Kann Statistik Spaß machen?

Eine kurze Einführung in die Beschreibende Statistik

Skript zur Statistik-Vorlesung

Fachhochschule Heilbronn, Standort Künzelsau

Wintersemester 2002/2003

Dr. Manuel Frondel[1]

Zentrum für Europäische Wirtschaftsforschung, Mannheim

[1]Der Autor dieses Skriptes ist Stefanie Göbel, Filmon Zerai und Dominik Zahrnt für Ihre Anregungen, Korrekturen und aufmunternden Worte, den eingeschlagenen Weg weiter zu gehen, zu besonderem Dank verpflichtet.

Inhaltsverzeichnis

1 Grundbegriffe

- **Statistische Einheit** oder **Merkmalsträger**: Dieser Begriff beinhaltet Personen oder Gegenstände, aber auch Ereignisse wie Geburten oder Sterbefälle.
- Die bei einer statistischen Einheit interessierenden Eigenschaften, z. B. Haarfarbe oder Augenfarbe bei Personen, werden **Merkmale** genannt.
- Die verschiedenen Alternativen für eine bei einer statistischen Einheit interessierenden Eigenschaft werden **Merkmalsausprägungen** genannt. Beim Merkmal Haarfarbe beispielsweise gibt es die klassischen Ausprägungen wie blond, rot- bzw. schwarzhaarig, welche heutzutage durch viele andere Alternativen, beispielsweise durch die Farbe lila, ergänzt werden.
- Die **statistische Masse** oder **Grundgesamtheit** ist die Menge aller relevanten statistischen Einheiten mit übereinstimmenden *sachlichen, räumlichen* und *zeitlichen Identifikationskriterien* (siehe dazu nachfolgendes Beispiel).
- **Bestandsmasse**: Die statistischen Einheiten einer Bestandsmasse besitzen *eine von null verschiedene Lebensdauer*. Beispielsweise stellt die Masse der Einwohner der Stadt Heidelberg eine Bestandsmasse dar. Ebenso ist die Masse der Touristen, die Heidelberg besuchen, eine Bestandsmasse. Wesentliches Charakteristikum einer Bestandsmasse ist: Die Erfassung der Zahl der zu dieser Masse gehörigen statistischen Einheiten erfolgt zu einem festgelegten *Zeitpunkt*, nicht jedoch über einen längeren Zeitraum hinweg.
- **Bewegungsmasse** oder **Ereignismasse**: Die statistischen Einheiten einer Ereignismasse treten nur punktuell auf und haben *keine von null verschiedene Lebensdauer*. Zum Beispiel bilden die Massen an Geburten, aber auch an Todesfällen innerhalb eines Jahres sowie die Zuzüge nach wie auch die Wegzüge aus Heidelberg Ereignis- bzw. Bewegungsmassen. Wesentliches Charakteristikum einer solchen statistischen Masse ist: Da die statistischen Einheiten keine Lebensdauer haben, erfolgt ihre Erfassung über einen längeren *Zeitraum* hinweg, *nicht* jedoch zu einem *Zeitpunkt*.

Beispiel: Bundestagswahl.
Die relevante statistische Masse bzw. Grundgesamtheit ist die Masse aller bei der Bundestagswahl am 22. 9. 2002 (zeitliches Identifikationskriterium) Wahlberechtigten (sachliches Identifikationskriterium), welche Staatsbürgerinnen und -bürger der Bundesrepublik Deutschland sind (räumliche Identifikationskriterium). Die Masse der Wahlberechtigten – die statistischen Einheiten – bildet eine Bestandsmasse.

1.1 Klassifizierung von Merkmalen

- **Nominale Merkmale**: Die Merkmalsausprägungen eines solchen Merkmals weisen *keine natürliche Rangfolge* auf. Zwei Merkmalsausprägungen können deshalb nur danach beurteilt werden, ob sie entweder gleich oder verschieden sind.

Beispiele: a) Familienstand mit den Ausprägungen ledig, verheiratet, geschieden, verwitwet, b) Geschlecht mit den Ausprägungen männlich und weiblich, c) Staatsangehörigkeit.

- **Ordinale Merkmale**: Die Merkmalsausprägungen eines solchen Merkmals weisen eine *natürliche Rangfolge* auf.

  Beispiele: a) Klausurnoten mit den Ausprägungen sehr gut, gut, befriedigend, ausreichend, mangelhaft, b) Hotelgüteklassen, c) Ratings der Qualität der Statistik-Vorlesung mit den Ausprägungen unter aller Sau, miserabel, erträglich, moderat, motivierend, inspirierend, mitreissend.

- **Kardinale** bzw. **metrische** oder **quantitative Merkmale**: Die Merkmalsausprägungen eines kardinalen Merkmals können in *reellen Zahlen* angegeben werden und weisen daher alle Ordnungseigenschaften reeller Zahlen auf.

  Beispiele: a) Körpergewicht gemessen in kg, b) Körpergröße gemessen in m.

Die Merkmale sind je nach Kategorie von unterschiedlicher Qualität, beispielsweise hinsichtlich der Frage, ob eine natürliche Rangfolge bei den Merkmalsausprägungen existiert oder nicht. Insbesondere können Abstände auch nur zwischen den Merkmalswerten kardinaler Merkmale, wie z. B. Körpergewicht, sinnvoll gemessen werden: Es hat Gewicht, wenn die Waage nach den Festtagen der Weihnachtszeit 3 Kilo mehr als vorher anzeigt.

Abstände zwischen den Merkmalsausprägungen ordinaler oder nominaler Merkmale können indes nicht ermittelt werden: Der Unterschied zwischen einem Zwei-Sterne Hotel und einem Ein-Sterne-Hotel ist nicht exakt quantifizierbar, sondern beruht auf mehr oder weniger genauen Einschätzungen: Es kann nur gesagt werden, dass das Zwei-Sterne Hotel besser ist, aber nicht um wieviel. Auch eine Diskussion über die Unterschiede zwischen beispielsweise den Haarfarben “schwarz” und “blond” ist unter statistischen Gesichtspunkten sinnlos.

Die unterschiedliche Qualität von Merkmalen dieser drei Kategorien hat zur Konsequenz, dass bestimmte Mittelwerte nicht für Merkmale aller drei Kategorien berechnet werden können. Durchschnittswerte wie das bekannte arithmetische Mittel können zum Beispiel nur bei kardinalen, nicht aber bei ordinalen oder nominalen Merkmalen ausgerechnet werden: Es macht beispielsweise keinen Sinn eine durchschnittliche Haarfarbe oder ein durchschnittliches Geschlecht bei einer Gruppe von Personen angeben zu wollen.

Kardinale Merkmale lassen sich weiter unterteilen in *diskrete* oder *stetige* Merkmale:

Diskrete Merkmale: Merkmale, deren Zahl an Merkmalsausprägungen entweder endlich (begrenzt) oder abzählbar unendlich ist. Dabei bedeutet abzählbar unendlich, dass die Merkmalsausprägungen mit Hilfe der natürlichen Zahlen durchnumeriert werden können.

Beispiele: a) Semesterzahl, b) Zahl der in einem Haushalt lebende Personen, c) Einwohnerzahl von Heidelberg, d) Zahl der Menschen auf der Erde.

Stetige Merkmale: Merkmale, deren Zahl an Merkmalsausprägungen überabzählbar unendlich ist.

Beispiele: a) Körpergewicht, b) Körpergröße, c) Alter.

Kardinale Merkmale können noch weiter unterteilt werden in *intervall-, verhältnis-* und *absolutskalierte* Merkmale. In einer knapp gehaltenden Einführung in die Beschreibende Statistik muss diese zusätzliche Einteilung nicht notwendigerweise diskutiert werden.

Beispiel: Erhebung des Energieverbrauchs privater Haushalte.
Das Bundesministerium für Wirtschaft (BMWi) gibt im Jahre 2002 eine statistische Erhebung in Auftrag, in der der Energieverbrauch der Sektoren "Private Haushalte" und "Gewerbe, Handel und Dienstleistungen" der BRD möglichst genau ermittelt werden soll. Dr. F., Wissenschaftlicher Mitarbeiter des Zentrums für Europäische Wirtschaftsforschung (ZEW), Mannheim, bewirbt sich beim BMWi um die Durchführung der Erhebung. Im Sektor "Private Haushalte" sind u. a. die folgenden Merkmale zu erheben:

Merkmal A: Zahl der Haushaltsmitglieder

Merkmal B: Der Haushalt wohnt in der eigenen Wohnung bzw. im eigenen Haus (ja/nein)

Merkmal C: Soziale Schicht des Haushaltsvorstandes (Ausprägungen: Ober-, Mittel- oder Unterschicht)

Merkmal D: Alter der Wohnung bzw. des Hauses (Ausprägungen: Neubau, Vor- bzw. Nachkriegsbau)

Merkmal E: Beruf des Haushaltsvorstandes

Merkmal F: Größe der Wohnung bzw. des Hauses in m^2

Merkmal G: Zentralheizung (ja/nein).

Merkmal H: Bildung des Haushaltsvorstandes.

Während die Merkmale B, E, F nominale Merkmale darstellen, sind die Merkmale C, D und H ordinalen Charakters. A sowie F sind kardinale Merkmale.

Natürlich wäre es in diesem Beispiel sehr aufwendig und kostenintensiv, eine *Totalerhebung* durchzuführen, das heißt alle Haushalte der BRD zu befragen. Totalerhebungen werden abgesehen von Bundes- und Landtagswahlen usw. nur noch selten durchgeführt – die Volkszählung vom 25. 5. 1987 war vielleicht eine der letzten Ausnahmen. Allerdings kann stattdessen eine *Teilerhebung* in Form einer Stichprobe genügen, wenn deren *Repräsentativät* gesichert und diese von einem entsprechenden Umfang ist.

Wie die aufgeführten Beispiele zeigen, sind die Merkmalsausprägungen kardinaler Merkmale reelle Zahlenwerte. Wir wollen daher im folgenden bei kardinalen Merkmalen aus-

schließlich von *Merkmalswerten* anstatt von Merkmalsausprägungen sprechen und für nominale und ordinale Merkmale werden wir ausschließlich den Begriff *Merkmalsausprägungen* verwenden.

Merkmalsausprägungen sollten eigentlich nicht in Form reeller Zahlen angegeben werden, denn dies könnte zur unzulässigen Bildung von Mittelwerten wie dem arithmetischen Mittel verführen. Im besonderen Falle von Klausur-Noten werden indes Merkmalsausprägungen wie "sehr gut" meist durch die Zahl 1, "gut" durch die Zahl 2 usw. abgekürzt. Die Berechnung eines arithmetischen Mittels ist allerdings rechtfertigbar durch die Existenz eines der Notenbildung zugrunde liegenden kardinalen Merkmals wie "bei der Klausur erreichte Punktzahl".

Zur Vereinfachung der Datenspeicherung ist es indes üblich, die Merkmalsausprägungen nominaler bzw. ordinaler Merkmale, wie beispielsweise Geschlecht, mit Hilfe reeller Zahlen zu kodieren. Zum Beispiel könnte die Ausprägung "männlich" des Merkmals Geschlecht durch die Zahl 0 und die Ausprägung "weiblich" durch die Zahl 1 kodiert werden oder auch durch 1 respektive 2. Postleitzahlen stellen praktische Beispiele für Kodierungen von Ortsnamen dar. So steht 69117 für die Heidelberger Altstadt. Trotz dieser Kodierung steckt hinter Postleitzahlen lediglich das Merkmal „Wohnort“, welches nominal ist: Es macht bei diesem Merkmal beispielsweise keinen Sinn das arithmetische Mittel von Altstadt (PLZ 69117) und Neuenheim (PLZ 69119) zu bilden, wenngleich rein arithmetisch dabei Ziegelhausen (PLZ 69118) herauskäme.

1.2 Zusammenfassung

Die folgende Tabelle faßt die drei hier diskutierten Kategorien von Merkmalen zusammen.

Merkmals-Kategorie	Charakteristika der Merkmalsausprägungen	Beispiele	zulässige Mittelwerte
Nominales Merkmal	keine Rangfolge	Wohnort	Modus
Ordinales Merkmal	natürliche Rangfolge	Noten	Modus und Median
Kardinales Merkmal	reelle Zahlenwerte	Temperatur	alle Mittelwerte

Entscheidend ist, die wesentlichen Qualitätsunterschiede der Merkmale dieser drei Kategorien und deren Konsequenzen zu kennen: Bei kardinalen Merkmalen können Abstände quantifiziert werden, was bei ordinalen und nominalen Merkmalen nicht der Fall ist, während bei ordinalen Merkmalen immerhin noch eine natürliche Rangfolge unter den Merkmalswerten existiert. Diese Unterschiede haben Konsequenzen u. a. hinsichtlich der Möglichkeit der Berechnung von Mittelwerten, aber auch anderer statistischer Maßzahlen. Mittelwerte sowie andere statistische Maßzahlen und deren Anwendbarkeit bei den drei Merkmalskategorien werden ausführlich in Kapitel 3 behandelt.

1.3 Übungsaufgaben

Übungsaufgabe: Pizza und Pasta.

Alfredo ist Chef zweier italienischer Steh-Restaurants R_1 und R_2 in Heidelberg, in denen man hervorragende Pasta, aber auch andere Gerichte sowohl mittags wie auch abends zu sich nehmen kann. Im Monat August wurden von 250 Gästen insgesamt 5.000 Gerichte bestellt.

	R_1		R_2		
	mittags	abends	mittags	abends	insgesamt
Pasta	400	600	600	800	2.400
Sonstige	700	1100	400	400	2.600
Summe	1.100	1.700	1.000	1.200	5.000

a) Welche Merkmale werden in diesem Beispiel erwähnt und wie lauten ihre Merkmalsausprägungen?

b) Den folgenden Sachverhalten sollen die Begriffe statistische Masse bzw. deren Umfang, statistische Einheit, Merkmal, Merkmalsausprägung und Identifkationskriterium korrekt zugeordnet werden.

- Die in obiger Statistik mitgezählten „Spaghetti Arrabiata“, die Prof. Dr. CMS (Ph. D.) am 25. August bei Alfredo im Restaurant R_1 zu Mittag genüßlich verspeist hat.
- Die Angabe „Spaghetti Arrabiata“.
- Die Angabe „Alfredo“.
- Die Zahl 5.000.
- Das Restaurant R_1.
- Der Monat August.

Übungsaufgabe: Piefkes in Austria.

Aus der Statistik des Landes Tirol ergibt sich, dass sich jedes Jahr viele Piefkes – Fachbegriff für den typischen deutschen Austria-Urlauber – beiderlei Geschlechts und jeglicher Bundesländer in den österreichischen Alpen verirren oder in Bergnot geraten und von Bergwachten gerettet werden müssen. Man trage die Begriffe statistische Masse, statistische Einheit, Merkmal, Merkmalsausprägung, Merkmalswert korrekt in die folgende Tabelle ein:

Bayer	Piefkes	Geschlecht	verirrter Piefke	weiblich	Zahl verirrter Piefkes

2 Darstellung statistischer Information

Die Darstellung statistischen Datenmaterials bezüglich eines einzigen Merkmals sowie die statistischen Konzepte zur Analyse dieser Daten werden im folgenden vor allem an zwei Beispielen erläutert. Das erste Beispiel verwendet *klassierte Daten*, wohingegen im zweiten Beispiel Individualdaten, also *Einzelwerte* aufgeführt werden. Die beiden Beispiele illustrieren die beiden grundlegenden Arten an Dateninformation, die in der Statistik üblicherweise aus einer Stichprobe gezogen wird: Information in Form klassierter oder unklassierter Daten (Einzelwerten).

2.1 Beispiel: Verteilung der Körpergröße

Die folgende Tabelle gibt Auskunft über die Verteilung der Körpergröße von 20 Studentinnen und Studenten eines Ausbildungskurses des Fachbereichs Spedition an der Berufsakademie Mannheim. Die individuellen Werte des Merkmals Körpergröße werden in $k = 5$ Größenklassen eingeteilt:

Größenklassen	absolute Häufigkeit	relative Häufigkeit	relative Häufigkeit in %
$[1,50; 1,65)$	2	2/20	10 %
$[1,65; 1,75)$	6	6/20	30 %
$[1,75; 1,85)$	7	7/20	35 %
$[1,85; 1,95)$	4	4/20	20 %
$[1,95; 2,10]$	1	1/20	5 %
Summe	20	1	100 %

Im Gegensatz zum folgenden Beispiel 2.2 ist zur Erstellung dieser Tabelle bereits Information in Form der einzelnen Körpergrößen vernichtet worden. Während bei einem Stichprobenumfang von $n = 20$ eine solche Informationsreduktion nicht notwendig gewesen wäre, ist dies bei großen Stichprobenumfängen absolut erforderlich: Man stelle sich nur einmal die Einkommensverteilung der Millionen von deutschen Haushalten vor, wenn diese nicht klassiert angegeben werden würde!

2.2 Beispiel: Altersverteilung

Die nachfolgende Tabelle informiert über das jeweilige Alter, das Geschlecht und die Note in der Statistik-Klausur der einzelnen Personen des in Beispiel 2.1 erwähnten BA-Kurses:

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Alter	19	21	26	20	22	19	20	19	23	21	22	19	21	20	23	22	21	21	20	20
Geschlecht	w	w	m	w	m	w	m	w	m	w	m	w	m	w	m	w	m	m	w	w
Note	1	1	2	1	3	2	4	3	5	2	3	1	3	2	4	4	3	2	1	1

Kürzer könnte die relevante Information bezüglich des Merkmals "Alter" durch eine sogenannte *Urliste*

$$(19, 21, 26, 20, 22, 19, 20, 19, 23, 21, 22, 19, 21, 20, 23, 22, 21, 21, 20, 20)$$

dargestellt werden. Dabei wird nicht auf Information verzichtet, solange die erste Komponente des Vektors stellvertretend für die erste Person steht, die zweite Komponente für die zweite Person usw. Stattdessen könnte ein Vektor von $n = 20$ Zahlen benutzt werden, der die Lebensalter in bereits geordneter Form wiedergibt:

$$(19, 19, 19, 19, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 22, 23, 23, 26).$$

Daraus geht aber nicht mehr hervor, welche Person wie alt ist. Dieselbe Information bietet eine Tabelle, welche die *absoluten Häufigkeiten* h_i bzw. die *relativen Häufigkeiten* $f_i := h_i/n$ für den i-ten in der Stichprobe auftretenden Merkmalswert enthält:

Alter	19	20	21	22	23	26	Summe
h_i	4	5	5	3	2	1	20
f_i	4/20	5/20	5/20	3/20	2/20	1/20	1
$f_i \cdot$ 100 %	20 %	25 %	25 %	15 %	10 %	5 %	100 %

Bei Angaben dieser Art spricht man von *absoluter* bzw. *relativer Häufigkeitsverteilung.*

2.3 Eindimensionale Häufigkeitsverteilung

Die in einer Stichprobe vorhandene Information bezüglich eines kardinalen Mermals X kann, falls der Stichprobenumfang n nicht allzu groß ist, allgemein in Form einer *Urliste* oder – meist ohne wesentlichen Verlust an Information – eines Vektors $(x_1, x_2, ..., x_n)$ mit bereits geordneten Werten angegeben werden. Häufig werden außerdem einzelne Merkmalswerte mehrfach in der Stichprobe beobachtet, so dass tatsächlich nur $m < n$ Merkmalswerte verschieden sind. Diese werden im neuen Vektor $(x_1, x_2, ..., x_m)$ zusammengefasst, der die verschiedenen tatsächlich auftretenden Merkmalswerte in *geordneter* Reihenfolge enthält. Man beachte: Beide Vektoren sind i. A. voneinander verschieden! Im Folgenden wird immer davon ausgegangen, dass die Merkmalswerte in bereits geordneter Form in den beiden Vektoren zusammengefasst sind.

Zur Charakterisierung einer Stichprobe genügt meist – siehe Beispiel 2.2 – die Angabe des Vektors $(x_1, x_2, ..., x_m)$ der auftretenden Merkmalswerte sowie der absoluten bzw. relativen Häufigkeitsverteilung $(h_1, h_2, h_3, ..., h_m)$ bzw. $(f_1, f_2, f_3, ..., f_m)$:

X	x_1	x_2	x_3	...	x_{m-1}	x_m	Summe
$h(X = x_i) = h_i$	h_1	h_2	h_3	...	h_{m-1}	h_m	n
$f(X = x_i) = f_i$	f_1	f_2	f_3	...	f_{m-1}	f_m	1

Anstatt auf solche Tabellen wird zur Veranschaulichung der in einer Stichprobe vorhandenen Information sehr oft auf grafische Illustrationen zurückgegriffen. Bei unklassierten Daten geschieht dies in Form von *Stab-*, *Balken-*, *Säulen-* oder *Kreis*diagrammen usw. , bei klassierten Daten in Form von Histogrammen.

Beispiel: Stab- und Balkendiagramm.

Die absoluten Häufigkeiten, mit denen die unterschiedlichen Lebensalter im BA-Kurs des Beispiels 2.2 auftreten, sind in Figur 1 durch ein Stabdiagramm und in Figur 2 durch ein Balkendiagramm dargestellt.

Fig. 1: Stabdiagramm einer Verteilung absoluter Häufigkeiten.

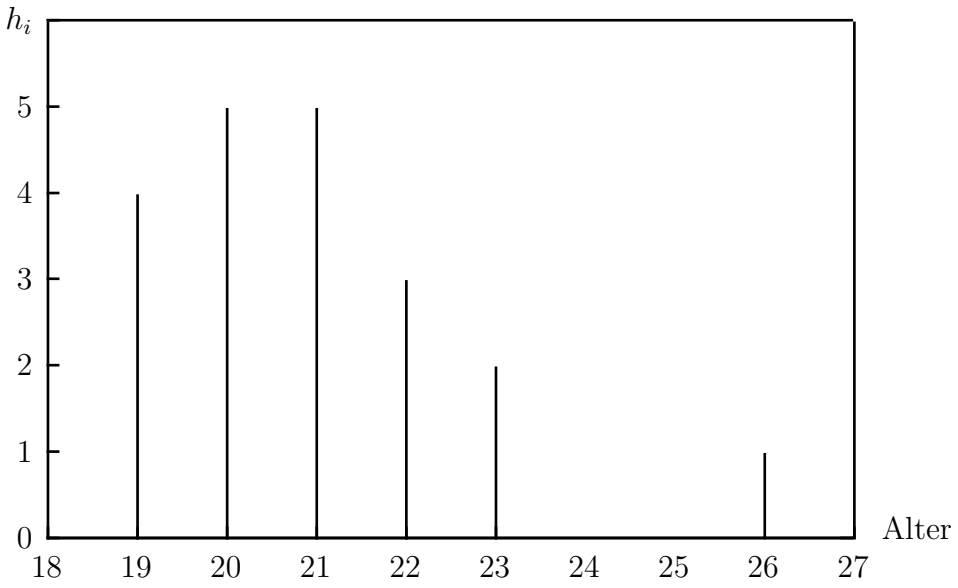
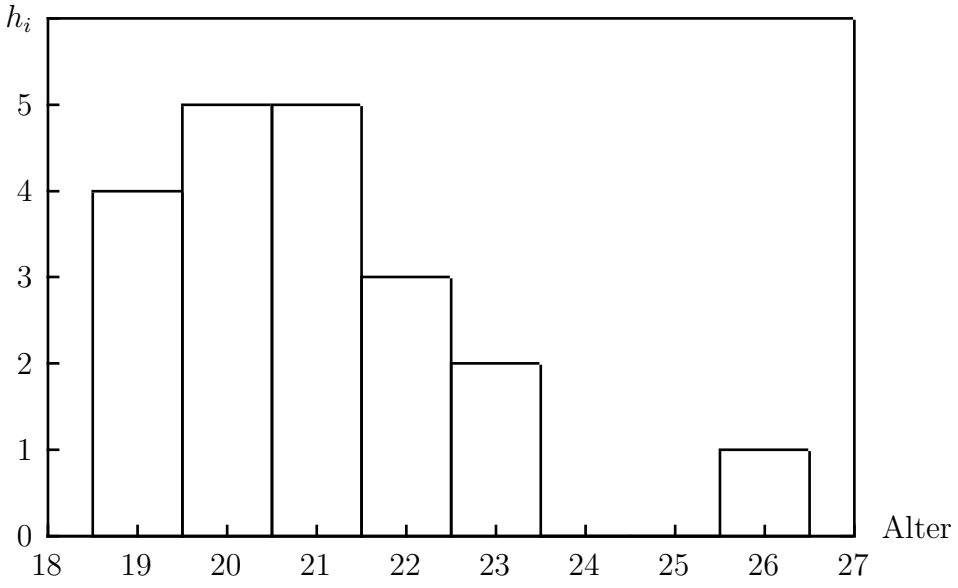


Fig. 2: Balkendiagramm einer Verteilung absoluter Häufigkeiten.



Beispiel: Histogramm zur Illustration klassierter kardinaler Daten.
Liegt die Stichprobeninformation in Form klassierter Daten vor, müssen Histogramme zur Illustration verwendet werden. Figur 3 zeigt ein *Histogramm*, das die relative Häufigkeitsverteilung für das Merkmal Körpergröße aus Beispiel 2.1 wiedergibt. Die *Flächen* der einzelnen Rechtecke entsprechen dabei *per Konstruktion* genau den *relativen Häufigkeiten* für die jeweilige Klasse. Dazu wird auf der Ordinate (der y-Achse) die sogenannte *Dichte* f_i^* abgetragen. Sie ist definiert ist als die relative Häufigkeit f_i, welche die i-te Klasse aufweist, dividiert durch die dazugehörige Klassenbreite Δx_i:

$$f_i^* := \frac{f_i}{\Delta x_i} \ . \tag{1}$$

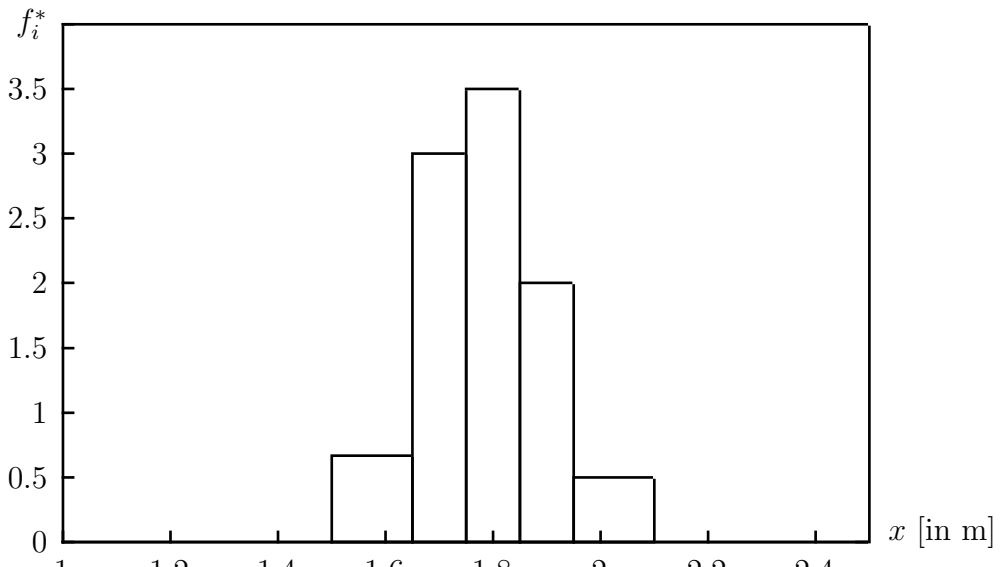
Die Fläche des i-ten Rechtecks, welche sich aus (Klassen-)Breite Δx_i mal Höhe f_i^* ergibt, ist somit immer identisch mit der relativen Häufigkeit f_i für diese Klasse:

$$\text{Fläche} = \text{Breite} \ \times \ \text{Höhe} = \Delta x_i \cdot f_i^* = \Delta x_i \cdot \frac{f_i}{\Delta x_i} = f_i \ .$$

Wären bei klassierten Daten smtliche Klassenbreiten immer allesamt gleich groß, könnte auf die Berechnung von Dichten verzichtet werden. Dies ist bei klassierten Daten in den allermeisten Fällen nicht so – insbesondere sind die Breiten der untersten und der obersten Klassen meist sehr verschieden von denen der restlichen Klassen. Zusammen mit den nach Definition (1) berechneten Dichten sind die absoluten und relativen Häufigkeiten für Beispiel 2.1 in folgender Tabelle dargestellt:

Größenklassen	abs. Häufigkeit h_i	rel. Häufigkeit f_i	Breite Δx_i	Dichte f_i^*
$[1,50; 1,65)$	2	2/20	0,15	$0,\bar{6}$
$[1,65; 1,75)$	6	6/20	0,10	3,0
$[1,75; 1,85)$	7	7/20	0,10	3,5
$[1,85; 1,95)$	4	4/20	0,10	2,0
$[1,95; 2,10]$	1	1/20	0,15	0,5

Fig. 3: Histogramm für eine Verteilung der Körpergröße.



Beispiel: Häufigkeitsverteilungen für ein nominales Merkmal.
Absolute und relative Häufigkeitsverteilungen ordinaler und nominaler Merkmale können ebenfalls durch Stab-, Balken- oder Kreis-Diagramme dargestellt werden. Bei lediglich zwei Merkmalsausprägungen wie beim Merkmal „Geschlecht“ des Beispiels 2.2 genügt aber auch eine einfache Tabelle:

Geschlecht	mnnlich	weiblich	Summe
h_i	9	11	20
f_i	9/20	11/20	1
$f_i \cdot$ 100 %	45 %	55 %	100 %

2.4 Kumulierte Häufigkeitsverteilungen

In einer Stichprobe vom Umfang n eines ordinalen Merkmals X mit $m \leq n$ verschiedenen, geordneten Merkmalsausprägungen, treten diese mit den relativen Häufigkeiten $(f_1, f_2, ..., f_m)$ auf. Unter der *kumulierten absoluten* bzw. *relativen Häufigkeit* H_i bzw. F_i wird die *Summe* der absoluten bzw. der relativen Häufigkeiten für all diejenigen Merkmalsausprägungen verstanden, welche das Niveau der i-ten Merkmalsausprägung höchstenfalls erreichen.

Beispiel: Kumulierte Häufigkeitsverteilungen für ein ordinales Merkmal.
Die absoluten bzw. relativen Häufigkeiten h_i bzw. f_i und kumulierten absoluten bzw. relativen Häufigkeiten H_i bzw. F_i der Noten für die Statistik-Klausur des BA-Kurses aus Beispiel 2.2 lauten:

Note	h_i	H_i	f_i	F_i
sehr gut	6	6	6/20	6/20
gut	5	11	5/20	11/20
befriedigend	5	16	5/20	16/20
ausreichend	3	19	3/20	19/20
mangelhaft	1	20	1/20	20/20
Summe	20	–	20/20	–

Die absolute Häufigkeit h_2 mit der beispielsweise die Note “gut” vorkommt beträgt 5. Das ist bei 20 Personen ein Anteil (eine relative Häufigkeit) von 5/20 bzw. 25 %. Die kumulierte absolute Häufigkeit $H_2 = 11$ besagt, dass 11 Personen die Note “gut” oder besser, also “sehr gut”, haben. Die kumulierte relative Häufigkeit $F_2 = 11/20 = 55$ % besagt, dass mehr als die Hälfte aller Personen gute und sehr gute Noten erzielt haben.

In einer Stichprobe vom Umfang 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. Unter der *kumulierten absoluten* bzw. *relativen Häufigkeit* H_i bzw. F_i wird die *Summe* der absoluten bzw. der relativen Häufigkeiten für all diejenigen Merkmalswerte verstanden, welche kleiner oder gleich dem Wert x_i sind:

$$H_i := \sum_{x_j \leq x_i} h_j \quad \text{bzw.} \quad F_i := \sum_{x_j \leq x_i} f_j. \tag{2}$$

Die damit gebildeten Vektoren $(H_1, H_2, ..., H_m)$ bzw. $(F_1, F_2, ..., F_m)$ geben die *kumulierte absolute* bzw. *relative Häufigkeitsverteilung* für den Vektor $(x_1, x_2, ..., x_m)$ der Merkmalswerte an.

Beispiel: Kumulierte Häufigkeitsverteilungen für ein kardinales Merkmal.
Die absoluten bzw. relativen und kumulierten absoluten bzw. relativen Häufigkeitsverteilungen der Lebensalter der Personen des BA-Kurses aus Beispiel 2.2 lauten:

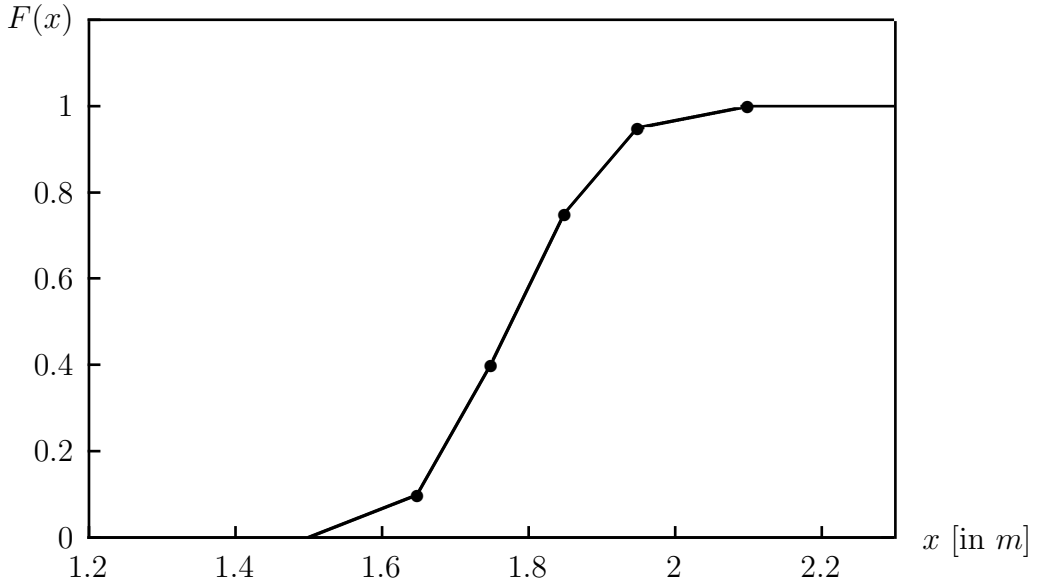
Alter	h_i	H_i	f_i	F_i
19	4	4	4/20	4/20
20	5	9	5/20	9/20
21	5	14	5/20	14/20
22	3	17	3/20	17/20
23	2	19	1/20	19/20
26	1	20	1/20	20/20
Summe	20	–	20/20	–

Liegt die Stichprobeninformation für ein kardinales Merkmal X in Form von *klassierten* Daten vor mit k Klassen vor, werden die kumulierten relativen Häufigkeiten F_i gebildet aus der *Summe* der relativen Häufigkeiten für die Klassen 1 bis i. Die kumulierte relative Häufigkeit F_i wird der oberen Grenze x_i^o der i-ten Klasse zugeordnet. Bei k Klassen muss die Summe der relativen Häufigkeiten für die Klassen 1 bis k ergo Eins ergeben: $F_k = 1$. Die Punkte $(x_1^o, F_1); (x_2^o, F_2); ...; (x_k^o, F_k)$ 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 wird vervollständigt durch eine Parallele zur Abszisse (zur x-Achse) in Höhe von 1, die beim letzten Eckpunkt $(x_k^o, F_k = 1)$ beginnt.

Beispiel: Verteilungspolygon bei klassierten kardinalen Daten.
Das Verteilungspolygon $F(x)$ für die klassierten Daten der Körpergrößen der Personen des BA-Kurses aus Beispiel 2.2 ist in Figur 4 skizziert. Die durch fette Punkte gekennzeichneten Eckpunkte (x_i^o, F_i) des Polygons können der folgenden Tabelle entnommen werden. Die Eckpunkte werden gebildet aus der jeweiligen oberen Klassengrenze x_i^o und der der jeweiligen Klasse zugeordneten kumulierten relativen Häufigkeit F_i. Die Funktion $F(x)$, das Verteilungspolygon, ergibt sich, indem die Eckpunkte durch Geraden verbunden werden.

Größenklassen $[x_i^o, x_i^u)$	absolute Häufigkeit h_i	relative Häufigkeit f_i	kumulierte relative Häufigkeit F_i	Eckpunkte (x_i^o, F_i)
$[1,50; 1,65)$	2	0,10	0,10	(1,65; 0,10)
$[1,65; 1,75)$	6	0,30	0,40	(1,75; 0,40)
$[1,75; 1,85)$	7	0,35	0,75	(1,85; 0,75)
$[1,85; 1,95)$	4	0,20	0,95	(1,95; 0,95)
$[1,95; 2,10)$	1	0,05	1,00	(2,10; 1,00)
Summe	20	1	–	–

Fig. 4: Verteilungspolygon $F(x)$ für eine Verteilung der Körpergröße.



2.5 Zusammenfassung

Stab-, Balken und Kreisdiagramme dienen der Illustration absoluter bzw. relativer Häufigkeitsverteilungen bei nominalen, ordinalen sowie kardinalen Merkmalen, bei denen die Stichprobeninformation als unklassierte Daten vorliegen. Bei stetigen kardinalen Merk-

malen, für welche die Stichprobeninformation in Form klassierter Daten vorliegt, werden in der Regel Histogramme verwendet, es sei denn die Klassenbreiten sind allesamt identisch.

2.6 Übungsaufgaben

Übungsaufgabe: Einkommensverteilungen im intertemporalen Vergleich.

Die Einkommens- und Verbrauchsstichprobe (EVS) wird alle 5 Jahre vom Statistischen Bundesamt erhoben, zuletzt 1998. Ca. 40.000 bis 50.000 Haushalte geben im Rahmen dieser Erhebung freiwillig und entgeltlos Auskunft über ihre monatlichen Einkünfte und Konsumgewohnheiten. Weil gerade die Bezieher niedriger und hohere Einkommen ungern über ihr Einkommen berichten, gibt die folgende Tabelle die Schichtung der Haushalte [in Prozent] nach ihrem monatlichen Haushaltnettoeinkommen in DM (Früheres Bundesgebiet) wohl nicht repräsentativ wieder. Dennoch lassen sich über die Zeit hinweg sinnvoll Vergleiche anstellen, wenn der Anteil der nicht vertretenen Haushalten niedrigen und hohen Einkommens über die Zeit relativ konstant geblieben ist.

von ... bis unter ...DM	1973	1978	1983	1988	1993
unter 2.500	72,9	50,4	40,5	36,8	22,7
2.500-5.000	24,7	41,7	42,3	41,2	39,5
5.000-10.000	2,3	7,3	16,0	20,1	31,7
10.000-15.000	0,2	0,4	0,9	1,6	4,9
15.000 und mehr	0,0	0,1	0,3	0,4	1,2

Man skizziere für jedes Jahr die zugehörigen Histogramme, um die Veränderung der Einkommenssituation in den alten Bundeländern über die Zeit zu veranschaulichen. Die Abschneidegrenze für Bezieher hohere Einkommen betrug 35.000 DM.

Übungsaufgabe: Histogramm einer Einkommensverteilung.

Bei einer Einkommensuntersuchung ergab sich für 100 Männer die folgende Verteilung des Nettojahreseinkommens in 1.000 DM (TDM). Wie sieht das Histogramm aus, das diese Einkommensverteilung widergibt? Warum könnte zur Illustration *kein* Balkendiagramm benutzt werden?

von ... bis unter ...TDM	absolute Häufigkeit h_i	relative Häufigkeit f_i
unter 10	5	0,05
10 - 20	15	0,15
20 - 25	20	0,20
25 - 30	25	0,25
30 - 40	20	0,20
40 - 60	10	0,10
60 - 85	5	0,05

3 Statistische Analyse eines Merkmals

Die statistische Analyse eines einzigen Merkmals wird häufig *univariate Analyse* genannt, während die gemeinsame Analyse zweier Merkmale *bivariate*, die von mehr als zwei Merkmalen *multivariate Analyse* heißt. Fundamentale Bedeutung bei der univariaten Analyse haben *statistische Maßzahlen*. Mit ihrer Hilfe wird die in einer Stichprobe enthaltene, häufig sehr umfangreiche Information bewußt reduziert und verdichtet. Durch wenige Maßzahlen wird das gesamte, ursprünglich vorhandene Datenmaterial charakterisiert. Dadurch geht viel Information verloren. Dies wird allerdings durch die Möglichkeit zur prägnanten Kurzbeschreibung einer Verteilung und der schnelleren Vergleichbarkeit mehrerer, sachlich ähnlicher Verteilungen in Kauf genommen.

Statistische Maßzahlen, auch Parameter genannt, stellen also charakteristische Werte zur einfachen Beschreibung der Gesamtheit aller Beobachtungswerte dar. Sie sollen außerdem spezifische Eigenschaften hervortreten lassen: Man unterscheidet bei statistischen Maßzahlen bespielsweise zwischen *Lagemaßen*, *Streuungsmaßen*, *Schiefemaßen* etc. Lagemaße, allen voran Durchschnitts- bzw. Mittelwerte, sollen für die Stichprobe repräsentative, *typische* bzw. *charakteristische* Werte eines betrachteten Merkmals für eine Stichprobe oder eine Grundgesamtheit wiedergeben. Streuungsmaße zeigen an, ob die Merkmalswerte dicht beim Mittelwert liegen oder mehr oder weniger stark davon abweichen. Schiefemaße informieren über Symmetrie oder Asymmetrie einer Häufigkeitsverteilung.

3.1 Arithmetisches Mittel

Das weitaus bekannteste Konzept von Mittelwerten ist das des *arithmetischen Mittels* $\bar{x}$. Es ist ausschließlich für *kardinale* Merkmale berechenbar.

Beispiel zur Berechnung des arithmetischen Mittels

Die Studenten und Studentinnen des BA-Kurses aus Beispiel 2.2 sind durchschnittlich

$$\bar{x} = (19+21+26+20+22+19+20+19+23+21+22+19+21+20+23+22+21+21+20+20)/20 = 20{,}95$$

Jahre alt. Hat man statt der Einzelwerte die absoluten bzw. relativen Häufigkeiten gegeben, gestaltet sich die Berechnung etwas kürzer:

$$\bar{x} = \frac{4 \cdot 19 + 5 \cdot 20 + 5 \cdot 21 + 3 \cdot 22 + 2 \cdot 23 + 1 \cdot 26}{20} = 20,95$$

bzw.

$$\bar{x} = \frac{4}{20} \cdot 19 + \frac{5}{20} \cdot 20 + \frac{5}{20} \cdot 21 + \frac{3}{20} \cdot 22 + \frac{2}{20} \cdot 23 + \frac{1}{20} \cdot 26 = 20,95 \ .$$

Definition des arithmetischen Mittels bei Einzelwerten

Stellt $(x_1, x_2, ..., x_n)$ den Vektor der beobachteten Merkmalsausprägungen eines **kardinalen** Merkmals X in einer Stichprobe vom Umfang n dar, so wird das arithmetische Mittel

$\bar{x}$ allgemein wie folgt berechnet:

$$\bar{x} := \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \ . \tag{3}$$

Eigenschaften des arithmetischen Mittels

- **Schwerpunkteigenschaft** des arithmetischen Mittels:

  Aus Formel (3) folgt unmittelbar, dass sich die Abweichungen der jeweiligen Einzelwerte x_i vom arithmetischen Mittel $\bar{x}$ in der Summe exakt aufheben:

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

  Kurzgefaßt lautet die Schwerpunktseigenschaft:

  $$\sum_{i=1}^{n}(\bar{x} - x_i) = 0. \tag{4}$$

- **Minimumseigenschaft** des arithmetische Mittels:

  Für eine gegebene Stichprobe $(x_1, x_2, ..., x_n)$ ist das arithmetische Mittel $\bar{x}$ die Lösung des folgenden Minimierungsproblems:

  $$\min_y \sum_{i=1}^{n}(x_i - y)^2, \tag{5}$$

  wobei bei gegebenem Vektor $(x_1, x_2, ..., x_n)$ die Summe $\sum(x_i - y)^2$ eine Funktion $f(y)$ allein der Variablen y ist. In Worten: Das arithmetische Mittel $\bar{x}$ ist genau derjenige Wert y, der die Summe der quadrierten Abweichungen der einzelnen Werte x_i von y minimiert.

  Das läßt sich mit Hilfe der Differentialrechnung zeigen: Das Minimum einer Funktion $f(y)$ wird berechnet, indem die 1. Ableitung gleich Null gesetzt wird:

  $$0 = \frac{d}{dy}f(y) = \frac{d}{dy}(\sum_{i=1}^{n}(x_i - y)^2) = \sum_{i=1}^{n}\frac{d}{dy}((x_i - y)^2) = \sum_{i=1}^{n}(-2)\cdot(x_i - y).$$

  Nach Division durch (-2) ergibt sich

  $$0 = \sum_{i=1}^{n}(x_i - y) \Longleftrightarrow 0 = \sum_{i=1}^{n} x_i - n\cdot y \Longleftrightarrow y = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \ .$$

  Die 2. Ableitung muss außerdem positiv sein,

  $$\frac{d^2}{(dy)^2}f(y) = \frac{d}{dy}(\sum_{i=1}^{n}(-2)\cdot(x_i - y)) = \sum_{i=1}^{n}\frac{d}{dy}((-2)\cdot(x_i - y)) = \sum_{i=1}^{n} 2 = 2n \ .$$

  In der Tat ist $2n$ immer positiv.

- Das arithmetische Mittel $\bar{x}$ ist *nicht robust* gegenüber Ausreißern bzw. extremen Werten, weil *jeder* einzelne Wert in die Berechnung des arithmetischen Mittels einfließt – siehe Formel (3). Beim Median hingegen, einem anderen, noch zu besprechenden Konzept zur Bildung von Mittelwerten, ist das nicht der Fall. Wegen dieser Eigenschaft der *Robustheit* des Medians wird dieser in manchen Fällen dem arithmetischen Mittel vorgezogen.

- **Transformationseigenschaft** des arithmetischen Mittels:

  Geht ein kardinales Merkmal Y durch eine allgemeine lineare Transformation

  $$Y = a + b \cdot X \tag{6}$$

  aus einem kardinalen Merkmal X hervor, dessen arithmetisches Mittel $\bar{x}$ bekannt ist, so ergibt sich das arithmetische Mittel $\bar{y}$ des Merkmals Y aus derselben linearen Transformation (6) des arithmetischen Mittels von X:

  $$\bar{y} = a + b \cdot \bar{x} \qquad \text{wobei } a, b \in \mathbb{R}. \tag{7}$$

  Das läßt sich zeigen: Jeder einzelne Wert x_i einer Stichprobe $(x_1, x_2, ..., x_n)$ ergibt durch die lineare Transformation (6) genau einen Merkmalswert des Merkmals Y: $y_i = a + b \cdot x_i$. Für die Stichprobe $(y_1, y_2, ..., y_n)$ ergibt sich dann das folgende arithmetische Mittel $\bar{y}$:

  $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + b \cdot x_i) = \frac{1}{n} \cdot n \cdot a + b \cdot \frac{1}{n}\sum_{i=1}^{n} x_i = a + b \cdot \bar{x}. \tag{8}$$

Beispiel: Lineare Transformation der Temperatur von Grad Celsius in Grad Fahrenheit.

Bezeichnet Y beispielsweise die Temperatur in Grad Fahrenheit und X die Temperatur in Grad Celsius, so lautet die Transformationsformel der Temperatur von Grad Celsius in Grad Fahrenheit:

$$Y = \frac{9}{5} \cdot X + 32.$$

Die Temperaturen $x_1 = 15, x_2 = 20, x_2 = 25$ Grad Celsius haben ein arithmetisches Mittel von $\bar{x} = 20$ Grad Celsius und lauten in Grad Fahrenheit gemessen: $y_1 = 59, y_2 = 68, y_2 = 77$. Mit Hilfe von Formel (7) ergibt sich das arithmetische Mittel $\bar{y}$ der Temperaturen in Grad Fahrenheit:

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

Dieser Wert hätte auch ohne die Transformationsformel (7) berechnet werden können:

$$\bar{y} = \frac{59 + 68 + 77}{3} = 68.$$

Alternative Formeln zur Berechnung des arithmetisches 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 wird das arithmetische Mittel wie folgt berechnet:

$$\bar{x} := \frac{h_1 x_1 + h_2 x_2 + h_3 x_3 + ... + h_m x_m}{n} = \frac{1}{n} \sum_{i=1}^{m} h_i x_i \tag{9}$$

bzw.

$$\bar{x} := \frac{h_1}{n} x_1 + \frac{h_2}{n} x_2 + \frac{h_3}{n} x_3 + ... + \frac{h_m}{n} x_m = \sum_{i=1}^{m} \frac{h_i}{n} \cdot x_i \ . \tag{10}$$

Sind statt der absoluten die relativen Häufigkeiten $(f_1, f_2, ..., f_m)$ bekannt, so folgt wegen $f_i = h_i/n$ für $i = 1, ..., m$ aus Berechnungsformel (10):

$$\bar{x} := f_1 x_1 + f_2 x_2 + f_3 x_3 + ... + f_m x_m = \sum_{i=1}^{m} f_i x_i \ . \tag{11}$$

Die Multiplikatoren f_i können als Gewichte aufgefasst werden, mit denen die Werte x_i multipliziert ("gewichtet") werden. Diese Gewichte sind im allgemeinen nicht identisch, sondern voneinander verschieden. Formel (11) macht plausibel, warum je nach Zusammenhang auch vom *gewogenen* arithmetischen Mittel gesprochen wird, obwohl das arithmetische Mittel, das für n vollkommen verschiedene Einzelwerte $(x_1, x_2, ..., x_n)$ gebildet wird, eigentlich ebenfalls ein "gewogenes Mittel" ist – allerdings mit völlig identischen Gewichten $1/n$:

$$\bar{x} := \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \cdot x_1 + \frac{1}{n} \cdot x_2 + ... + \frac{1}{n} \cdot x_n \ . \tag{12}$$

Die Berechnung des arithmetisches Mittels bei klassierten Daten

Während Formel (3) für das arithmetische Mittel bei klassiertem Datenmaterial wegen der Unkenntnis der Einzelwerte nicht angewandt werden kann, finden die beiden Formeln (9) und (11) auch bei klassierten Daten Anwendung. Dabei werden an Stelle der einzelnen Merkmalswerte die **Klassenmitten** verwendet.

Beispiel: Berechnung des arithmetisches Mittels bei klassierten Daten.
Für die klassierten Daten aus Beispiel 2.1 ergibt sich unter Verwendung von Formel (12) und der jeweiligen Klassenmitten eine mittlere Körpergröße [in m] der 20 Studentinnen und Studenten von

$$\bar{x} = \frac{2 \cdot 1,575 + 6 \cdot 1,7 + 7 \cdot 1,8 + 4 \cdot 1,9 + 1 \cdot 2,025}{20} = 1,78$$

bzw.

$$\bar{x} = \frac{2}{20} \cdot 1,575 + \frac{6}{20} \cdot 1,7 + \frac{7}{20} \cdot 1,8 + \frac{4}{20} \cdot 1,9 + \frac{1}{20} \cdot 2,025 = 1,78.$$

Beispiel: Gewogenes arithmetisches Mittel.
Der Primärenergieverbrauch (PEV) pro Kopf [in t Steinkohle-Einheiten (SKE)] im Jahre 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
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	58,84 %	3.564

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 (11) 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:

$$\frac{4,5 \cdot 576 + 4,5 \cdot 291 + 11,4 \cdot 314 + 1,4 \cdot 519 + 0,5 \cdot 794 + 1,1 \cdot 3.564}{6.058} = 2,15.$$

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.

3.2 Geometrisches Mittel

Beispiel: Die mittlere Rendite eines Sparbriefs
Ein Sparbrief verspricht bei Anlage 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 $q_3 = 8\%$. Die Rückzahlung der Geldanlage inklusive Zins und Zinseszins erfolgt am Ende der Laufzeit, also nach exakt 3 Jahren. Der hypothetische Kapitalbetrag K_1 nach Ende des 1. Jahres lautet

$$K_1 = K_0 + q_1 \cdot K_0 = 10.000 + 0,06 \cdot 10.000 = 10.600 \ .$$

Er ergäbe sich nach Ausklammern von K_0 auch mit Hilfe des sogenannten *Kapitalwachstumsfaktors* $(1 + q_1)$:

$$K_1 = (1 + q_1) \cdot K_0 = 1,06 \cdot 10.000 = 10.600 \ .$$

Die (geometrische) Folge der Kapitalbeträge K_1, K_2 und K_3 lautet:

$$\begin{aligned}
K_1 &= (1 + q_1) \cdot K_0 = 10.600, \\
K_2 &= (1 + q_2) \cdot K_1 = 1,07 \cdot 10.600 = 11.342 \\
&= (1 + q_2)(1 + q_1) \cdot K_0 = 1,07 \cdot 1,06 \cdot 10.000 = 11.342, \\
K_3 &= (1 + q_3) \cdot K_2 = 1.08 \cdot 11.342 = 12.249,36 \\
&= (1 + q_3) \cdot (1 + q_2) \cdot K_1 = 1,08 \cdot 1,07 \cdot 10.600 = 12.249,36 \\
&= (1 + q_3) \cdot (1 + q_2) \cdot (1 + q_1) \cdot K_0 = 1,08 \cdot 1,07 \cdot 1,06 \cdot 10.000 = 12.249,36.
\end{aligned}$$

Um den mittleren Zinssatz zu ermitteln, stellt sich die Frage nach demjenigen einheitlichen Zinssatz q, der ausgehend vom selben Anfangskapitalbetrag K_0 nach drei Jahren denselben Endkapitalbetrag K_3 ergibt:

$$K_0 \cdot (1 + q) \cdot (1 + q) \cdot (1 + q) = K_3.$$

Einsetzen von $K_3 = (1 + q_3) \cdot (1 + q_2) \cdot (1 + q_1) \cdot K_0$ und Kürzen von K_0 liefert:

$$(1+q)^3 = (1+q_3) \cdot (1+q_2) \cdot (1+q_1) \Longleftrightarrow q = \sqrt[3]{(1 + q_3) \cdot (1 + q_2) \cdot (1 + q_1)} - 1 \ . \tag{13}$$

Konkret ergibt sich exakt

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

wohingegen das arithmetische Mittel genau 7 % ergäbe:

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

Natürlich ist der Unterschied von 0,003 % marginal. Bei höheren Zinssätzen oder starken Wertveränderungen wie bei einer Aktie hat es allerdings größere Konsequenzen, wenn fälschlicherweise das arithmetische Mittel zur Berechnung der durchschnittlichen Rendite benutzt wird.

Der durchschnittliche Kapitalwachstumsfaktor $\bar{x}_G := 1 + q$ resultiert nach Gleichung (13) aus einer speziellen Mittelung der Kapitalwachstumsfaktoren $x_1 := 1 + q_1, x_2 := 1 + q_2, x_3 := 1 + q_3$ für die einzelnen Jahre:

$$\bar{x}_G := 1 + q = \sqrt[3]{(1 + q_3) \cdot (1 + q_2) \cdot (1 + q_1)} = \sqrt[3]{x_3 \cdot x_2 \cdot x_1}. \tag{14}$$

Den aus einer solchen Mittelwertbildung resultierenden Wert nennt man *geometrisches Mittel*.

Definition des geometrischen Mittels

Durch Verallgemeinerung der Formel (14) erhält man die für n Einzelwerte $(x_1, x_2, ..., x_n)$ gültige Formel für das geometrische Mittel $\bar{x}_G$:

$$\bar{x}_G := \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} \ . \tag{15}$$

Kommen einzelne Merkmalswerte mehrfach vor 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. relativen Häufigkeiten $(f_1, f_2, ..., f_m)$ auftreten, so berechnet sich das geometrische Mittel durch

$$\bar{x}_G = \sqrt[n]{(x_1)^{h_1} \cdot (x_2)^{h_2} ... \cdot (x_m)^{h_m}} = (x_1)^{h_1/n} \cdot (x_2)^{h_2/n} \cdot ... \cdot (x_m)^{h_m/n} \tag{16}$$

bzw.

$$\bar{x}_G = (x_1)^{f_1} \cdot (x_2)^{f_2} ... \cdot (x_m)^{f_m}. \tag{17}$$

Beispiel: Die mittlere Rendite einer Investition in eine Aktie.
Eine Aktie wurde zum Kurs von $K_0 = 100$ gekauft. Genau ein Jahr später ist sie nur noch die Hälfte wert, also $K_1 = 50$. Damit betrug die Rendite im 1. Jahr $q_1 = -50\%$. Im zweiten Jahr stieg der Kurs der Aktie wieder um $q_2 = 100\%$, so dass er exakt 2 Jahre später wieder auf dem ursprünglichen Niveau angekommen ist: $K_2 = 100$. Würde man die durchschnittliche Rendite mit Hilfe des arithmetischen Mittels berechnen, ergäbe sich

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

Während scharze Schafe unter den professionellen Kapital-Anlegern ihren Kunden diese Rendite-Formel weismachen, kann dies nicht die richtige durchschnittliche Rendite sein: Die Aktie ist nach exakt zwei Jahren nur genauso viel wert wie vorher – die Gesamtrendite beträgt 0 % und dies ist auch die tatsächliche durchschnittliche Rendite. Ursache für den Trugschluss auf Basis des arithmetischen Mittels sind die unterschiedlichen Kursniveaus, auf die sich die Kursveränderungen q_1 und q_2 beziehen. Mit Hilfe des geometrischen Mittels der Kapitalwachstumsfaktoren ergäbe sich dagegen der korrekte Wert:

$$\bar{q}_G = \sqrt{(1 + q_1) \cdot (1 + q_2)} - 1 = \sqrt{0.5 \cdot 2} - 1 = 0.$$

Dieses Beispiel zeigt deutlich, warum durchschnittliche Renditen, ob aus Zins-Papieren oder aus anderen Geldanlagen resultierend, korrekterweise nicht mittels des arithmetischen Mittels berechnet werden sollten.

3.3 Harmonisches Mittel

Bei kardinalen Merkmalen besteht zur Mittelwertbildung nicht nur die Auswahl zwischen arithmetischem oder geometrischem Mittel. Im folgenden Beispiel ist es sogar keinesfalls sinnvoll, einen dieser beiden Mittelwerte anzuwenden.

Beispiel: Berechnung mittlerer Geschwindigkeiten mittels des harmonischen Mittels.
Ein PKW-Fahrer legt eine Strecke s von 300 km zurück. Auf den ersten 100 km erreicht er eine durchschnittliche Geschwindigkeit von $v_1 = 120$ km/h, auf den zweiten 100 km immerhin noch $v_2 = 100$ km/h und im letzten Streckenabschnitt lediglich $v_3 = 80$ km/h. Wie hoch war die durchschnittliche Geschwindigkeit auf der gesamten Strecke?
Fälschlicherweise könnte man auf die Idee kommen, die hier gesuchte durchschnittliche Geschwindigkeit mit Hilfe des arithmetischen Mittels $\bar{v}$ zu berechnen:

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

Indem nach der intuitiven Formel

$$\text{Durchschnittliche Geschwindigkeit} = \text{Gesamtweg dividiert Gesamtzeit}$$

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

$$T = \frac{100 \text{ km}}{120 \text{ km/h}} + \frac{100 \text{ km}}{100 \text{ km/h}} + \frac{100 \text{ km}}{80 \text{ km/h}} = 5/6 \text{ h } + 1 \text{ h } + 5/4 \text{ h } \approx 3 \text{ h}$$

dividiert wird, erhält man die richtige Durchschnittsgeschwindigkeit:

$$\begin{aligned} \bar{v}_H &= \frac{s}{T} = \frac{300 \text{ km}}{\frac{100 \text{ km}}{120 \text{ km/h}} + \frac{100 \text{ km}}{100 \text{ km/h}} + \frac{100 \text{ km}}{80 \text{ km/h}}} \\ &= \frac{1}{\frac{100}{300}(\frac{1}{120 \text{ km/h}} + \frac{1}{100 \text{ km/h}} + \frac{1}{80 \text{ km/h}})} = 97,30 \text{ km/h}. \end{aligned}$$

Natürlich sind die Unterschiede in den beiden Geschwindigkeitswerten 100 km/h und 97,30 km/h eigentlich nicht der Rede wert – jeder Verkehrstau, in den man gerät, läßt diese Differenz zur Nichtigkeit werden. Das Beispiel soll lediglich illustrieren, welche Kategorie von Mittelwerten sachlogisch richtig ist. Mit Hilfe der allgemeinen Bezeichnungen v_1, v_2 und v_3 lautet die analoge Formel für die korrekte Durchschnittsgeschwindigkeit:

$$\bar{v}_H = \frac{1}{\frac{1}{3}(\frac{1}{v_1} + \frac{1}{v_2} + \frac{1}{v_3})}. \tag{18}$$

Durch Verallgemeinerung von Formel (18) gelangen wir zur Definition des harmonischen Mittels.

Definition des harmonischen Mittels

Für die n Einzelwerte $(x_1, x_2, ..., x_n)$ lautet die Formel für das harmonische Mittel $\bar{x}_H$:

$$\bar{x}_H := \frac{1}{\frac{1}{n}(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n})}. \tag{19}$$

Sind die n Einzelwerte nicht allesamt verschieden, sondern kommen einzelne Merkmalswerte mehrfach vor, wobei in Wahrheit nur $m < n$ Merkmalswerte $(x_1, x_2, ..., x_m)$ verschieden sind, die mit den absoluten Häufigkeiten $(h_1, h_2, ..., h_m)$ bzw. relativen Häufigkeiten $(f_1, f_2, ..., f_m)$ auftreten, so berechnet sich das harmonische Mittel durch

$$\bar{x}_H = \frac{1}{\frac{1}{n}(\frac{h_1}{x_1} + \frac{h_2}{x_2} + ... + \frac{h_m}{x_m})} \tag{20}$$

bzw.

$$\bar{x}_H = \frac{1}{(\frac{f_1}{x_1} + \frac{f_2}{x_2} + ... + \frac{f_m}{x_m})}. \tag{21}$$

Beispiel: Berechnung mittlerer Geschwindigkeiten mittels des arithmetischen Mittels

Ein PKW-Fahrer fährt anstatt mit dem Zug insgesamt 3 Stunden im Stop-and-Go auf einer verstopften Autobahn. In der ersten Stunde erreicht er eine durchschnittliche Geschwindigkeit von $v_1 = 30$ km/h, in der zweiten Stunde immerhin $v_2 = 40$ km/h und in der letzten Stunde lediglich $v_3 = 20$ km/h. Wie hoch war die durchschnittliche Geschwindigkeit während der gesamten Fahrtzeit?

Indem zuerst die gesamte zurückgelegte Strecke

$$s = v_1 \cdot T_1 + v_2 \cdot T_2 + v_3 \cdot T_3$$

berechnet wird, kann durch Division des Gesamtweges s durch die Gesamtzeit T die durchschnittliche Geschwindigkeit berechnet werden:

$$\frac{s}{T} = \frac{v_1 \cdot T_1 + v_2 \cdot T_2 + v_3 \cdot T_3}{T} = v_1 \cdot \frac{T_1}{T} + v_2 \cdot \frac{T_2}{T} + v_3 \cdot \frac{T_3}{T}.$$

Dieser Ausdruck ist ein Spezialfall der Formel (11) für das gewogene arithmetische Mittel:

$$\bar{x} := f_1 x_1 + f_2 x_2 + f_3 x_3 + ... + f_m x_m = \sum_{i=1}^{m} f_i x_i \ .$$

In der Tat ergibt sich mit $T_1 = T_2 = T_3 = T/3$ gerade das arithmetische Mittel:

$$v_1 \cdot \frac{1}{3} + v_2 \cdot \frac{1}{3} + v_3 \cdot \frac{1}{3} = \frac{v_1 + v_2 + v_3}{3} = \bar{v} = \frac{30 + 40 + 20}{3} \text{ km/h} = 30 \text{ km/h}.$$

3.4 Median oder Zentralwert

Bei *ordinalen* Merkmalen kann anstatt dem arithmetischen, geometrischen oder dem harmonischen Mittel das Konzept des *Medians* benutzt werden. Der Median, auch Zentralwert genannt, kann immer dann angewandt werden, wenn eine Rangfolge unter den Merkmalswerten bzw. Merkmalsausprägungen existiert. Das ist bei ordinalen Merkmalen der Fall, aber auch bei kardinalen Merkmalen. Bei nominalen Merkmalen ist indes das Konzept des Medians aus diesem Grund prinzipiell nicht anwendbar.

Mit Hilfes des Medians will man bei kardinalen Merkmalen denjenigen Wert $\bar{x}_Z$ bzw. bei ordinalen Merkmalen diejenige Merkmalsausprägung ermitteln, welche gestattet, das vorhandene Datenmaterial in zwei möglichst gleich große Hälften aufzuteilen: Daher ist mit dem Konzept des Medians ebenfalls die Ermittlung eines mittleren Wertes 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 mittels einer Stamm-Blatt-Darstellung.
In Beispiel 2.2 der Altersverteilung von 20 Personen läßt sich der Median mit Hilfe einer sogenannten *Stamm-Blatt-Darstellung* bestimmen, die die Merkmalswerte in eine **geordnete** Reihenfolge bringt:

2	6
2	3 3
2	2 2 2
2	1 1 1 1 1
2	0 0 0 0 0
1	9 9 9 9

Links des Trennstriches werden die Zehnerstellen des Lebensalters aufgetragen. Dort befindet sich der *Stamm* der Stamm-Blatt-Darstellung. Rechts des Trennstriches repräsentiert jede einzelne Ziffer – das *Blatt* – die Einerstelle der jeweiligen Lebensalter. Das Alter von 21 teilt die Datenmenge in zwei Teilmengen auf – leider nicht gleichen Umfanges, so dass es nicht gerechtfertigt wäre, von Hälften zu sprechen: In der einen Teilmenge sind 14 von 20 Personen, also 70 % aller Personen, enthalten, deren Lebensalter kleiner oder gleich dem Wert 21 ist, während in der anderen Teilmenge alle Personen zusammengefaßt sind, deren Lebensalter größer oder gleich dem Wert 21 ist. Das sind genau 11 von 20, sprich 55 % aller Personen. Der Merkmalswert 21 stellt in diesem Beispiel den Median – auch 50 %-Quantil genannt – dar. Andere Werte, zum Beispiel 20 oder 22, würden nicht dafür in Frage kommen: Es sind nur 9 von 20 Werten, also weniger als 50 %, kleiner oder gleich dem Lebensalter 20 und es sind nur 6 von 20, sprich wesentlich weniger als 50 % aller Werte größer oder gleich dem Alter von 22.

Beispiel: Berechnung des Medians bei kardinalen Merkmalen.
Hätte man beispielsweise einen Vektor $(x_1, x_2, ..., x_{11})$ von $n = 11$ **geordneten** Einzelwerten vorliegen, so ist ungeachtet des tatsächlichen Aussehens der Merkmalswerte der 6. Wert der Median, denn beim 6. einer geordneten Reihe von 11 Werten sind 6 von 11, folglich über 50 % aller Werte, kleiner oder gleich dem Wert x_6 und auch 6 von 11 Werten sind größer oder gleich x_6:

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

Sind diese Überlegungen anhand des abstrakten Zahlenvektors $(x_1, x_2, ..., x_{11})$ zu wenig konkret, dient beispielsweise der folgende Vektor mit 11 Zahlen zur Veranschaulichung:

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

Der 6. Wert in dieser geordneten Reihe von Zahlen ist bei 11 Werten der Median: $\bar{x}_Z = x_6 = 4$. Der Wert 4 teilt die Menge von 11 Zahlen in zwei gleich große Hälften – in die Menge $\{0, 1, 1, 2, 2, 4, 4\}$, deren Werte kleiner oder gleich dem Median 4 sind, und in die Menge $\{4, 4, 6, 9, 13, 17\}$, deren Werte größer oder gleich 4 sind.

Die beschreibende Definition des Medians legt diesen nicht immer eindeutig fest, wie aus folgender Überlegung hervorgeht. Hätten statt 11 nun 12 geordnete Einzelwerte im Vektor $(x_1, x_2, ..., x_{11}, x_{12})$ vorgelegen, kämen nach der beschreibenden Definition zwei Werte als Median in Frage: Der 6. und der 7. Wert in der geordneten Reihenfolge – beide Werte würden die genannten Bedingungen erfüllen. Um diese Zweideutigkeit zu beseitigen, wird der Median in diesem Fall in dann eindeutiger Weise als arithmetisches Mittel der beiden in Frage kommenden Werte festgelegt:

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

Konkret könnte beispielsweise folgender Vektor mit 12 geordneten Zahlen vorgelegen haben:

$$(0, 1, 1, 2, 2, \mathbf{3}, 4, 4, 6, 9, 13, 17).$$

Dabei ist die fett markierte 3 diejenige Zahl, welche im Vergleich zu obigem Vektor mit 11 Zahlen hinzugekommen ist. Der Median lautet in diesem Fall $\bar{x}_Z = \frac{3+4}{2} = 3,5$. Dieser Wert teilt die Menge von 12 Zahlen in zwei genau gleich große Hälften (Teilmengen gleichen Umfangs) auf:

$$\{0, 1, 1, 2, 2, 3\} \qquad \text{und} \qquad \{4, 4, 6, 9, 13, 17\}.$$

Das Beispiel macht deutlich: Der konkrete Wert des Medians muss, ebenso wie die Werte anderer Mittelwertkonzepte auch, nicht ein Wert des vorliegenden Datenmaterials sein. Schließlich wird klar, dass beim Median – im Gegensatz zum arithmetischen, geometrischen und auch harmonischen Mittel – i. A. nicht alle Einzelwerte in dessen Berechnung einfließen. Insbesondere beeinflussen der kleinste und der größte Datenwert den konkreten Wert des Medians normalerweise nicht. Dies macht den Median robust gegenüber

positiven wie negativen Ausreißern bzw. dem Auftreten von extremen Werten.

Definition des Medians bei n Einzelwerten eines kardinalen Merkmals

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

$$\bar{x}_Z := \begin{cases} x_i \ , & \text{wobei } i = (n+1)/2, \text{ falls } n \text{ ungerade ist,} \\ \\ \dfrac{x_i + x_{i+1}}{2} \ , & \text{wobei } i = n/2, \text{ falls } n \text{ gerade ist.} \end{cases} \tag{22}$$

Eigenschaften des Medians

- **Minimumseigenschaft** des Medians:

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

  $$\min_y \sum_{i=1}^{n} |x_i - y|, \tag{23}$$

  wobei bei gegebenen $(x_1, x_2, ..., x_n)$ die Summe $\sum |x_i - y|$ eine Funktion $f(y)$ allein der Variablen y ist. In Worten: Der Median $\bar{x}_Z$ ist genau derjenige Wert y, der die Summe der *absoluten Abweichungen*, das heißt der *Beträge* der Abweichungen der einzelnen Werte x_i von y minimiert. Diese Eigenschaft läßt sich nicht mit Hilfe der Differentialrechnung beweisen, da die Betragsfunktion nicht differenzierbar ist. Sie soll daher an dieser Stelle unbewiesen bleiben.

- **Robustheit** des Medians:

  Der Wert des Medians ist, im Gegensatz zu dem des arithmetischen Mittels, gegenüber Ausreißern robust: Er wird durch das Auftreten von extremen Werten *nicht* beeinflußt.

- **Anwendbarkeit** des Medians:

  Das Konzept des Medians ist nicht nur bei kardinalen, sondern auch bei ordinalen Merkmalen anwendbar. Der Median ist jedoch nicht bei nominalen Merkmalen anwendbar, da das Konzept eine Rangfolge (Ordnung) unter den Merkmalswerten bzw. - ausprägungen voraussetzt.

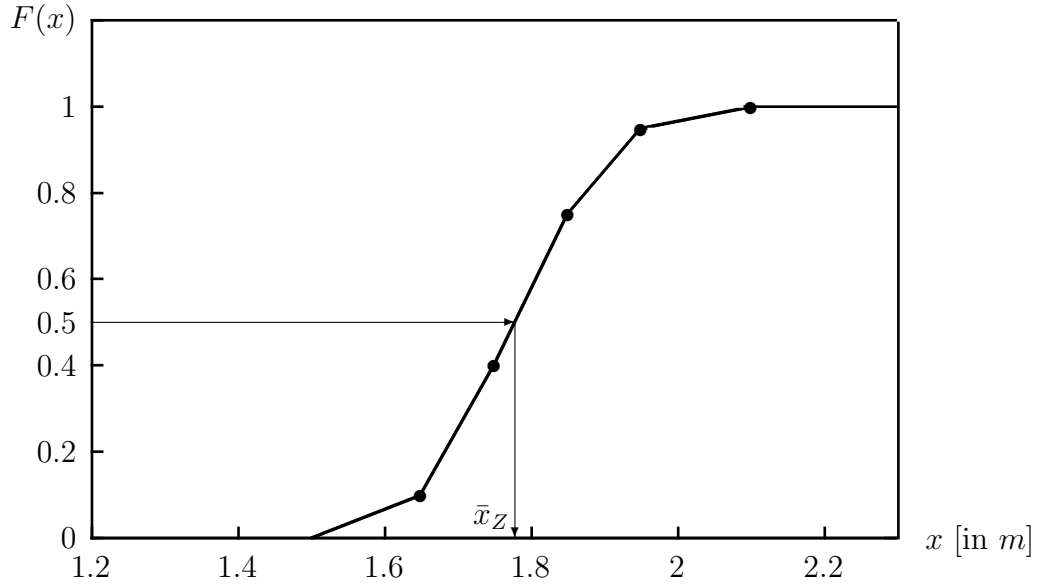
Ermittlung des Medians bei klassierten Daten

Bei klassierten Daten kann der Median graphisch mit Hilfe des Verteilungsploygons $F(x)$ ermittelt werden. Der Median $\bar{x}_Z$, das 50 % – Quantil, ist der zum Funktionswert $F = 0.5 = 50$ % gehörige x-Wert:

$$F(\bar{x}_Z) = 0.5. \tag{24}$$

Beispiel: Graphische Bestimmung des Medians bei klassierten Daten.
Mit Hilfe des in Figur 5 skizzierten Verteilungspolygons $F(x)$ für die klassierten Daten der Körpergrößen aus Beispiel 2.2 läßt sich der Median $\bar{x}_Z$ bestimmen, indem der zum Funktionswert $F(x) = 0{,}5$ gehörige x-Wert ermittelt wird.

Fig. 5: Graphische Bestimmung des Medians bei klassierten Daten



Offenbar liegt der Median in diesem Beispiel innerhalb der 3. Körpergrößenklasse: $[x_3^u; x_3^o) = [1,75; 1,85)$. Die beiden zur 3. Klasse gehörigen Eckpunkte $(1,75; 0,40)$ und $(1,85; 0,75)$ werden beim Verteilungspolygon durch eine Gerade verbunden. Deshalb sind die Verhältnisse aus den Differenzen zweier beliebiger x-Werte innerhalb dieser Klasse und den Differenzen der zugehörigen Funktionswerte allesamt gleich.
Beispielsweise gilt:

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

Mit $F(\bar{x}_Z) = 0{,}5$ ergibt sich daraus die genaue Lage des Medians:

$$\bar{x}_Z = x_3^u + (x_3^o - x_3^u) \cdot \frac{F(\bar{x}_Z) - F(x_3^u)}{F(x_3^o) - F(x_3^u)} = 1,75 + 0,1 \cdot \frac{0,5 - 0,4}{0,75 - 0,4} = 1,78.$$

Wie aus der Figur 5 ersichtlich, liegt der so bestimmte Median knapp unterhalb der Körpergröße von 1,8 m bei 1,78 m und stimmt damit mit dem arithmetischen Mittel $\bar{x}$ überein.

Durch Verallgemeinerung der im letzten Beispiel zur Berechnung des Medians benutzen Gleichung ergibt sich die allgemeine Formel zur Berechnung des Medians bei klassierten Daten:

$$\bar{x}_Z = x_j^u + (x_j^o - x_j^u) \cdot \frac{F(\bar{x}_Z) - F(x_j^u)}{F(x_j^o) - F(x_j^u)} = x_j^u + (x_j^o - x_j^u) \cdot \frac{0,5 - F(x_j^u)}{F(x_j^o) - F(x_j^u)}. \qquad (25)$$

Zur Rekapitulation: Diese Formel resultiert aus der Bedingung $F(\bar{x}_Z) = 0,5$ sowie 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. Zur konkreten Berechnung des Medians mit Hilfe dieser Formel muss allerdings erst festgestellt werden, zum Beispiel grafisch oder mit Hilfe einer Tabelle der kumulierten relativen Häufigkeiten, auf welche Klasse j der Median entfällt.

Definition des Medians bei ordinalen Merkmalen

Der Median ist diejenige Merkmals*ausprägung* eines ordinalen Merkmals, deren Niveau mindestens 50 % aller Merkmalsausprägungen unterschreiten oder höchstenfalls erreichen **und** deren Niveau von mindestens 50 % aller Merkmalsausprägungen überschritten oder zumindest erreicht wird.

Beispiel: Berechnung des Medians bei ordinalen Merkmalen.
Eine Studentin der Universität Heidelberg soll angeben, wie regelmäßig sie die 12 für sie relevanten Veranstaltungen des Sommersemesters 2002 besucht hat. Ihre Antworten bilden das Merkmal X: „Häufigkeit der Teilnahme an den Veranstaltungen“, wobei die Ausprägungen wie folgt kodiert sind: 0 (nie), 1 (sehr selten), 2 (selten), 3 (häufig), 4 (fast immer), 5 (immer). Die absoluten, relativen und kumulierten relativen Häufigkeiten für ihre 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 (fast immer)	3	3/12	11/12
3 (häufig)	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

Die Studentin besucht also beispielsweise 3 von 12 Veranstaltungen nur selten. Die kumulierte relative Häufigkeit von 6/12 besagt: 6 von 12 Veranstaltungen hat sie 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”. Diese Zweideutigkeit kann nicht wie bei kardinalen Merkmalen beseitigt werden: Die Bildung des arithmetischen Mittels von “selten” und “häufig” macht keinen Sinn! Jemand, der der Studentin geneigt ist, definiert die Ausprägung “häufig” als Median. Andererseits könnte genauso gut die Ausprägung “selten” als Median festgelegt werden. Auf eine eindeutige Festlegung kann aber auch verzichtet werden.

3.5 Modus

Liegt ein nominales Merkmal vor, so kann keines der bisher diskutierten Mittelwertskonzepte angewandt werden. Für diesen Fall existiert das Konzept des Modus bzw. der modalen Klasse.

Definition des Modus und der modalen Klasse

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

Beispiele zum Modus und zur modalen Klasse.
In Beispiel 2.1 ist die Größenklasse $[1,75; 1,85)$ die modale Klasse, denn 7 von 20 – und damit die meisten – Studentinnen und Studenten weisen eine Körpergröße auf, welche in diese Klasse entfällt. In Beispiel 2.2 gibt es keinen eindeutigen Modus: Die Lebensalter 20 und 21 sind beide gleich häufig unter den 20 Personen vertreten. Hinsichtlich der Noten der Statistik-Klausur dieses Kurses stellt die Note „sehr gut" den Modus dar, beim Geschlecht die Ausprägung „weiblich".
Auch im obigen Beispiel des ordinalen Merkmals „Häufigkeit der Teilnahme an den Veranstaltungen" gibt es keinen eindeutigen Modus: „fast immer" bzw. „selten" sind die beiden gleichermaßen häufig auftretenden Merkmalsausprägungen, die für die Studentin und ihre Besuchshäufigkeit der Veranstaltungen charakteristisch sind.

Eigenschaften des Modus

- Das Konzept des Modus ist durchweg sowohl für nominale, ordinale wie auch kardinale Merkmale anwendbar. Bei klassierten Daten spricht man anstatt von Modus von modaler Klasse.
- Robustheit: Der Wert des Modus wird bei kardinalen Merkmalen nicht durch Ausreißer oder extreme Werte beinflusst. (Bei nominalen und ordinalen Merkmalen existiert der Begriff der Robustheit nicht!)
- Das Konzept des Modus liefert nicht zwingend eine eindeutig als Modus zu identifizierende Merkmalsausprägung bzw. einen eindeutig als Modus zu definierenden Merkmalswert. Ebenso wenig

  muss Eindeutigkeit darüber herrschen, welche Klasse bei klassierten Daten die modale Klasse ist.

3.6 Quantile

Neben den Mittelwerten sind noch andere Lageparameter in der Statistik von wesentlicher Bedeutung, die sogenannten *Quantile*. Das Konzept des Quantils ist eine Verallgemeinerung des Konzeptes des Medians: Der Median stellt nichts anderes als ein spezielles Quantil dar, nämlich das 50%-Quantil. Von Interesse könnten daneben beispielsweise auch das 25 %-Quantil oder das 75 %-Quantil sein, respektive 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.

Beispiel: Berechnung des unteren Quintils und des unteren Quartils.
In Beispiel 2.2 käme das Lebensalter von 19 Jahren als 20 %-Quantil in Frage: 4 von 20 Personen, und damit genau 20 % aller Personen haben ein Alter von 19 Jahren und definitiv mehr als 80 %, nämlich alle Personen, haben ein Lebensalter von 19 und mehr Jahren. Aber auch das Lebensalter von 20 Jahren bildet nach der beschreibenden Definition ein 20 %-Quantil: 9 von 20, und damit mehr als 20 % aller Personen, haben ein Alter von 20 Jahren und weniger, während 16 von 20 Personen ein Alter von 20 und älter haben. Ebenso wie beim Median wird diese Zweideutigkeit per Definition beseitigt, indem das 20 %-Quantil in eindeutiger Weise als arithmetisches Mittel der beiden in Frage kommenden Werte festgelegt wird.
Das 20 %-Quantil, auch unteres Quintil genannt, lautet damit:

$$x_{0.2} = \frac{19 + 20}{2} = 19,5.$$

Das Lebensalter von 19,5 Jahren teilt die Masse der Studentinnen und Studenten dieses BA-Kurses in zwei Teilmassen auf: Eine Teilmasse enthält die 19-Jährigen, das sind genau 20 % aller Personen dieses Kurses, während die andere Teilmasse die über 19-Jährigen umfasst – dies ist genau ein Anteil von 80 % des Kurses.

Für das untere Quartil $x_{0.25}$, dem 25 % - Quantil, kommt dagegen in diesem Beispiel lediglich ein Wert in Frage: $x_{0.25} = 20$. Der Wert von 19 Jahren würde die Bedingungen für ein 25 % - Quantil nicht erfüllen: Nur 4 von 20, und damit lediglich 20 % anstatt 25 % aller Personen, haben ein Alter von 19. Auch der Wert von 21 Jahren kann kein 25 % - Quantil darstellen: Zwar haben 70 %, nämlich 14 von 20 Personen, und damit weit mehr als 25 % ein Alter von 21 und weniger, aber nur 11 von 20, also lediglich 55 % anstatt der geforderten 75 % aller statistischen Einheiten, haben ein Alter von 21 und älter.

Wie beim Median demonstriert, können Quantile aus einer geordneten Reihe von Einzelwerten ermittelt werden, ohne dass konkrete Werte bekannt sind.

Beispiel: Intuitive Berechnung des unteren Quartils.
Hätte man beispielsweise $n = 11$ geordnete Einzelwerte im Vektor $(x_1, x_2, ..., x_{11})$ vorliegen, so ist – ungeachtet des tatsächlichen Aussehens der Merkmalswerte – der 3. Wert das 25 % -Quantil. Der Wert x_3 erfüllt tatsächlich die Bedingungen, die an ein 0.25-Quantil gestellt werden: Das 25 % -Quantil soll die Stichprobe in zwei Teilmengen aufteilen, 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 $\{x_3, x_4, ..., x_{11}\}$ 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

$$\{0, 1, 1\} \qquad \text{und} \qquad \{1, 1, 2, 2, 4, 4, 6, 9, 13, 17\}.$$

Durch Division der Gesamtzahl von 11 Elementen durch 4 ergibt sich ein Hinweis auf die Aufteilung der zwei Teilmengen. Leider liefert die Division aber keine ganze Zahl, sondern den Wert 2,75. Dennoch ist damit klar, dass die kleinere Teilmenge mindestens 3 Elemente – die nächstgrößerere ganze Zahl bezogen auf 2,75 – der Stichprobe enthalten muss, während die andere Teilmenge mindestens 9 Elemente enthalten muss – die nächstgrößere ganze Zahl bezogen auf $8{,}25 = 3/4 \cdot 11$.
12 anstatt 11 geordnete Einzelwerte wie beispielsweise

$$(0, 1, 1, 2, 2, \mathbf{3}, 4, 4, 6, 9, 13, 17),$$

lassen sich mathematisch gesehen leichter 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, wird wie bei der Definition des Medians das arithmetische Mittel der beiden in Frage kommenden Werte als unteres Quartil festgelegt:

$$x_{0.25} = \frac{x_3 + x_4}{2} = \frac{1 + 2}{2} = 1, 5.$$

Der Wert 1,5 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 nötige Notation bereits eine gewisse Gedächtnisleistung erfordert, ist es kaum vorstellbar, dass jemand diese Definition reproduzieren kann, ohne das dahinter liegende Prinzip verstanden zu haben. Ist das Prinzip aber verstanden, ist die Definition überflüssig.

Definition des p-Quantils x_p 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

$$x_p := \begin{cases} x_i \ , & \text{wobei } i = [n \cdot p] + 1, \text{ falls } n \cdot p \text{ nicht ganzzahlig ist,} \\ \dfrac{x_i + x_{i+1}}{2} \ , & \text{wobei } i = [n \cdot p], \text{ falls } n \cdot p \text{ ganzzahlig ist.} \end{cases} \tag{26}$$

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. Man vergewissere sich, dass sich für den Median $\bar{x}_Z$, dem 50 %-Quantil $x_{0.5}$, aus der Definition (26) des p-Quantils mit $p = 0.5 = 1/2$ die Definition (22) des Medians bei Individualdaten eines kardinalen Merkmals ergibt.

Beispiel: Berechnung des unteren Quartils nach Definition (26).
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 mit Hilfe dieser Definition $x_{0.25} = x_3 = 1$, denn $n \cdot p = 11 \cdot 0,25 = 2{,}75$, so dass der erste Teil der Defintion anzuwenden ist, 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 $x_{0.25} = (x_3+x_4)/2 = (1+2)/2 = 1, 5$. Wegen $n \cdot p = 12 \cdot 0, 25 = 3$ ist der zweite Teil der Definition anzuwenden, wobei $[n \cdot p] = [3] = 3$.

Ermittlung des p-Quantils bei klassierten Daten

Bei klassierten Daten kann das p-Quantil x_p graphisch mit Hilfe des Verteilungsploygons $F(x)$ ermittelt werden. Das p-Quantil x_p ist der zum Funktionswert $F(x) = p$ gehörige x-Wert:

$$F(x_p) = p. \tag{27}$$

Rechnerisch kann das p-Quantil durch eine, zu der des Medians analoge, allgemeine Formel ermittelt werden:

$$x_p = x_j^u + (x_j^o - x_j^u) \cdot \frac{F(x_p) - F(x_j^u)}{F(x_j^o) - F(x_j^u)} = x_j^u + (x_j^o - x_j^u) \cdot \frac{p - F(x_j^u)}{F(x_j^o) - F(x_j^u)}. \qquad (28)$$

Diese Formel resultiert aus der Bedingung (27) sowie 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. Zur konkreten Berechnung des Medians mit Hilfe (28) muss erst grafisch oder mit Hilfe einer Tabelle der kumulierten relativen Häufigkeiten festgestellt werden, auf welche Klasse j das p-Quantil entfällt. Für $p = 0.5 = 1/2$ ergibt sich aus Formel (28) die Formel (25) des Medians für klassierte Daten.

Definition des p-Quantils bei ordinalen Merkmalen

Das P %-Quantil bzw. p-Quantil ist diejenige Merkmals*ausprägung* eines ordinalen Merkmals, deren Niveau mindestens P % aller Merkmalsausprägungen unterschreiten oder höchstenfalls erreichen **und** deren Niveau von mindestens (1-P)% aller Merkmalsausprägungen überschritten oder zumindest erreicht wird.

Beispiel: Berechnung von Quantilen bei ordinalen Merkmalen.
Anhand der Tabelle der absoluten, relativen und kumulierten relativen Häufigkeiten für die Teilnahme einer Studentin der Universität Heidelberg an 12 Veranstaltungen im Sommer 2002 können das untere und obere Quartil leicht bestimmt werden:

Ausprägung	absolute Häufigkeit h_i	relative Häufigkeit f_i	kum. relative Häufigkeit F_i
5 (immer)	1	1/12	12/12
4 (fast immer)	3	3/12	11/12
3 (häufig)	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

Das untere Quartil oder 25 %-Quantil ist in diesem Fall die Ausprägung „sehr selten": Damit lassen sich die 12 Veranstaltungen in zwei Teilmengen mit ungefähr gleichem Umfang von 1/4 bzw. 3/4 aller Veranstaltungen aufteilen, wobei eine Teilmenge die Veranstaltungen mit einer Teilnahmehäufigkeit von höchstens „sehr selten" umfasst – das sind 3 von 12 Veranstaltungen. Die andere Teilmenge beinhaltet alle Veranstaltungen, die eine Teilnahmehäufigkeit von mindestens „sehr selten" aufweisen – das sind 10 von 12 Veranstaltungen. Das obere Quartil oder 25 %-Quantil ist die Ausprägung „fast immer": 11 von 12, also mehr als 75 % aller Veranstaltungen weisen eine Teilnahmehäufgikeit von höchstensfalls „fast immer" auf und 4 von 12, also mehr als 25 % sind mindestens „fast immer" besucht worden.

3.7 Zusammenfassung zu den Lagemaßen

Bei nominalen (qualitativen) Merkmalen beschränkte sich unsere Diskussion der dafür in Frage kommenden Lageparameter auf den Modus. Der Modus ist die, nicht notwendigerweise eindeutig bestimmte, am häufigsten auftretende Merkmalsausprägung. Bei ordinalen Merkmalen können neben dem Modus zusätzlich noch der Median, das 50 %-Quantil, sowie alle anderen Quantile ermittelt werden. Bei kardinalen Merkmalen schließlich sind darüber hinaus die Konzepte des arithmetischen, geometrischen bzw. harmonischen Mittels anwendbar. Andere Konzepte als die hier angesprochenden populärsten Lageparameter sollen nicht erwähnt werden.

Für die Mittelwertbildung bei kardinalen Merkmalen ist die folgende Faustregel festzuhalten: Bei *additiven* Zusammenhängen findet das arithmetische Mittel $\bar{x}$, bei *multiplikativen* Zusammenhängen das geometrische Mittel $\bar{x}_G$ Anwendung. Das harmonische Mittel $\bar{x}_H$ kann bei der Mittelwertbildung von Quotienten verwendet werden – nicht immer allerdings, wie ein Beispiel uns gezeigt hat. Daher ist dies lediglich eine Faustregel, jedoch keine immer und allgemein gültige Aussage. Allgemein gültig ist indessen: Keines dieser drei Mittelwertkonzepte kann und darf bei ordinalen oder nominalen Merkmalen angewandt werden.

Die folgende Tabelle fasst die Anwendbarkeit der besprochenen Mittelwertskonzepte für die verschiedenen Kategorien von Merkmalen zusammen. Ein Kreuz gekennzeichnet die Anwendbarkeit des jeweiligen Konzepts bei der entsprechenden Merkmals-Kategorie:

	nominal	ordinal	kardinal
Modus	x	x	x
Median	–	x	x
$\bar{x}$, $\bar{x}_G$, $\bar{x}_H$	–	–	x

Oft wird die folgende, für beliebige Stichproben gültige Ungleichungskette angegeben:

$$\bar{x} \geq \bar{x}_G \geq \bar{x}_H.$$

Diese Kette von Ungleichungen ist allerdings ein Muster ohne Wert: Sachlogisch kann für einen bestimmten Zusammenhang im Allgemeinen nur eines der drei Konzepte zur Mittelwertbildung bei kardinalen Merkmalen richtig sein – Vergleiche mit den Werten, welche die beiden anderen Mittelwert-Konzepte liefern würden, erübrigen sich, wenn deren Anwendung unangemessen ist.

3.8 Übungsaufgaben

Übungsaufgabe: Börsencrash auf Raten.

Herr F. ist selbstständiger Statistiker und erwirtschaftet im Jahr 2000 im Zuge der allgemeinen Börsen-Euphorie mit statistischen Aktienanalysen ein Ergebnisplus von 30 % verglichen mit dem Vorjahr. Im Jahr 2001 wendet sich das Blatt: Kaum jemand möchte

sich ab dem Beginn des 2. Halbjahres noch für Aktien begeistern, geschweige denn Portfolioanalysen erstellen lassen. Deshalb verzeichnet Herr F. im Jahr 2001 im Vergleich zum Vorjahr 2000 ein Ergebnisminus von 30 %. Ist sein Ergebnis aus der selbständigen Statistik-Tätigkeit im Jahr 2001 gleich hoch als das Ergebnis vor dem Aktienboom? Wie hoch ist die durchschnittliche Ergebnisentwicklung innerhalb dieser zwei Jahre?

Übungsaufgabe: Heidelberg – Wien und zurück.

Herr F. ist mit seinem Citroen 2CV, im Volksmund „Ente“ genannt, auf großer Fahrt von Heidelberg nach Wien (Distanz ca. 660 km). Aufgrund des Rückenwindes (Westwind) erreicht er auf dem Hinweg eine durchschnittliche Geschwindigkeit von 110 km/h. Auf dem Rückweg bläst ihm der Wind vehement entgegen. Das reduziert die Durchschnittsgeschwindigkeit drastisch – sie beträgt auf dem Rückweg nur noch 80 km/h. Mit welcher Durchschnittsgeschwindigkeit war Herr F. insgesamt auf Hin- und Rückweg unterwegs?

Übungsaufgabe: Der optimale Standort eines Heidelberger Obdachdachlosen.

Der Obdachlose F. möchte seinen Standort in der Heidelberger Hauptstrasse – der längsten Fußgängerzone Europas – für seine hauptberufliche Tätigkeit möglichst so wählen, dass er seine Spenden-Einnahmen maximieren kann. Natürlich würde er, trotz seiner genauen Kenntnis über die Spender und deren Wohnsitze, keinen seiner Spender verraten – weder an seine Konkurrenz noch an sonst jemanden, Politiker beispielweise. Seine 15 Hauptspender wohnen allesamt in der Hauptstrasse. Im k-ten Haus, das x_k Meter vom Anfang der Hauptstraße entfernt ist, wohnen n_k seiner Hauptspender:

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

In welcher Entfernung vom Anfang der Hauptstrasse sollte der Obdachlose F. seinen Standort optimalerweise wählen?

Übungsaufgabe: Durchschnittliche Arbeitslosenquote in Schröder*land.*

Schröderland besteht aus 16 Bundesländern, welche die folgenden Arbeitslosenquoten [in %] bzw. Bevölkerungszahlen [in 1.000] aufweisen:

Land	1	2	3	4	5	6	7	8
Quote	4,9	5,3	16,1	17,4	12,4	8,3	6,6	18,3
Personen	10.524	12.230	3.382	2.602	660	1.715	6068	1776
Land	9	10	11	12	13	14	15	16
Quote	9,1	8,8	6,8	9,0	17,5	19,7	8,4	15,3
Personen	7.926	18.010	4.035	1.069	4.426	2.615	2.790	2.431

Wie hoch ist die durchschnittliche Arbeitslosenquote in Schröderland? Wie hoch ist absolute Zahl an Arbeitslosen?

Übungsaufgabe: Durchschnittliche Mietpreise in Heidelberg.

Vor der Einführung des Euro betrug der durchschnittliche Mietpreis in DM in Heidelberg $\bar{x} = 18DM/m^2$ bei einer Varianz von $s_X^2 = 3.24DM^2/m^4$. Wie lauten der durchschnittlicher Mietpreis $\bar{y}$ in EUR sowie die empirische Varianz und Standardabweichung s_Y^2 und s_Y, wenn von 1 EUR = 1,95583 DM und der Tatsache ausgegangen wird, dass die Mietpreise unverändert geblieben und lediglich umgerechnet wurden?

Übungsaufgabe: Spaghetti bei Alfredo in Heidelberg.

Alfredo macht seine Spaghetti noch selbst, wie es sich für ein richtiges italienisches Lokal gehört. Sie haben daher keine Normlänge. Die relative Häufigkeitsverteilung für die Länge X der Spaghetti [in *cm*] lautet:

X	$20 \leq x < 40$	$40 \leq x < 60$	$60 \leq x < 80$	$80 \leq x < 100$	100 und mehr
f_i	0,15	0,20	0,30	0,10	0,05

Dottore Michele F. bekommt zufälligerweise eine Portion Spaghetti, bei der alle länger als 80 *cm* sind. Das bereitet ihm beim Essen erhebliche Schwierigkeiten, weshalb er sich fragt, warum er eigentlich nicht wie sonst auch die Gnocchi genommen hat.

a) Wie häufig kommt es bei Alfredo eigentlich vor, fragt sich Dottore Michele F. , dass die Spaghettilänge 80 cm und mehr beträgt?

b) Wie lautet das arithmetische Mittel $\bar{x}$ der Länge der Spaghetti, wenn die maximale Länge 120 cm beträgt?

c) Wie lautet der Median $x_{0.5}$ bzw. das obere Quartil $x_{0.75}$ der Länge der Spaghetti, wenn innerhalb der Längenklassen interpoliert wird? Wie sind beide Werte zu interpretieren?

d) Muß ein Merkmal nominaler, ordinaler oder kardinaler Natur sein, damit dafür ein Median bestimmt werden kann?

Bei Alfredo gibt es neben Spaghetti auch Pizza, Tortellini und Gnocchi, die von den Gästen wie folgt geschätzt werden:

Art des Gerichts	relative Häufigkeit
Spaghetti	60 %
Tortellini	25 %
Pizza	10 %
Gnocchi	5 %

e) Wie könnte man dieses Ergebnis in einer Skizze geeignet grafisch darstellen?

f) Ist der einzelne Gast das statistische Merkmal, die statistische Einheit oder die Merkmalsausprägung? Wie steht es damit bei der Art des Gerichts?

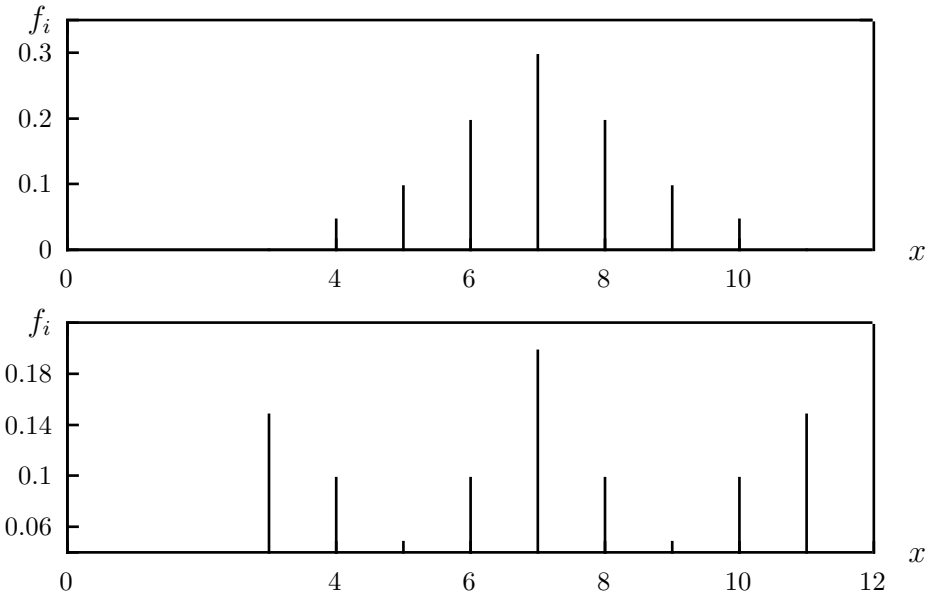
g) Wer Art sind die folgenden Merkmale: „Anzahl der Gäste“, „Art des Gerichts“, „Anzahl der Bestellungen“ und „Länge der Spaghetti“?

3.9 Streuung

Während Mittelwerte typische Werte einer Stichprobe repräsentieren, sollen Streuungsmaße anzeigen, ob die Merkmalswerte dicht beim Mittelwert liegen oder sich in mehr oder weniger großen Abständen davon befinden. Abstände können aber nur für kardinale Merkmale gemessen werden. Daher sind Streuungsmaße vor allem für kardinale Merkmale sinnvoll zu definieren.

Die beiden folgenden Diagramme zeigen zwei relative Häufigkeitsverteilungen, die dasselbe arithmetische Mittel aufweisen, sich aber in der Streuung der Werte um den Mittelwert $\bar{x} = 7$ unterscheiden: Bei der unteren Verteilung streuen die Werte stärker als bei der oberen.

Fig. 6: Vergleich der Streuungen zweier Verteilungen mit demselben arithmetischen Mittel.



Im folgenden werden Maßzahlen diskutiert, welche die Streuung unter den Merkmalswerten eines kardinalen Merkmals mehr oder weniger gut reflektieren. Das primitivste Streuungsmaß ist die sogenannte *Spannweite*.

3.10 Spannweite

Die Spannweite s_W gibt die Differenz zwischen kleinstem und größtem Merkmalswert einer Stichprobe $(x_1, x_2, ..., x_n)$ eines kardinalen Merkmals X an:

$$s_W := \max\{x_1, x_2, ..., x_n\} - \min\{x_1, x_2, ..., x_n\}. \tag{29}$$

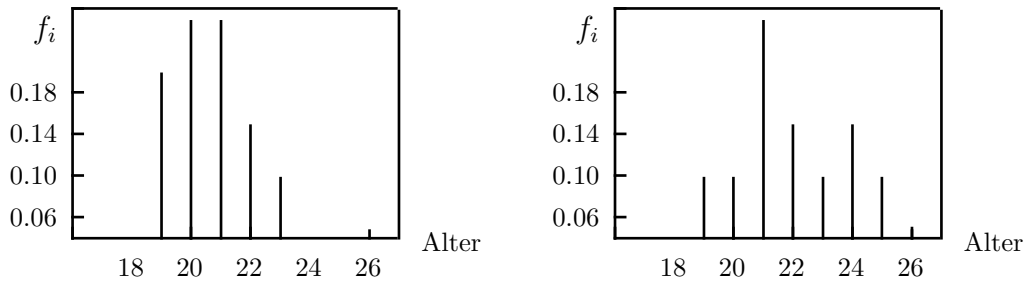
Bei klassierten Daten wird die Spannweite s_W als Differenz zwischen der oberen Klassengrenze x_k^o der obersten von k Klassen und der unteren Klassengrenze x_1^u der ersten Klasse definiert:

$$s_W := x_k^o - x_1^u. \tag{30}$$

Eigenschaften der Spannweite:

- Die Aussagekraft der Spannweite für eine Stichprobe $(x_1, x_2, ..., x_n)$ ist sehr eingeschränkt, da s_W nur aus **zwei** Werten der Stichprobe berechnet wird: Zugunsten einer einfachen Berechnung ist ein hoher Informationsverlust in Kauf zu nehmen. Die Spannweite gibt daher lediglich die Größe des Bereichs an, aus der 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.

Fig. 7: Stabdiagramme zweier Altersverteilungen gleicher Spannweite.



Das linke Stabdiagramm zeigt die relativen Häufigkeiten f_i mit denen verschiedene Lebensalter im BA-Kurs aus Beispiel 2.2 auftreten, während das rechte Diagramm eine davon verschiedene Altersverteilung jedoch mit derselben Spannweite von s_W = 26 - 19 = 7 Jahren zeigt. Ohne den Ausreißer von 26 Jahren bei der linken Verteilung betrüge die Spannweite nur 4 Jahre, was die geringere Streuung der Lebensalter dieser Personen besser wiedergeben würde.

- Spannweiten verschiedener Stichproben unterschiedlichen Umfangs können nicht miteinander verglichen werden, da bei der Berechnung der Spannweite nicht die Größe des Stichprobenumfangs berücksichtigt wird.

Anstatt mittels der Spannweite liegt es nahe, die Streuung unter den Merkmalswerten einer Stichprobe $(x_1, x_2, ..., x_n)$ eines kardinalen Merkmals X mittels der Summe der Abweichungen der Einzelwerte x_i von einem Mittelwert, zum Beispiel dem arithmetischen Mittel $\bar{x}$, messen zu wollen:

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

Diese Summe ist jedoch für beliebige Stichproben $(x_1, x_2, ..., x_n)$ stets Null – dies ist durch die Definition des arithmetischen Mittels und der daraus resultierenden Schwerpunktseigenschaft (4) impliziert: Positive und negative Abweichungen $x_i - \bar{x} > 0$ bzw. $x_j - \bar{x} < 0$ der Einzelwerte vom arithmetischen Mittel heben sich in der Summe gegenseitig exakt auf.

3.11 Mittlere absolute Abweichungen

Um zu verhindern, dass sich positive und negative Abweichungen der Einzelwerte von einem Mittelwert gegenseitig aufheben, können die Absolutbeträge dieser Abweichungen

benutzt werden. Dies geschieht bei der *mittleren absoluten Abweichung* der Merkmalswerte einer Stichprobe $(x_1, x_2, ..., x_n)$ eines kardinalen Merkmals X vom *arithmetischen Mittel*,

$$d_{\bar{x}} := \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|, \tag{31}$$

bzw. bei der *mittleren absoluten Abweichung* der Einzelwerte vom *Median*:

$$d_{\bar{x}_Z} := \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}_Z|. \tag{32}$$

Das letztere Maß wird häufig bevorzugt, weil der Median $\bar{x}_Z$ die Summe der absoluten Abweichungen der einzelnen Werte x_i von ihm selbst minimiert (siehe Eigenschaften des Medians).

Die Werte beider Maße können wegen der Beträge nicht negativ werden. Sie nehmen beide nur dann den Wert Null an, wenn alle Einzelwerte identisch sind:

$$x_1 = x_2 = ... = x_n.$$

In diesem Fall gibt es tatsächlich keine Streuung unter den Merkmalswerten, was durch den Wert Null beider Streuungsmaße dokumentiert wird: Das arithmetische Mittel sowie der Median stimmen in diesem Falle mit dem einen, n-fach auftretenden Merkmalswert überein, so dass alle Differenzen in den beiden Definitionen (31) und (32) verschwinden und damit beide Maße den Wert Null aufweisen.

Beispiel: Berechnung der mittleren absoluten Abweichungen vom Median und arithmetischen Mittel.
Der Median der Altersverteilung des BA-Kurses aus Beispiel 2.2 beträgt 21 Jahre. Die mittlere absolute Abweichung der Lebensalter vom Median lautet

$$d_{\bar{x}_Z} = \frac{1}{20}(|26-21| + 2 \cdot |23-21| + 3 \cdot |22-21| + 5 \cdot |20-21| + 4 \cdot |19-21|) = \frac{1}{20} \cdot 25 = 5/4,$$

also 1.25 Jahre. Die mittlere absolute Abweichung der Lebensalter vom arithmetischen Mittel $\bar{x} = 20{,}95$ ist mit 1,255 Jahren sehr ähnlich, aber leicht größer:

$$\begin{aligned} d_{\bar{x}} &= \frac{1}{20}(|26 - 20,95| + 2 \cdot |23 - 20,95| + 3 \cdot |22 - 20,95| \\ &\quad +5 \cdot |21 - 20,95| + 5 \cdot |20 - 20,95| + 4 \cdot |19 - 20,95|) = 1,255. \end{aligned}$$

Sind für die $m < n$ tatsächlich verschiedenen Einzelwerte $(x_1, x_2, ..., x_m)$ die absoluten bzw. relativen Häufigkeiten $(h_1, h_2, ..., h_m)$ bzw. $(f_1, f_2, ..., f_m)$ gegeben, können die mittleren absoluten Abweichungen mittels

$$d_{\bar{x}} = \frac{1}{n} \sum_{i=1}^{m} |x_i - \bar{x}| \cdot h_i \quad \text{bzw.} \quad d_{\bar{x}} = \sum_{i=1}^{m} |x_i - \bar{x}| \cdot f_i \tag{33}$$

und

$$d_{\bar{x}_Z} = \frac{1}{n}\sum_{i=1}^{m}|x_i - \bar{x}_Z| \cdot h_i \quad \text{bzw.} \quad d_{\bar{x}_Z} = \sum_{i=1}^{m}|x_i - \bar{x}_Z| \cdot f_i \tag{34}$$

berechnet werden.

Zur Berechnung sowohl der mittleren absoluten Abweichungen vom arithmetischen Mittel wie auch vom Median werden **alle** Werte der Stichprobe $(x_1, x_2, ..., x_n)$ verwendet. Bei der Berechnung beider Maße wird also keine Information mißachtet, wie das bei der Berechnung der Spannweite der Fall ist. Außerdem berücksichtigt ein zum Vergleich der Streuungen verschiedener Stichproben konzipiertes Streuungsmaß sinnvollerweise den Umfang der Stichprobe. Das ist bei den beiden Maßen (31) und (32) der Fall. Das weitaus am meisten verwendete Streuungsmaß ist allerdings die *empirische Standardabweichung*. Sie basiert ganz wesentlich auf der *empirische Varianz*.

3.12 Empirische Varianz und empirische Standardabweichung

Definition der empirischen Varianz

Die empirischen Varianz s^2 einer Stichprobe $(x_1, x_2, ..., x_n)$ von Einzelwerten eines kardinalen Merkmals X ist die *durchschnittliche quadratische Abweichung* vom arithmetischen Mittel $\bar{x}$:

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{35}$$

Oftmals einfacher zu berechnen – vor allem, wenn dies ohne Taschenrechner oder Computer geschehen muss – ist die empirische Varianz nach der Formel

$$s^2 = (\frac{1}{n}\sum_{i=1}^{n}x_i^2) - \bar{x}^2. \tag{36}$$

Beide Formeln (35) und (36) sind äquivalent: Aus (35) folgt mit Hilfe der 2. binomischen Formel

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2 = \frac{1}{n}\sum_{i=1}^{n}[x_i^2 - 2 \cdot x_i \cdot \bar{x} + \bar{x}^2] = (\frac{1}{n}\sum_{i=1}^{n}x_i^2) - 2 \cdot \bar{x} \cdot (\frac{1}{n}\sum_{i=1}^{n}x_i) + \frac{1}{n} \cdot n \cdot \bar{x}^2,$$

wobei wegen $1/n \sum x_i = \bar{x}$ der mittlere Term $-2 \cdot \bar{x}^2$ ergibt, während der letzte Term sich auf $\bar{x}^2$ reduziert. Der mittlere und der letzte Term können addiert werden, so dass als Ergebnis die Formel (36) für die empirische Varianz resultiert.

Die empirische Varianz s^2 nimmt genau dann den Wert Null an, wenn jede einzelne quadratische Differenz 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 Wirklichkeit 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 die Situation, bei der es keine Streuung

unter den Merkmalswerten gibt. Dies stellt aber eine höchst seltene und unter statistischen Gesichtspunkten eher langweilige Situation dar.

Sind für die $m < n$ tatsächlich verschiedenen Einzelwerte $(x_1, x_2, ..., x_m)$ die absoluten bzw. relativen Häufigkeiten $(h_1, h_2, ..., h_m)$ bzw. $(f_1, f_2, ..., f_m)$ gegeben, kann die empirische Varianz mittels

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

bzw.

$$s^2 := \sum_{i=1}^{m}(x_i - \bar{x})^2 \cdot f_i \tag{38}$$

berechnet werden.

Würde mit dem kardinalen Merkmal X zum Beispiel die Körpergröße in m bezeichnet werden, 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 aber die Wurzel aus der empirischen Varianz ein interpretierbares Maß dar, das *empirische Standardabweichung* genannt wird.

Beispiel: Berechnung der empirischen Varianz und Standardabweichung.
Das arithmetische Mittel der Altersverteilung des BA-Kurses aus Beispiel 2.2 beträgt $\bar{x}$ = 20,95 Jahre. Mit Hilfe von Formel (35) errechnet sich die empirischen Varianz s^2 zu

$$\begin{aligned} s^2 &= \frac{1}{20}[(26 - 20,95)^2 + 2 \cdot (23 - 20,95)^2 + 3 \cdot (22 - 20,95)^2 \\ &\quad +5 \cdot (21 - 20,95)^2 + 5 \cdot (20 - 20,95)^2 + 4 \cdot (19 - 20,95)^2] = 2,8475. \end{aligned}$$

Zur Überprüfung dieses Wertes wird die Varianz nach der alternativen Formel (36) berechnet:

$$\begin{aligned} s^2 &= \frac{1}{20}(26^2 + 2 \cdot 23^2 + 3 \cdot 22^2 + 5 \cdot 20^2 + 5 \cdot 20^2 + 4 \cdot 19^2) - 20,95^2 \\ &= \frac{1}{20} \cdot 8,835 - 20,95^2 = 441,75 - 20,95^2 = 2,8475. \end{aligned}$$

Die empirische Varianz beträgt also 2.8475 Jahre im Quadrat, was immer dieser Wert aussagen soll. Interpretierbar ist dagegen die Wurzel daraus: Die empirische Standardabweichung, also die Wurzel aus der durchschnittlichen quadratischen Abweichung der einzelnen Lebensalter vom durchschnittlichen Lebensalter von $\bar{x} = 20,95$ Jahren, beträgt $s = 1.69$, sprich rund 2 Jahre.

Definition der empirischen Standardabweichung

Die empirische Standardabweichung s ist definiert als Wurzel aus der empirischen Varianz:

$$s := \sqrt{s^2}. \tag{39}$$

Beispiel: Lineare Transformation der Körpergröße von Zoll in Zentimeter.
Bezeichnet Y die Körpergröße in Zentimeter und X die Körpergröße in Zoll, so lautet eine vereinfachte Transformationsformel, die Größenangaben von Zoll in Zentimeter umzurechnen gestattet:

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

Das Absolutglied a der allgemeinen linearen Transformation $Y = a + b \cdot X$ verschwindet also in diesem Falle, während $b = 2,5$ ist.

5 Kinder haben Körpergrößen wie sie in der folgenden Tabelle aufgelistet sind:

Y [in cm]	120	130	125	130	135
X [in Zoll]	48	52	50	52	54

Das arithmetisches Mittel der Körpergrößen dieser 5 Kinder in Zoll beträgt $\bar{x} = 51,2$ bei einer empirischen Standardabweichung von $s_X = 2,04$ Zoll. Ohne mit Hilfe der einzelnen Körpergrößen in cm das arithmetische Mittel $\bar{y}$ bzw. die Standardabweichung s_Y ausrechnen zu müssen, erhält man beides durch Anwendung der entsprechenden Transformationsformeln (7) und (41) aus den bereits bekannten Werten $\bar{x}$ und s_X:

$$\bar{y} = a + b \cdot \bar{x} = 2,5 \cdot 51,2 = 128 \qquad \text{bzw.} \qquad s_Y = |b| \cdot s_X = 2,5 \cdot 2,03961 = 5,1.$$

Mit der Definition von s_Y hätte man den Wert 5,1 cm umständlicher erhalten:

$$s_Y = \sqrt{\frac{1}{5}[(120-128)^2 + (130-128)^2 + (125-128)^2 + (130-128)^2 + (135-128)^2]}.$$

Eigenschaften von empirischer Varianz und Standardabweichung

- *Nicht-Negativität*:

  Sowohl die empirische Varianz als Summe lauter quadratischer Terme wie auch die Wurzel daraus, die empirische Standardabweichung, sind **immer 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. Nomalerweise sind indes die Werte der empirischen Varianz und Standardabweichung von Null verschieden, weil üblicherweise eine gewisse Streuung unter den Merkmalswerten einer Stichprobe vorhanden ist.

- *Transformationseigenschaften* von empirischer Varianz und Standardabweichung:

  Geht ein kardinales Merkmal Y durch eine allgemeine lineare Transformation aus einem kardinalen Merkmal X hervor, dessen arithmetisches Mittel $\bar{x}$ und dessen empirische Varianz s_X^2 bekannt sind, so berechnet sich die empirische Varianz s_Y^2 des

Merkmals Y aus der empirischen Varianz s_X^2 des Merkmals X, ohne dass vorher jeder einzelne Wert x_i einer Stichprobe $(x_1, x_2, ..., x_n)$ durch die lineare Transformation $y_i = a + b \cdot x_i$ in genau einen Merkmalswert des Merkmals Y umgerechnet wird:

$$s_Y^2 = b^2 \cdot s_X^2. \tag{40}$$

Das läßt sich leicht zeigen: Für die Stichprobe $(y_1, y_2, ..., y_n)$ ergibt sich nach der linearen

Transformation die folgende empirische Varianz s_Y^2,

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

wobei $\bar{y} = a + b \cdot \bar{x}$, die Transformationseigenschaft (7) für das arithmetische Mittel, und die Definition (35) der empirischen Varianz verwendet wurde. Aus (40) ergibt sich durch Wurzelziehen:

$$s_Y = \sqrt{s_Y^2} = \sqrt{b^2 \cdot s_X^2} = \sqrt{b^2} \cdot \sqrt{s_X^2} = |b| \cdot s_X. \tag{41}$$

3.13 Variationskoeffizient

Die empirische Standardabweichung ist, ebenso wie die Spannweite und die mittleren absoluten Abweichungen vom arithmetischen Mittel oder vom Median, ein Maß für die *absolute Streuung*. Diese sind im Allgemeinen *dimensionsbehaftete* Maße, die von der Einheit abhängen, in der ein Merkmal gemessen wird. Relative Streuungsmaße sind dagegen dimensionslos. Ein Beispiel eines solchen relativen Maßes ist der sogenannte *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

$$v_X := \frac{s_X}{\bar{x}}. \tag{42}$$

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

Beispiel: Körpergröße in Zoll und in Zentimeter.

Obwohl die Streuung der Körpergrößen der 5 Kinder des obigen Beispiels natürlich dieselbe ist, gleich ob die Körpergröße in Zoll oder in Zentimetern gemessen wird, signalisieren die unterschiedlichen Werte $s_X = 2,04$ bzw. $s_Y = 5,1$ der empirischen Standardabweichungen etwas anderes. Das liegt aber schlicht daran, dass die empirische Standardabweichung ein Maß für die absolute Streuung ist, dessen Wert offensichtlich von der Einheit abhängt, in welcher das zugehörige Merkmal „Körpergröße“ gemessen wird. Der Wert $s_Y = 5,1$ der empirischen Standardabweichung der Körpergröße in *cm* gemessen ist größer als jener, der in Zoll gemessenen Körpergrößen, weil die Körpergrößen in *cm* um ca. den Faktor 2,5 größere Zahlenwerte aufweisen. Dieser Skaleneffekt wird in der Statistik dadurch berücksichtigt, dass anstatt absoluter, relative Streuungsmaße verwendet werden.

Relative Streuungsmaße setzen allgemein absolute Streuungsmaße ins Verhältnis zum durchschnittlichen Niveau, welches ein Merkmal aufweist. Ein Beispiel dieser Art ist der Variationskoeffizient. Er ist in diesem Beispiel identisch für beide Merkmale X und Y, da es sich bei beiden Merkmalen um denselben Sachverhalt handelt:

$$v_X = \frac{s_X}{\bar{x}} = \frac{2,04}{51,2} = 0,04 = 4\%, \qquad v_Y = \frac{s_Y}{\bar{y}} = \frac{5,1}{128} = 0,04 = 4\%.$$

Diese Zahl besagt nichts anderes als: Die empirische Standardabweichung der Körpergrößen der 5 Kinder beträgt 4 % des Mittelwertes, gleichgültig ob die Körpergröße in Zoll wie beim Merkmal X oder in Zentimeter wie beim Merkmal Y gemessen wird.

3.14 Schiefe

Ob eine Verteilung als *symmetrisch* oder unsymmetrisch bzw. *schief* zu bezeichnen ist, läßt sich anhand von Stab- und Balken-Diagrammen bzw. Histogrammen sehr leicht erkennen. Beispielsweise sind die beiden in Figur 6 dargestellten Verteilungen symmetrisch. Die in Figur 1 gezeigte Altersverteilung wird als schief und insbesondere als *linkssteil* oder synonym als *rechtsschief* bezeichnet.

Eine einfache Möglichkeit zur Einschätzung der Schiefe *eingipfliger* Verteilungen, welche durch die Eindeutigkeit des Modus charakterisiert sind, bietet die folgende Faustregel bezüglich des arithmetisches Mittels $\bar{x}$, des Medians $\bar{x}_Z$ und des Modus $\bar{x}_M$:

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 sogar allgemein gültigen Regel $\bar{x} = \bar{x}_Z = \bar{x}_M$ bei symmetrischen, eingipfligen Verteilungen sind bei Kenntnis des arithmetischen Mittels auch Modus und Median bekannt.

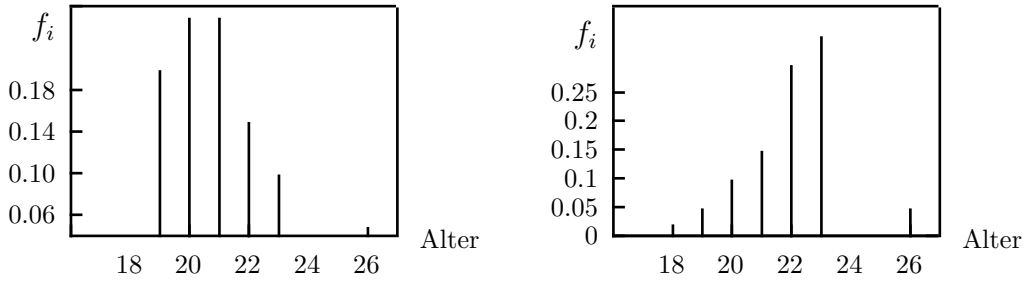
Linkssteile bzw. rechtsschiefe Verteilungen sind von erheblicher empirischer Bedeutung. Beispielweise sind Verteilungen, welche die Einkommens- und Vermögensverhältnisse verschiedener Länder wiedergeben, typischerweise linkssteil. Charakteristisch für diese Verteilungen ist insbesondere, dass der Median $\bar{x}_Z$ 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 dadurch charakterisiert, dass der Median $\bar{x}_Z$ größer als das arithmetische Mittel $\bar{x}$ ist.

Beispiel: Links- und rechtsschiefe Verteilungen.
Für die nicht sehr ausgeprägte linkssteile (=rechtsschiefe) Altersverteilung aus Beispiel 2.2, welche in Figur 8 dargestellt ist, ist bestenfalls eine tendenzielle Bestätigung der obigen Faustregel zu erkennen:

$$\bar{x} = 20,95, \bar{x}_Z = 21, \bar{x}_M = 20 \text{ bzw. } 21 \ .$$

Eigentlich verlangt die FECHNERsche *Lageregel* bei linkssteilen Verteilungen, dass das arithmetische Mittel größer als der Median ist. Das trifft in diesem Falle allerdings nicht zu. Außerdem ist der Modus ist nicht eindeutig, wie das bei ausgeprägten, eingipfligen Verteilungen der Fall wäre. In Figur 8 ist schließlich noch eine rechtssteile (=linksschiefe) Verteilung dargestellt.

Fig. 8: Stabdiagramme links- und rechtssteiler Altersverteilungen.



Maßzahlen zur Quantifizierung der Schiefe einer Verteilung stützen sich auf das dritte zentrale Moment. Richtig gelesen: Auch Statistiker haben ihre Momente! Während auf Maßzahlen zur Schiefe nicht weiter eingegangen werden muß, verdienen die Begriffe *gewöhnliche* und *zentrale Momente* allerdings noch einen Augenblick der Aufmerksamkeit.

3.15 Statistische Momente

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

$$m_r(a) := \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r. \tag{43}$$

Beim Bezugspunkt $a = 0$ spricht man vom *gewöhnlichen Moment* r-ter Ordnung:

$$m_r(0) := \frac{1}{n}\sum_{i=1}^{n} x_i^r \ . \tag{44}$$

Das arithmetische Mittel $\bar{x}$ kann hiernach als ein spezielles gewöhnliches Moment eingeordnet werden – es ist das gewöhnliche Moment 1. Ordnung:

$$m_1(0) := \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}. \tag{45}$$

Von allen gewöhnlichen Momenten hat das arithmetische Mittel die größte Bedeutung.

Stellt das arithmetische Mittel $\bar{x}$ den Bezugspunkt a dar, spricht man vom *zentralen Moment* r-ter Ordnung:

$$m_r(\bar{x}) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r. \tag{46}$$

Die empirische Varianz s^2 entpuppt sich gemäß dieser Einordnung als das zweite zentrale Moment:

$$m_2(\bar{x}) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2. \tag{47}$$

Von allen zentralen Momenten hat das zweite zentrale Moment, das heißt die empirische Varianz s^2, die größte Bedeutung.

3.16 Zusammenfassung zu Streuungs- und Schiefemaßen

Streuungsmaße sollen Aussagen über die Variabilität von Merkmalswerten innerhalb einer Stichprobe treffen. Die Frage, welches Streuungsmaß im Einzelnen heranzuziehen ist, läßt sich pauschal nicht beantworten. Prinzipiell wird unterschieden zwischen *absoluten Streuungsmaßen*, beispielsweise der Spannweite, den mittleren absoluten Abweichungen vom Median bzw. dem arithmetischen Mittel oder der empirischen Varianz bzw. Standardabweichung, und *relativen Streuungsmaßen*, zum Beispiel dem Variationskoeffizienten.

Während absolute Streuungsmaße im Allgemeinen dimensions- und einheitenbehaftet sind, besitzen relative Streuungsmaße weder Dimension noch Einheit, da bei diesen immer ein absolutes Streuungsmaß in Beziehung zu einem Lageparameter gesetzt wird. Beim Variationskoeffizienten beispielsweise wird das Verhältnis von empirischer Standardabweichung zu 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 eine falschen Eindruck vermittelt.

Allgemein gilt die folgende Kette von Ungleichungen zwischen mittlerer absoluter Abweichung vom Median, welche die Minimumseigenschaft 23 besitzt, der empirischen Standardabweichung und der Spannweite:

$$d_{\bar{x}Z} \leq s \leq s_W. \tag{48}$$

Hieraus ergibt sich ein schnell zu überprüfender Hinweis, ob die empirische 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 es keine Streuung unter den Merkmalswerten gibt, das heißt alle Beobachtungen ein und denselben Wert haben. Dies kommt indes in der Praxis normalerweise nicht vor. Üblicherweise weisen die Beobachtungen in einer Stichprobe mehr oder weniger stark voneinander abweichende Werte auf: Diese Streuung der Werte wird je nach Maß quantifiziert durch die absoluten oder quadratischen Abstände der Einzelwerte von einem Bezugspunkt wie dem arithmetischen Mittel oder dem Median.

Einfache Differenzen der Merkmalswerte speziell vom arithmetischen Mittel sind hingegen nicht geeignet: Die Summe der einfachen Abweichungen der Merkmalswerte vom arithmetischen Mittel sind per Definition des arithmetischen Mittels immer Null:

$$\sum^{n}(x_i - \bar{x}) = 0.$$

Streng genommen setzt die Berechnung von Streuungsmaßen die Kardinalität von Merkmalen voraus. Bei ordinalen Merkmalen versucht man sich mit „Krücken“ wie zum Beispiel dem *Quartilsabstand* zu behelfen, also dem „Abstand“ zwischen oberen und unterem Quartil. Bei kardinalen Merkmalen hat der Quartilsabstand gegenüber der Spannweite den Vorteil, nicht anfällig gegenüber Ausreißern zu sein.

3.17 Übungsaufgaben zu Streuungs- und Schiefemaßen

Übungsaufgabe: 18 % – Illusionen einer kleinen Partei.

Landtagswahlen in 7 Bundesländern brachten den Parteien A und B folgende Prozentzahlen:

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 Parteiergebnisse der Parteien A bzw. B lauten $\overline{x} = 6,6$ % bzw. $\overline{y} = 43,4$ %. Silversterwelle, der Sprecher der Partei A, die sich des geringeren

Wählerzuspruchs gegenübersieht, behauptet trotzig: „Unser Ergebnis ist in allen Ländern ziemlich gleich, während die Wahlergebnisse der Partei B wesentlich weniger stabil sind.“ Überprüfen Sie diese Behauptung.

Übungsaufgabe: Lohnerhöhung ohne Gewerkschaften und Betriebsrat?

Die 200 Beschäftigten eines Software-Unternehmens erhalten einen monatlichen Durchschnittslohn von $\bar{x} = 2.200$ Euro bei einer Standardabweichung $s_X = 800$ Euro. Aufgrund der boomenden Wirtschaft ist eine Lohnverhandlung der weder gewerkschaftlich organisierten, noch durch einem Betriebsrat vertretenen Beschäftigten von Erfolg gekrönt: Das Monatsgehalt jedes Beschäftigten soll für die Dauer des Booms um 10 % angehoben werden und es soll in Zukunft jedes Jahr 960 Euro Urlaubsgeld gewährt werden. Wie ändern sich Mittelwert, empirische Standardabweichung und Varianz der Gehälter der Beschäftigten sowie der Varaitionskoeffizient?

Übungsaufgabe: Lage- und Streuungsparameter.

Man bestimme für die folgende relative Häufigkeitsverteilung eines kardinalen Merkmals X sämtliche hier diskutierten Lage- und Streuungsmaße.

Merkmalsausprägung	0	2	4	6	8	10
relative Häufigkeit	0,15	0,20	0,05	0,30	0,20	0,10

Übungsaufgabe: Variation des Bruttomonatsverdienstes männlicher Angestellter vor der Wiedervereinigung.

Für die 11 „alten“ Bundesländer berechne man den durchschnittlichen Bruttomonatsverdienst männlicher Angestellter im Jahre 1989 in Form des arithmetischen Mittels sowie die empirische Varianz und Standardabweichung.

Bundesland	Verdienst	Bundesland	Verdienst
Schleswig Holstein	3.986	Berlin (West)	4.348
Niedersachen	4.081	Nordrhein Westfalen	4.408
Saarland	4.158	Hessen	4.428
Bayern	4.246	Baden-Württemberg	4.509
Bremen	4.254	Hamburg	4.766
Rheinland-Pfalz	4.285		
Quelle: Statistisches Jahrbuch 1989, Seite 490.			

Übungsaufgabe: Der Variationskoeffizient und linearen Transformationen.

Das Merkmal X besitze das arithmetische Mittel $\bar{x}$ und die Standardabweichung s_X. Zeigen Sie, dass der Variationskoeffizient v_Y für das Merkmal Y, das aus X und der linearen Transformation $Y = b \cdot X$ resultiert, 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 Konzentration und Disparität

Konzentration im ökonomischen Sinne kann zweierlei bedeuten:

1. Die Konzentration von beispielsweise Marktanteilen, also von ökonomischer Macht, auf einzig und allein eine Wirtschaftseinheit (Monopol) oder nur auf einige wenige Wirtschaftseinheiten (Oligopol).

2. Die Existenz erheblicher Unterschiede zwischen Wirtschaftseinheiten bezüglich deren Anteile am Gesamtbetrag eines relevanten Merkmals wie beispielsweise Umsatz.

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

Beispiele zur absoluten und relativen Konzentration

Eine Aussage im Sinne der relativen Konzentration wäre beispielsweise: 2 % der Bevölkerung lateinamerikanischer Staaten haben mehr als 90 % des Geldvermögens dieser Staaten. In dieser Aussage tauchen ausschließlich relative Werte – angegeben in Prozenten – auf: Diese relativen Werte geben den Anteil am Gesamtwert des betrachteten Merkmals – hier des Geldvermögens – an, den ein gewisser Anteil von Merkmalsträgern aufweist.

Eine Aussage im Sinne der absoluten Konzentration wäre hingegen: Auf dem deutschen Energiemarkt haben nur 2 Konzerne zusammen einen Marktanteil von ca. 80 %. Die Merkmalsträger sind in absoluter Anzahl angegeben, deren Zahl zudem sehr gering ist.

Der Unterschied zwischen absoluter und relativer Konzentration wird besonders deutlich bei der sogenannten *Gleichverteilung*, bei welcher der Gesamtwert eines Merkmals wie beispielsweise Geldvermögen völlig gleichmäßig auf alle Merkmalsträger verteilt ist. Unabhängig davon wie groß die Zahl der Merkmalsträger ist, existiert bei völliger Gleichverteilung per definitionem keine relative Konzentration. Je kleiner indes dabei die Zahl der Merkmalsträger ist, desto größer ist die absolute Konzentration.

Der nächste Abschnitt beginnt mit der Herleitung des wohl bekanntesten Ungleichheitsmaßes, dem Gini-*Koeffizienten*. Dessen Interpretation basiert auf der sogenannten *Lorenzkurve*, mit deren Hilfe Ungleichheit-Situationen illustriert werden. Danach wird der Herfindahl-Index[2] diskutiert, welcher das populärste Maß zur Messung der absoluten Konzentration darstellt. Statistische Maße zur Messung relativer Konzentration berücksichtigen nur den Aspekt der Ungleichheit (Disparität), wohingegen Maße zur Messung der absoluten Konzentration beide Aspekte erfassen, den Aspekt der absoluten Anzahl sowie den der Disparität.

[2]Dieser Index ist benannt nach dem Energie-Ökonom Orris C. Herfindahl, 1918-1972.

4.1 Lorenzkurve

Mit dem Begriff der Ungleichheit bzw. der relativen Konzentration ist die Frage verbunden, ob ein *großer Anteil* am Gesamtwert eines Merkmals wie beispielsweise dem Weltenergieverbrauch, um den es im folgenden Beispiel geht, auf einen *geringen Anteil* aller Merkmalsträger entfällt.

Beispiel: Stilisierte Fakten zum Weltenergieverbrauch.
Die Weltbevölkerung (WB) 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 Weltenergieverbrauch (WEV) teilt sich auf diese drei „Welten" in etwa wie folgt auf:

	Anteil an der WB f_i	kum. Anteil an der WB F_i	Anteil am WEV q_i	kum. Anteil am WEV 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 Weltbevölkerung von etwa 10 % der in den Industrieländern lebenden Menschen beanspruchen demnach einen großen Anteil von ca. 60 % der weltweit jährlich verbrauchten Energie, während ca. 60 % der Weltbevölkerung mit nur ca. 10 % auskommt.

Die Eckpunkte $(F_i;Q_i)$, mit deren Hilfe eine *Lorenzkurve* L gezeichnet wird, werden allgemein gebildet aus den *kumulierten* relativen Anteilen

$$F_i := \sum_{k=1}^{i} f_k = f_1 + f_2 + ... + f_i, \tag{49}$$

einer bestimmten Gruppe von Merkmalsträgern an der Grundgesamtheit (Population) – im obigen Beispiel die Weltbevölkerung – und derem *kumulierten* Anteil

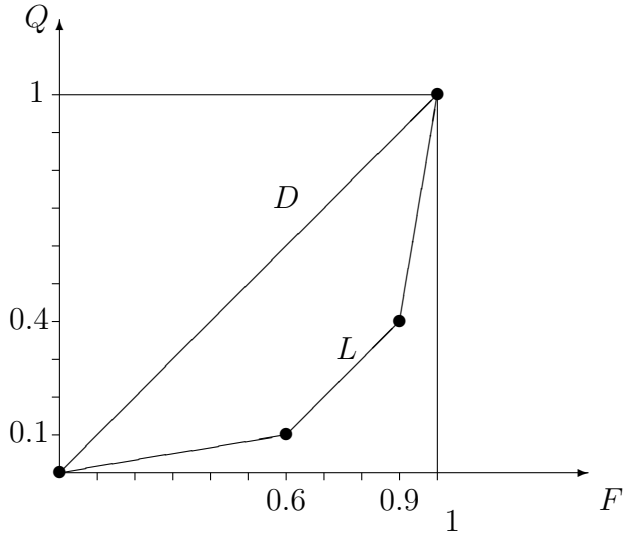
$$Q_i := \sum_{k=1}^{i} q_k = q_1 + q_2 + ... + q_i \tag{50}$$

am Gesamtwert des betrachteten Merkmals. Ergänzt werden diese Punkte noch um den Ursprung $(0;0) = (F_0;Q_0)$, welcher den Ausgangspunkt der Lorenzkurve bildet. Die Lorenzkurve 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 Weltenergieverbrauchs.

Für das obige Beispiel des stilisierten Weltenergieverbrauchs 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 diese Eckpunkte durch Geraden.

Fig. 9: Die Lorenzkurve zur Illustration des stilisierten Weltenergieverbrauchs.



Die Diagonale D ist diejenige Kurve, welche den Zustand der Gleichverteilung dargestellt. Je stärker eine Lorenzkurve L, welche die Ungleichheit bezüglich eines betrachteten Merkmals grafisch veranschaulichen soll, von der Diagonalen D abweicht, desto größer ist die Ungleichheit und desto stärker ist – mit anderen Worten – die relative Konzentration bezüglich dieses Merkmals innerhalb der betrachteten Grundgesamtheit (Population).

4.2 Gini-Koeffizient

Ein Maß für die Abweichung der Lorenzkurve L von der Diagonalen D – salopp gesprochen für den Bauch der Lorenzkurve – ist der sogenannte Gini-*Koeffizient*. Im extremen Grenzfall, welcher in Realität allerdings nicht auftreten kann, entspricht dieser Bauch gerade der kompletten Fläche unter der Diagonalen, und damit der Fläche eines Dreiecks.

Genau genommen mißt der Gini-Koeffizient G die Fläche zwischen Diagonale D und Lorenzkurve L und setzt sie ins Verhältnis zur Fläche des Dreiecks unter der Diagonalen, das eine Fläche von 1/2 aufweist:

$$\begin{aligned} G &:= \frac{\text{Fläche zwischen } D \text{ und } L}{\text{Dreiecksfläche unter } D} = \frac{\text{Fläche zwischen } D \text{ und } L}{1/2} \\ &= 2 \cdot \text{Fläche zwischen } D \text{ und } L. \end{aligned} \tag{51}$$

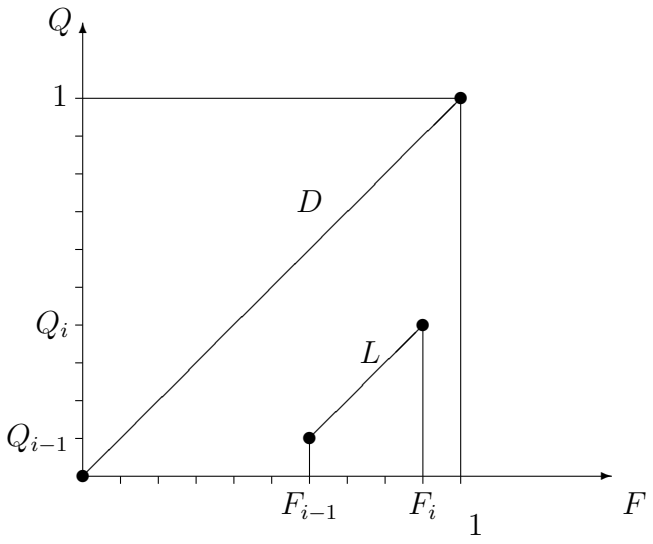
Bei dieser Definition bewirkt die Division durch die Zahl 1/2, dass der Wert des GINI-Koeffizienten immer unterhalb der Zahl 1 bleibt.

Im Falle einer völligen Gleichverteilung, bei dem in anderen Worten überhaupt keine Tendenz zur Konzentration vorliegt, weicht die Lorenzkurve nicht von der Diagonalen ab. Daher ist die Fläche zwischen beiden Null – der Wert des GINI-Koeffizienten G beträgt ebenfalls 0. Im konträren Fall einer extremen Ungleichverteilung kommt diese Fläche zwischen D und L der Dreicksfläche unter der Diagonalen sehr nahe, ohne sie jemals zu erreichen. Insgesamt ergibt sich die folgende Bandbreite für den GINI-Koeffizienten:

$$0 \leq G < 1. \tag{52}$$

Die Fläche zwischen D und L gewinnt man aus der Subtraktion der Fläche des Dreiecks unter der Diagonalen D und der Summe der Flächen aller Trapeze, welche unterhalb der Lorenzkurve liegen. In Figur 10 ist ein Teil einer Lorenzkurve sowie ein Trapez exemplarisch dargestellt.

Fig. 10: Illustration der Herleitung einer Formel für den GINI-Koeffizienten.



Die Fläche des dargestellten Trapezes errechnet sich aus der Länge der Grundseite, $F_i - F_{i-1} = f_i$, multipliziert mit der durchschnittlichen Höhe, $(Q_{i-1} + Q_i)/2$, welche sich aus dem arithmetischen Mittel, der beiden Höhen Q_{i-1} und Q_i ergibt:

$$\begin{aligned} G &= 2 \cdot (\frac{1}{2} - \text{Summe der Flächen der Trapeze}) = 1 - 2 \cdot \sum_{i=1}^{n} f_i \frac{(Q_{i-1} + Q_i)}{2} \\ &= 1 - \sum_{i=1}^{n} f_i (Q_{i-1} + Q_i). \end{aligned} \tag{53}$$

Man beachte, dass $Q_0 = 0, F_0 = 0$ und $Q_n = 1, F_n = 1$.

Beispiel: Messung der Ungleichheit beim stilisierten Weltenergieverbrauch.
Für das Beispiel des stilisierten Weltenergieverbrauchs ergibt sich nach Formel (53) ein GINI-Koeffizient von

$$G = 1 - \sum_{i=1}^{n} f_i(Q_{i-1} + Q_i) = 1 - 0,6 \cdot 0,1 - 0,3 \cdot (0,1 + 0,4) - 0,1 \cdot (0,4 + 1) = 0,65 \ .$$

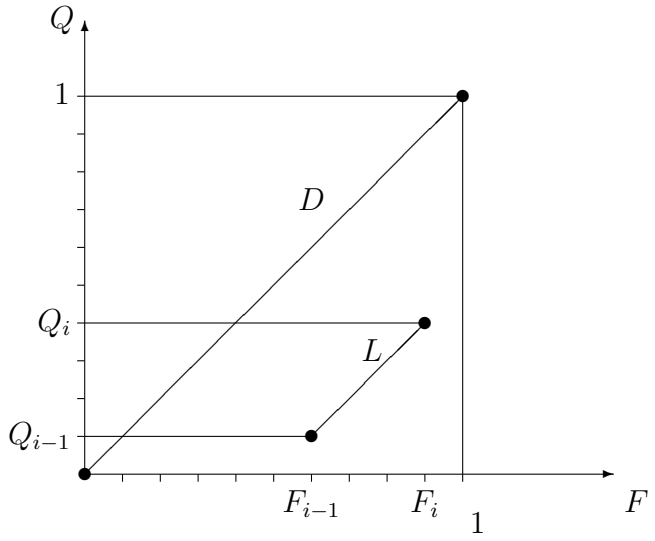
Wie ist diese Zahl zu bewerten und zu interpretieren? Angesichts der Bandbreite von $[0, 1)$ für den GINI-Koeffizienten, für den 1 die unerreichbare Obergrenze darstellt, kann von einer mittelstarken Konzentration des Weltenergieverbrauchs auf die Industrie- und Schwellenländer gesprochen werden.

Alternativ kann der GINI-Koeffizient G berechnet werden, indem von der Summe aller Trapez-Flächen die Fläche des Dreiecks oberhalb der Diagonalen D abgezogen wird (siehe Figur 11). Die Fläche des in Figur 11 dargestellten Trapezes errechnet sich aus der Länge der Grundseite, $Q_i - Q_{i-1} = q_i$, multipliziert mit der durchschnittlichen Höhe, $(F_{i-1} + F_i)/2$, welche sich aus dem arithmetischen Mittel der beiden Höhen F_{i-1} und F_i ergibt:

$$\begin{aligned} G &= 2 \cdot (\text{Summe der Flächen der Trapeze} - \frac{1}{2}) = 2 \cdot \sum_{i=1}^{n} q_i \frac{(F_{i-1} + F_i)}{2} - 1 \\ &= \sum_{i=1}^{n} q_i (F_{i-1} + F_i) - 1. \end{aligned} \tag{54}$$

Diese Formel ist der Vollständigkeit halber abgeleitet worden. Sie wird nicht benötigt: Zur Berechnung des GINI-Koeffizienten genügt einzig und allein Formel (53).

Fig. 11: Illustration zur Herleitung einer alternativen Formel für den GINI-Koeffizienten.



Beispiel: Messung der Marktmacht innerhalb einer Branche.
In einer Branche sind fünf Unternehmen $U_1, ..., U_5$ tätig, die in einem Jahr folgende Umsätze (in Millionen Euro) erzielt haben:

Unternehmen	U_1	U_2	U_3	U_4	U_5	Branche
Umsatz	600	1.500	900	1.800	1.200	6.000

Zur Ermittlung der Lorenzkurve zur graphischen Darstellung der Konzentration der Umsätze in dieser Branche wird die folgende Tabelle erstellt, in der die Unternehmen nach der Größe ihres Umsatzes geordnet sind – denn entscheidend bei der Ermittlung der Lorenzkurve ist, dass mit der „kleinsten“ statistischen Einheit, die hinsichtlich des Anteils am Gesamtwert des betrachteten Merkmals den kleinsten Wert aufweist, begonnen wird. Sukzessive werden dann die nächst „größeren“ statistischen Einheiten hinzugezogen.

Unternehmen	Umsatz	$f_i = 1/n$	$F_i = i \cdot 1/n$	Marktanteile q_i	kum. Marktanteile Q_i	$(F_i; Q_i)$
U_1 (i = 1)	600	20%	20%	10%	10%	(0,2; 0,10)
U_3 (i = 2)	900	20%	40%	15%	25%	(0,4; 0,25)
U_5 (i = 3)	1200	20%	60%	20%	45%	(0,6; 0,45)
U_2 (i = 4)	1500	20%	80%	25%	70%	(0,8; 0,70)
U_4 (i = 5)	1800	20%	100%	30%	100%	(1,0; 1,00)

Der Wert des GINI-Koeffizientes G, mit dem die Konzentration des Umsatzes in dieser Branche beurteilt werden soll, errechnet sich nach Formel (53) zu:

$$\begin{aligned} G &= 1 - \sum_{i=1}^{n} f_i(Q_{i-1} + Q_i) = 1 - 0,2 \cdot 0,1 - 0,2 \cdot (0,1 + 0,25) \\ &\quad -0,2 \cdot (0,25 + 0,45) - 0,2 \cdot (0,45 + 0,7) - 0,2 \cdot (0,7 + 1) \\ &= 1 - 0,2 \cdot 2 \cdot (0,1 + 0,25 + 0,45 + 0,7 + 1 - \frac{1}{2}) = 0,2 \ . \end{aligned}$$

Diese Berechnung suggeriert übrigens die alternative Formel $G = 1 - 1/n \cdot 2 \cdot (\sum Q_i - 1/2)$. Wie bereits betont, genügt allerdings zur Berechnung des GINI-Koeffizienten Formel (53) vollkommen.

Der Wert $G = 0,2$ deutet auf eine relativ geringe Konzentration des Umsatzes in dieser Branche hin. Allerdings muss berücksichtigt werden, dass der maximale Wert des GINI-Koeffizienten, der im folgenden Abschnitt hergeleitet wird, mit $G_{max}(5) = 0,8$ deutlich geringer als 1 ist. Diesen maximalen Wert würde man erhalten, wenn der gesamte Umsatz der Branche sich auf ein einziges Unternehmen konzentrierte, dieses also faktisch Monopolist in dieser Branche ist, während die anderen Unternehmen tatenlos zusehen und langfristig aus dem Markt ausscheiden müßten.

Aus Formel (54) ergäbe sich für n statistische Einheiten wegen $F_i = \sum_{k=1}^{i} f_k = i \cdot \frac{1}{n}$, $F_i + F_{i-1} = 2 \cdot i \cdot \frac{1}{n} - \frac{1}{n} = \frac{1}{n}(2i-1)$ und $\sum_{i=1}^{n} q_i = 1$ außerdem noch die Formel:

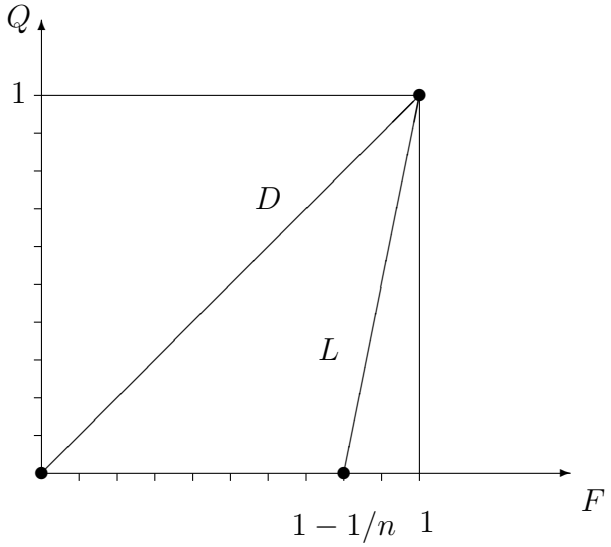
$$G = \sum_{i=1}^{n} q_i(F_i + F_{i-1}) - 1 = \frac{1}{n}\sum_{i=1}^{n} q_i(2i-1) - \frac{n}{n}\sum_{i=1}^{n} q_i = \frac{1}{n}\sum_{i=1}^{n} q_i(2i-n-1).$$

Auch diese Formel dient ausschließlich der Vollständigkeit, nicht des Erkenntnisgewinns. Man kann leichten Herzens auf sie verzichten.

4.2.1 Der Maximalwert des GINI-Koeffizienten

Figur 12 zeigt die Lorenzkurve für den wenig realistischen Fall, bei dem sich der Gesamtwert eines Merkmals auf eine einzige von insgesamt n statistische Einheiten konzentriert.

Fig. 12: Konzentration auf eine einzige von n statistischen Einheiten.



Der GINI-Koeffizient berechnet sich für diesen Fall besonders leicht, indem von der Fläche 1/2 des Dreiecks unter der Diagonalen D, die Fläche des Dreiecks unter der Lorenzkurve L abgezogen wird:

$$G_{max}(n) := 2 \cdot (\frac{1}{2} - \frac{1}{2} \cdot \frac{1}{n} \cdot 1) = 1 - \frac{1}{n} = \frac{n-1}{n}. \tag{55}$$

Damit lautet die theoretische Bandbreite für den GINI-Koeffizienten

$$0 \leq G \leq 1 - \frac{1}{n}. \tag{56}$$

Es ist leicht ersichtlich, dass

$$G_{max}(n) = 1 - \frac{1}{n} = \frac{n-1}{n} < 1,$$

weshalb das bereits in (52) angegebene Intervall

$$0 \leq G < 1$$

bestätigt bzw. durch die Bandbreite (55) präzisiert wird.

4.2.2 Der normierte GINI-Koeffizient

Im Beispiel des Marktes mit 5 Unternehmen $U_1, ..., U_5$ ergäbe sich bei einer Konzentration des gesamten Umsatzes auf ein einziges der 5 Unternehmen der maximale Wert $G_{max}(5) = 4/5 = 0,8$. In Relation zu diesem Maximalwert deutet der ermittelte Wert von $G = 0,2$ auf eine größere Konzentration hin als aufgrund des absoluten Wertes von 0,2 anzunehmen ist:

$$\frac{0,2}{4/5} = 5/4 \cdot 0,2 = 0,25 \ .$$

Diese Überlegung suggeriert die Definition eines *normierten* GINI-Koeffizienten $G_{norm}(n)$, welcher dadurch gebildet wird, dass der GINI-Koeffizient G in Relation gesetzt wird zum jeweils maximalen Wert $G_{max}(n)$:

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

Damit gilt

$$0 \leq G_{norm}(n) \leq 1$$

und

$G_{norm}(n) = 1$ bei vollständiger Konzentration,
$G_{norm}(n) = 0$ bei gleichmäßiger Verteilung der Merkmalssumme.

Mit Hilfe des normierten GINI-Koeffizienten $G_{norm}(n)$ lässt sich der Grad der relativen Konzentration bzw. der Ungleichheit für zwei verschiedenen Stichproben mit unterschiedlicher Anzahl an statistischen Einheiten miteinander vergleichen. Für Stichproben sehr großen Umfangs n oder bei Betrachtung einer Grundgesamtheit ist die Berechnung von $G_{norm}(n)$ allerdings nicht notwendig, denn

$$\frac{n}{n-1} \to 1 \quad \text{für} \quad n \to \infty \ ,$$

so dass damit

$$G_{norm}(n) \to G \quad \text{für} \quad n \to \infty. \tag{58}$$

Tatsächlich wirkt sich der Normierungsfaktor bereits bei einer relativ geringen Anzahl n an statistischen Einheiten nicht mehr wesentlich aus. Eine Normierung ist außerdem nur dann möglich, wenn der Stichprobenumfang n bekannt ist – auf Basis beispielsweise einer relativen Häufigkeitsverteilung alleine kann der normierte GINI-Koeffizient nicht berechnet werden.

4.3 Absolute Konzentration

Das bekannteste Maß zur Messung absoluter Konzentration ist der HERFINDAHL-Index. Bei n verschiedenen statistischen Einheiten, die bezüglich eines Merkmals X die Werte

$x_1, x_2, ..., x_n$ aufweisen, ist der HERFINDAHL-Index definiert durch

$$H := q_1^2 + q_2^2 + ... + q_n^2 = \sum_{i=1}^{n} q_i^2, \tag{59}$$

wobei q_i der Anteil der statistische Einheit i am Gesamtwert des Merkmals X ist:

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

Beispiel: Konzentrationsmessung mittels HERFINDAHL-Index und GINI-Koeffizient.

In einem Markt wie beispielsweise dem zumindest offiziell liberalisierten deutschen Gasmarkt sind z. B. fünf Unternehmen tätig, deren Marktanteile 60 %, 10 %, 5 %, 20 % bzw. 5% betragen.
Der HERFINDAHL-Index zur Messung der Konzentration in diesem Markt lautet

$$H = (0,6)^2 + (0,1)^2 + (0,05)^2 + (0,2)^2 + (0,05)^2 = 0,415 \ .$$

Der GINI-Koeffizient weist den Wert 0,5 auf und berechnet sich nach Formel (51),

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

wie folgt:

$$G = 1 - 0,2 \cdot [(0 + 0,05) + (0,1 + 0,05) + (0,2 + 0,1) + (0,2 + 0,4) + (0,4 + 1)] = 0,5 \ .$$

Dabei beträgt der relative Anteil jedes einzelnen der 5 Unternehmen $f_i = 0,02 = 20$ %. Die beiden Unternehmen mit den geringsten Marktanteilen beschließen eine Fusion, um sich gegenüber den anderen Unternehmen besser im Markt zu positionieren. Nach einer Fusion wären nur noch 4 anstatt 5 Unternehmen am Markt. Der HERFINDAHL-Index betrüge dann

$$H = (0,6)^2 + 2 \cdot (0,1)^2 + (0,2)^2 = 0,42 \ .$$

Er hat sich damit gegenüber der Situation vor der Fusion leicht erhöht. Der GINI-Koeffizient weist dagegen nach der Fusion einen geringfügig kleineren Wert auf:

$$G = 1 - 0,25 \cdot [(0 + 0,1) + (0,1 + 0,2) + (0,2 + 0,4) + (0,4 + 1)] = 0,4 \ .$$

Durch das Verschmelzen der zwei kleinsten Unternehmen gäbe es 2 Unternehmen mit einem Marktanteil von 10 %. Das würde die Ungleichheit leicht reduzieren, wenngleich das Unternehmen mit 60 % Marktanteil weiterhin ein deutliches Übergewicht besitzt.

Wenn die Anteile von n statistischen Einheiten am Gesamtwert eines Merkmals X allesamt gleich wären, nimmt der HERFINDAHL-Index den Wert 1/n an. Eine beliebige

statistische Einheit i hätte dabei einen Anteil von $q_i = 1/n$, der HERFINDAHL-Index H lautete:

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

Je kleiner dabei die absolute Anzahl an statistischen Einheiten, desto größer ist der HERFINDAHL-Index H. Bei gleichmäßiger Aufteilung auf 3 statistische Einheiten wäre $H = 1/3$, bei gleichmäßiger Aufteilung auf 2 Einheiten wäre $H = 1/2$. Im Extremfall vereinigt eine einzige statistische Einheit den gesamten Merkmalswert auf sich: $H = 1$. Der HERFINDAHL-Index besitzt demnach den folgenden Wertebereich:

$$1/n \leq H \leq 1. \tag{60}$$

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

Beispiel: Die Wirkung von Fusionen auf den HERFINDAHL-Index.
In einem Markt mit n verschiedenen Unternehmen, welche die Marktanteile $(q_1, q_2, ..., q_n)$ besitzen, fusioniert das Unternehem i mit dem Unternehmen $i + 1$. Durch die Fusion hat die absolute Konzentration zugenommen – nicht notwendigerweise die relative Konzentration, wie das letzte Beispiel beweist. Die Zunahme der absoluten Konzentration spiegelt sich in der Zunahme des HERFINDAHL-Index wider,

$$H = q_1^2 + ... + (q_i + q_{i+1})^2 + ... + q_n^2 = q_1^2 + ... + q_i^2 + 2 \cdot q_i \cdot q_{i+1} + q_{i+1}^2 + ... + q_n^2.$$

Dieser Wert ist görßer als jener vor der Fusion,

$$H = q_1^2 + ... + q_i^2 + q_{i+1}^2 + ... + q_n^2,$$

denn $2 \cdot q_i \cdot q_{i+1}$ ist immer positiv, wenn die Marktanteile q_i und q_{i+1} beider Unternehmen i und $i + 1$ positiv sind.
Waren insbesondere die Marktanteile aller Unternehmen vor der Fusion allesamt gleich, $(q_1, q_2, ..., q_n) = (1/n, 1/n, ..., 1/n)$, und war damit $H = 1/n$, ergibt sich nach der Fusion von Unternehmen i mit Unternehmen $i + 1$ ein HERFINDAHL-Index von

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

4.4 Konzentration und Disparität in Zusammenfassung

Im Gegensatz zur relativen Konzentration, bei welcher die Abweichung von der Gleichverteilung, oder in anderen Worten die Disparität, von Interesse ist, unabhängig davon wie hoch die Zahl der Merkmalsträger ist, spielt bei der absoluten Konzentration die absolute Anzahl an Merkmalsträgern die wesentliche Rolle: Beispielsweise bei gleichmäßiger Aufteilung des gesamten Umsatzes einer Branche auf z. B. 5 Unternehmen würde

ein Disparitäts-Maß wie der GINI-Koeffizient ebenso den Wert 0 aufweisen wie bei einer gleichmäßigen Aufteilung auf nur 2 oder auf 10 Unternehmen. Während in allen drei Fällen die relative Konzentration gleich ist, unterscheiden diese sich in der *absoluten Konzentration*. Bei einer gleichmäßigen Aufteilung sollte ein Maß für die absolute Konzentration um so höher sein, je kleiner die Zahl der statistischen Einheiten ist, auf die sich der Gesamtwert eines Merkmals verteilt.

Wegen ihrer besonderen Bedeutung wurde den beiden populärsten statistischen Maßen für die relative bzw. absolute Konzentration, dem GINI-Koeffizienten bzw. dem HERFINDAHL-Index, ein eigenes Kapitel gewidmet. Beide Maße sind, wie Lage- oder Streuungsparameter auch, lediglich statistische Parameter zur Beschreibung eines gewissen Aspektes eines einzigen Merkmals und stehen sogar in engem Zusammenhang zu den im letzten Kapitel besprochenen Streuungsparametern – siehe dazu die folgende Übungsaufgabe zum Zusammenhang des HERFINDAHL-Index und des Variationskoeffizienten. Der GINI-Koeffizient verdichtet gegenüber der Lorenzkurve Information in einem einzigen Zahlenwert. Aufgrund dieses Informationsverlustes ist es insbesondere möglich, bei zwei verschiedenen Fällen denselben Wert für den GINI-Koeffizienten zu erhalten, obwohl tatsächlich wesentliche Unterschiede existieren. Der nächste Abschnitt enthält eine Übungsaufgabe, in der zwei ökonomisch sehr unterschiedliche Marktsituationen dargestellt sind, die denselben GINI-Koeffizienten aufweisen, obwohl einer der beiden Märkte intuitiv deutlich konzentrierter erscheint.

4.5 Übungsaufgaben

Übungsaufgabe: Der GINI-*Koeffizient – ein fragwürdiges Maß.*

Zwei unterschiedliche Märkte werden jeweils von 10 Firmen beliefert:

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

a) Man zeichne für beide Märkte die Lorenzkurven.

b) Welcher Markt könnte im ökonomischen Sinne als konzentrierter bezeichnet werden?

c) Bei Berechnung der GINI-Koeffizienten werden Sie feststellen, dass sich für beide Märkte derselbe Wert ergibt. Welcher Schluß kann daraus gezogen werden?

Übungsaufgabe: Der HERFINDAHL-*Index – ein Maß sowohl der absoluten wie auch der relativen Konzentration.*

In einem Markt mit insgesamt 1.000 Unternehmen entfällt auf 999 Unternehmen ein jeweils gleicher, aber verschwindend geringer Anteil, so dass diese 999 Unternehmen insgesamt ein Marktanteil von lediglich 1 % haben, während sich ein einziges Unternehmen den Löwenanteil von 99 % gesichert hat. Zeigen Sie, dass obwohl bei der Anzahl von insgesamt 1.000 Unternehmen nicht von absoluter Konzentration gesprochen werden kann,

der HERFINDAHL-Index einen Wert nahe 1 aufweist und damit eine große (relative) Konzentration anzeigt.

Übungsaufgabe: Das Verhalten des GINI-*Koeffizienten und des* HERFINDAHL-*Index bei Konkurs eines Unternehmens eines paritätischen Marktes.*

In einem speziellen Markt, bei dem die Marktanteile aller n Unternehmen dieses Marktes allesamt gleich sind, geht ein Unternehmen in Konkurs. Wie lautet der HERFINDAHL-Index vor und nach dem Konkurs, wenn die Marktanteiles des insolventen Unternehmens sich ebenfalls gleichmäßig auf die anderen Unternehmen aufteilen? Wie verhält sich der GINI-Koeffizient?

Übungsaufgabe: Der Zusammenhang zwischen dem HERFINDAHL-*Index und dem Variationskoeffizienten.*

Für n verschiedene statistische Einheiten, die bezüglich eines Merkmals X die Werte $x_1, x_2, ..., x_n$ aufweisen, besteht zwischen dem Variationskoeffizienten $v_X = s_X/\bar{x}$ und dem HERFINDAHL-Index H der folgende Zusammenhang:

$$H = \frac{v_X^2 + 1}{n}.$$

Dieser Zusammenhang soll mit Hilfe des Ausdrucks

$$s_X^2 = (\frac{1}{n}\sum_{i=1}^{n} x_i^2) - \bar{x}^2$$

für die empirische Standardabweichung s_X – siehe Formel (36) – bewiesen werden, in dem von $v_X^2 = s_X^2/\bar{x}^2$ ausgegangen und dieser Ausdruck eingesetzt wird.

Übungsaufgabe: Vererben nach dem Koran.

In einer arabischen Familie wird das Vermögen von 240 Kamelen an die 2 Söhne und 4 Töchter getreu dem Koran vererbt: „Und wenn die Geschwister Männer und Frauen sind, so soll ein Mann so viel erhalten wie zwei Frauen“(Sure 4, Vers 175). Man skizziere die Lorenz-Kurve, die sich bei Befolgung dieser Empfehlung des Koran ergeben würde und berechne den GINI-Koeffizienten, der die Ungleichverteilung des Erbes an Kamelen beschreiben soll.

Übungsaufgabe: Tonga – ein Paradies für Aussteiger?

Die Vava'u Insel-Gruppe des Königreichs Tonga liegt in der Südsee und besteht aus 200 Inseln, wovon nicht weniger als 120 unbewohnt sind. Neben den 10.159 zumeist sehr korpulenten Eingeborenen, die vom vorbildlich korpulenten König Taufa'ahau Tupou IV. regiert werden, leben auf der Inselgruppe noch 40 Aussteiger vor allem aus Europa, aber auch aus anderen Kontinenten. Diese Aussteiger leben alle als Eremiten jeweils allein auf einer Insel. Die Verteilung der Bevölkerung sieht insgesamt folgendermaßen aus:

Einwohnerzahl	Zahl der Inseln
0	120
1	40
2-100	20
über 100	20

a) Man bestimme die mittlere Einwohnerzahl je Insel in Form des arithmetischen Mittels und des Medians.

b) Wie groß ist die mittlere Einwohnerzahl der 20 größten Inseln mit jeweils über 100 Einwohnern?

c) Man berechne die Eckpunkte der Lorenz-Kurve und skizziere diese.

d) Würde sich die relative Konzentration vergrößern oder verringern, wenn man die 120 unbewohnten Inseln bei der Berechnung der Lorenz-Kurve wegließe? Dürfen die 120 Inseln einfach ignoriert werden, weil auf diesen Inseln niemand wohnt?

e) Wenn man die Größenklasse mit über 100 Bewohnern weiter untergliedern ürde, würde sich dann die relative Konzentration in der Regel vergrößern oder verringern? Wann könnte die relative Konzentration gleichbleiben?

f) Wie würde sich die Konzentration ändern, wenn alle Aussteiger beschließen würden, auf der Hauptinsel der Inselgruppe zu leben? Man berechne für diesen Fall die Eckpunkte der Lorenz-Kurve und skizziere diese.