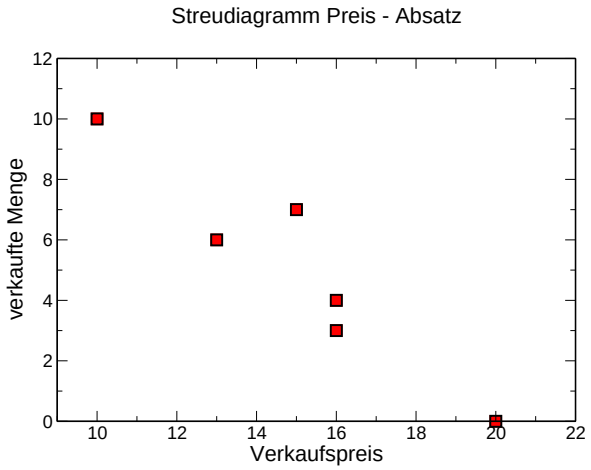

Abbildung 14: 'Messpunkte' zum Beispiel der Hersteller von Kolbenrückholfedern.

5.2.1 Die Kovarianz

Wir haben die Varianz einer Häufigkeitsverteilung bereits in Abschnitt 4.2.3 kennengelernt. Sie baut auf den quadrierten Abweichungen $(x_i - \bar{x})^2$ einzelner Daten vom Schwerpunkt des Datensatzes, dem arithmetischen Mittel $\bar{x}$, auf.

Die Kovarianz basiert auf einem ähnlichen Prinzip, sie berechnet sich aber aus den Abständen der Beobachtungspaare (x_i, y_i) von den jeweiligen arithmetischen Mitteln. An die Stelle der Abweichungsquadrate, die wir von der

Berechnung der empirischen Varianz kennen, treten die Produkte der Abweichungen der einzelnen Merkmale, die Ergebnisse aller Ausprägungspaare werden schließlich addiert und durch die Anzahl n geteilt:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{60}$$

Die Größe s_{xy} bezeichnet man als *Kovarianz*, sie lässt sich natürlich nur berechnen, wenn beide Merkmale kardinal skaliert sind.

Vier-Quadranten-Schema

Zur Interpretation der Kovarianz betrachten wir das sog. Vier-Quadranten-Schema. Dieses Schema ermöglicht es, auf einfache Art einen tendenziellen Zusammenhang zwischen zwei vorliegenden Merkmalen zu erkennen. Man hat beim Betrachten der Daten in Abb. 14 den Eindruck, eine fallende Tendenz zu erkennen. Wir wollen diesen bildlichen Eindruck präzisieren. Dazu werden in das Streudiagramm zusätzliche Parallelen zu den Achsen, jeweils durch das arithmetische Mittel der betrachteten Merkmale eingezeichnet. Den Schnittpunkt dieser beiden Geraden mit den Koordinaten $(\bar{x}, \bar{y})$ bezeichnen wir als Schwerpunkt der bivariaten Verteilung. Die dadurch entstandenen vier Teilbereiche werden gegen den Uhrzeigersinn - rechts oben beginnend - als I. bis IV. Quadrant bezeichnet.

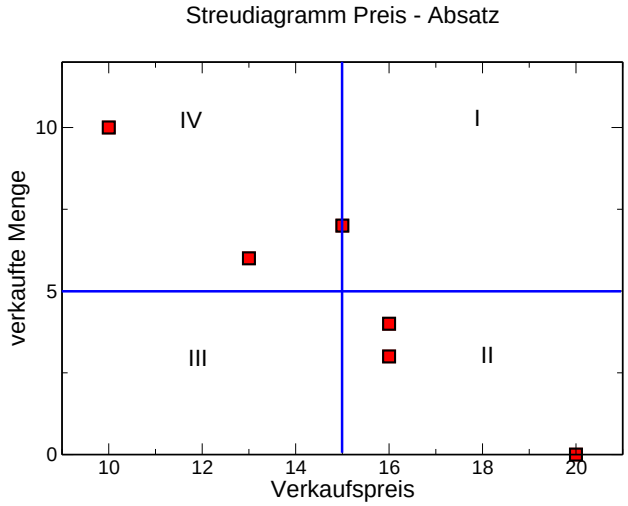

Abbildung 15: Vier-Quadranten-Schema am Beispiel der Hersteller von Kolbenrückholfedern. Man erkennt die Häufung der Daten in den Quadranten II und IV.

Wir können anhand der Hilfslinien feststellen, dass im Beispiel der Hersteller von Kolbenrückholfedern der überwiegende Teil der Punkte im zweiten und vierten Quadranten liegt. Dies ist offensichtlich Ausdruck der 'abwärtsgerichteten' Tendenz.

Allgemein gilt: liegen die Punkte des Streudiagramms hauptsächlich im

- I. und III. Quadranten, so deutet dies auf einen positiven Zusammenhang zwischen den Merkmalen hin; mit steigenden Werten von x steigen auch die Werte von y. In diesem Fall ist die Kovarianz $s_{xy} > 0$.
- II. und IV. Quadranten, so deutet dies auf einen negativen Zusammenhang zwischen den Merkmalen hin; mit steigenden Werten von x fallen auch die Werte von y. In diesem Fall ist $s_{xy} < 0$.
- Verteilen sich die Punkte in etwa gleichmäßig auf alle vier Quadranten, so deutet das daraufhin, dass kein Zusammenhang zwischen den Merkmalen besteht. Die Punktwolke hat eine diffuse Gestalt, es gilt $s_{xy} \approx 0$.

Die Kovarianz gibt an, welche Tendenz der Zusammenhang zwischen den beiden Merkmalen besitzt. Sie ist aber allgemein dimensionsbehaftet und damit abhängig von der gewählten Skala, weshalb sie sich nicht vergleichen lässt.

Auch kann aus $s_{xy} = 0$ nicht geschlossen werden, dass kein Zusammenhang zwischen den beiden untersuchten Größen besteht. Einerseits hängt der Betrag der Kovarianz stark von der jeweiligen Skalierung der Merkmale ab, andererseits gibt es durchaus Zusammenhänge zwischen zwei Merkmalen, die sich nicht unbedingt durch die Kovarianz erfassen lassen (beispielsweise nichtlineare Zusammenhänge).

5.2.2 Lineare Regression

Wir gehen nun davon aus, dass wir eine Datenreihe zweier Merkmale vorliegen haben, bei denen wir annehmen, dass ein Zusammenhang zwischen den Merkmalen existiert. Es liegen also mehrere unabhängige Angaben zum selben Sachverhalt vor, aus denen zunächst ein Schätzwert des tatsächlichen

Werts bestimmt werden muss, da jedes einzelne Datenpaar mit einer Abweichung vom realen Wert behaftet ist. Gedanklich könnte jedes einzelne Paar von Punkten im Streudiagramm mit einer Funktion (im einfachsten Fall einer Geraden) verbunden werden. Man erhält eine Anzahl von Funktionen, die sich ähnlich, aber nicht identisch sein werden. Es sind im Prinzip mehrere Möglichkeiten des 'Ausgleichens' solcher redundanten Messungen denkbar. Aus historischen Gründen hat sich die sogenannte *Methode der kleinsten Quadrate* durchgesetzt: in Abb. 14 gegeben sind Punkte in einem System der Variablen x, y. Diese Punktwolke soll durch eine Gerade repräsentiert werden. Gesucht sind die Parameter jener Geraden, die für diese Approximation 'am besten' geeignet ist. Wir nennen diese Gerade auch *Ausgleichsgerade*.

Die Punkte in Abb. 14 könnten beispielsweise die grafische Darstellung einer Messreihe sein, wobei auf der Abszisse eine Steuergröße aufgetragen wurde und auf der Ordinate die untersuchte Messgröße. Wir nehmen ferner an, dass die wahren Werte der Messgrößen auf einer Geraden liegen, d.h. zwischen der Variablen x und den wahren Werten der beobachteten Größe y soll ein linearer Zusammenhang bestehen.

$$y = f(x) = a + b \cdot x$$

Mathematisch können wir zur Bestimmung der beiden Geradenparameter Steigung b und Ordinatenabschnitt a ein Gleichungssystem aufstellen, wobei jeder Punkt eine Gleichung beisteuert (wir setzen dabei voraus, dass mehr Datenpunkte vorliegen, als zur Bestimmung der Unbekannten notwendig sind) $\Rightarrow$ es handelt sich bei unserem Beispiel um ein Gleichungssystem mit zwei Unbekannten und 6 Gleichungen. Ein solches System wird als überbestimmt bezeichnet, wegen der leichten Abweichungen realer Messwerte kann es nicht eindeutig gelöst werden. Eine Gerade ist in der Ebene durch zwei Punkte definiert. Haben wir mehr als zwei Punkte, die auf der Geraden liegen sollen, so können wir im Allgemeinen keine eindeutige Lösung angeben. Es muss ein Kriterium dafür gefunden werden, welche Gerade der Punktwolke 'möglichst gut angepasst' ist. Wir fordern also, dass die Abweichung der Funktionswerte der gesuchten Funktion von den gemessenen Werten so klein wie möglich ist.

Die Abweichung kann hierbei unterschiedlich definiert werden - plausibel erscheint eine Definition, in der ein größerer Fehler überproportional mehr wiegt als ein kleiner. Zudem darf das Fehlermaß nicht vorzeichenbehaftet sein (sonst würde ein negativer Fehler einen betragsgleichen positiven Fehler ausgleichen). Die einfachen Quadrate der Differenzen zwischen Mess- und Funk-

tionswert erfüllen diese beiden Forderungen. $(f(x) - y)^2$ ist also ein geeignetes Maß für die Abweichung der gemessenen Größe y[6].

Gesucht wird nun eine Funktion, für die die Summe

$$S = \sum (f(x_i) - y_i)^2$$

der Quadrate der einzelnen Differenzen zwischen Funktions- und Messwerten minimal wird. Die notwendige Bedingung dafür, dass diese Summe minimal ist, ist dass ihre erste Ableitung 0 wird. Vermutet man einen linearen Zusammenhang $f(x_i) = y_i = a + b \cdot x_i$, kann die Summe wie folgt umgeschrieben werden:

$$\sum_{i=1}^{n} (f(x_i) - y_i)^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \longrightarrow \min!$$

Durch partielles Differenzieren und Nullsetzen der Ableitungen erster Ordnung erhält man ein System von Normalgleichungen. Die gesuchten Regressionskoeffizienten sind die Lösungen

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{61}$$

und

$$a = \bar{y} - b\bar{x} \tag{62}$$

Beispiel: lineare Regression.
Für die Daten des Herstellers von Kolbenrückholfedern

	Preis	Menge				
i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	20	0	5	-5	-25	25
2	16	3	1	-2	-2	1
3	15	7	0	2	0	0
4	16	4	1	-1	-1	1
5	13	6	-2	1	-2	4
6	10	10	-5	5	-25	25
Summe	90	30	0	0	-55	56
Mittelwert	15	5				

[6]Wir gehen hier - wie in der Literatur üblich - davon aus, dass die Größe x exakt bestimmbar ist und suchen die Abweichung in Richtung der Abszisse

Tabelle 14: lineare Regression am Beispiel von Kolbenrückholfedern

ergeben sich für die beiden gesuchten Regressionsparameter die Werte

$$b = \frac{S_{xy}}{S_{xx}} = \frac{-55}{56} = -0,98$$

und

$$a = \bar{y} - b \cdot \bar{x} = 5 + 0,98 \cdot 15 = 19,73$$

Die jeweilige verkaufte Menge an Produkten y hängt also mit dem Preis x angenähert wie $y = a + b \cdot x$ zusammen. Mit einer Preiserhöhung um eine Einheit sinkt der Absatz um etwa eins.

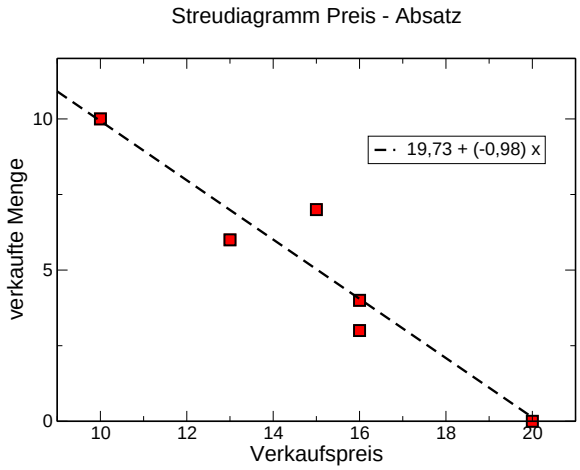

Abbildung 16: 'Messpunkte' zum Beispiel der Hersteller von Kolbenrückholfedern mit eingezeichneter Ausgleichsgerade.

5.3 Der Korrelationskoeffizient

Der Pearson-Korrelationskoeffizient oder Korrelationswert[7] ist ein dimensionsloses Maß, das den Grad eines linearen Zusammenhangs zwischen zwei kardinalen Merkmalen angibt. Er eignet sich zur Untersuchung der Kreuzkorrelation (Korrelation zeitgleicher Messwerte zweier Merkmale) und der Autokor-

[7] von Bravais und Pearson

relation (Korrelation zeitlich verschiedener Messwerte eines einzelnen Merkmals).

Der Korrelationswert kann Werte zwischen -1 und +1 annehmen. Bei einem Wert von +1 (bzw. -1) besteht ein vollständig positiver (negativer) linearer Zusammenhang zwischen den betrachteten Merkmalen. Weist der Korrelationskoeffizient den Wert 0 auf, sind die betrachteten Merkmale nicht linear voneinander abhängig (sie können aber in nicht-linearer Weise voneinander abhängen - der Korrelationskoeffizient ist kein geeignetes Maß für die reine stochastische Abhängigkeit von Merkmalen).

Der Pearson-Korrelationskoeffizient ist für zwei Zufallsvariablen X, Y mit positiver (d. h. von Null verschiedener) Standardabweichung $\varsigma(X), \varsigma(Y)$ und der Kovarianz COV(X, Y) definiert durch

$$\rho_{xy} = \frac{\text{COV}(X,Y)}{\varsigma(X)\varsigma(Y)}$$

Wir kennen bereits die empirische Standardabweichung und die empirische Kovarianz für eine Messreihe gepaarter Merkmalsausprägungen $(x_1; y_1), (x_2; y_2), \ldots (x_n; y_n)$ - das erlaubt uns natürlich die Berechnung eines empirischen Pearson-Korrelationskoeffizienten.

5.3.1 empirischer Korrelationskoeffizient

Bezeichnen $\bar{x}$ und $\bar{y}$ die arithmetischen Mittel der Merkmale X und Y in einer Messreihe gepaarter Merkmalsausprägungen $(x_1; y_1), (x_2; y_2), \ldots (x_n; y_n)$, so bezeichnen wir die Größe

$$\begin{aligned} r_{xy} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \end{aligned} \tag{63}$$

als den Korrelationskoeffizienten der Messreihe.

5.3.2 Interpretation des Korrelationskoeffizienten

Aus der Definition des Korrelationskoeffizienten (63) ist der Wert des Koeffizienten für perfekte Korrelation sofort ersichtlich. Der perfekte lineare Zusammenhang ist sicher gegeben, wenn man eine Variable als abhängig von sich

selbst betrachtet, also:

$$r_{xx} := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1.$$

Für Wertepaare, die auf einer perfekten Geraden liegen (perfekte lineare Abhängigkeit), ist der Korrelationskoeffizient offensichtlich 1, falls die Gerade ansteigt (sonst -1).

Je mehr sich die Gestalt der durch die Wertepaare gebildeten Punktwolke von einer Geraden abweicht, desto kleiner wir der Wert r_{xy}, bei $r_{xy} = 0$ kann der Zusammenhang zwischen den Merkmalen nicht mehr durch eine eindeutig steigende oder fallende Gerade dargestellt werden. Dies bedeutet, dass die Werte nicht mehr verlässlich an eine Gerade angepasst werden können (oder natürlich, dass eines der beiden Merkmale konstant ist - auch dann ist kein Zusammenhang gegeben).

Beispiel:
Für unser Beispiel des Herstellers der unterschiedlichen Kolbenrückholfedern sind die Werte, die zur Berechnung der empirischen Kovarianz und der Standardabweichung s_x des Merkmals x bereits bekannt. Wir benötigen lediglich die Werte für die Standardabweichung s_y:

	Preis	Menge				
x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
20	0	5	-5	-25	25	25
16	3	1	-2	-2	1	4
15	7	0	2	0	0	4
16	4	1	-1	-1	1	1
13	6	-2	1	-2	4	1
10	10	-5	5	-25	25	25
90	30	0	0	-55	56	60
Mittelwert	15	5				

Tabelle 15: Bestimmung des Pearson-Korrelationskoeffizienten

Der Korrelationskoeffizient nach Pearson und Bravais kann jetzt einfach berechnet werden:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{-55}{\sqrt{56} \cdot \sqrt{60}} = -0,95$$

Der Wert nahe -1 legt einen relativ starken linearen Zusammenhang nahe, das Vorzeichen bedeutet, dass die Gerade abfällt.

5.3.3 Das Bestimmtheitsmaß R^2

Die lineare Regression beschreibt den Zusammenhang zwischen einer oder mehreren unabhängigen (oder erklärenden) Variablen und einer abhängigen Variablen. Handelt es sich um eine Regression mit einer unabhängigen Variablen, so spricht man von einer einfachen Regression, bei mehreren unabhängigen Variablen von einer multiplen Regression. Das Bestimmtheitsmaß gibt an, wie gut die unabhängigen Variablen geeignet sind, die Varianz (und damit die typische Abweichung von einem einfachen Lageparameter) zu erklären. Das (empirische) Bestimmtheitsmaß, auch als Determinationskoeffizient bezeichnet, ist eine Kennzahl zur Beurteilung der Anpassungsgüte einer Regression. Sie liegt als statistischer Parameter zwischen Null (bei einem vollkommen ungeeigneten Modell) und 1 (bei perfekter Anpassung des Modells an eine Stichprobe), beispielsweise um zu bewerten, wie gut gemessene Werte zu einem Modell passen. FIXME: Bravais-Pearson?? Varianz und Variation

Bei der Behandlung bivariater Verteilungen gibt es meist eine Größe (die abhängige Variable y), deren Schwankung (Varianz) mit Hilfe anderer Größen (hier die unabhängige Variable x) erklärt werden soll. Das ist beispielweise sinnvoll, um über den ermittelten Zusammenhang eine Vorhersage (oder Prognose) für die abhängige Variable zu erstellen. Sind n Wertepaare (x_i, y_i) gegeben, so wird jedem Wert x_i ein Merkmals- oder Messwert y_i zugeordnet. Die Modellfunktion, in der Statistik auch als Schätzer bezeichnet, ordnet jedem Wert x_i einen Schätzwert $\hat{y}_i$ zu, für den linearen Zusammenhang aus unserer linearen Regression gilt also

$$\hat{y}_i = b \cdot x_i + a.$$

mit dem Schätzer $\hat{y}(x) = a \cdot x + b$. Jede einzelne Beobachtung y_i lässt sich dann berechnen als die Summe aus dem vorhergesagten Wert (der Schätzung) $\hat{y}_i$ und der Abweichung zwischen beobachtetem und vom Modell vorhergesagten Wert e_i (Fehler bzw. Residuum):

$$y_i = \hat{y}_i + e_i$$

Jetzt lässt sich aus den geschätzten Werten ganz analog zur empirischen Varianz bzw. Variation (Summe der quadrierten Abweichungen der Merkmalswerte y_i vom arithmetischen Mittel) natürlich auch eine Varianz oder Variation der

Schätzwerte bestimmen, indem die Abweichungen *der Schätzwerte* $\hat{y}_i$ *vom arithmetischen Mittel quadriert und summiert werden. Die Gesamtvariation enthält die Beiträge der Residuen, sie ist*

$$\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Die Variation der Schätzwerte bezeichnet man als erklärte Variation. Sie ist typischerweise geringer als die Gesamtvariation, weil sie zusätzlich die Beiträge der Residuen enthält

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

Den Quotient aus erklärter Variation und Gesamtvariation bezeichnet man als Bestimmtheitsmaß

$$R^2 := \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Zusätzlich gilt mit den Residuen e_i *(die Gleichheit ist nicht trivial zu zeigen, deshalb wird hier darauf verzichtet)*

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(e_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

mit der unerklärten Variation

$$\sum_{i=1}^{n}(e_i)^2$$

. Was bedeutet das Bestimmtheitsmaß anschaulich? Die naive Näherung für die Werte y *in der Stichprobe ist das arithmetische Mittel*

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Durch Einbeziehen der unabhängigen Größe x *übe die lineare Regression kann diese sehr simple Näherung deutlich verbessert werden - um den Beitrag der erklärten Variation (die Abweichungen der Schätzwerte* $\hat{y}_i$ *vom arithmetischen Mittel - kennt man den Schätzwert in Abhängigkeit von* x*, hat man genau diese Abweichungen eliminiert). Übrig bleiben aber noch immer die Beiträge der Residuen, der Abweichungen zwischen den Schätzwerten und den gemessenen Werten* $e_i = y_i - \hat{y}_i$*. Sie werden durch die Regression nicht erklärt, man bezeichnet sie deshalb auch als unerklärte Variation. Das Bestimmtheitsmaß lässt sich jetzt über eine Betrachtung zweier Grenzfälle interpretieren:*

- *für eine vollkommen ungeeignete Modellfunktion erwarten wir, dass sich die Abweichungen nicht über das Modell erklären lassen. In diesem Fall wird die erklärte Variation verschwinden (die Residuen werden dabei maximal). Damit muss sich für R^2 als Quotient aus erklärter Variation und Gesamtvariation der Wert Null ergeben.*
- *für eine perfekt Modellfunktion verschwinden die Residuen (weil jeder Wert y_i perfekt durch den Schätzwert $\hat{y}_i$ modelliert wird. In diesem Fall verschwindet die unerklärte Variation und es ist $R^2 = 1$.*

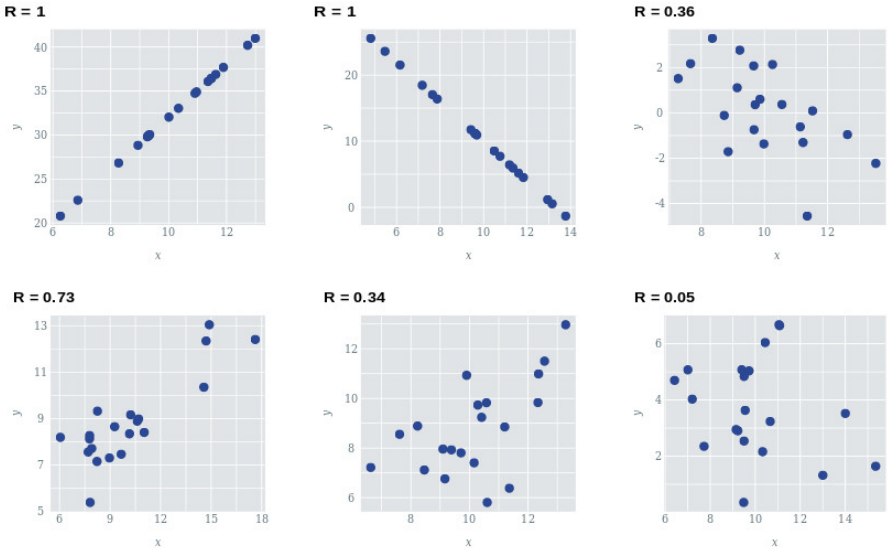

Abbildung 17: Streudiagramme für unterschiedliche Werte von R^2. Je besser die Datenpunkte auf einer Linie liegen, desto höher ist das Bestimmtheitsmaß. Streuen die Datenpunkte ohne Zusammenhang im Raum, liegt R^2 nahe Null.

R^2 kann also als Maß für die Qualität des Modells gesehen werden, kleine Werte bedeuten eine schlechte Modellierung der gemessenen Daten durch den Schätzer.

Zusammenhang mit dem Korrelationskoeffizienten

Bei einer einfachen linearen Regression mit einer erklärenden Variablen ist $y_i = b \cdot x_i + a + e_i$, das Bestimmtheitsmaß lässt sich als

$$R^2 := \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

ausdrücken. Mit der Steigung der Regressionsgeraden

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

folgt

$$\begin{aligned} R^2 &:= \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}} \right)^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})} \right)^2 = r_{xy}^2 \end{aligned}$$

Das Bestimmtheitsmaß entspricht bei einer einfachen Regression dem Quadrat des Bravais-Pearson-Korrelationskoeffizienten.

Tabellenverzeichnis

1 *die zu Beginn erhobenen Daten.* 19

2 *absolute Häufigkeiten h_i und relative Häufigkeiten f_i zum Alter.* 20

3 *die Größenverteilung im Kurs in klassierter Form. Die Dichte f_k^* ist der Quotient f_k/Δ_k.* 24

4 *kumulierte Häufigkeiten H_i und F_i für die Altersverteilung.* . . . 26

5 *Klassengrenzen und kumulierte Häufigkeiten F_k zur Konstruktion des Verteilungspolygons.* 27

6 *Altersverteilung im Kurs.* 31

7 *Primärenergieverbrauch pro Kopf im Jahr 2000.* 35

8 *Stamm-Blatt-Darstellung zur Ermittlung des Medians.* 41

9 *Teilnahmehäufigkeiten des Herrn Dent.* 46

10 *kumulierte Teilnahmehäufigkeiten des Herrn Dent.* 52

11 *Felgengrößen.* 61

12 *Mathenoten und Vertiefungsfach* 83

13 *Preise verschiedener Modelle von Kolbenrückholfedern* 85

14 *lineare Regression am Beispiel von Kolbenrückholfedern* 89

15 *Bestimmung des Pearson-Korrelationskoeffizienten* 92

Abbildungsverzeichnis

1 *Altersverteilung im Stabdiagramm: absolute Häufigkeiten* . . . 22

2 *Altersverteilung im Balkendiagramm: relative Häufigkeiten, angegeben in %.* 23

3 *Größenverteilung im Histogramm: die relativen Häufigkeiten ergeben sich durch Multiplikation der aufgetragenen Dichte mit der Klassenbreite Δ_i* 25

4 Größenverteilung im Verteilungspolygon 28

5 Bestimmung des Medians im Verteilungspolygon 45

6 Vergleich zweier Verteilungen mit dem gleichen arithmetischen Mittel 54

7 Stabdiagramme zweier Verteilungen ähnlicher Spannweite . . . 55

8 die Altersverteilung im Kurs ist linkssteil bzw. rechtsschief. Mit den bereits bestimmten Werten $\bar{x} = 21.4, \bar{x}_Z = 20$ und $\bar{x}_M = 20$ ergibt sich die Ungleichung $\bar{x} \geq \bar{x}_Z \geq \bar{x}_M$. 64

9 Die Lorenzkurve zur Illustration des fiktiven Energieverbrauchs. 71

10 a) Herleitung einer Formel für den GINI*-Koeffizienten. b) Alternative Berechnung. 73*

11 zur Bestimmung des Maximalwerts des GINI*-Koeffizienten . . . 76*

12 Beispiel der Anwendung des Gini-Koeffizienten: die Disparität der Einkommensverteilung pro Familie [2](CIA Factbook, 2009 [3]) 77

13 Messwerte des Geysirs 'Old faithful' zur Pause zwischen den Ausbrüchen und der Dauer der Ausbrüche [1]. Die Verteilung der Daten legt einen Zusammenhang zwischen Dauer und Pause nahe. 84

14 'Messpunkte' zum Beispiel der Hersteller von Kolbenrückholfedern. 85

15 Vier-Quadranten-Schema am Beispiel der Hersteller von Kolbenrückholfedern. Man erkennt die Häufung der Daten in den Quadranten II und IV. 86

16 'Messpunkte' zum Beispiel der Hersteller von Kolbenrückholfedern mit eingezeichneter Ausgleichsgerade. 90

17 Streudiagramme für unterschiedliche Werte von R^2. Je besser die Datenpunkte auf einer Linie liegen, desto höher ist das Bestimmtheitsmaß. Streuen die Datenpunkte ohne Zusammenhang im Raum, liegt R^2 nahe Null. 95

Literatur

[1] http://en.wikipedia.org/wiki/File:Oldfaithful3.png, public domain

[2] http://commons.wikimedia.org/wiki/File:Gini_Coefficient_World_CIA_Report_2009.png, public domain

[3] The CIA world factbook, 2009: https://www.cia.gov/library/publications/the-world-factbook/fields/2172.html

A Lösungen zu den Übungsaufgaben im Skript

A.1 Lageparameter

- *Im ersten Jahr steigt der Gewinn um 35%, also* $q_1 = 0,35$*, im zweiten Jahr sinkt er aber um 35%,* $q_2 = -0,35$*. Der korrekte Lageparameter ist das geometrische Mittel*

$$\bar{q} = \sqrt{(1+q_1)\cdot(1+q_2)} - 1$$

$$= \sqrt{1,35 \cdot 0,65} - 1 = -0,063.$$

Im Mittel ist der Gewinn des Herrn B. also um 6,3% geschrumpft.

- *Manfred M. wird seinen Standort sinnvollerweise beim arithmetischen Mittel der Abstände (in m) seiner insgesamt* 15 *Stammkunden vom Anfang der Ostfriesenstraße wählen, also bei*

$$\frac{1}{15}\left(3\cdot 0 + 4\cdot 10 + 1\cdot 20 + 2\cdot 30 + 3\cdot 35 + 2\cdot 50\right) = 21\frac{2}{3}.$$

A.2 Streuungsmaße

- *Landtagswahlen: die Aussage des Herrn Osterwelle kann durch Vergleich der Variationskoeffizienten überprüft werden. Benötigt werden die arithmetischen Mittel (in %)*

$$\bar{x}_A = \frac{1}{7}\left(5,6 + 6,3 + 6,6 + 6,9 + 7,1 + 7,6 + 6,1\right) = 6,6$$

$$\bar{x}_B = \frac{1}{7}\left(40,4 + 41,9 + 47,9 + 40,4 + 48,9 + 41,4 + 42,9\right) = 43,4$$

sowie die Varianzen und die daraus berechneten Standardabweichungen

$$s_A^2 = \frac{1}{7}\left(5,6^2 + 6,3^2 + 6,6^2 + 6,9^2 + 7,1^2 + 7,6^2 + 6,1^2\right) - 6,6^2$$

$$= 0,383$$

$$\Rightarrow s_A = \sqrt{s_A^2} = 0,619$$

$$s_B^2 = \frac{1}{7}\left(40,4^2 + 41,9^2 + 47,9^2 + 40,4^2 + 48,9^2 + 41,4^2 + 42,9^2\right) - 43,4^2$$

$$= 10,714$$

$$\Rightarrow s_B = \sqrt{s_B^2} = 3,273.$$

Die Variationskoeffizienten $v_x = \frac{s_x}{\bar{x}}$ für die beiden Stichproben sind

$$v_A = \frac{s_A}{\bar{x}_A} = \frac{0,619}{6,6} = 0,094$$

und

$$v_B = \frac{s_B}{\bar{x}_B} = \frac{3,273}{43,4} = 0,075.$$

Damit wird klar, dass Herr Osterwelle falsch liegt, die Verteilung der Stimmenanteile für seine Partei A ist deutlich breiter.

- *Bekannt sind der Mittelwert $\bar{x} = 2200$ EUR/Monat und die Standardabweichung $s_x = 800$ EUR. Der Lohn wird um 10% angehoben, zusätzlich erhält jeder Mitarbeiter eine Einmalzahlung von 960 EUR/Jahr oder 80 EUR/Monat. Das Gehalt wird über eine lineare Transformation*

$$Y = 80 + 1,1 \cdot x$$

berechnet. Die neuen Werte (in EUR bzw. EUR^2) sind also

$$\bar{y} = 80 + 1,1 \cdot \bar{x} = 80 + 1,1 \cdot 2200 = 2500,$$

$$s_y^2 = (1,1)^2 \cdot s_x^2 = (1,1 \cdot 800)^2 = 774400$$

und

$$s_y = 1,1 \cdot s_x = 1,1 \cdot 800 = 880.$$

A.3 Konzentration

- *Die beiden Märkte M_1 und M_2 in tabellarischer Form dargestellt:*

M_1	h_i	f_i	F_i	q_i	Q_i
	9	*0,9*	*0,9*	*0,5*	*0,5*
	1	*0,1*	*1*	*0,5*	*1*
M_2	h_i	f_i	F_i	q_i	Q_i
	5	*0,5*	*0,5*	*0,1*	*0,1*
	5	*0,5*	*1*	*0,9*	*1*

Der Gini-Koeffizient G berechnet sich na ch (54) wie folgt:

$$M_1 : G_1 = 0,9 \cdot (0,5 + 0) + 0,1 \cdot (1 + 0,5) = 0,4$$

$$M_1 : G_2 = 0,5 \cdot (0,1 + 0) + 0,5 \cdot (1 + 0,1) = 0,4$$

Die Werte der Koeffizienten sind exakt gleich, die Ungleichverteilung bzw. die relative Konzentration ist in beiden Fällen dieselbe.

- *Der Herfindahl-Index berechnet sich nach (58) aus den Anteilen der einzelnen Merkmalsträger am Markt:*

$$q_i = \frac{x_i}{\sum_{i=1}^{n} x_i},$$

Der Anteil der einzelnen Unternehmen beträgt für 999 Unternehmen jeweils $q_i = 1\%/999$, in einem Fall 99%, also ist der Herfindahl-Index

$$H = 999 \cdot \left(\frac{0,01}{999}\right)^2 + (0,99)^2 = 0,98$$