Großes Lehrbuch der Statistik

Von
Professor Dr. Karl Bosch
o. Professor für angewandte Mathematik
und Statistik an der
Universität Stuttgart-Hohenheim

R. Oldenbourg Verlag München Wien

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Bosch, Karl:
Großes Lehrbuch der Statistik / von Karl Bosch. - München ;
Wien : Oldenbourg, 1996
ISBN 3-486-23350-5

Druck: Grafik + Druck, München
Bindung: R. Oldenbourg Graphische Betriebe GmbH, München

ISBN 3-486-23350-5

Inhaltsverzeichnis

Teil III: Beurteilende (induktive) Statistik

Vorwort

Dieses Buch wendet sich an alle Studierenden, die während ihres Studiums Vorlesungen über Statistik oder Wahrscheinlichkeitsrechnung hören bzw. in wissenschaftlichen Arbeiten statistische Methoden anwenden müssen. Der Rahmen des Buches geht über eine kurze Einführung in die Statistik hinaus. Neben der ausführlichen Behandlung der wichtigsten Grundlagen der Statistik und Wahrscheinlichkeitsrechnung soll das Buch auch einen Einblick in Gebiete geben, die im späteren Studium und im Berufsleben Anwendung finden.

Ziel des Autors ist es, Interesse zu wecken, den Stoff möglichst klar und verständlich darzustellen. Dabei sollen viele Beispiele zum besseren Verständnis beitragen. Zur Übung des Stoffes sind am Ende eines jeden Kapitels zahlreiche Übungsaufgaben angegeben. Um vor allem die Anwender anzusprechen, wurde bei vielen Sätzen auf die Beweise verzichtet. Gelegentlich wurden Beweise durch Plausibilitätsbetrachtungen ersetzt. Quellenhinweise auf vollständige Beweise werden im Text angegeben. Im Literaturverzeichnis sind die zitierten Werke zusammengestellt.

Das Buch gliedert sich in drei Teile. Im ersten Teil wird die beschreibende (deskriptive) Statistik behandelt. Der zweite Teil beschäftigt sich mit Wahrscheinlichkeitsrechnung, ohne die keine sinnvolle Statistik möglich ist. In der beurteilenden (induktiven) Statistik im Teil III werden schließlich die statistischen Verfahren behandelt. Zur Aufstellung der entsprechenden Formeln und vor allem für die Interpretation der damit gewonnenen Ergebnisse ist die Wahrscheinlichkeitsrechnung unentbehrlich. Gleichzeitig werden dabei Grundlagen aus der beschreibenden Statistik benutzt.

Es dürfte kaum möglich sein, in einer zweisemestrigen Vorlesung den gesamten Stoff dieses Buches zu behandeln. Aus diesem Grunde müssen Schwerpunkte gesetzt werden. Als Schwerpunkt könnte z. B. die beschreibende Statistik oder die Wahrscheinlichkeitsrechnung gewählt werden. Der eine oder andere Abschnitt kann durchaus übersprungen oder erst später nachgearbeitet werden. Es ist nicht möglich, auch in einem ausführlichen Buch sämtliche Verfahren der Statistik zu behandeln. Das Buch erhebt also keineswegs den Anspruch auf Vollständigkeit. In der Anwendung werden immer wieder Verfahren benötigt, die hier nicht behandelt wurden. In einem solchen Fall muß auf die Spezialliteratur zurückgegriffen werden.

Zur Vorbereitung auf Prüfungen sei auf das ebenfalls im Oldenbourg-Verlag erschienene Buch KLAUSURTRAINING STATISTIK (Bosch) verwiesen. Dort sind zahlreiche prüfungsrelevante Aufgaben mit vollständigen Lösungen zu finden.

Bezüglich der benötigten Mathematik sei auf zwei im selben Verlag erschienene Bücher hingewiesen:
Eine elementare Darstellung mathematischer Grundlagen ist in Bosch K.: MATHEMATIK FÜR WIRTSCHAFTSWISSENSCHFTLER, 10., erweiterte Auflage, 1995 zu finden.
Eine ausführlichere Darstellung enthält das Lehrbuch Bosch K./Jensen U.: GROSSES LEHRBUCH DER MATHEMATIK FÜR ÖKONOMEN.

Für die sorgfältige Durchsicht des Manuskripts sowie die wertvollen Hinweise und Verbesserungsvorschläge bedanke ich mich bei meinen Mitarbeitern Herrn Dipl. math. oec. D. Reepschläger, Herrn Dipl. math. T. Severin und Herrn Diplom-Betriebswirt (BA) C. Frank.

Karl Bosch

Teil I:

Beschreibende (deskriptive) Statistik

Ziel der beschreibenden Statistik ist es, umfangreiches Datenmaterial aus statistischen Erhebungen übersichtlich darzustellen. Dazu werden oft graphische Darstellungen benutzt, die eine "optische Information" über das gesamte Datenmaterial ergeben. Ferner werden aus dem Datenmaterial Kenngrößen berechnet, welche über das gesamte Stichprobenmaterial möglichst viel Informationen liefern sollen. Durch die Angabe solcher Kenngrößen findet allerdings im allgemeinen eine Datenreduktion statt. In der Regel gehen dabei Informationen über das in der statistischen Erhebung gewonnene Datenmaterial (Urmaterial) verloren.
Mit Hilfe dieser Kenngrößen (Parameter) können zunächst nur Aussagen über die Grundgesamtheit gemacht werden, welche im vorliegenden Datenmaterial untersucht wurde. Aus diesem Datenmaterial abgeleitete Aussagen dürfen nicht ohne weiteres auf größere Grundgesamtheiten übertragen werden. Dazu müssen bestimmte Voraussetzungen bezüglich der Stichprobenentnahme erfüllt sein. Es muß sich um sogenannte repräsentative Stichproben handeln. Diese Thematik wird in der beurteilenden Statistik (Teil III) behandelt.

Nach der Kassifikation von Merkmalen in Kapitel 1 beschäftigt sich Kapitel 2 mit eindimensionalen Stichproben bzw. Häufigkeitsverteilungen. Dabei werden verschiedene Lage- und Streuungsparameter angegeben, die über die Häufigkeitverteilung möglichst viel Information liefern sollen. Ferner werden in Abschnitt 2.6 Konzentrationsmaße und in Abschnitt 2.7 Indexzahlen untersucht.

In Kapitel 3 werden zweidimensionale Beobachtungsreihen behandelt. In der Korrelationsrechnung (Abschnitt 3.4) wird nur der Zusammenhang zweier Merkmale untersucht, während man in der Regressionsrechnung (Abschnitt 3.4) von einem Merkmal auf das andere schließem möchte. Hier interessiert also die Ursache der Abhängigkeit.

Abschnitt 3.5 beschäftigt sich mit Zeitreihen. Neben der Trendbestimmung werden dort Komponenten geschätzt und Saisonbereinigungen durchgeführt.

In Kapitel 4 werden schließlich p-dimensionale Stichproben untersucht. Im Vordergrund steht dabei die multiple lineare Regression sowie die Korrelationsrechnung.

Kapitel 1:
Merkmale und Skalierung

1.1 Merkmale

In einer statistischen Erhebung werden in der Regel bei verschiedenen Merkmalsträgern (Individuen oder statistischen Einheiten) ein oder auch mehrere **Merkmale** gleichzeitig festgestellt. Die verschiedenen Ergebnisse, die bei der Beobachtung eines bestimmten Merkmals auftreten können, nennt man **Merkmalsausprägungen**.

Beispiel 1.1:

a) Das Merkmal "Geschlecht" besitzt die beiden Merkmalsausprägungen männlich und weiblich.

b) Bei einer Qualitätskontrolle interessiere nur, ob ein untersuchtes Werkstück fehlerhaft oder brauchbar ist. Bezüglich dieses Beobachtungsmerkmals gibt es nur die beiden Ausprägungen fehlerhaft und brauchbar.

c) Das Merkmal "Farbe eines Gegenstands" besitzt die Ausprägungen rot, weiß, grün, blau, schwarz,... . Als mögliche Merkmalsausprägungen kommen sämtliche Farben in Frage.

d) Das Merkmal "Beruf einer Person" besitzt sehr viele verschiedene Ausprägungen, z. B. Verkäuferin, Kaufmann, Automechaniker oder Lehrer. Sämtliche Berufe vollständig aufzuzählen, dürfte kaum möglich sein.

e) Der Handelspreis für eine bestimmte Obst- oder Gemüsesorte hängt im allgemeinen von der Handelsklasse ab. Falls vier Handelsklassen zugelassen werden, bezeichnet man diese mit I, II, III, IV bzw. mit A, B, C und D. Dabei ist I (A) die beste und IV (D) die schlechteste Klasse. Hier gibt es also vier Merkmalsausprägungen.

f) Auf verschiedenen Feldern werde der Ernteertrag pro Hektar festgestellt. Die Merkmalsausprägungen sind reelle Zahlen in einem bestimmten Bereich (Intervall).

In diesem Beispiel wird bereits deutlich, daß bei den Beobachtungsmerkmalen verschiedene Typen auftreten können. Beim Zählen, Messen oder Wiegen sind die Merkmalsausprägungen unmittelbar reelle Zahlen, die in bestimmten Einheiten gemessen werden. Die einzelnen Merkmalsausprägungen unterscheiden sich durch ihre Größe. Durch die vier Handelsklassen in Beispiel e) kommt nur ein Qualitätsunterschied zum Ausdruck. Dabei wird jedoch keine Aussage darüber gemacht, um wieviel eine Handelsklasse besser oder schlechter ist als eine andere. Im Gegensatz zu e) und f) können in a) bis d) die einzelnen Merkmalsausprägungen nicht miteinander verglichen werden. Zwischen den einzelnen Ausprägungen gibt es keine Rangordnung.

Wir wollen nun verschiedene Merkmalstypen klassifizieren. Unterschieden wird dabei nach der Art des Merkmals und nach der Anzahl der möglichen Merkmalsausprägungen.

Quantitative (zahlenmäßige) Merkmale sind solche, deren Ausprägungen in bestimmten Einheiten gemessen werden können. Sie werden durch reelle Zahlen dargestellt. Zwischen verschiedenen Ausprägungen eines quantitativen Merkmals besteht immer eine Rangordnung (Reihenfolge), also eine Größer-Kleiner-Beziehung. Die Ausprägungen unterscheiden sich durch ihre Größe. Bei quantitativen Merkmalen muß der Unterschied zwischen zwei Merkmalsausprägungen stets quantifiziert (gemessen) werden können. Beim Zählen, Messen oder Wiegen werden Ausprägungen quantitativer Merkmale festgestellt.

Qualitative (artmäßige) Merkmale sind Merkmale, welche nicht quantitativ sind. Solche Merkmale können nur qualitativ (verbal) beschrieben werden. Sie lassen sich nicht direkt durch Zahlen kennzeichnen, zwischen denen eine natürliche Reihenfolge (Größer-Kleiner-Beziehung) besteht. Die Ausprägungen eines qualitativen Merkmals unterscheiden sich nur durch ihre Art. Der Unterschied zwischen zwei Ausprägungen eines qualitativen Merkmals kann nicht gemessen werden. Qualitative Merkmale sind z. B. Geschlecht, Familienstand, Beruf, Konfession, Haarfarbe, Handelsklasse oder Steuerklasse. Formal könnte man zwar allen Ausprägungen eines qualitativen Merkmals Zahlen zuordnen. Durch eine solche formale Quantifizierung geht das qualitative Merkmal jedoch keineswegs in ein quantitatives über, es bleibt weiterhin qualitativ. Nur die Bezeichnungen für die Ausprägungen werden geändert.

Beispiel 1.2:
Bei den üblichen Zensuren für Leistungen in der Schule oder Universität "sehr gut, gut, befriedigend, ausreichend, mangelhaft, ungenügend" handelt es sich um ein qualitatives Merkmal. Dabei ist zwischen den Ausprägungen zwar eine Rangordnung vorgegeben, denn "sehr gut" ist besser als "gut", "gut" besser als "befriedigend" usw. Die genauen Unterschiede zwischen den einzelnen Noten liegen im allgemeinen aber nicht fest und sind meistens auch nicht gleich. Insbesondere gilt dies bei der Bewertung von Aufsätzen in Deutsch oder Geschichtsarbeiten. In der Regel werden den Zensuren zwar die Zahlen $1, 2, 3, 4, 5, 6$ zugeordnet. Dadurch findet eine Quantifizierung statt. Das Merkmal wird also formal quantifiziert. Durch diese Quantifizierung entsteht allerdings der Eindruck, daß die Unterschiede zwischen zwei aufeinanderfolgenden Zensuren jeweils gleich sind, was im allgemeinen aber nicht der Fall ist.

Diskrete Merkmale sind solche, die nur endlich viele oder höchstens abzählbar unendlich viele verschiedene Merkmalsausprägungen besitzen. "Endlich

viele" bedeutet dabei, daß die Merkmalsausprägungen von 1 an bis zu einer endlichen ganzen Zahl durchnumeriert werden können. "Abzählbar unendlich" bedeutet, daß es zwar unendlich viele verschiedene Merkmalsausprägungen gibt, die jedoch wie die natürlichen Zahlen von 1 an ohne Ende durchnumeriert werden können. Beim Zählen werden diskrete Merkmale untersucht.

Stetige Merkmale sind Merkmale, deren Ausprägungen ein ganzes Intervall der Zahlengeraden bilden. Ihre Ausprägungen gehen also im Gegensatz zu diskreten Merkmalen fließend ineinander über. Beim Messen oder Wiegen werden im allgemeinen Ausprägungen stetiger Merkmale festgestellt.

1.2 Skalierung

Um die verschiedenen Ausprägungen eines Merkmals nach den gleichen Kriterien angeben oder messen zu können, muß zuerst eine **Skala** vorgegeben werden. Durch die **Skalierung** werden den Merkmalsausprägungen einzelne Werte (Plätze) der Skala zugeordnet. Die jeweilige Skala hängt dabei vom Typ des Merkmals ab.

Nominalskala: Eine Nominalskala liegt vor, wenn durch sie nur die Verschiedenheit der Ausprägungen eines Merkmals zum Ausdruck gebracht werden kann. Merkmale, deren Ausprägungen nur in einer solchen Skala dargestellt werden könnnen, heißen **nominale Merkmale**. Nominalskalen sind Skalen qualitativer Merkmale, bei denen es keine natürliche Rangordnung gibt. Nominalskalen sagen am wenigsten über die Merkmalsausprägungen aus. Sie stellen die niedrigste Stufe einer Skala dar.

Beispiel 1.3:
Die Ausprägungen der Merkmale Geschlecht, Konfession, Beruf, Farbe oder Steuerklasse sind nicht miteinander vergleichbar. Es handelt es sich um nominale Merkmale. Durch die Zuordnung: männlich$\leftrightarrow$0; weiblich$\leftrightarrow$1 entsteht auch nur eine Nominalskala. Durch diese Zuordnung wird das Merkmal Geschlecht zwar formal quantifiziert, es bleibt trotzdem nur qualitativ. Man hätte auch die Zuordnung: weiblich$\leftrightarrow$0; männlich$\leftrightarrow$1 oder eine andere Zahlenzuordnung wählen können.

Ordinalskala (Rangskala): Eine Ordinalskala (Rangskala) liegt vor, wenn die unterscheidbaren Merkmalsausprägungen in eine natürliche Rangordnung (Reihenfolge) gebracht werden können. Ordinal skalierte Merkmale heißen **ordinale Merkmale**. Abstände zwischen verschiedenen Ausprägungen ordinaler Merkmale sind jedoch nicht quantifizierbar (nicht interpretierbar). Durch die Rangordnung können den Ausprägungen zwar Zahlen zuge-

ordnet werden, doch sagen diese Zuordnungszahlen nichts über die Abstände der einzelnen Merkmalsausprägungen aus.

Im Gegensatz zu qualitativen können quantitative Merkmale immer angeordnet werden. So besteht bei den Merkmalen Güteklasse bei Lebensmitteln, Tabellenplatz einer Fußballiga oder Intelligenzquotient eine natürliche Rangordnung. Ihre Ausprägungen lassen sich anordnen, obwohl es sich um kein quantitatives Merkmal handelt.

Metrische Skala (Kardinalskala): Man spricht von einer metrischen Skala oder Kardinalskala, wenn zwischen den Merkmalsausprägungen nicht nur eine Reihenfolge (Rangordnung) besteht, sondern auch die Abstände zwischen den Merkmalsausprägungen miteinander verglichen werden können. Metrische Skalen sind Skalen quantitativer Merkmale. Merkmale mit einer metrischen Skala nennt man **metrisch skaliert** oder **kardinal**.

Beispiele für metrisch skalierte Merkmale sind: Börsenkurse, Gewinne, Verluste, Erträge, Längen, Gewichte, monetäre und physikalische Größen. Die metrischen Skalen sind im allgemeinen bis auf die Wahl der Maßeinheit eindeutig bestimmt.

Beispiel 1.4:
In einer statistischen Erhebung sind verschiedene Fragen zu beantworten. Für einige Merkmale sollen die entsprechenden Typen angegeben werden:
Wohnort: nominales diskretes Merkmal. Ausprägungen sind alle Städte und Ortschaften des Landes.
Geschlecht: nominales diskretes Merkmal mit den beiden Ausprägungen "weiblich" oder "männlich".
Beruf: nominales diskretes Merkmal mit den Angaben "nicht berufstätig", "angestellt", "Arbeiter", "Landwirt", "selbständig" oder "freiberuflich".
Konfession: nominales diskretes Merkmal. Die Ausprägungen sind die verschiedenen Religionsgemeinschaften.
Alter (ganze vollendete Jahre): metrisch skaliertes diskretisiertes Merkmal mit den Ausprägungen $0, 1, 2, \ldots$.
Körpergröße in cm: metrisch skaliertes diskretisiertes Merkmal.
Körpergewicht in kg: metrisch skaliertes diskretisiertes Merkmal.

Bemerkung: Alter, Körpergröße und Körpergewicht sind zunächst stetige Merkmale, da jede reelle Zahl aus einem bestimmten Intervall als Merkmalswert möglich ist. Durch die Angaben "in ganzen Jahren, in cm und in kg" findet eine Diskretisierung statt. Im Gegensatz zur Altersangabe ist die Rundung bei der Körpergröße und beim Gewicht nicht eindeutig festgelegt. Manche Personen werden im mathematischen Sinne korrekt runden, andere prinzipiell ab- bzw. aufrunden.
Bei der Feststellung eines stetigen Merkmals (z. B. beim Messen oder Wieen) findet durch das Runden eine Diskretisierung statt.

Kapitel 2 :
Eindimensionale Darstellungen

In diesem Abschnitt soll nur ein einziges Merkmal untersucht werden. An n Merkmalsträgern aus einer bestimmten Grundgesamtheit wird jeweils die Ausprägung des Merkmals festgestellt. Die Merkmalsausprägung beim i-ten Merkmalsträger bezeichnen wir mit x_i für $i = 1, 2, \ldots, n$. Man nennt x_i die i-te **Beobachtungseinheit**. Alle n Merkmalswerte zusammen bilden das n-Tupel $x = (x_1, x_2, \ldots, x_n)$. Dieses n-Tupel heißt **Beobachtungsreihe** (**Urliste** oder **Stichprobe**) vom **Umfang** n. Falls die Merkmalswerte sämtlicher Individuen einer Grundgesamtheit festgestellt werden, spricht man von einer **Total-** oder **Vollerhebung**, andernfalls von einer **Teilerhebung**. Bei Volkszählungen finden in der Regel Totalerhebungen, bei Meinungsumfragen Teilerhebungen statt.

2.1 Häufigkeitsverteilungen bei diskreten Merkmalen

Ein diskretes Merkmal besitze die verschiedenen Ausprägungen $a_1, a_2, \ldots$. Die Anzahl der verschiedenen Ausprägungen kann dabei endlich oder abzählbar unendlich sein. Der Merkmalswert x_i stimmt dann mit einem dieser Merkmalsausprägungen überein. Falls es nur m verschiedene Ausprägungen gibt, können in der **Beobachtungsreihe** höchstens m verschiedene Werte x_i auftreten.

2.1.1 Absolute und relative Häufigkeiten

Die Anzahl derjenigen Beobachtungseinheiten aus der Urliste vom Umfang n, welche die Merkmalsausprägung a_j besitzen, nennt man die **absolute Häufigkeit** von a_j. Wir bezeichnen sie mit $h_n(a_j)$. Dabei gibt der Index n den Umfang der Urliste an. Es ist also

$$h_j = h_n(a_j) = \text{Anzahl der Beobachtungswerte, die gleich } a_j \text{ sind.} \quad (2.1)$$

Die absolute Häufigkeit 28 ist bei einem Versuchsumfang $n = 30$ groß, während sie bei einem Versuchsumfang $n = 100$ klein ist. Aus diesem Grund muß die absolute Häufigkeit in Relation zum Versuchsumfang n gesetzt werden. Division der absoluten Häufigkeit $h_n(a_j)$ durch den Stichprobenumfang n ergibt eine Größe, die vom Versuchsumfang n unabhängig ist. Den so erhaltenen Wert

$$r_j = r_n(a_j) = \frac{1}{n} \cdot h_n(a_j), \quad j = 1, 2, \ldots, m \tag{2.2}$$

nennt man die **relative Häufigkeit** von a_j in der Urliste.

Weil $100\, r_n(a_j)\,\%$ der Beobachtungswerte die Ausprägung a_j besitzen, beschreibt die relative Häufigkeit den prozentualen Anteil (**prozentuale Häufigkeit**) der Merkmalsausprägung a_j. Die relative Häufigjkeit liegt immer zwischen Null und Eins unabhängig vom Stichprobenumfang n. Je größer sie ist, um so öfter ist der Merkmalswert eingetreten. Die relative Häufigkeit beschreibt damit die absolute Häufigkeit unabhängig vom Versuchsumfang n. Die prozentuale Häufigkeit liegt zwischen 0 und 100.

Allgemein gelten für die absoluten und die relativen Häufigkeiten die Eigenschaften:

$$\begin{aligned} &0 \le h_n(a_j) \le n \quad \text{für jedes } j\,; \quad \sum_{j=1}^{m} h_n(a_j) = n\,; \\ &0 \le r_n(a_j) \le 1 \quad \text{für jedes } j\,; \quad \sum_{j=1}^{m} r_n(a_j) = 1\,. \end{aligned} \tag{2.3}$$

Die absoluten und relative Häufigkeiten können sowohl bei qualitativen als auch bei quantitativen Merkmalen bestimmt werden.

Definition 2.1 (Häufigkeitsverteilung):
In einer Stichprobe vom Umfang n sollen die Merkmalsausprägungen $a_1, a_2, \ldots$ die absoluten Häufigkeiten $h_n(a_1), h_n(a_2), \ldots$ und die relativen Häufigkeiten $r_n(a_1), r_n(a_2), \ldots$ besitzen. Dann heißt die Gesamtheit der Paare

$$(a_j,\ h_n(a_j)),\quad j = 1, 2, \ldots$$

die **absolute Häufigkeitsverteilung** und

$$(a_j,\ r_n(a_j)),\quad j = 1, 2, \ldots$$

die **relative Häufigkeitsverteilung** des diskreten Merkmals.

2.1.2 Strichliste und Häufigkeitstabelle

In der Urliste sind die Beobachtungswerte im allgemeinen völlig ungeordnet und damit - vor allem bei großen Stichprobenumfängen n - nicht übersichtlich. Aus diesem Grund versucht man, die Beobachtungswerte in einer Tabelle übersichtlich darzustellen.
Dazu trägt man in der ersten Spalte der Häufigkeitstabelle (vgl. Tabelle 2.1) alle Merkmalsausprägungen ein. Falls es sehr viele oder gar abzählbar unendlich viele verschiedene Merkmalswerte gibt, müssen Merkmalswerte zusammengefaßt werden, am besten solche, die in der Urliste selten vorkommen.

Für jeden Beobachtungswert der Urliste wird in die zweite Spalte hinter dem entsprechenden Merkmalswert ein senkrechter Strich | eingetragen. Der Übersicht halber werden fünf Striche durch den Block 𝍸 dargestellt. Jeweils der fünfte Strich wird waagrecht durch die vorangehenden vier Striche gezogen. Dadurch entstehen Fünferblöcke mit einem Rest.
In zwei weiteren Spalten werden die absoluten Häufigkeiten (Anzahl der Striche) und die relativen Häufigkeiten der jeweiligen Merkmalswerte eingetragen. Die Häufigkeitstabelle enthält also die absolute und die relative Häufigkeitsverteilung.

Durch die Übertragung der Urliste in eine Häufigkeitstabelle gehen allerdings wesentliche Informationen über die Urliste verloren, da die Reihenfolge, in der die Beobachtungswerte auftreten, in der Tabelle nicht mehr feststellbar ist.

Beispiel 2.1:
In einem Verein kandidierten die drei Personen A, B und C für den Posten des ersten Vorstands. Bei der Abstimmung waren 75 Personen stimmberechtigt. Nach der Satzung ist derjenige Kandidat gewählt, welcher die meisten Stimmen erhält. Bei der Auszählung der Stimmzettel wird für jede Stimme in der Tabelle 2.1 an der entsprechenden Stelle ein | eingetragen (2. Spalte). In der dritten Spalte sind die absoluten Häufigkeiten als Anzahl der Stimmen (Striche) aufgeführt. Der Kandidat B erhielt die Mehrheit der Stimmen und wurde damit gewählt. Division der absoluten Häufigkeiten durch den Stichprobenumfang n = 75 ergibt die relativen Häufigkeiten in der 4. Spalte. Multiplikation der relativen Häufigkeiten mit 100 ergibt die prozentualen Stimmenanteile.

Kandidat	abgegebene Stimmen	absolute Häufigkeit	relative Häufigkeit	prozentualer Anteil
Kandidat A	𝍸 𝍸 𝍸 |	16	0,213	21,3
Kandidat B	𝍸 𝍸 𝍸 𝍸 ||||	24	0,320	32,0
Kandidat C	𝍸 𝍸 𝍸 𝍸 ||	22	0,293	29,3
Enthaltungen	𝍸 𝍸 |||	13	0,173	17,3
ungültig		0	0	0
Summe		n = 75	0,999	99,9

Tab. 2.1: Strichliste und Häufigkeitstabelle

In der letzten Zeile stehen die Spaltensummen gebildet. Die Summe aller relativen Häufigkeiten müßte eigentlich 1 ergeben. Die hier aufgetretene Abweichung ist auf das Runden der relativen Häufigkeiten zurückzuführen. Genauso müßten alle prozentualen Anteile zusammen gleich 100 sein.

Beispiel 2.2:
In 50 Familien wurde jeweils die Anzahl der Kinder festgestellt und in der Tabelle 2.2 eingetragen.

Anzahl der Kinder $a_j = j$	Anzahl der Familien	absolute Häufigkeit $h_{50}(a_j)$	relative Häufigkeit $r_{50}(a_j)$	prozentualer Anteil $100 \cdot r_{50}(a_j)$
0	𝍸 𝍸 ||	12	0,24	24
1	𝍸 𝍸 𝍸 ||	17	0,34	34
2	𝍸 ||||	9	0,18	18
3	𝍸 |	6	0,12	12
4	||||	4	0,08	8
5	||	2	0,04	4
mehr als 5		0	0	0
Summe		n = 50	1,00	100

Tab. 2.2: Strichliste und Häufigkeitstabelle

2.1.3 Graphische Darstellungen

Eine in einer Häufigkeitstabelle dargestellte Beobachtungsreihe kann in einer graphischen Darstellung übersichtlicher dargestellt werden. Bei der Wahl der graphischen Darstellung wird zwischen quantitativen und qualitativen Merkmalen unterschieden.

2.1.3.1 Graphische Darstellungen von Häufigkeitsverteilungen quantitativer diskreter Merkmale

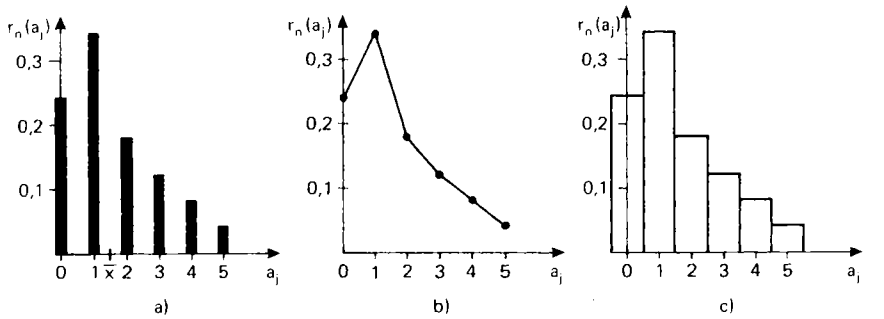
Bei einem quantitativen Merkmal sind die Ausprägungen reelle Zahlen und können somit auf dem Zahlenstrahl (x-Achse) dargestellt werden. Senkrecht nach oben trägt man die absoluten bzw. relativen Häufigkeiten ab.

In einem **Stabdiagramm (Balkendiagramm)** werden über den einzelnen Merkmalswerten senkrecht nach oben Stäbe angetragen, deren Längen die absoluten bzw. relativen Häufigkeiten sind. Im Stabdiagramm der absoluten Häufigkeiten haben alle Stäbe zusammen die Länge n (Anzahl der Stichprobenwerte). Diese Eigenschaft muß bei der Maßstabsfestsetzung berücksichtigt werden. Im Stabdiagramm der relativen Häufigkeiten ist die Gesamtlänge aller Stäbe zusammen immer gleich Eins unabhängig vom Stichprobenumfang n. Aus diesem Grund kann bei Stabdiagrammen für die relativen Häufigkeiten immer der gleiche Maßstab gewählt werden.

In einem **Häufigkeitspolygon** werden die (oberen) Endpunkte der einzelnen Stäbe geradlinig miteinander verbunden.

In einem **Histogramm (Säulendiagramm)** stellt man die absoluten bzw. relativen Häufigkeiten durch Flächeninhalte von Rechtecken senkrecht über den einzelnen Merkmalsausprägungen dar. Nur wenn alle Rechtecke gleich breit sind, können als Höhen jeweils die Häufigkeiten bzw. das gleiche Vielfache davon benutzt werden (**flächenproportionale Darstellung**).

In Bild 2.1 ist die relative Häufigkeitsverteilung aus Tab. 2.1 (Beispiel 2.2) in einem Stabdiagramm, Häufigkeitspolygon und Histogramm graphisch dargestellt. Weil jeweils zwei benachbarte Merkmalsausprägungen (Anzahl der Kinder) voneinander den Anstand 1 besitzen, kann im Histogramm als Höhe direkt die relative Häufigkeit gewählt werden. Die Bilder für die absoluten und relativen Häufigkeiten unterscheiden sich nur durch den Maßstab auf der y-Achse.



a) Stabdiagramm b) Häufigkeitspolygon c) Histogramm

Bild 2.1: Verteilung der relativen Häufigkeiten eines quantitativen Merkmals

2.1.3.2 Graphische Darstellungen von Häufigkeitsverteilungen qualitativer diskreter Merkmale

Bei qualitativen Merkmalen sind die Ausprägungen im allgemeinen keine reellen Zahlen. Formal könnte man die abstrakten Ausprägungen zwar auf der Zahlengeraden darstellen und die Graphiken wie bei quantitativen Merkmalen anfertigen. Dieses Vorgehen ist jedoch nicht sinnvoll. Bei einer Darstellung auf dem Zahlenstrahl besteht nämlich die Gefahr, daß durch

die willkürlich gewählte Anordnung fälschlicherweise eine Rangordnung zwischen den Ausprägungen hineininterpretiert wird. Aus diesem Grund benutzt man hier andere graphische Darstellungen.

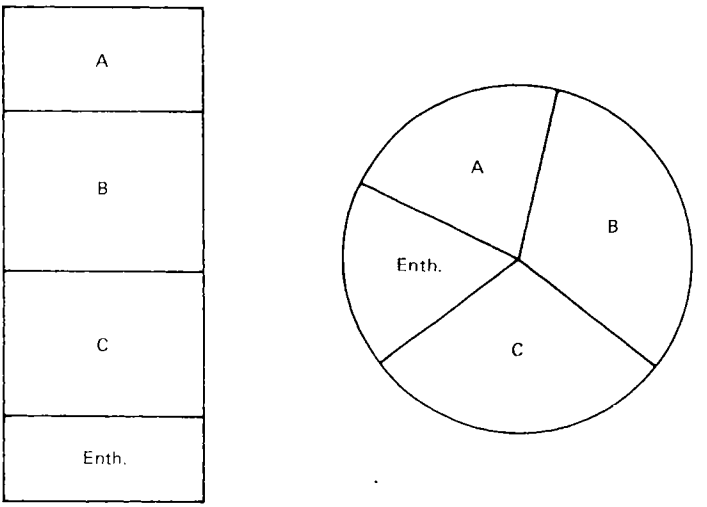
In einem **Rechteckdiagramm** wird die Fläche eines Rechtecks proportional zu den Häufigkeiten aufgeteilt. Bei dieser flächenproportionalen Darstellung verhalten sich die Häufigkeiten zweier Merkmalswerte wie die Inhalte der ihnen zugeordneten Flächen.

Oft benutzt man ein **Kreisdiagramm**. In einem Kreis wird zu jeder Merkmalsausprägung ein Kreissektor gebildet, wobei die Flächen der Sektoren und damit auch die Innenwinkel proportional zu den Häufigkeiten gewählt werden. Auch hier verhalten sich die Flächeninhalte der Kreissektoren wie die Häufigkeiten der den Sektoren zugeordneten Merkmalsausprägungen.

Anstelle eines Kreises könnte man aber auch eine **beliebige Fläche** oder einen **dreidimensionalen Körper** wählen und die Fläche (das Volumen) im Verhältnis der Häufigkeiten aufteilen.

In Bild 2.2 ist die Häufigkeitsverteilung des in Tab. 2.1 (Beispiel 2.1) dargestellten qualitativen Merkmals skizziert. Da die Innenwinkel proportional zu den Stimmenzahlen sind, entspricht jeder einzelnen Stimme ein Winkel von $\frac{360}{75} = 4{,}8°$. Damit erhält man folgende Winkel:

Kandidat A: $16 \cdot 4{,}8 = 76{,}8°$; Kandidat B: $115{,}2°$; Kandidat C: $105{,}6°$; Enthaltungen: $62{,}4°$.



a) Rechteckdiagramm b) Kreisdiagramm

Bild 2.2: Häufigkeitsverteilungen eines qualitativen Merkmals

2.2 Häufigkeitsverteilungen bei Klassenbildungen

Falls ein stetiges Merkmal erhoben wird, sind die in der Urliste vorkommenden Beobachtungswerte in der Regel alle voneinander verschieden, wenn nur genau genug gemessen wird. Die Häufigkeitsverteilungen sind dann nicht übersichtlich. Das gleiche Problem tritt bei diskreten Merkmalen mit sehr vielen verschiedenen Ausprägungen auf. In einem solchen Fall ist es sinnvoll, Merkmalswerte zusammenzufassen.
Falls bei qualitativen Merkmalen Werte zusammengefaßt werden, sind die so entstehenden Ausprägungen (Klassen) wieder qualitativ. Dann können die Häufigkeitsverteilungen dieser Merkmalsklassen wie in Abschnitt 2.1 dargestellt werden.

Bei quantitativen stetigen Merkmalen wird die **Klasseneinteilung** auf einem Intervall vorgenommen, das alle Beobachtungswerte enthält. Dazu wird das Intervall in mehrere Teilintervalle zerlegt. Diese Teilintervalle nennt man **Klassen** oder **Gruppen**. Jede Klasse ist durch eine linke und eine rechte Klassengrenze bestimmt, wobei eindeutig festgelegt sein muß, zu welcher der beiden angrenzenden Klassen der entsprechende Grenzpunkt gehört. Als Klassenintervalle wählt man im allgemeinen halboffene Intervalle. Eine ideale Klasseneinteilung wäre eine mit lauter gleichen Klassenbreiten. In einem solchen Fall sind die Klassengrenzen äquidistant. Oft sind jedoch bei einer äquidistanten Einteilung Klassen, vor allem die Klassen an den Rändern sehr schwach besetzt. Dann ist es sinnvoll, diese Randklassen breiter zu machen. Die Anzahl der Klassen bezeichnen wir mit m und die einzelnen Klassen der Reihe nach mit

$$K_1, K_2, K_3, \ldots, K_{m-1}, K_m .$$

Die zugehörigen **Klassenbreiten** seien $b_1, b_2, \ldots, b_m$.

Aus einer Klasseneinteilung lassen sich allerdings die Beobachtungswerte nicht mehr genau feststellen. Man weiß nur, zwischen welchen Grenzen sie liegen. Daher ist eine Klassenbildung mit einem **Informationsverlust** verbunden. Man kann nur noch feststellen, wie viele Beobachtungswerte in der jeweiligen Klasse liegen. Die genauen Zahlenwerte können aus der Klasseneinteilung nicht mehr abgelesen werden.
Ein einfaches Beispiel einer Klasseneinteilung findet man bei den Portokosten eines Briefes. Für sämtliche Briefe bis zu 20 Gramm müssen die gleichen Portokosten entrichtet werden. Die nächste Klasse geht von 20 bis 50 Gramm, danach von 50 bis 100 Gramm usw. Die einzelnen Gewichtsklassen mit konstanten Portokosten sind verschieden breit.

Die Anzahl der Beobachtungswerte, welche in der Klasse K_j enthalten sind, heißt die **absolute Klassenhäufigkeit**. Wir bezeichnen sie mit

$h_j = h_n(K_j)$ = Anzahl der Beobachtungswerte in der Klasse K_j.

Division durch den Versuchsumfang $n = h_1 + h_2 + \ldots + h_m$
ergibt die **relative Klassenhäufigkeit**

$$r_j = r_n(K_j) = \frac{1}{n} \cdot h_j \quad \text{mit} \quad \sum_{j=1}^{m} r_j = 1.$$

Die Klasseneinteilung wird in einem **Histogramm** graphisch dargestellt. Dazu wird über jeder Klasse ein Rechteck gebildet, dessen Flächeninhalt proportional zur absoluten bzw. relativen Klassenhäufigkeit ist. Nur wenn sämtliche Klassen die gleiche Breite besitzen, dürfen als Höhen unmittelbar die Klassenhäufigkeiten benutzt werden. Sonst müssen andere Höhen gewählt werden. Für die relativen Klassenhäufigkeiten erhält man die

$$\text{Rechteckshöhe für die Klasse } K_j: \quad \frac{r_j}{b_j} = \frac{\text{relative Klassenhäufigkeit}}{\text{Klassenbreite}}.$$

Oft ist man gezwungen, auf beiden Achsen verschiedene Maßstäbe zu wählen. Das gesamte Histogramm besitzt dann den Flächeninhalt Eins.

Beispiel 2.3:
Bei 50 Aggregaten des gleichen Typs wurde die Betriebsdauer in Stunden festgestellt:

1366	647	864	1815	33	2503	1	2790	5	1040
1207	740	27	1618	476	702	566	1625	1562	1902
483	2128	238	1095	205	374	750	180	1153	1673
270	1228	118	206	194	1675	1558	671	363	27
2936	1855	550	362	48	1244	299	579	255	291

Für diese Meßwerte ist in Tabelle 2.3 eine Klasseneinteilung angegeben mit den Klassengrenzen 200, 400, 800, 1200, 1600, 2000, 3000.

Klasse	h_i = absolute Klassenhäufigkeit	r_i = relative Klassenhäufigkeit
$K_1 = (0;200]$	9	0,18
$K_2 = (200;400]$	10	0,20
$K_3 = (400;800]$	10	0,20
$K_4 = (800;1200]$	4	0,08
$K_5 = (1200;1600]$	6	0,12
$K_6 = (1600;2000]$	7	0,14
$K_7 = (2000;3000]$	4	0,08
Summe	$n = 50$	1,00

Tab. 2.3: Klasseneinteilung

Im flächenproportionalen Histogramm in Bild 2.3 für die relativen Klassenhäufigkeiten dürfen als Höhen der Rechtecke nicht unmittelbar die relativen Klassenhäufigkeiten gewählt werden, weil die Klassenbreiten verschieden sind. Die relativen Häufigkeiten müssen durch die Klassenbreiten dividiert werden. Dadurch erhält man der Reihe nach die Rechteckshöhen

$$\frac{0{,}18}{200} = 0{,}0009\,;\ \frac{0{,}20}{200} = 0{,}001\,;\ \frac{0{,}20}{400} = 0{,}0005\,;\ \frac{0{,}08}{400} = 0{,}0002\,;\ \frac{0{,}12}{400} = 0{,}0003\,;$$

$$\frac{0{,}14}{400} = 0{,}00035\,;\ \frac{0{,}08}{1000} = 0{,}00008\,.$$

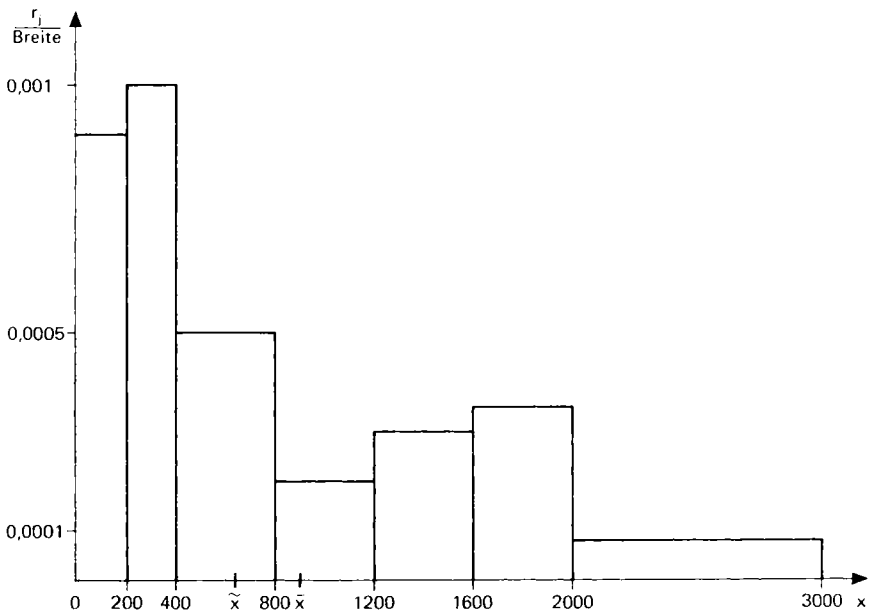


Bild 2.3: Histogramm einer Klasseneinteilung

Bemerkung: Würde man in diesem Histogramm als Rechteckshöhen unmittelbar die absoluten Häufigkeiten der jeweiligen Klassen wählen, so hätte das zweite und dritte Rechteck die gleiche Höhe 10. Das dritte Rechteck ist aber doppelt so breit wie das zweite. Dann wäre der Flächeninhalt des dritten Rechtecks aber doppelt so groß wie der des zweiten Rechtecks. Da man aber unweigerlich Flächeninhalte mit den Häufigkeiten in Zusammenhang bringt, würde man daraus vermutlich den falschen Schluß ziehen, die Häufigkeit der dritten Klasse sei doppelt so groß wie die der zweiten. Nur wenn alle Klassen gleich breit sind, dürfen als Höhen unmittelbar die Häufigkeiten benutzt werden.

2.3 Die empirische Verteilungsfunktion

Bei vielen Problemen möchte man wissen, wie viele der Beobachtungswerte eine bestimmte Grenze nicht überschreiten. Beispiele dafür sind: Die Anzahl der Betriebe, die einen Jahresumsatz von höchstens einer Milliarde DM haben, die Anzahl der Studierenden, die für das Studium nicht mehr als 10 Semester benötigen oder der Bevölkerungsanteil, dessen Monatseinkommen höchstens 5 000 DM beträgt. Eine solche allgemeine Fragestellung ist nur bei **quantitativen Merkmalen** sinnvoll, deren Ausprägungen reelle Zahlen sind.

Ausgangspunkt ist ein **quantitatives Merkmal**, dessen Ausprägungen der Größe nach geordnet werden können.

2.3.1 Die empirische Verteilungsfunktion einer Stichprobe

Wir bezeichen mit $H_n(x)$ die Anzahl der Beobachtungswerte x_i, die kleiner oder höchstens gleich dem festen Zahlenwert x sind, also

$$H_n(x) = \text{Anzahl der Stichprobenwerte } x_i \text{ mit } x_i \leq x\,. \tag{2.4}$$

Die für jedes $x \in \mathbb{R}$ definierte Funktion H_n heißt die **absolute Summenhäufigkeitsfunktion** der Stichprobe.

Division des Funktionswertes $H_n(x)$ durch n ergibt die Funktion

$$F_n(x) = \frac{1}{n} \cdot H_n(x) = \frac{\text{Anzahl der Stichprobenwerte } x_i \text{ mit } x_i \leq x}{n}. \tag{2.5}$$

Die Funktion F_n nennt man die **relative Summenhäufigkeitsfunktion** oder **empirische Verteilungsfunktion** der Stichprobe (Beobachtungsreihe).

An jeder Stelle $x \in \mathbb{R}$ stellt der Funktionswert $F_n(x)$ den relativen Anteil derjenigen Stichprobenwerte dar, die kleiner oder gleich, also höchstens gleich x sind. Zur Bestimmung der Funktionen H_n und F_n und für deren graphische Darstellungen müssen die n Stichprobenwerte der Größe nach geordnet werden.

H_n und F_n sind monoton wachsende (nicht fallende) Treppenfunktionen. Sie springen an den einzelnen Stichprobenwerten um die absolute bzw. relative Häufigkeit des Stichprobenwertes nach oben. Die empirische Verteilungsfunktion F_n steigt von Null auf Eins an. Links vom kleinsten Stichprobenwert verschwindet F_n, vom größten Stichprobenwert an nimmt sie immer den Wert 1 an. Die Summenhäufigkeitsfunktion steigt von Null auf den Stichprobenumfang n an. Beide Treppenfunktionen sind an den jeweiligen Sprungstellen (Unstetigkeitsstellen) noch rechtsseitig stetig.

2.3.2 Die empirische Verteilungsfunktion bei diskreten Merkmalen

Wir betrachten ein diskretes Merkmal mit den Merkmalsausprägungen a_j, $j = 1, 2, \ldots$. Mit Hilfe einer Häufigkeitstabelle können die Häufigkeitsfunktion und empirische Verteilungsfunktion sehr einfach bestimmt werden. Dazu benötigt man nur absolute bzw. relative Häufigkeiten. Es gilt

$$H_n(x) = \sum_{j:\, a_j \le x} h_n(a_j); \quad F_n(x) = \sum_{j:\, a_j \le x} r_n(a_j). \tag{2.6}$$

Nur an den Merkmalsausprägungen können die Funktionen H_n und F_n Sprünge haben. Sprungshöhen sind die absoluten bzw. relativen Häufigkeiten der zugehörigen Ausprägungen.

Beispiel 2.4 (vgl. Beispiel 2.1):
In Beispiel 2.1 gibt die empirische Verteilungsfunktion $F_{50}(x)$ an der ganzzahligen Stelle j den relativen Anteil derjenigen Familien an, die höchstens j Kinder haben für $j = 0, 1, \ldots, 5$. Bis zur nächsten Sprungstelle ist die Treppenfunktion konstant. Aus der Tabelle 2.2 erhält man die Verteilungsfunktion:

$F_{50}(x) = 0$ für $x < 0$; $F_{50}(x) = 0{,}88$ für $3 \le x < 4$;

$F_{50}(x) = 0{,}24$ für $0 \le x < 1$; $F_{50}(x) = 0{,}96$ für $4 \le x < 5$;

$F_{50}(x) = 0{,}58$ für $1 \le x < 2$; $F_{50}(x) = 1$ für $x \ge 5$.

$F_{50}(x) = 0{,}76$ für $2 \le x < 3$;

Die empirische Verteilungsfunktion ist in Bild 2.4 graphisch dargestellt.

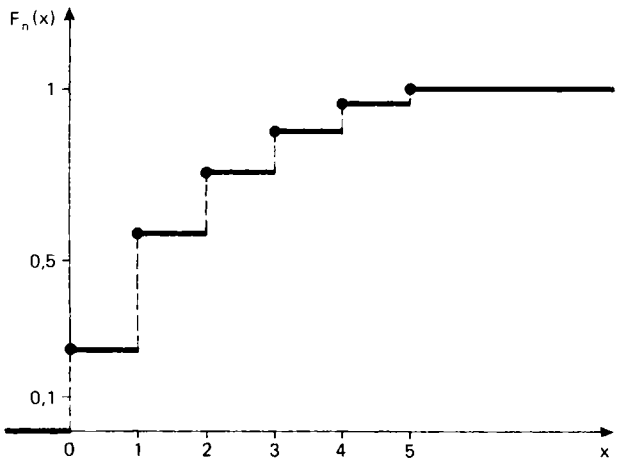


Bild 2.4: Empirische Verteilungsfunktion

2.3.3 Die klassierte empirische Verteilungsfunktion

Aus einer Klassenbildung können die Beobachtungswerte x_i nicht mehr genau festgestellt werden. Man sieht nur, wie viele der Werte in den einzelnen Klassen liegen. Daher kann der vollständige Verlauf der empirischen Verteilungsfunktion nicht exakt angegeben werden.
Man kann allerdings die genauen Werte der empirischen Verteilungsfunktion an den Klassengrenzen berechnen, weil aus der Klasseneinteilung ja feststellbar ist, wie viele Stichprobenwerte die rechte Klassengrenze nicht übersteigen. Wir nehmen an, alle m Klassen seien links offen und rechts abgeschlossen mit der Darstellung

$$K_j = (u_{j-1}, u_j] = \{x \mid u_{j-1} < x \leq u_j\}; \; j = 1, 2, \ldots, m.$$

Die Klasseneinteilung wird also durch die Randpunkte $u_0, u_1, \ldots, u_m$ festgelegt. Die Anzahl der Beobachtungswerte, welche die Grenze u_k nicht übersteigen, ist dann gleich der Summe der absoluten Häufigkeiten aller Klassen bis zur Stelle u_k. Mit den relativen Klassenhäufigkeiten r_j kann der Wert der empirischen Verteilungsfunktion an der Grenzstelle u_k berechnet werden durch

$$F_n(u_0) = 0; \; F_n(u_k) = \sum_{j=1}^{k} r_j \quad \text{für } k = 1, 2, \ldots, m. \tag{2.7}$$

Setzt man die Funktion zwischen diesen Stellen konstant fort, so erhält man zwar eine Treppenfunktion als Näherung für die empirische Verteilungsfunktion. Diese Näherung ist umso besser, je feiner die Klasseneinteilung gewählt wird. Bei diesem Vorgehen legt man allerdings sämtliche Werte einer Klasse an den rechten Randpunkt.
Die Verteilungsfunktion der nicht klassifizierten Stichprobe stimmt an allen Klassengrenzen u_k mit dieser Treppenfunktion überein. Zwischen zwei Klassengrenzen treten in der exakten Verteilungsfunktion allerdings weitere Sprungstellen auf. Man könnte nun die in einer Klasse liegenden Stichprobenwerte auf den gesamten Klassenbereich gleichmäßig verteilen und davon die empirische Verteilungsfunktion berechnen. Die so erhaltene Funktion dürfte im allgemeinen besser mit der empirischen Verteilungsfunktion der Urliste übereinstimmen. Die Konstruktion dieser Näherungsfunktion ist jedoch bei großen Klassenbesetzungen sehr mühsam. Aus diesem Grund ist es naheliegend, die exakten Punkte auf der Verteilungsfunktion an den Klassengrenzen geradlinig miteinander zu verbinden. Dadurch entsteht eine stetige Funktion. Man nennt sie **klassierte empirische Verteilungsfunktion.** Verschiedene Klasseneinteilungen ergeben verschiedene klassierte Verteilungsfunktionen. Bei einer sehr feinen Klasseneinteilung stellt die klassierte Verteilungsfunktion eine gute Näherung der empirischen Verteilungsfunktion der Urliste dar. Die klassierte empirische Verteilungsfunktion ist diejenige Integralfunktion des Histogramms der entsprechenden Klasseneinteilung, welche am linken Grenzpunkt u_0 verschwindet.

Für die Klasseneinteilung in Tab. 2.3 ist die empirische Verteilungsfunktion in Bild 2.5 dargestellt. Sie ist die Integralfunktion des in Bild 2.3 dargestellten Histogramms, welche an der Stelle 0 verschwindet.

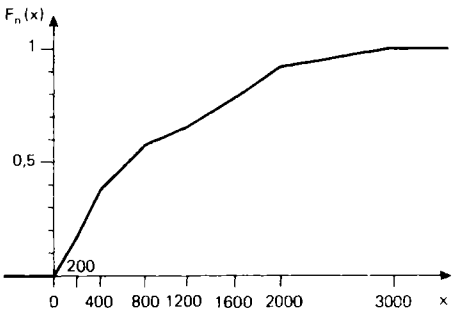


Bild 2.5: Klassierte empirische Verteilungsfunktion

2.4 Lageparameter von Häufigkeitsverteilungen

Bei vielen Problemstellungen ist es nicht sinnvoll, die ganze Beobachtungsreihe bzw. deren Häufigkeitsverteilung anzugeben. Man wird daher versuchen nur eine oder wenige Größen zu berechnen, welche die Beobachtungsreihe gut beschreiben. Solche Größen nennt man **Parameter** oder **Kenngrössen** der Stichprobe. In diesem Abschnitt werden Parameter angegeben, welche nur die Lage der Beobachtungsreihe beschreiben.

2.4.1 Der Modalwert (häufigster Wert)

Eine Merkmalsausprägung, welche in der Beobachtungsreihe die größte absolute Häufigkeit besitzt, heißt **Modalwert** (**Modus** oder **häufigster Wert**). Wir bezeichnen einen Modalwert mit x_{Mod}. Falls die maximale Häufigkeit gleichzeitig von mehreren Ausprägungen angenommen wird, gibt es mehrere Modalwerte. Für die absoluten Häufigkeiten des Modus gilt

$$h_n(x_{Mod}) \geq h_n(a_j) \quad \text{für alle Merkmalsausprägungen } a_j \,. \tag{2.8}$$

In Beispiel 2.1 hat der Kandidat B am meisten Stimmen bekommen. Er stellt also gewissermaßen bei dem Wahlvorgang den Modalwert dar. In Beispiel 2.2 ist "ein einziges Kind" der Modalwert. In Beispiel 2.3 sind die beiden Klassen K_2 und K_3 Modalklassen.

Der Modalwert kann bei sämtlichen Merkmalen bestimmt werden. Er ist bei nominal skalierten Merkmalen, bei denen es keine Rangordnung gibt, der einzige sinnvolle Lageparameter. Falls es nur einen einzigen Modus gibt, nennt man die Verteilung **eingipflig**.

2.4.2 Das arithmetische Mittel (Mittelwert)

Das Gesamteinkommen einer bestimmten Personenschicht allein enthält nicht viel Information, falls nicht gleichzeitig mitgeteilt wird, um wie viele Personen es sich handelt. Division des Gesamteinkommens durch die Anzahl der entsprechenden Personen ergibt das Durchschnitts- oder das Pro-Kopf-Einkommen, das wesentlich mehr Information enthält. Bei der Berechnung des durchschnittlichen Zuckerverbrauchs wird der gesamte Zuckerverbrauch durch die Anzahl der Personen dividiert. Dieser Durchschnittswert allein läßt jedoch keine Aussage über den Verbrauch der einzelnen Personen zu. Manche werden viel mehr, manche wesentlich weniger Zucker konsumiert haben.

Bei metrisch skalierten Merkmalen heißt der Zahlenwert

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot a_j = \sum_{j=1}^{m} r_j \cdot a_j$$

das **arithmetische Mittel (Mittelwert** oder **Durchschnittswert)** der Beobachtungsreihe. Oft nennt man $\bar{x}$ *den* Mittelwert und läßt den Zusatz arithmetisch weg.
Falls die Werte nur in Form einer Urliste gegeben sind, wird zur Berechnung des Mittelwertes die erste Gleichung benutzt. Die zweite oder dritte Darstellung verwendet man bei Häufigkeitsverteilungen. Wegen

$$n \cdot \bar{x} = x_1 + x_2 + \ldots + x_n = \sum_{i=1}^{n} x_i \tag{2.9}$$

beschreibt das arithmetische Mittel immer die Gesamtsumme. Bei vielen Problemstellungen wird nur der Durchschnittswert $\bar{x}$ angegeben, z.B. der Pro-Kopf-Verbrauch oder das Durchschnittseinkommen. Multipliziert man diesen Durchschnittswert mit der Anzahl, bezüglich derer der Durchschnitt gebildet wurde, so erhält man den Gesamtverbrauch bzw. das Gesamteinkommen. Würde z. B. die gesamte Lohnsumme eines Betriebs unter allen Betriebsangehörigen gleichmäßig aufgeteilt werden, so müßte jede Person diesen Durchschnittswert erhalten.

Beispiel 2.5 (vgl. Beispiel 2.2):
Für die Anzahl der Kinder pro Familie in Beispiel 2.2 erhält man das arithmetische Mittel

$$\bar{x} = \frac{1}{50}(12 \cdot 0 + 17 \cdot 1 + 9 \cdot 2 + 6 \cdot 3 + 4 \cdot 4 + 2 \cdot 5) = 1{,}58\,.$$

Bei diesen 50 Familien beträgt die mittlere Kinderzahl 1,58. Im Stabdiagramm in Bild 2.1 ist der Mittelwert $\bar{x} = 1{,}58$ ebenfalls eingetragen.

Allgemein stellt in einem Stabdiagramm das arithmetische Mittel den Abszissenwert des **Schwerpunkts** der Stäbe dar.

Beispiel 2.6 (vgl. Beispiel 2.3):
Die 50 Aggregate aus Beispiel 2.3 hatten zusammen eine Betriebsdauer von 44 497 Stunden. Division durch 50 ergibt die mittlere Betriebsdauer

$$\bar{x} = \frac{44\,497}{50} = 889{,}94 \approx 890 \text{ Stunden.}$$

Beispiel 2.7 (Durchschnittspreis beim Kauf gleicher Mengen zu verschiedenen Preisen):
Von einer Ware werde n-mal die gleiche Menge M, allerdings zu verschiedenen Preisen gekauft. Die Preise je Mengeneinheit seien der Reihe nach $p_1, p_2, \ldots, p_n$. Der Preis für die Gesamtmenge $n \cdot M$ beträgt damit

$$G = M \cdot \sum_{i=1}^{n} p_i .$$

Den Durchschnittspreis pro Mengeneinheit bezeichnen wir mit $\bar{p}$. Die Bedingung $G = n \cdot M \cdot \bar{p}$ ergibt den Durchschnittspreis als

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} p_i .$$

Falls zu verschiedenen Preisen jeweils die gleiche Menge gekauft wird, ist das arithmetische Mittel der Einzelpreise der Durchschnittspreis.

Der Mittelwert ist sehr empfindlich gegenüber sog. **Ausreißern**, die entweder viel größer oder viel kleiner als die übrigen Stichprobenwerte sind. Durch den Mittelwert werden die Beobachtungswerte zwar in zwei Gruppen zerlegt, die im speziellen Fall verschieden groß sein können. Im Extremfall kann es sogar vorkommen, daß auf der einen Seite vom Mittelwert nur ein einziger Beobachtungswert liegt, während alle übrigen auf der anderen Seite sind (vgl. Beispiel 2.10).

Lineare Transformation

Die Beobachtungswerte x_i werden durch

$$y_i = a + b x_i, \quad a, b \in \mathbb{R}$$

linear transformiert. Dann lautet der Mittelwert der transformierten Stichprobe $y = (y_1, y_2, \ldots, y_n)$

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

Damit transformiert sich auch der Mittelwert $\bar{y}$ nach dem gleichen Gesetz:

$$\overline{a + bx} = a + b\bar{x} \; ; \; a, b \in \mathbb{R}. \tag{2.10}$$

Aus einer **Klasseneinteilung** allein kann der Mittelwert nicht mehr exakt berechnet werden. In diesem Fall identifiziert man alle Werte einer Klasse mit der Klassenmitte und berechnet davon den Mittelwert. Dadurch erhält man einen Näherungswert für den tatsächlichen Mittelwert.

Beispiel 2.8 (vgl. Beispiele 2.3 und 2.6):
Für die mittlere Betriebsdauer der 50 Aggregate soll ein Näherungswert allein aus der Klasseneinteilung berechnet werden. Mit den Klassenmitten erhält man

$$\bar{x} \approx \frac{1}{50}(9 \cdot 100 + 10 \cdot 300 + 10 \cdot 600 + 4 \cdot 1000 + 6 \cdot 1400 + 7 \cdot 1800 + 4 \cdot 2500)$$
$$= 898 \text{ Stunden.}$$

Nach Beispiel 2.6 ist der genaue gerundete Wert $\bar{x} = 890$ Stunden. Im Histogramm in Bild 2.3 ist dieser Näherungswert für $\bar{x} = 898$ ebenfalls eingetragen.

Im Histogramm einer Klasseneinteilung stellt der aus den Klassenmitten berechnete Näherungswert für $\bar{x}$ den Abszissenwert vom **Schwerpunkt** des Histogramms dar.

Bemerkung: In einer Stichprobe eines diskreten Merkmals mit m verschiedenen Merkmalsausprägungen stimmt das Stichprobenmittel nur dann mit dem arithmetischen Mittel der m Ausprägungen überein, wenn alle m Häufigkeiten gleich groß sind. Es gilt also

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{a} = \frac{1}{m}\sum_{j=1}^{m} a_j \Leftrightarrow h_n(a_1) = h_n(a_2) = \ldots = h_n(a_m). \quad (2.11)$$

2.4.3 Gewichtete (gewogene) arithmetische Mittel

Gibt man sich allgemein n Zahlen (**Gewichte**) $w_1, w_2, \ldots, w_n$ vor mit

$$0 \le w_i \le 1 \text{ für alle i und } \sum_{i=1}^{n} w_i = 1,$$

so heißt

$$\bar{x}^w = \sum_{i=1}^{n} w_i \cdot x_i$$

ein **gewichtetes (gewogenes) arithmetisches Mittel** der Stichprobe.

Mit $w_1 = w_2 = \ldots = w_n = \frac{1}{n}$ erhält man das gewöhnliche arithmetische Mittel. Das gewöhnliche arithmetische Mittel ist also ein gewichtetes arithmetisches Mittel, bei dem alle n Gewichte gleich sind.

Der **Mittelwert** $\bar{x}$ ist das arithmetische Mittel aller n Beobachtungswerte $x_1, x_2, \ldots, x_n$, jedoch das mit den einzelnen relativen Häufigkeiten gewichtete arithmetische Mittel der m Merkmalsausprägungen $a_1, a_2, \ldots, a_m$.

Beispiel 2.9 (Durchschnittspreis beim Kauf verschiedener Mengen zu verschiedenen Preisen):
Von einer Ware werden M_i Einheiten zum Preis p_i pro Mengeneinheit gekauft für $i = 1, 2, \ldots, n$. Dann erhält man

Gesamtmenge: $\sum_{i=1}^{n} M_i$; Gesamtpreis: $G = \sum_{i=1}^{n} M_i \cdot p_i$.

Division des Gesamtpreises durch die Gesamtmenge ergibt den Durchschnittspreis

$$\bar{p}^w = \frac{\sum_{i=1}^{n} M_i \cdot p_i}{\sum_{k=1}^{n} M_k} = \sum_{i=1}^{n} w_i \cdot p_i \,.$$

Der Durchschnittspreis ist hier das gewichtete arithmetische Mittel der Einzelpreise mit den Gewichten

$$w_i = \frac{M_i}{\sum_{k=1}^{n} M_k} \quad \text{für } i = 1, 2, \ldots, n \,; \quad \sum_{i=1}^{n} w_i = 1.$$

Die Gewichte sind also proportional zu den Mengen.

2.4.4 Der Median (Zentralwert)

Beispiel 2.10:
Neun Personen erhalten folgende Gehälter in DM:

2 200; 2 250; 2 480; 2 700; **2 750**; 2 930; 3 000; 3 100; 16 480.

Die Gehälter sind also bereits der Größe nach geordnet. Der Mittelwert $\bar{x} = \frac{37\,890}{9} = 4\,210$ liegt nicht im Zentrum der Stichprobenwerte. Links von ihm befinden sich 8 Werte, rechts davon nur ein einziger. Der **Ausreißer** 16 480 zieht den Mittelwert stark nach oben.

Wir suchen nach einem Wert, der die Stichprobenwerte in zwei ungefähr gleich große Gruppen zerlegt.
Weil der Stichprobenumfang ungerade ist, gibt es genau einen Stichprobenwert, welcher in der Mitte der (der Größe nach) geordneten Stichprobenwerte liegt, nämlich der fünfte Wert 2 750. Dieser Wert ist der sog. **Median** oder **Zentralwert** $\tilde{x}$ der Stichprobe, also $\tilde{x} = 2\,750$.

Wir nehmen noch einen weiteren Wert dazu und erhalten die Stichprobe

2 150; 2 200; 2 250; 2 480; **2 700; 2 750**; 2 930; 3 000; 3 100; 16 480

vom Umfang $n = 10$ (gerade). Bei geradem Stichprobenumfang n gibt es keinen Einzelwert, sondern gleichzeitig zwei Stichprobenwerte, die in der Mitte der geordneten Stichprobe liegen.
Bei geradem Stichprobenumfang nennt man die beiden in der Mitte der geordneten Stichprobe stehenden Stichprobenwerte **Mediane** (**Zentralwerte**). Man kann aber auch jeden zwischen diesen beiden Stichprobenwerten liegenden Zahlenwert als Median bezeichnen. Dann spricht man vom **Medianintervall** [2 700; 2 750].
Um den Median eindeutig festzulegen, gibt man oft die Intervallmitte an, hier also den Wert $\tilde{x} = 2\,725$.

Der **Median** oder **Zentralwert** $\tilde{x}$ einer Beobachtungsreihe kann jeweils durch eine der beiden nachfolgenden gleichwertigen Eigenschaften erklärt werden:

a) Mindestens die Hälfte der Beobachtungswerte sind kleiner oder gleich und mindestens die Hälfte größer oder gleich dem Median $\tilde{x}$.
b) Höchstens die Hälfte der Beobachtungswerte sind kleiner und höchstens die Hälfte größer als der Median $\tilde{x}$.

Der Median kann nur von **ordinal** oder **metrisch skalierten Merkmalen** berechnet werden. Weil die Merkmalsausprägungen zur Bestimmung des Medians in einer Reihenfolgen angeordnet werden, muß eine Rangordnung (Größer-Kleiner-Beziehung) vorgegeben sein.

Zunächst werden die Beobachtungswerte der Größe nach geordnet. Diese geordneten Werte bezeichnet man der Reihe nach mit

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(n)} .$$

Bei **ungeradem n** ist der **Median (Zentralwert)** $\tilde{x}$ der in der Mitte der geordneten Reihe stehende Beobachtungswert, also

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} , \text{ falls n ungerade ist.}$$

Bei **geradem n** erfüllt jeder Wert zwischen $x_{\left(\frac{n}{2}\right)}$ und $x_{\left(\frac{n}{2}+1\right)}$, die Grenzen eingeschlossen, die Bedingung eines Medians. Dann ist jeder Wert zwischen diesen beiden Merkmalswerten **Median (Zentralwert)**.
Bei **metrisch skalierten** Merkmalen wählt man häufig das arithmetische Mittel der beiden mittleren Stichprobenwerte als Median, also

$$\tilde{x} = \frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right) \quad \text{für gerades n.}$$

Diese Mittelwertbildung ist allerdings bei ordinalen Merkmalen nicht möglich.

Bestimmung des Medians aus der Häufigkeitstabelle

a) Springt die relative Summenhäufigkeit bei einem Merkmalswert von unter 0,5 auf über 0,5, so ist dieser Merkmalswert der Median.
b) Ist die relative Summenhäufigkeit eines Merkmalswerts gleich 0,5, so ist jeder Wert zwischen diesem und dem nächst größeren Merkmalswert Median.

Beispiel 2.11 (vgl. Beispiel 2.2):
In der nachfolgenden Tabelle 2.4 springt beim Merkmalswert 1 die relative Summenhäufigkeit erstmals auf über 0,5. Daher lautet der Median $\tilde{x} = 1$.

Anzahl der Kinder	relative Häufigkeit	relative Summenhäufigkeit	
0	0,24	0,24	
1	0,34	0,58	← Median
2	0,18	0,76	
3	0,12	0,88	
4	0,08	0,96	
5	0,04	1,00	

Tab. 2.4: Bestimmung des Medians aus einer Häufigkeitstabelle

Beispiel 2.12:
In der nachfolgenden Häufigkeitstabelle ist beim Merkmalswert 20 die relative Summenhäufigkeit gleich 0,5. Daher sind 20 und 25 gleichzeitig Mediane.

a_j	relative Häufigkeit	relative Summenhäufigkeit	
10	0,18	0,18	
20	0,32	0,50	← Median
25	0,41	0,91	← Median
30	0,09	1,00	

Tab. 2.5: Bestimmung des Medians aus einer Häufigkeitstabelle

Bestimmung des Medians aus der empirischen Verteilungsfunktion

Die Bestimmung des Medians aus der Häufigkeitstabelle ergibt unmittelbar die folgende Eigenschaft:

a) Falls die empirische Verteilungsfunktion auf einer Treppenstufe den Wert 0,5 annimmt, sind dieser Merkmalswert und der nächst größere Mediane.
b) Wenn die empirische Verteilungsfunktion den Wert 0,5 nicht annimmt, ist der Median gleich dem kleinsten Merkmalswert, an dem die Verteilungsfunktion größer als 0,5 ist.

Beispiel 2.13:
In der nachfolgenden empirischen Verteilungsfunktion auf der linken Seite erhält man die Mediane 3 und 4. Rechts ist der Median $\tilde{x} = 3$ eindeutig bestimmt.

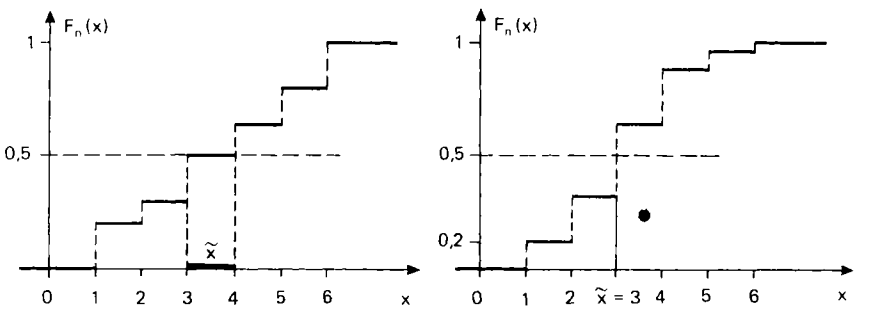


Bild 2.6: Bestimmung des Medians aus der Verteilungsfunktion

Median einer transformierten Stichprobe

Bei einem **metrisch skalierten** Merkmal werden die Stichprobenwerte x_i durch die Funktion g abgebildet auf $y_i = g(x_i)$ für $i = 1, 2, \ldots, n$. Falls die Funktion g **streng monoton wachsend** oder **fallend** ist, besitzt die transformierte Stichprobe $\left(y_1 = g(x_1), y_2 = g(x_2), \ldots, y_n = g(x_n)\right)$ den Median

$$\tilde{y} = g(\tilde{x}) . \tag{2.12}$$

Diese Eigenschaft ist allerdings nur dann richtig, wenn man bei geradem n sämtliche Werte zwischen den beiden in der Mitte der geordneten Stichprobenwerte oder nur diese beiden als Mediane zuläßt. Benutzt man jedoch als Median nur das arithmetische Mittel dieser beiden mittleren Stichprobenwerte, so gilt die Eigenschaft (2.12) im allgemeinen nicht mehr.

Jede lineare Transformation $y = a + bx$ mit $a, b \in \mathbb{R}$ ist streng monoton.

$$\text{Aus } y_i = a + bx_i,\ a, b \in \mathbb{R} \quad \text{folgt} \quad \tilde{y} = a + b\tilde{x} . \tag{2.13}$$

Diese Eigenschaft gilt bei geradem n auch für das arithemische Mittel aus den beiden mittleren Stichprobenwerten.

Bemerkung: Nach (2.10) wird bei linearen Transformationen der Mittelwert ebenfalls linear transformiert. Bei streng monotonen Funktionen g ist im allgemeinen $g(\bar{x})$ nicht mehr der Mittelwert der transformierten Stichprobe. Falls man z. B. die Stichprobenwerte quadriert, besitzt die Stichprobe $y_i = x_i^2$ im allgemeinen nicht mehr den Mittelwert $\bar{x}^2$, obwohl die Quadratfunktion streng monoton wachsend ist.

Aus einer Klasseneinteilung allein läßt sich der Median nicht mehr exakt bestimmen. Man kann nur diejenige Klasse K_l feststellen, in welcher der Median enthalten ist. Als Näherungswert für den Median wählen wir denjenigen Wert, an dem die klassierte Verteilungsfunktion (vgl. Abschnitt 2.3.3) den Wert $\frac{1}{2}$ annimmt. Das ist diejenige Stelle, die das Histogramm der relativen Klassenhäufigkeiten in zwei gleich große Teile mit dem jeweiligen Flächeninhalt 0,5 teilt.

Die Medianklasse K_l besitze den linken Randpunkt $u_l - 1$ und die Breite b_l. Der gesuchte Flächenhalbierungspunkt sei $u_{l-1} + c \cdot b_l$. Im Histogramm beträgt dann der Fächeninhalt links von diesem Punkt

$$\sum_{j=1}^{l-1} r_j + c \cdot r_l = 0{,}5 \quad \text{mit} \quad c = \frac{0{,}5 - \sum_{j=1}^{l-1} r_j}{r_l} \,. \tag{2.14}$$

Damit erhält man für den Median den Näherungswert

$$\tilde{x} \approx u_{l-1} + \frac{0{,}5 - \sum_{j=1}^{l-1} r_j}{r_l} \cdot b_l \,. \tag{2.15}$$

Für $\sum_{j=1}^{l} r_j = 0{,}5$ erhält man $\tilde{x} \approx u_l =$ rechte Klassengrenze. In diesem Fall liegen in der l-ten Klasse oder links davon die Hälfte der Stichprobenwerte.

Beispiel 2.14 (vgl. Beispiel 2.3):
Allein aus der Klasseneinteilung von Beispiel 2.3 soll ein Näherungswert für den Median bestimmt werden. Der Median liegt in der dritten Klasse. Einen Näherungswert erhält man nach (2.14) und (2.15) als

$$\tilde{x} \approx 400 + \frac{\frac{1}{2} - 0{,}38}{0{,}2} \cdot 400 = 640 \,.$$

Diesen Näherungswert erhält man auch aus der in Bild 2.5 dargestellten klassierten empirischen Verteilungsfunktion. Sie nimmt an der Stelle 640 den Wert $\frac{1}{2}$ an. Im Histogramm in Bild 2.5 ist dieser Näherungswert $\tilde{x} = 640$ ebenfalls eingetragen. Links und rechts von dieser Stelle liegt jeweils eine Fläche mit dem Inhalt 0,5. In der sortierten Orginalstichprobe vom Umfang $n = 50$ lautet der 25. Wert 647 und der 26. ist 671. Alle Zahlen zwischen diesen beiden sind die tatsächlichen Mediane.

Bemerkungen:
Im Gegensatz zum Mittelwert liegt der Median immer im Zentrum der geordneten Stichprobenwerte. Er ist unempfindlich gegenüber Ausreißern.
Der Median kann nicht nur bei metrisch skalierten, sondern auch bei **ordinalen** qualitativen Merkmalen berechnet werden, bei denen die Berechnung des arithmetischen Mittels gar nicht möglich ist. Zur Bestimmung des Medians benötigt man nur eine Anordnung (Rangreihenfolge).

2.4.5 Quantile und Quartile

Der Median einer Stichprobe ist dadurch charakterisiert, daß mindestens 50 % aller Beobachtungswerte diesen Wert nicht übersteigen und mindestens 50 % der Werte diesen nicht unterschreiten. Bei vielen Problemen möchte man jedoch wissen, wie viele der Beobachtungswerte zu den 10 oder 20 % kleinsten bzw. größten Beobachtungswerten gehören. Falls jemand in einer Prüfung zu den 10 % besten gehört, liegt dessen Leistung unter den 10 % größten Werten der Zensuren. Eine schlechtere Note müssen dann mindestens 90 % der Teilnehmer haben. Bei der Untersuchung der Studiendauer interessiert man sich oft für die maximale Semesteranzahl der 90 % Studierenden, die das Studium zuerst beenden, also für die maximale Studienzeit der 90 % "am schnellsten Studierenden".

Allgemein werden Zerlegungen der geordneten Beobachtungsreihen durch sogenannte **Quantile** oder **Fraktile** beschrieben.
Ein **α-Quantil** (**α-Fraktil**) $\tilde{x}_\alpha$ wird durch die beiden gleichwertigen Eigenschaften definiert:

a) Mindestens $100 \cdot \alpha$ % der Beobachtungswerte sind kleiner oder gleich $\tilde{x}_\alpha$ und mindestens $100 \cdot (1-\alpha)$ % größer oder gleich $\tilde{x}_\alpha$.

b) Höchstens $100 \cdot \alpha$ % der Beobachtungswerte sind kleiner als $\tilde{x}_\alpha$ und höchstens $100 \cdot (1-\alpha)$ % größer als $\tilde{x}_\alpha$.

Falls ein α-Quantil mit keinem Beobachtungswert übereinstimmt, teilt es die aufsteigend geordnete Beobachtungsreihe im Verhältnis α zu $1-\alpha$.

Wie der Median können Quantile nur von **ordinal** oder **metrisch skalierten** Merkmalen berechnet werden.

Die Beobachtungswerte seien der Größe nach geordnet durch

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)} \cdot$$

Dann kann das **α-Quantil** oder das **$100\,\alpha$ %-Quantil** $\tilde{x}_\alpha$ folgendermaßen bestimmt werden:

1. Fall: $n \cdot \alpha$ sei **nicht ganzzahlig**. Es sei k die auf $n \cdot \alpha$ folgende ganze Zahl, d. h. die kleinste ganze Zahl, welche größer als $n \cdot \alpha$ ist. Dann gilt

$$\tilde{x}_\alpha = x_{(k)}\,;\ k = \text{kleinste ganze Zahl mit } k < n \cdot \alpha \text{ (nicht ganzzahlig)}.$$

2. Fall: $n \cdot \alpha = k$ sei **ganzzahlig**. Dann sind sowohl $x_{(k)}$ als auch $x_{(k+1)}$ und jeder dazwischen liegende Wert α-Quantile. Bei **metrisch skalierten** Merkmalen benutzt man oft das arithmetische Mittel dieser beiden Stichprobenwerte, um das Quantil eindeutig festzulegen, also

$$\tilde{x}_\alpha = \frac{1}{2} \cdot \left(x_{(k)} + x_{(k+1)}\right) \quad \text{für} \quad k = n \cdot \alpha \ \text{(ganzzahlig)}.$$

Bei **ordinalen** Merkmalen ist die Mittelwertsbildung nicht möglich.

Bestimmung der Quantile aus der empirischen Verteilugsfunktion

1. Fall: Die empirische Verteilungsfunktion F_n nimmt den Wert α nicht an. Dann ist das α-Quantil $\tilde{x}_\alpha$ der kleinste Merkmalswert, an dem die empirische Verteilungsfunktion größer als α ist.

2. Fall: Die empirische Verteilungsfunktion F_n ist auf einer ganzen Treppenstufe gleich α. Dann sind alle Abszissenwerte dieser Treppenstufe α-Quantile. Zur eindeutigen Darstellung wird manchmal der Abszissenwert der Stufenmitte als Median gewählt.

Das 0,5-Quantil ist der Median, es gilt also

$$\tilde{x}_{0,5} = \tilde{x}. \tag{2.16}$$

Das 0,25-Quantil $\tilde{x}_{0,25}$ nennt man auch **unteres Quartil** und $\tilde{x}_{0,75}$ **oberes Quartil.**

Beispiel 2.15 (vgl. Beispiele 2.2 und 2.4):
Für die in Beispiel 2.2 (Tab. 2.2) dargestellte Stichprobe sollen die 0,25-; 0,5- (Median); 0,75- und 0,96-Quantile bestimmt werden.

$\alpha = 0{,}25$: $50 \cdot 0{,}25 = 12{,}5$ ist nicht ganzzahlig. Damit ist in der geordneten Stichprobe der 13. Wert das Quantil, also $\tilde{x}_{0,25} = x_{(13)} = 1$.

$\alpha = 0{,}5$ (Median): $50 \cdot 0{,}5 = 25$ ist ganzzahlig mit

$\tilde{x}_{0,5} = \tilde{x} = x_{(25)} = x_{(26)} = 1$.

$\alpha = 0{,}75$: $50 \cdot 0{,}75 = 37{,}5$, also $\tilde{x}_{0,75} = x_{(38)} = 2$.

$\alpha = 0{,}96$: $50 \cdot 0{,}96 = 48$. Dann sind $x_{(48)} = 4$ und $x_{(49)} = 5$ α-Quantile.

Diese Quantile können auch mit Hilfe der Verteilungsfunktion in Bild 2.7 (vgl. Bild 2.4) bestimmt werden.

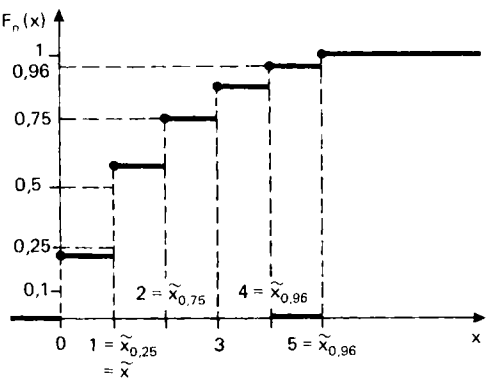


Bild 2.7: Quantilbestimmung aus der Verteilungsfunktion

Bei einer **Klasseneinteilung** lassen sich die Quantile nicht mehr genau feststellen, sondern nur noch diejenige Klasse K_l, in der das entsprechende Quantil liegt. Für die Quantilklasse K_l gilt

$$\sum_{j=1}^{l-1} r_j < \alpha \quad \text{und} \quad \sum_{j=1}^{l} r_j \geq \alpha \quad \text{mit} \quad \sum_{j=1}^{0} r_j = 0 \,. \tag{2.17}$$

Als Näherungswert kann ähnlich wie beim Median diejenige Stelle der Medianklasse bestimmt werden, von welcher die Gesamtfläche des Histogramms der relativen Häufigkeiten so zerlegt wird, daß links von ihr der Flächeninhalt α und rechts davon der Flächeninhalt $1-\alpha$ entsteht. Mit der linken Klassengrenze u_{l-1}, der Klassenbreite b_l und der relativen Klassenhäufigkeit erhält man für das α-Quantil den Näherungswert

$$\tilde{x}_\alpha \approx u_{l-1} + \frac{\alpha - \sum_{j=1}^{l-1} r_j}{r_l} \cdot b_l \,. \tag{2.18}$$

Für $\sum_{j=1}^{l} r_j = \alpha$ ergibt $\tilde{x}_\alpha \approx u_l =$ rechte Klassengrenze. In diesem Fall liegen in der l-ten Klasse oder links davon genau $100 \cdot \alpha$ % der Stichprobenwerte.

Aus der klassierten empirischen Verteilungsfunktion können Quantile sehr einfach graphisch bestimmt werden. Diejenige Stelle ist der Näherungswert für $\tilde{x}_\alpha$, an der die klassierte Verteilungsfunktion den Wert α annimmt.

Beispiel 2.16 (vgl. Beispiele 2.3 und 2.4):
Allein aus der Klasseneinteilung in Tab. 2.3 sollen Näherungswerte für die 0,25-, 0,75- und 0,92-Quantile bestimmt werden.

$\underline{\alpha = 0{,}25:}$ aus (2.17) und (2.18) erhält man $l = 2$ und

$$\tilde{x}_{0,25} \approx 200 + \frac{0{,}25 - 0{,}18}{0{,}20} \cdot 200 = 270 \,;$$

$\underline{\alpha = 0{,}75:}$ $l = 5$; $\tilde{x}_{0,75} \approx 1\,200 + \frac{0{,}75 - 0{,}66}{0{,}12} \cdot 400 = 1\,500 \,;$

$\underline{\alpha = 0{,}92:}$ $\sum_{j=1}^{6} r_j = 0{,}92$ ergibt $\tilde{x}_{0,92} \approx 2\,000$ (rechte Klassengrenze).

Die Näherungswerte für die Quantile $\tilde{x}_\alpha$ können auch direkt aus der in Bild 2.5 skizzierten klassierten Verteilungsfunktion abgelesen werden. Es sind diejenigen Stellen, an denen die klassierte Verteilungsfunktion jeweils den Wert α annimmt.
Sortiert man die Werte der Urliste in aufsteigender Reihenfolge, so erhält man nach dem oben beschriebenen Verfahren folgende exakte Quantile
$\tilde{x}_{0,25} = x_{(13)} = 255$; $\tilde{x}_{0,75} = x_{(38)} = 1\,558$.

Wegen $50 \cdot 0{,}92 = 46$ sind $x_{(46)} = 1\,902$ und $x_{(47)} = 2\,128$ und alle dazwischen liegenden Werte 0,92-Quantile. Mit dem arithmetischen Mittel dieser beiden Stichprobenwerte erhält man $\tilde{x}_{0,92} = 2\,015$.

2.4.6 Das harmonische Mittel

Falls ein Autofahrer immer die gleichen Zeiten mit jeweils konstanten Geschwindigkeiten fährt, ist die Durchschnittsgeschwindigkeit das arithmetische Mittel der Einzelgeschwindigkeiten. Diese Mittelwertbildung darf jedoch nicht mehr benutzt werden, wenn gleich oder gar verschieden lange Strecken mit verschiedenen Geschwindigkeiten gefahren werden. Dazu das

Beispiel 2.17:
Ein Autofahrer möchte eine Strecke von 450 km fahren. Für die Zeitplanung geht er von folgender Vorstellung aus: jeweils ein Drittel der Strecke möchte er mit den konstanten Geschwindigkeiten (in km/h) $x_1 = 150$, $x_2 = 100$ und $x_3 = 75$ fahren. Gesucht ist die Durchschnittsgeschwindigkeit bei Einhaltung dieser Bedingungen. In der nachfolgenden Tabelle 6 sind die für die einzelnen Strecken benötigten Zeiten angegeben.

Streckenlänge	Durchschnittsgeschwindigkeit in $\frac{\text{km}}{\text{h}}$	benötigte Zeit in Stunden
150	150	1
150	100	1,5
150	75	2

Tab. 2.6: Tabelle zur Berechnung des harmonischen Mittels

Für die Gesamtstrecke 450 km werden 4,5 Stunden benötigt. Daraus erhält man die Durchschnittsgeschwindigkeit

$$\bar{x}_h = \frac{450}{4{,}5} = 100 \text{ km/h}.$$

Dieser Durchschnittswert ist kleiner als das arithmetische Mittel der drei Einzelgeschwindigkeiten

$$\bar{x} = \frac{1}{3}(150 + 100 + 75) = \frac{325}{3} \approx 108{,}33\,.$$

Die Durchschnittsgeschwindigkeit kann folgendermaßen dergestellt werden:

$$\bar{x}_h = \frac{450}{4{,}5} = \frac{450}{\frac{150}{150} + \frac{150}{100} + \frac{150}{75}} = \frac{1}{\frac{150}{450}\left(\frac{1}{150} + \frac{1}{100} + \frac{1}{75}\right)}$$

$$= \frac{1}{\frac{1}{3}\left(\frac{1}{150} + \frac{1}{100} + \frac{1}{75}\right)}\,.$$

Im Nenner des letzten Bruches steht der Mittelwert der reziproken Stichprobenwerte $\frac{1}{x_1}$, $\frac{1}{x_2}$ und $\frac{1}{x_3}$.

Man nennt $\bar{x}_h$ das **harmonische Mittel** der Beobachtungswerte.

Das **harmonische Mittel** der Stichprobe $(x_1, x_2, \ldots, x_n)$ soll (kann) nur berechnet werden, wenn die n Beobachtungswerte entweder alle positiv oder alle negativ sind. Es ist erklärt durch

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}}$$

$$= \frac{n}{\sum_{j=1}^{m} \frac{h_n(a_j)}{a_j}} = \frac{1}{\sum_{j=1}^{m} \frac{r_n(a_j)}{a_j}}.$$

Das harmonische Mittel ist der Kehrwert (reziproke Wert) des arithmetischen Mittels der reziproken Beobachtungswerte $\frac{1}{x_i}$, $i = 1, 2, \ldots, n$.

Beispiel 2.18 (Durchschnittspreis beim Kauf für gleiche Beträge zu verschiedenen Preisen):
Von einer Ware werde n-mal zu verschiedenen Preisen für den gleichen Betrag c gekauft. Zwischen den gekauften Mengen M_i und den zugehörigen Preisen p_i pro Mengeneinheit gilt also die Beziehung $M_i \cdot p_i = c$ (konstant). In Abhängigkeit vom Preis betragen die Kaufmengen $M_i = \frac{c}{p_i}$.
Damit gilt:

Gesamtpreis: $n \cdot c$

Gesamtmenge: $M = \sum_{i=1}^{n} M_i = \sum_{i=1}^{n} \frac{c}{p_i}$.

Hieraus erhält man den Durchschnittspreis

$$\frac{n \cdot c}{M} = \frac{n}{\frac{1}{p_1} + \frac{1}{p_2} + \ldots + \frac{1}{p_n}} = \frac{n}{\frac{1}{n}\left(\frac{1}{p_1} + \frac{1}{p_2} + \ldots + \frac{1}{p_n}\right)} = \bar{p}_h.$$

Beim Kauf zu verschiedenen Preisen für **jeweils gleiche Beträge** ist der Durchschnittspreis das **harmonische Mittel** der n Einzelpreise.

2.4.7 Gewichtete harmonische Mittel

Beispiel 2.19 (vgl. Beispiel 2.17): ein Fahrzeug fahre 150, 200 und 100 km jeweils mit den konstanten Geschwindigkeiten (in km/h): $x_1 = 150$, $x_2 = 100$, $x_3 = 50$. Gesucht ist die Durchschnittsgeschwindigkeit (in km/h) für die gesamte Strecke. Für die Strecke 450 km werden

$$\frac{150}{150} + \frac{200}{100} + \frac{100}{50} = 4 \text{ Stunden}$$

benötigt. Die Durschnittsgeschwindigkeit $\frac{450}{4} = 112{,}5$ kann berechnet werden durch

$$\bar{x}_h^w = \frac{450}{\frac{150}{150} + \frac{200}{100} + \frac{100}{50}} = \frac{1}{\frac{150}{450} \cdot \frac{1}{150} + \frac{200}{450} \cdot \frac{1}{100} + \frac{100}{450} \cdot \frac{1}{50}}.$$

Auf der rechten Seite steht im Nenner das gewichtete arithmetische Mittel der reziproken Stichprobenwerte $\frac{1}{x_1}, \frac{1}{x_2}, \frac{1}{x_3}$. Die Gewichte sind die relativen Streckenanteile. Man nennt diesen Wert **gewichtetes harmonisches Mittel.**

Mit den Gewichten w_i, $0 \leq w_i \leq 1$ für alle i und $\sum_{i=1}^{n} w_i = 1$ heißt

$$\bar{x}_h^w = \frac{1}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

ein **gewichtetes (gewogenes) harmonisches Mittel.**

$w_1 = w_2 = \ldots = w_n = \frac{1}{n}$ ergibt das gewöhnliche harmonische Mittel.

Das harmonische Mittel aller n Beobachtungswerte $x_1, x_2, \ldots, x_n$ ist nach Abschnitt 2.4.6 das mit den relativen Häufigkeiten gewichtete harmonischen Mittel der m verschiedenen Merkmalsausprägungen $a_1, a_2, \ldots, a_m$.

Beispiel 2.20 (Durchschnittspreis beim Kauf für verschiedene Beträge zu verschiedenen Preisen, vgl. Beispiel 2.18):
Von der gleichen Ware werden für die verschiedenen Beträge b_i zu den jeweiligen Preisen p_i (je ME) für $i = 1, 2, \ldots, n$ gekauft. Dann erhält man:

$$\text{Gesamtpreis:} \quad b = \sum_{i=1}^{n} b_i; \qquad \text{Gesamtmenge} : M = \sum_{i=1}^{n} \frac{b_i}{p_i}.$$

Der Durchschnittspreis

$$\bar{x}_h^w = \frac{\sum_{k=1}^{n} b_k}{\sum_{i=1}^{n} \frac{b_i}{p_i}} = \frac{1}{\sum_{i=1}^{n} \frac{w_i}{p_i}} \quad \text{mit} \quad w_i = \frac{b_i}{\sum_{k=1}^{n} b_k}; \quad \sum_{i=1}^{n} b_i = 1$$

ist das gewichtete harmonische Mittel der Einzelpreise, wobei die Gewichte die relativen Anteile der einzelnen Beträge am Gesamtbetrag sind.

2.4.8 Das geometrische Mittel

Beispiel 2.21 (mittlere Preissteigerung):
Während n Jahren stiegen die Preise für eine bestimmte Ware der Reihe nach um $p_1, p_2, \ldots, p_n\%$. Prozentuale Preissteigerung bedeutet dabei, daß der zu Beginn des i-ten Jahres gültige Preis am Ende des Jahres mit dem Preissteigerungsfaktor $q_i = 1 + p_i/100$ multipliziert werden muß. Mit dem Ausgangspreis A erhält man damit nach n Jahren den Endpreis $E_1 = A \cdot q_1 \cdot q_2 \cdot \ldots \cdot q_n$. Die durchschnittliche (mittlere) Preissteigerung p ist diejenige jährlich konstante Preissteigerung, die nach n Jahren zum gleichen Endpreis geführt hätte wie die verschiedenen Preissteigerungen. Mit dem Steigerungsfaktor $q = 1 + p/100$ erhält man den Endpreis

$$E_2 = A \cdot q^n = A \cdot \left(1 + \frac{p}{100}\right)^n .$$

Gleichsetzen von E_1 und E_2 ergibt

$$q^n = q_1 \cdot q_2 \cdot \ldots \cdot q_n ; \quad q = \sqrt[n]{q_1 \cdot q_2 \cdot \ldots \cdot q_n} .$$

Der mittlere Preissteigerungsfaktor q ist das sog. **geometrische Mittel** der einzelnen Preissteigerungsfaktoren. Hieraus erhält man die mittlere prozentuale Preissteigerung als $100 \cdot (q - 1)$ %.

Allgemein kann das geometrische Mittel nur für positive Beobachtungswerte erklärt werden, d. h. $x_i > 0$ für $i = 1, 2, \ldots, n$.

Das **geometrische Mittel** der n positiven Beobachtungswerte $x_1, x_2, \ldots, x_n$ ist erklärt durch

$$\begin{aligned} \bar{x}_g &= \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \\ &= \sqrt[n]{a_1^{h_1} \cdot a_2^{h_2} \cdot \ldots \cdot a_m^{h_m}} = a_1^{r_1} \cdot a_2^{r_2} \cdot \ldots \cdot a_m^{r_m} . \end{aligned}$$

Hier sind $a_1, a_2, \ldots, a_m$ die Merkmalsausprägungen, h_i die absoluten und r_i die relativen Häufigkeiten der jeweiligen Ausprägungen.
Mit Hilfe der ersten Zeile wird das geometrische Mittel aus einer Urliste, mit der zweiten aus einer Häufigkeitstabelle berechnet.

Mit Hilfe des geometrischen Mittels können durchschnittliche Wachstumsfaktoren berechnet werden.

2.4.9 Gewichtete (gewogene) geometrische Mittel

Beispiel 2.22 (mittlere Preissteigerung; vgl. Beispiel 2.21):
Für ein bestimmtes Produkt betrug in 4 Jahren die mittlere Preissteigerung 3,1 %, während der nachfolgenden 5 Jahre 3,45 % und in den letzten 6 Jahren 3,61 %. Gesucht ist die mittlere Preissteigerung in den gesamten 15 Jahren.
Für den gesuchten mittleren Preissteigerungsfaktor erhält man die Bedingung

$$q^{15} = 1{,}031^4 \cdot 1{,}0345^5 \cdot 1{,}0361^6 \approx 1{,}65614$$

mit der Lösung

$$\begin{aligned} q_g^w &= \sqrt[15]{1{,}031^4 \cdot 1{,}0345^5 \cdot 1{,}0361^6} \\ &= 1{,}031^{\frac{4}{15}} \cdot 1{,}0345^{\frac{5}{15}} \cdot 1{,}0361^{\frac{6}{15}} \approx 1{,}0342. \end{aligned}$$

Die mittlere Preissteigerung betrug somit während der 15 Jahre ungefähr 3,42 % pro Jahr.

q_g^w ist das **gewichtete geometrische Mittel** der drei Werte $q_1 = 1{,}031$; $q_2 = 1{,}0345$; $q_3 = 1{,}0361$. Die Gewichte sind die relativen Anteile der jeweiligen Jahre an der Gesamtzeit.

Mit den Gewichten w_i, $0 \leq w_i \leq 1$ für alle i und $\sum_{i=1}^{n} w_i = 1$ erhält man das **gewichtete (gewogene) geometrische Mittel** der Beobachtungsreihe als

$$\bar{x}_g^w = x_1^{w_1} \cdot x_2^{w_2} \cdot \ldots \cdot x_n^{w_n} \quad \text{für } x_i > 0 \text{ für alle i.}$$

$w_1 = w_2 = \ldots = w_n = \frac{1}{n}$ ergibt das gewöhnliche geometrische Mittel.

Das gewöhnliche geometrische Mittel $\bar{x}_g$ sämtlicher n Werte $x_1, x_2, \ldots, x_n$ der Urliste stimmt überein mit dem gewichteten geometrischen Mittel der verschiedenen Merkmalsausprägungen $a_1, a_2, \ldots, a_m$. Die Gewichte sind dabei die relativen Häufigkeiten.

2.4.10 Vergleich der verschiedenen Mittelwerte

Das arithmetische Mittel und der Median können nicht miteinander verglichen werden, einmal kann der eine Wert, ein anderes Mal der andere größer sein. Der Grund dafür ist die Empfindlichkeit des Medians gegenüber Ausreißern.

Falls alle Stichprobenwerte positiv sind, können das arithmetische, das geometrische und das harmonische Mittel miteinander verglichen werden.

Wenn alle n Stichprobenwerte übereinstimmen, sind diese drei Mittelwerte gleich, also

$$\bar{x}_h = \bar{x}_g = \bar{x} = x_1 \quad \text{für } x_1 = x_2 = \ldots = x_n > 0\,. \tag{2.19}$$

Falls nicht alle n Werte der Beobachtungsreihe gleich, also mindestens zwei voneinander verschieden sind, gilt allgemein

$$\bar{x}_h < \bar{x}_g < \bar{x}, \tag{2.20}$$

falls nicht alle Beobachtungswerte gleich sind mit $x_i > 0$ für alle i.

Die Beziehungen (2.19) und (2.20) sind auch für die entsprechenden gewichteten Mittelwerte richtig, falls bei allen drei Mittelwertsbildungen jeweils dieselben Gewichte benutzt werden.

2.5 Streuungsmaße (Streuungsparameter) von Häufigkeitsverteilungen

Die Lageparameter aus Abschnitt 2.4 liefern ohne Zusatzinformationen nicht genügend Information über die Häufigkeitsverteilung der Beobachtungsreihe. Bei quantitativen metrisch skalierten Merkmalen lassen die Mittelwerte allein keine Aussage darüber zu, ob alle oder wenigstens die meisten der Beobachtungswerte in ihrer Nähe oder weiter weg liegen. Oft möchte man gerne wissen, wie stark die Werte der Beobachtungsreihe um diese Lageparameter streuen. Die Abweichungen der Beobachtungswerte von einem Lageparameter werden durch sogenannte **Streuungsparameter** (**Streuungsmaße**) beschrieben. Diese können allerdings nur von Beobachtungswerten quantitativer Merkmale berechnet werden, deren Ausprägungen metrisch skaliert sind (reelle Zahlen). Je kleiner diese Streuungsmaße sind, umso besser wird die Häufigkeitsverteilung durch den entsprechenden Lageparameter beschrieben.

2.5.1 Die Spannweite

Die **Spannweite (Range)** R einer Beobachtungsreihe ist der Abstand des größten vom kleinsten Beobachtungswert, also

$$R = x_{(n)} - x_{(1)} = \text{größter minus kleinster Beobachtungswert.}$$

Die Spannweite beschreibt den gesamten Streubereich der Beobachtungsreihe und ist für Maßstabsbetrachtungen auf der x-Achse wichtig. Sie hängt nur vom kleinsten und größten Beobachtungswert ab. Daher ist die Spannweite sehr empfindlich gegenüber Ausreißern.

In Beispiel 2.2 lautet die Spannweite $R = 5 - 0 = 5$. In Beispiel 2.3 erhält man die Spannweite $2\,936 - 1 = 2\,935$; aus der Klasseneinteilung allein ergibt sich für die Spannweite nur der Näherungswert $3\,000 - 0 = 3\,000$ (Länge des Gesamtintervalls).

2.5.2 Der Quartilsabstand und Quantilsabstände

Zwischen dem oberen Quartil, dem 75 %-Quantil und dem unteren Quartil, dem 25 %-Quantil befinden sich mindestens 50 % aller Beobachtungswerte. Daher beschreibt der **Quartilsabstand**

$$\tilde{x}_{0,75} - \tilde{x}_{0,25}$$

die Länge des Bereichs, der mindestens die Hälfte aller Beobachtungswerte enthält.

Zwischen dem 5 %- und dem 95 %-Quantil liegen mindestens 90 % der Beobachtungswerte. Daher beschreibt die Differenz

$$\tilde{x}_{0,95} - \tilde{x}_{0,05}$$

die Länge eines Bereichs, in dem mindestens 90 % der Werte liegen.

$$\tilde{x}_{1-\alpha} - \tilde{x}_{\alpha} \quad \text{mit} \quad 0 < \alpha < 0{,}5$$

stellt als **Quantilsdifferenz** die Länge eines Bereichs dar, der mindestens $100 \cdot (1 - 2\alpha)$ % der Beobachtungswerte enthält.

2.5.3 Mittlere Abstände

Mittlere Abstände lassen sich nur bei kardinalen Merkmalen bestimmen. Von einem festen Zahlenwert c hat der Beobachtungswert x_i den Abstand $|x_i - c|$. Der **mittlere Abstand** (**mittlere absolute Abweichung**) von c ist

$$d_c = \frac{1}{n} \sum_{i=1}^{n} |x_i - c| = \frac{1}{n} \sum_{j=1}^{m} h_j \cdot |a_j - c| = \sum_{j=1}^{m} r_j \cdot |a_j - c|$$

mit h_j = absolute, r_j = relative Häufigkeit der Merkmalsausprägung a_j.

Für $c = \bar{x}$ erhält man den **mittleren Abstand vom Mittelwert** $\bar{x}$ als

$$d_{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| = \frac{1}{n} \sum_{j=1}^{m} h_j \cdot |a_j - \bar{x}| = \sum_{j=1}^{m} r_j \cdot |a_j - \bar{x}| .$$

$c = \tilde{x}$ ergibt den **mittleren Abstand vom Median**

$$d_{\tilde{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}| = \frac{1}{n} \sum_{j=1}^{m} h_j \cdot |a_j - \tilde{x}| = \sum_{j=1}^{m} r_j \cdot |a_j - \tilde{x}| .$$

Allgemein kann folgende Ungleichung bewiesen werden

$$d_{\tilde{x}} \leq d_c \quad \text{für jedes } c \in \mathbb{R} . \tag{2.21}$$

Der mittlere Abstand ist also bezüglich des Medians am kleinsten. Insbesondere gilt

$$d_{\tilde{x}} \leq d_{\bar{x}} . \tag{2.22}$$

Beispiel 2.23 (vgl. Beispiele 2.2, 2.5 und 2.11):
Für die Kinderzahl in den 50 Familien erhält man den mittleren Abstand vom Mittelwert $\bar{x} = 1{,}58$

$$d_{\bar{x}} = \frac{1}{50} (12 \cdot |0 - 1{,}58| + 17 \cdot |1 - 1{,}58| + 9 \cdot |2 - 1{,}58| + 6 \cdot |3 - 1{,}58| + 4 \cdot |4 - 1{,}58| + 2 \cdot |5 - 1{,}58| = 1{,}1528 .$$

Der mittlere Abstand vom Median $\tilde{x} = 1$ beträgt

$$d_{\tilde{x}} = \frac{1}{50} (12 \cdot |0 - 1| + 17 \cdot |1 - 1| + 9 \cdot |2 - 1| + 6 \cdot |3 - 1| + 4 \cdot |4 - 1| + 2 \cdot |5 - 1|) = 1{,}06 .$$

2.5.4 Varianz und Standardabweichung

Die mittleren absoluten Abweichungen $d_{\bar{x}}$ und $d_{\tilde{x}}$ lassen sich zwar einfach berechnen und beschreiben die Abweichungen der Beobachtungswerte vom jeweiligen Mittel auch ganz gut. In der beurteilenden Statistik sind diese Parameter jedoch für Hochrechnungen auf umfangreichere Grundgesamtheiten nicht geeignet. Aus diesem Grund ist das wohl am häufigsten benutzte Streuungsmaß die Varianz bzw. die Standardabweichung. Diese Streuungsparameter haben in der beurteilenden Statistik große Bedeutung.

Die **Varianz** s^2 einer Beobachtungsreihe $x_1, x_2, \ldots, x_n$ ist erklärt durch

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

$$= \frac{1}{n-1}\sum_{j=1}^{m} h_j \cdot (a_j - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{j=1}^{m} h_j \cdot a_j^2 - n\bar{x}^2\right),$$

$\bar{x}$ = Mittelwert; h_j = absolute Häufigkeit von a_j.

Die Gleichheit der beiden Ausdrücke der ersten Zeile läßt sich durch Ausquadrieren und Zusammenfassen beweisen. Der Vorteil bei der Benutzung des Ausdrucks auf der rechten Seite besteht darin, daß zur Berechnung der Varianz mit einem Rechner die Beobachtungswerte nicht gespeichert werden müssen. Es genügt, die Werte der Reihe nach einzugeben und nur ihre Summen und Quadratsummen zu bilden. Die Summe der Beobachtungswerte wird anschließend durch n dividiert, was den Mittelwert $\bar{x}$ ergibt. Mit Hilfe der Quadratsummen berechnet man dann nach der rechten Seite der ersten Zeile die Varianz. Die Berechnung nach der zweiten Zeile erfolgt bei Häufigkeitsverteilungen.

Die Varianz verschwindet nur dann, wenn alle n Beobachtungswerte übereinstimmen, also nur für $x_1 = x_2 = \ldots = x_n$.

Bemerkung: Zunächst wäre es naheliegend, im Ausdruck für die Varianz nicht durch $n-1$, sondern durch n zu dividieren, also die mittlere quadratische Abweichung

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}s^2 < s^2 \qquad (2.23)$$

zu benutzen. In der beurteilenden Statistik hat jedoch s^2 eine größere Anwendungsmöglichkeit als $\hat{s}^2$. Wegen $s^2 > \hat{s}^2$ verwendet man zwar einen Ausdruck, der etwas größer ist als die mittlere quadratische Abweichung. Bei umfangreichen Beobachtungsreihen ist der Unterschied allerdings gering.

Anstelle der Abstandsquadrate vom Mittelwert $\bar{x}$ könnte man auch Abweichungsquadrate bezüglich einer beliebigen reellen Zahl c wählen, also den Ausdruck

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - c)^2. \tag{2.24}$$

Für jede beliebige Zahl c gilt allgemein

$$\sum_{i=1}^{n}(x_i-c)^2 = \sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-c)]^2$$
$$= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(\bar{x}-c)}_{=0} + n\cdot(\bar{x}-c)^2 .$$

Es gilt also der

Steinersche Verschiebungssatz

$$\sum_{i=1}^{n}(x_i-c)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n\cdot(\bar{x}-c)^2 \quad \text{für jede Konstante c.} \tag{2.25}$$

Für $c = \tilde{x}$ erhält man hieraus

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\tilde{x})^2 > \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = s^2 \quad \text{für } \tilde{x}\neq\bar{x}. \tag{2.26}$$

Die mittleren quadratischen Abweichungen sind also bezüglich des Mittelwerts $\bar{x}$ minimal im Gegensatz zu den mittleren absoluten Abweichungen, bei denen das Minimum beim Median $\tilde{x}$ angenommen wird.

Die **Standardabweichung (Streuung)**

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

ist die Quadratwurzel aus der Varianz. Sie besitzt die gleiche Dimension wie die Beobachtungswerte x_i und der Mittelwert $\bar{x}$.

Beispiel 2.24 (vgl. Beispiele 2.2 und 2.5):
In Beispiel 2.2 erhält man die Varianz

$$s^2 = \frac{1}{49}(12\cdot 0 + 17\cdot 1^2 + 9\cdot 2^2 + 6\cdot 3^2 + 4\cdot 4^2 + 2\cdot 5^2 - 50\cdot 1{,}58^2)$$
$$\approx 1{,}9629;\ s \approx 1{,}401.$$

Lineare Transformation

Die Beobachtungsreihe $x_1, x_2, \ldots, x_n$ besitze den Mittelwert $\bar{x}$ und die Varianz s_x^2. Die linear transformierte Reihe $y_i = a + b x_i$ für $i = 1, 2, \ldots, n$ hat dann wegen $\bar{y} = a + b\bar{x}$ die Varianz

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a+bx_i-a-b\bar{x})^2$$
$$= b^2\cdot\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = b^2\cdot s_x^2 .$$

Damit gilt allgemein

$$s_{a+bx}^2 = b^2\cdot s_x^2;\quad s_{a+bx} = |b|\cdot s_x;\ a,b\in\mathbb{R}. \tag{2.27}$$

Eine Parallelverschiebung (a beliebig, b = 1) ändert also die Varianz und Standardabweichung nicht. Falls alle Werte mit b multipliziert werden, ändert sich die Varianz um den Faktor b^2 und die Standardabweichung um den Faktor $|b|$, also um den Betrag von b.

Aus einer **Klasseneinteilung** allein läßt sich die Varianz nicht mehr exakt bestimmen. Wie bei der Mittelwertbildung könnte man für sämtliche Werte einer Klasse die Klassenmitte wählen und die Varianz $\tilde{s}^2$ dieser Werte als Näherungswert für die Varianz der Urliste benutzen. Im allgemeinen erhält man bei dieser Näherung jedoch zu große Werte, d.h. die Varianz der Urliste wird hierdurch überschätzt. Bei gleichen Klassenbreiten b kann diese Überschätzung jedoch korrigiert werden durch die sogenannte

Sheppardsche Korrektur

$$s^2_{korr} = \tilde{s}^2 - \frac{b^2}{12}, \tag{2.28}$$

b = konstante Klassenbreite, $\tilde{s}^2$ = Varianz mit den Klassenmitten.

Mit diesen korrigierten Varianzen dürfen jedoch **keine statistischen Tests** durchgeführt werden.

Allgemein kann man zeigen, daß für $s > 0$ die mittlere absolute Abweichung $d_{\bar{x}}$ kleiner als die Sandardabweichung s ist, also

$$d_{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| < \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = s\,. \tag{2.29}$$

Deswegen stellt die Standardabweichung s ein geeignetes Abweichungsmaß dar. Nur wenn alle Stichprobenwerte gleich sind, verschwinden s und $d_{\bar{x}}$.

2.5.5 Der Variationskoeffizient

Die Standardabweichung s wird bezüglich des Mittelwertes $\bar{x}$ berechnet. Sie beschreibt, wie stark die Beobachtungswerte um den Mittelwert schwanken. Die tatsächliche Größe des Mittelwertes spielt dabei keine Rolle. Jede Parallelverschiebung der Beobachtungswerte ergibt die gleiche Standardabweichung, auch wenn dabei der Mittelwert noch so groß wird. Daher ist es manchmal sinnvoll, die Standardabweichung in Relation zum Mittelwert zu setzen. Für positive Merkmalswerte nennt man den Quotienten

$$v = \frac{s}{\bar{x}}$$

den **Variationskoeffizienten** der Stichprobe.

Der Variationskoeffizient bleibt als dimensionslose Größe von Maßstabsänderungen unberührt.

Für Beispiel 2.2 erhält man den Variationskoeffizienten $v \approx \frac{1{,}401}{1{,}58} \approx 0{,}887$.

2.5.6 Die Momente einer Verteilung

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot a_j^r = \sum_{j=1}^{m} r_j \cdot a_j^r \quad , r = 1,2,\ldots$$

heißt das r-te (**empirische**) **Anfangsmoment** und

$$m_r = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^r = \frac{1}{n}\sum_{j=1}^{m} h_j \cdot (a_j - \bar{x})^r = \sum_{j=1}^{m} r_j \cdot (a_j - \bar{x})^r, r = 1,2,\ldots$$

das r-te (**empirische**) **zentrale Moment** der Verteilung.

Wegen $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^1 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = 0$ verschwindet das erste zentrale Moment immer, für jede Beobachtungsreihe gilt also $m_1 = 0$.

Beispiel 2.25:
Das Stabdiagramm der relativen Häufigkeitsverteilung

a_j	1	2	3	4	5	6
r_j	0,1	0,15	0,25	0,25	0,15	0,1

ist in Bild 2.8 graphisch dargestellt. Es ist achsensymmetrisch. Die vertikale Symmetrie-Achse geht durch den Mittelwert $\bar{x} = 3{,}5$. Jeweils die beiden von der Symmetrie-Achse gleich weit entfernten Merkmalsausprägungen besitzen die gleiche relative Häufigkeit.

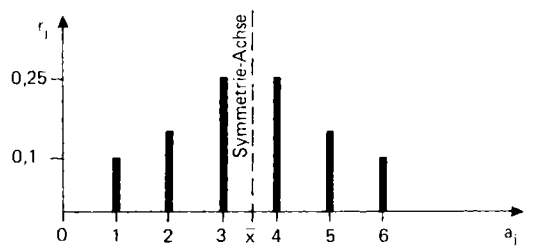


Bild 2.8: Symmetrische Verteilung

Bei **symmetrischen Verteilungen** ist die Symmetrie-Stelle der Mittelwert (Schwerpunkt) und gleichzeitig der Median. Sind in einer symmetrischen Verteilung die beiden Merkmalswerte a_k und a_l von $\bar{x}$ gleich weit entfernt, so ist $r_k = r_l$. Für ungerades r gilt ferner $(a_k - \bar{x})^r = -(a_l - \bar{x})^r$. Dann heben sich in der Summe $\sum r_j \cdot (a_j - \bar{x})^r$ jeweils 2 Summanden paarweise auf, weshalb die Summe verschwindet.

Bei symmetrischen Verteilungen verschwinden alle zentralen Momente ungerader Ordnung r, d.h.

$$m_r = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^r = 0 \quad \text{für} \quad r = 1,3,5,\ldots .$$

Bemerkung: Momente können auch bezüglich einer beliebigen Zahl $c \in \mathbb{R}$ erklärt werden durch

$$m_r(c) = \frac{1}{n}\sum_{i=1}^{n} (x_i - c)^r \qquad \text{mit} \quad m'_r = m_r(0) \; ; \; m_r = m_r(\bar{x}) .$$

2.5.7 Die Schiefe einer Häufigkeitsverteilung

Bei symmetrischen Häufigkeitsverteilungen verschwinden alle zentralen Momente ungerader Ordnung. Das erste zentrale Moment verschwindet auch bei asymmetrischen Verteilungen. Daher könnte man das dritte zentrale Moment m_3 als Maß für die Abweichung von der Symmetrie, also als Maß für die "Schiefe einer Verteilung" benutzen. Diese Größe ist jedoch maßstabsabhängig. Division durch

$$\hat{s}^3 = \left(\sqrt{\frac{n-1}{n}}\, s\right)^3 = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^3}$$

ergibt eine vom Maßstab unabhängige Größe

$$v_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\hat{s}^3} = \frac{m_3}{\hat{s}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^3}}\,.$$

Man nennt v_1 die **Schiefe** der Verteilung. Bei symmetrischen Verteilungen verschwindet die Schiefe. Je größer der Betrag $|v_1|$ ist, umso "schiefer" ist die Verteilung.

Für $v_1 < 0$ nennt man die Verteilung **linksschief**, für $v_1 > 0$ **rechtsschief**. Bei linksschiefen Verteilungen hängt das Stabdiagramm nach links, bei rechtsschiefen nach rechts durch.

Beispiel 2.26 (vgl. Beispiel 2.2):
Die Häufigkeitsverteilung aus Beispiel 2.2 besitzt nach Beispiel 2.24 die Varianz $s^2 \approx 1{,}9629$; nach (2.23) erhält man

$$\hat{s}^2 = \frac{49}{50}s^2 \approx 1{,}9236\,.$$

Das dritte zentrale Moment lautet

$$\sum_{j=1}^{m} r_j\cdot(a_j-\bar{x})^3 = 0{,}24\cdot(0-1{,}58)^3 + 0{,}34\cdot(1-1{,}58)^3 + 0{,}18\cdot(2-1{,}58)^3$$
$$+ 0{,}12\cdot(3-1{,}58)^3 + 0{,}08\cdot(4-1{,}58)^3 + 0{,}04\cdot(5-1{,}58)^3 \approx 2{,}0778.$$

Hieraus erhält man die Schiefe $v_1 = \dfrac{2{,}0778}{\sqrt{1{,}9236^3}} = 0{,}7788.$

Die Verteilung ist also rechtsschief, das in Bild 2.1 dargestellte Stabdiagramm "hängt nach rechts durch".

Bei **symmetrischen** Verteilungen verschwindet die Schiefe. Es gilt $\bar{x} = \tilde{x}$. Falls die Verteilung auch noch eingipflig ist, gilt auch noch $\bar{x} = \tilde{x} = x_{Mod}$.

Eingipflige Häufigkeitsverteilungen sind
rechtsschief (linkssteil), falls $\bar{x} > \tilde{x} > x_{Mod}$,
linksschief (rechtssteil), falls $\bar{x} < \tilde{x} < x_{Mod}$,
symmetrisch, falls $\bar{x} = \tilde{x} = x_{Mod}$.

2.6 Konzentrationsmaße

Falls zu einem bestimmten Zeitpunkt ein kleiner Anteil der Bevölkerung einen hohen Anteil an einem Gesamtbestand (z. B. Einkommen, Vermögen oder Wertpapierbesitz) hat, spricht man von einer Konzentration. In diesem Abschnitt untersuchen wir die Aufteilung der Summe von n Beobachtungswerten $x_1, x_2, \ldots, x_n$ auf die verschiedenen Merkmalsträger. Dabei interessiert uns, ob die Gesamtsumme ungefähr gleichmäßig verteilt oder auf wenige Merkmalsträger konzentriert ist. Die einzelnen Beobachtungswerte dürfen dabei nicht negativ sein.

Die m Ausprägungen eines metrisch skalierten Merkmals seien geordnet:

$$0 \leq a_1 < a_2 < \ldots < a_m .$$

In der Beobachtungsreihe $(x_1, x_2, \ldots, x_n)$ seien die Werte bereits der Größe nach geordnet, d.h.

$$0 \leq x_1 \leq x_2 \leq \ldots \leq x_n .$$

In der (der Größe nach) geordneten Beobachtungsreihe besitze die Merkmalsausprägung a_j die absolute Häufigkeit h_j und die relative Häufigkeit r_j . Die Gesamtsumme soll positiv sein, also

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{m} h_j \cdot a_j > 0 .$$

2.6.1 Die Lorenzkurve

Ein wichtiges graphisches Hilfsmittel zur Feststellung einer Konzentration ist die sogenannte **Lorenzkurve.** Sie soll in den nachfolgenden Unterabschnitten für eine Beobachtungsreihe, Häufigkeitsverteilung und Klasseneinteilung behandelt werden.

2.6.1.1 Die Lorenzkurve bei Einzelwerten (einer Beobachtungsreihe)

Die Beobachtungswerte seien bereits der Größe nach geordnet mit

$$0 \leq x_1 \leq x_2 \leq \ldots \leq x_n \quad \text{mit} \quad \sum_{i=1}^{n} x_i > 0 .$$

Die Träger der ersten k Beobachtungswerte besitzen an der Gesamtmenge der n Merkmalsträger den kumulierten relativen Anteil

$$u_k = \frac{k}{n} \quad \text{für} \quad k = 1, 2, \ldots, n .$$

Die Merkmalsträger mit den ersten k Beobachtungswerten haben an der Gesamtsumme $\sum_{i=1}^{n} x_i$ den kumulierten relativen Anteil

$$v_k = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{n} x_i} \quad \text{für} \quad k = 1, 2, \ldots, n .$$