Einführung in die Statistik für Wirtschafts-wissenschaftler

Von
Universitätsprofessor
Dr. Götz Uebe
und
Dr. Martin Schäfer

R. Oldenbourg Verlag München Wien

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Uebe, Götz:
Einführung in die Statistik für Wirtschaftswissenschaftler / von Götz Uebe und Martin Schäfer. - München ; Wien : Oldenbourg, 1991
ISBN 3-486-21759-3
NE: Schäfer, Martin:

Gesamtherstellung: R. Oldenbourg Graphische Betriebe GmbH, München

ISBN 3-486-21759-3

Vorwort

Das n-te Statistikbuch

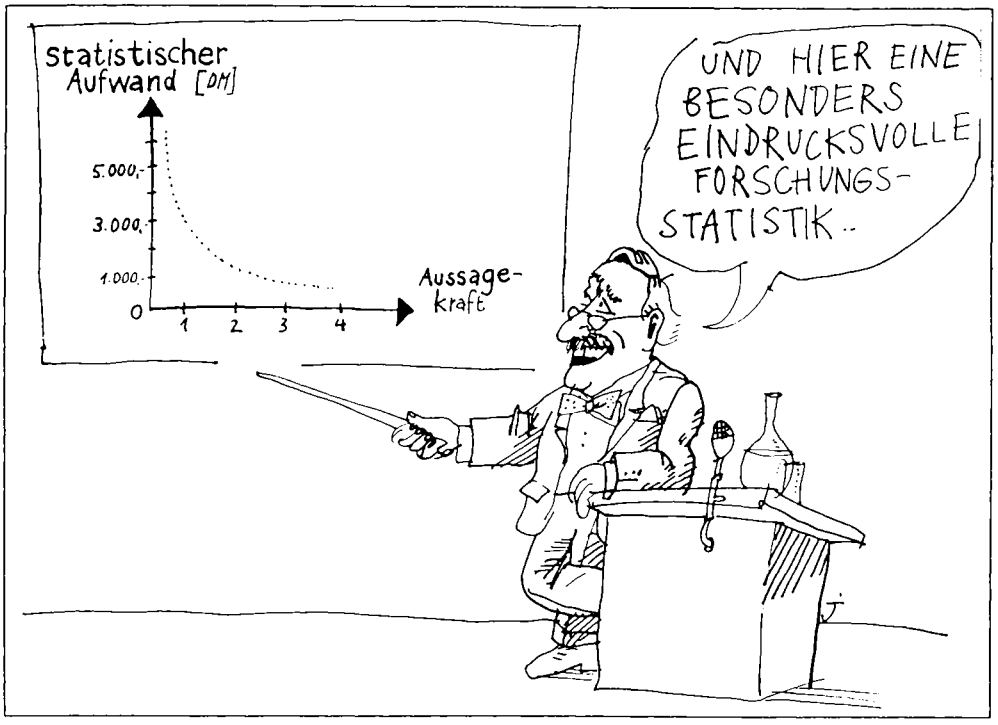
Zum Erscheinen eines weiteren Lehrbuchs der Statistik müssen sich die Autoren selbstverständlich fragen lassen, weshalb nun noch ein solches Werk geschrieben worden ist. Selbst in der Beschränkung auf die Veröffentlichungen der letzten Jahrzehnte strebt ihre Zahl (in Anlehnung an die Gesetze der großen Zahl) über alle überschaubaren Größen. Daß dies im exponentiellen Wachstum des statistischen Wissenstoffs seine Rechtfertigung findet, ist kaum zu behaupten, denn die meisten Werke sind wie diese Einführung auch und nehmen an der Wissensexplosion nur sehr gedämpft teil.
Hauptgrund ist wohl die Schwierigkeit der Übermittlung des statistischen Wissensstoffes, den der amerikanische Mathematiker J.A. Paulos (International Herald Tribune, 'You dolts are wrong about math', 25.4.1991, 7) in schöner Anschaulichkeit wie folgt charakterisiert: "Most students (and most adults) cannot interpret graphs, do not understand statistical notions, are unable to model situations mathematically, seldom estimate or compare magnitudes, are immune to mathematical beauty and, most distressing of all in a democracy, hardly ever develop a critical, skeptical attitude toward numerical, spatial and quantitative data or conclusions."
Aufgrund dieser Erkenntnis ist das Hauptanliegen ein pädagogisches: Zugeschnitten auf den Studiengang des Wirtschaftswissenschaftlers an der Universität der Bundeswehr Hamburg, sei es für einen Volkswirt, Betriebswirt oder Wirtschaftsingenieur, wird Statistik in einer zweitrimestrigen Vorlesung angeboten, nämlich als

(1) eine Einführung in die formale Untersuchung von Massenerscheinungen, für die es keine ausreichende substanzwissenschaftliche Erklärung gibt,
(2) ein Zweig der Mathematik, und vor allem
(3) ein System von Verfahren, Techniken, Vereinbarungen, Erfahrungen und Beispielen.

Statistik, Wissen und Wahrheit

Grundlegend dabei ist die Überzeugung, daß es ein Irrglaube ist, daß Daten, Zahlen oder bloße Tatsachen für sich selbst sprechen. Im Regelfall ist das nicht der Fall, wie z.B. die bekannte Aussage: "das Glas ist halb leer" *oder* "das Glas ist halb voll" illustriert. Fakten müssen aufbereitet und analysiert werden. Dazu kommt, daß über Statistik, Wissen und Wahrheit viel Schwachsinn palavert wird, z.B. "es gibt die einfache Lüge und es gibt die statistische Lüge". Solche flotten Sprüche machen das Fach noch schwieriger, als es dem Studenten ohnehin erscheint.
Der Standpunkt in dieser Vorlesung ist der wie in der Chirurgie: so wie man mit dem Messer heilen kann, so kann man auch damit töten. In allen Anwendungen, bei denen es um materielle Interessen geht, ist es selbstverständlich, daß auch diese Wissenschaft als Streithilfe herangezogen wird.
So wie in der Karikatur von Dr. Tomaschoff sollte unser Fach zumindest

nicht gesehen werden:



Mit einigem Ernst und etwas Fleiß sollte selbst der Student, der in seinem späteren Leben Statistik nicht hauptberuflich nutzt, einigen Gewinn davontragen. Sollte der Leser das nicht glauben, so wollen wir ihm zumindest unsere Überzeugung kundtun:
Für Statistiker gilt zumindest, was für Astrologen gilt "We are not" says Florida astrologer Jan Walsek, "(wo)men who hang onto superstitions and watch soap operas all day. We are professionals with a body of knowledge that enables us to render advice." (Newsweek 22.1.1990).
Dieses zu lernen, soll dieses Buch eine Hilfe sein.

Dank

Bei der Erstellung des Buchmanuskripts sind wir unseren Mitarbeitern Thomas Bradtke, Anke Frier, Günter Köpp, Uta Lieberum, Christian Schnack, Silke Voß und Yanqing Xia zu Dank verpflichtet. Rainer Dyckerhoff verdanken wir das Programm für die t-Verteilungstafel. Herrn M. Weigert vom Verlag Oldenbourg danken wir für die gute und freundliche Zusammenarbeit in der Erstellung des Buches.
Alle Fehler gehen selbstverständlich zu unseren Lasten.

Inhaltsverzeichnis

Deskriptive Statistik

1 Grundbegriffe der deskriptiven Statistik

1.1 Einführende Begriffe

Gegenstand jeder statistischen Untersuchung ist eine Gesamtheit von statistischen Elementen, die **Grundgesamtheit** oder **Population**. Betrachtet man in einer statistischen Erhebung die ganze Population, dann spricht man von einer Gesamterhebung. Ein Beispiel hierfür sind Volkszählungen, z.B. die für die Bundesrepublik vom Mai 1987. Im Regelfall sind solche Gesamterhebungen aber aus Kostengründen zu aufwendig. Damit ist im allgemeinen der Gegenstand einer statistischen Untersuchung nur eine **Teilerhebung** oder **Stichprobe**. Dabei ist besonders zu beachten, wie die Stichprobe gezogen wird, d.h. wie die zu untersuchenden (auszuwählenden) Elemente der Stichprobe aus der Grundgesamtheit ausgewählt werden. Werden die zu betrachtenden Elemente zufällig bestimmt, so spricht man von einer **Zufallsstichprobe**. Was dabei Zufall ist, ist keineswegs selbstverständlich und wird noch erläutert. Im folgenden werden wir stets von einer Zufallsstichprobe ausgehen, wenn wir von einer Stichprobe sprechen. Ein verwandter, jedoch nicht unbedingt übereinstimmender Begriff ist der der **repräsentativen Stichprobe**. Für sie werden die Elemente so ausgewählt, daß sie im Blick auf die untersuchte Frage die Grundgesamtheit getreu widerspiegeln. Dies Erfordernis ist ebenfalls nicht selbstverständlich, obwohl es für zahllose statistische Untersuchungen als unerläßlich oder wünschenswert vorauszusetzen ist. Man denke nur an politische Meinungsumfragen oder Marketingstudien.
Jede statistische Untersuchung bezieht sich auf **Merkmale**. Die Elemente der Erhebung, die **Urliste** (ob Voll- oder Teilerhebung) sind Träger dieser Merkmale, und es wird untersucht, welche **Merkmalsausprägung** bei jedem einzelnen Element, dem **Merkmalsträger**, vorkommt. Dafür ist es besonders wichtig, die verschiedenen Arten von Merkmalen zu unterscheiden.

1.2 Klassifikation von Merkmalen

Die verschiedenen Arten, Typen, Klassen von Merkmalen werden dadurch gekennzeichnet, wie ihre Ausprägungen dargestellt und geordnet werden können.
Beginnend mit dem *schwächsten* Merkmalstyp sind dies die folgenden:

(A) **Nominale Merkmale**
Ein Merkmal heißt **nominal**, wenn seine Ausprägungen nicht in eine angfolge gebracht werden können, z.B. Vornamen (Hilde, Klaus, Peter, Susanne,...); gesellschaftsrechtliche Bezeichnungen (AG, GmbH, KG); Glaubensbekenntnis (katholisch, evangelisch, jüdisch, muslimisch,...); Parteien (CDU, CSU, FDP, GRÜNE, SPD,...); Krankheiten (TB, Krebs, MS, Kinderlähmung,...).

(B) **Qualitative Merkmale**
Ein Merkmal heißt **qualitativ** oder **ordinal**, wenn es zwar in eineRangfolge gebracht werden kann, aber die *Rangunterschiede* nicht gemessen werden können, d.h. das Merkmal läßt sich *nicht natürlich* einer reellen Zahl zuordnen. Beispiele hierfür sind: Lebensstil (auf großem Fuß, mittel, bescheiden, dürftig), charakterliche Eigenschaften (gutherzig, gleichgültiger Typ,..., Charakterschwein), Examensnoten (1.0,1.3, 1.7, 2.0,...), Konjunkturverlauf (auf, gleich, ab), Geburtstage (15.April, 11.Mai, 12.Mai, 22. September), Popularitätsskalen in Punkten (4 Pluspunkte sind *nicht* doppelt so gut wie 2 Pluspunkte!).

(C) **Quantitative Merkmale**
Ein Merkmal heißt **quantitativ** oder **kardinal**, wenn es sich *natürlich* einer reellen Zahl zuordnen läßt, d.h. die Ausprägung ist eine reelle Zahl.

Das heißt insbesondere auch, daß Rangunterschiede gemessen werden können (z.B.Temperaturen, Flutmarken, Erträge, Zeitpunkte u.ä.). Für die folgenden Überlegungen ist dies die wichtigste Klasse. Sie hat die größte Vielfalt der Analyseverfahren und Zahl der Anwendungen und steht im Mittelpunkt der Betrachtung.
Unter den quantitativen Merkmalen ist insbesondere noch die Unterteilung in stetige und diskrete zu sehen:

(C.1) **Diskrete** und **stetige Merkmale**

Ein Merkmal heißt **diskret**, wenn es höchstens abzählbar viele Ausprägungen hat. Dabei sind zwei Fälle zu unterscheiden:
endlich viele Ausprägungen: z.B. Geschlecht (weiblich, männlich); die Anzahl der täglichen Geburten in einer Klinik; die Tage des Jahres (365 im Regelfall), Möglichkeiten einer Lottozahl (49), Examensnoten (1.0, 1.3, 1.7,..., 4.0, 4.3, 4.7, 5.0); die Verkaufszahlen eines bestimmten

Autohändlers.
Abzählbar unendlich viele Ausprägungen sind mehr theoretisch von Interesse, z.B. die Zahl der Sterne im Universum, Anzahl der Mißerfolge bis zum ersten Erfolg bei einem Glücksspiel.

Ein Merkmal heißt **stetig (kontinuierlich)**, wenn es *nicht diskret* ist, d.h. überabzählbar viele Ausprägungen hat, z.B. die Zeit, Einkommens- und Umsatzzahlen.
Bei der Beobachtung stetiger Merkmale erhält man durch entsprechende Maßeinheiten (für die Zeit etwa: Sekunden oder Minuten, für Längen: Meter oder Zentimeter) diskrete Daten. Wegen ihrer Vielfalt faßt man sie dann aber zu Klassen zusammen.

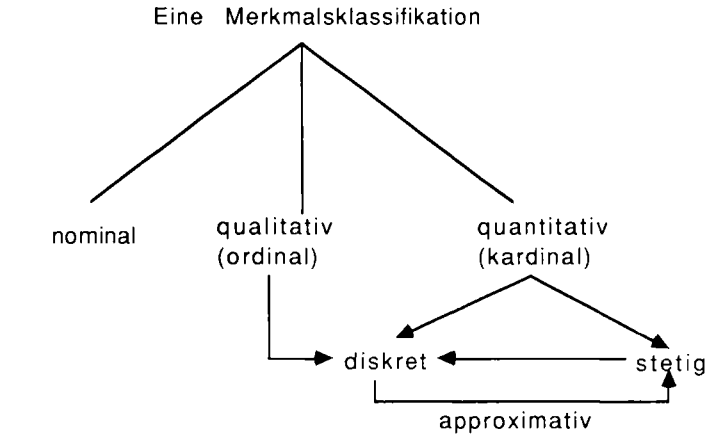
(C.2) **Quantifizierung von qualitativen Merkmalen**
In vielen Situationen werden nominale oder ordinale/qualitative Merkmale **quantifiziert**, z.B. nominalen Skalen werden Zahlen zugeordnet:

nominale Skala:	+++	++	+	0	-	- -	- - -	
quantitative Skala:	3	2	1	0	-1	-2	-3	oder
	100	50	25	0	-25	-50	-100	

Hier ist besondere Vorsicht in den Schlußfolgerungen angebracht.

(C.3) **Diskretisierung von stetigen Merkmalen**
In vielen Situationen werden stetige Merkmale **diskretisiert:**
Stetige Skalen werden in Intervalle eingeteilt, z.B. das Messen von Firmengrößen nach Umsatzhöhe (unter 1 Mio DM, 1 Mio bis unter 2 Mio DM, 2 Mio DM und darüber; monatliche Arbeitnehmereinkommen: unter 1000 DM, von 1000 DM bis 2000 DM, über 2000 DM). In diesem Fall spricht man von **klassierten** oder **gruppierten** Daten.
Umgekehrt werden diskrete Merkmale, die mit *sehr* vielen Ausprägungen auftreten, wie stetige behandelt. Sie werden als **quasistetige** oder **approximativ stetige** Merkmale bezeichnet (z.B. Verkaufszahlen pro Tag von LP's in einem Schallplattengeschäft).

Diagramm zur Merkmalsklassifikation



Legende: ⟶ mögliche Übergänge

1.3 Aufgaben

Aufgabe 1.3.1
Zur Verbesserung der Personalplanung wird ein Mitarbeiter beauftragt, Daten über alle Beschäftigten zusammenzutragen, die sich unter anderem auf das Alter, das Geschlecht, die Stellung im Unternehmen, die Dauer der Unternehmenszugehörigkeit und das Gehalt beziehen sollen.

a) Was ist die statistische Einheit der Untersuchung?
b) Was ist die statistische Gesamtheit (Masse, Population)?
c) Welcher Art sind die oben erwähnten Merkmale?
(nominal, qualitativ, quantitativ; diskret, stetig)
d) Welches sind mögliche Ausprägungen dieser Merkmale?

Aufgabe 1.3.2
Um die Auswirkungen der kommenden Tarifabschlüsse auf die eigenen Lohn- und Gehaltszahlungen abschätzen zu können, führt die Firma Nagel, Holz & Co. bei 100 ihrer 500 Beschäftigten eine Erhebung durch, bei der Alter, Tarifklasse, außertarifliche Zahlungen und Geschlecht festgestellt werden.

a) Geben Sie die Grundgesamtheit der Erhebung an.
b) Welcher Art sind die angeführten Merkmale?
c) Geben Sie mögliche Merkmalsausprägungen dieser Merkmale an.

2 Häufigkeitsverteilungen

2.1 Diskrete Merkmale

Nach Durchführung einer statistischen Erhebung - und dies wird im folgenden fast stets als erfolgt angesehen - ist das Datenmaterial (die Urliste) so aufzubereiten, daß man die Fülle der Beobachtungen intellektuell aufnehmen und verarbeiten kann. Als erstes werden einige einfache graphische Darstellungen vorgeführt, die sich in der Praxis bewährt haben. Sie finden sich z.B. in Zeitungen, Firmenberichten, Fernsehnachrichten. Dazu wird der Begriff der Häufigkeitsverteilung eingeführt:

Seien $\{b_1,b_2,\ldots,b_n\}$ die beobachteten Ausprägungen eines *beliebigen* Merkmals X aus einer Erhebung der Länge n. Nicht alle b_i ($i\in\{1,2,\ldots,n\}$) sind notwendigerweise verschieden, so daß die verschiedenen Ausprägungen $\{x_1,x_2,\ldots,x_K\}$, ($K\leq n$) mit den **absoluten Häufigkeiten** n_k ($k=1,2,\ldots,K$) mehrfach beobachtet wurden. Offensichtlich gilt:

$$\sum_{k=1}^{K} n_k = n\ ;$$

K=n trifft nur dann zu, falls alle Beobachtungen verschieden sind.

2.1.1 Beispiel

Die Menge der n Beobachtungen $\{b_1,b_2,\ldots,b_n\}$ seien die Ergebnisse von n=100 Würfen eines Würfels. Als Merkmalsausprägungen gibt es die sechs verschiedenen Augenzahlen {1,2,3,4,5,6}. Für die der Größe nach sortierten Merkmalsausprägungen seien die folgenden Häufigkeiten n_k (k=1,...,6) beobachtet worden:

$$\{\ x_1, x_2, x_3, x_4, x_5, x_6\ \}$$
$$\{\ 15, 20, 20, 10, 15, 20\ \},\ n=100,\ K=6$$

bzw. als Tabelle:

x_k	1	2	3	4	5	6
n_k	15	20	20	10	15	20

2.1.2 Definition

Sei $M = \{x_1, x_2, \ldots, x_k\}$ die Menge von beobachteten verschiedenen Merkmalsausprägungen; sei weiter

$$g: \begin{cases} M \to \mathbf{R}_+ \\ x \to g(x) := \begin{cases} n_k, \text{ falls } x = x_k \ (k=1,\ldots,K) \\ 0, \text{ sonst} \end{cases} \end{cases}$$

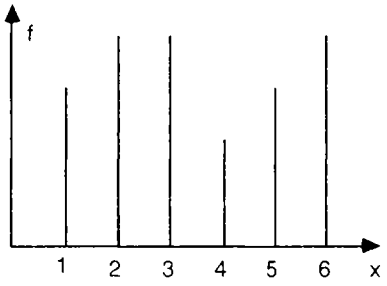
und

$$f: \begin{cases} M \to [0,1] \\ x \to f(x) := \frac{1}{n} g(x) \end{cases}$$

Die Funktion f heißt **empirische Häufigkeitsfunktion**.
Der Graph dieser Abbildung heißt **Stabdiagramm.**

Offensichtlich lassen sich solche Stabdiagramme für *alle* Merkmale erstellen. Für das Beispiel 2.1.1 ergibt sich:

Das Schema eines Stabdiagramms



2.1.3 Bemerkung

Es ist

$$f: M \to [0,1] \text{ mit } f(x) = \begin{cases} 0.15, & x=1 \\ 0.20, & x=2 \\ 0.20, & x=3 \\ 0.10, & x=4 \\ 0.15, & x=5 \\ 0.20, & x=6 \\ 0.00, & \text{sonst} \end{cases}$$

sowie

$$g: M \to \mathbf{R}_+ \text{ mit } g(x) = \begin{cases} 15, & x=1 \\ 20, & x=2 \\ 20, & x=3 \\ 10, & x=4 \\ 15, & x=5 \\ 20, & x=6 \\ 0, & \text{sonst} \end{cases}$$

Offensichtlich unterscheidet sich der Graph von g vom Graphen von f nur durch die Skalierung.

Da es sich um *Empirie* (=Beobachtungen aus der Praxis) handelt, spricht man auch von empirischen Häufigkeiten (d.h. empirische absolute Häufigkeiten bzw. empirische relative Häufigkeiten).

Sei X ein quantifiziertes ordinales oder ein kardinales Merkmal mit K verschiedenen Merkmalsausprägungen zu jeweils n_k (k=1,2,...,K) und insgesamt n Beobachtungen. Mit der Definition der **relativen Häufigkeiten**

$$f_k := f(x_k) = \frac{n_k}{n}$$

läßt sich aus den $f_1, f_2, \ldots, f_K$ der Begriff der **kumulierten** (empirischen) **Häufigkeitsverteilung** konstruieren:

$$F_1 := f$$
$$F_2 := f_1 + f_2 \qquad = F_1 + f_2$$
$$F_3 := f_1 + f_2 + f_3 \qquad = F_2 + f_3$$
$$\vdots$$
$$F_K := f_1 + f_2 + \ldots + f_K = F_{K-1} + f_K$$

Umgekehrt lassen sich die relativen Häufigkeiten aus der kumulierten Häufigkeitsverteilung bestimmen bzw. zurückgewinnen:

$$f_1 = F_1$$
$$f_i = F_i - F_{i-1} \qquad (i=2,3,\ldots,K)$$

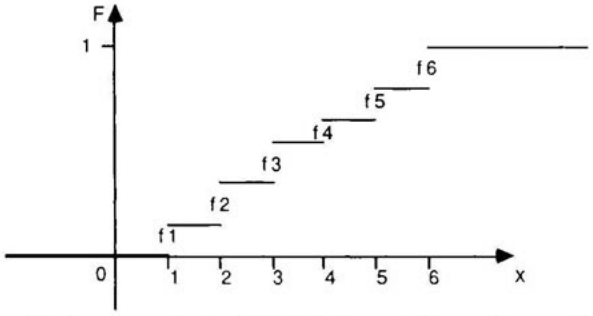
2.1.4 Definition

Seien die Ausprägungen $\{x_1, x_2, \ldots, x_k\}$ der Größe nach geordnet, d.h. $x_1 < x_2 < \ldots < x_k$; dann heißt die Funktion

$$F: \begin{cases} \mathbf{M} \to [0,1] \\ x \to F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ \sum\limits_{i:x_i \le x} f_i = F_i & \text{für } x \ge x_1 \end{cases} \end{cases}$$

empirische Verteilungsfunktion.

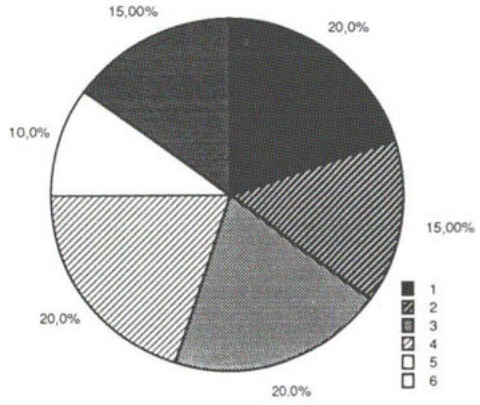
Der Graph von F sieht typischerweise folgendermaßen aus (Werte aus Beispiel 2.1.1):



Für die empirische kumulierte Häufigkeitsverteilung (kurz: Verteilung) gilt: Die Verteilung ist eine monoton *nichtfallende Treppenfunktion.*

Ein weiteres, sehr beliebtes Darstellungsverfahren ist das **Kuchendiagramm** (*pie-chart,* das auf die berühmte englische Wohltäterin Florence Nightingale zurückgeht, die es als erste benutzte, um dem britischen Parlament die Häufigkeiten verschiedener Kriegsleiden vor Augen zu führen).
Die relativen Häufigkeiten n_k werden dazu benutzt, eine Kreisscheibe in proportionale Kreissegmente zu zerlegen, d.h. man teilt die 360° des Kreises in Winkel α_k entspreched $f_k = n_k/n = \alpha_k/2\pi$.

Zu Beispiel 2.1.1 ergibt sich:



Solche Kuchendiagramme eignen sich insbesondere auch zur Aufbereitung und Darstellung von nominalen Merkmalen.

2.2 Stetige Merkmale

Bei stetigen Merkmalen faßt man die Beobachtungen zu schon vor Beginn der Erhebung festzulegenden Klassen K_i zusammen. Wenn möglich sollten diese Klassen alle die gleiche Breite haben. Oftmals empfiehlt es sich dennoch, dort, wo die meisten Beobachtungen zu erwarten sind, diese Klassenbreiten kleiner zu halten als in den Bereichen, in denen relativ wenige Beobachtungen zu erwarten sind.

2.2.1 Beispiel

Zur Schätzung zukünftiger Kosten führt eine Krankenversicherung bei ihren Mitgliedern eine Umfrage durch, bei der auch das Merkmal *Körpergewicht* (in Kilogramm) betrachtet wird. Unter den männlichen Befragten gab es dazu folgendes Ergebnis:

K_i:	[50,60)	[60,70)	[70,75)	[75,80)	[80,90)	[90,110)
n_i:	5	25	30	25	10	5

Dabei bedeutet n_i die Häufigkeit von Beobachtungen in der Klasse K_i.

2.2.2 Definition

Die Funktion

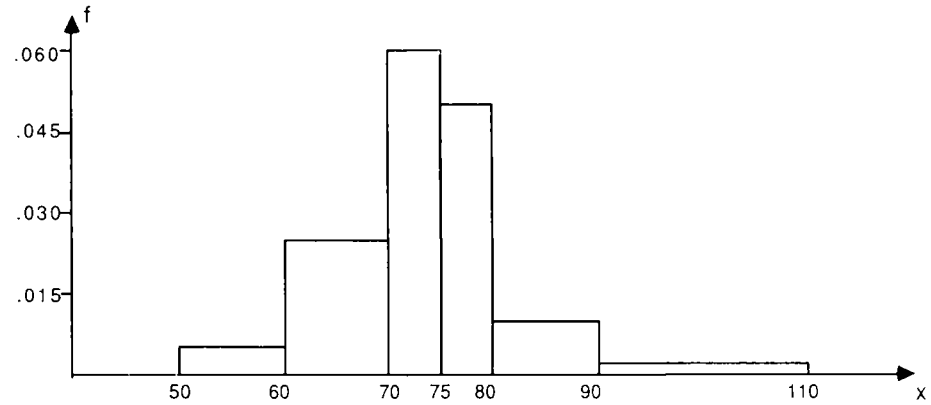
$$f: \begin{cases} R \to R_+ \\ x \to f(x) = \begin{cases} \dfrac{\frac{n_i}{n}}{\Delta K_i}, \text{falls } x \in K_i \\ 0 \quad , \text{sonst} \end{cases} \end{cases}$$

heißt **empirische Dichtefunktion**. Dabei bedeutet ΔK_i die Breite der Klassen K_i. Der Graph der Dichtefunktion heißt Histogramm.

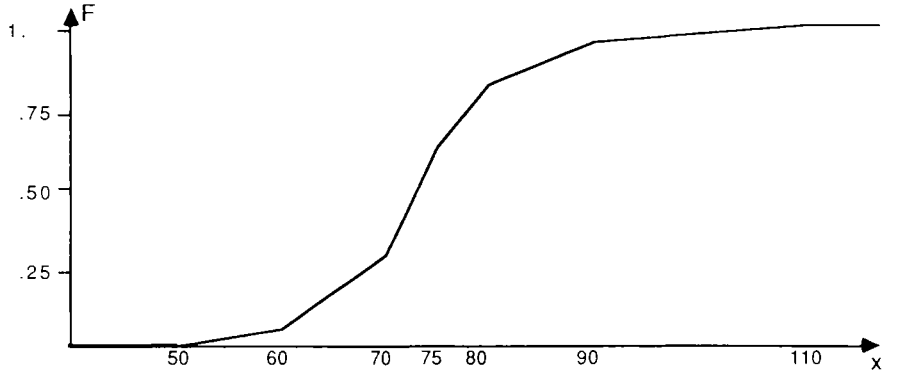
Für das Beispiel 2.2.1. erhält man für f folgende Funktionswerte:

$$f(x) = \begin{cases} 0.0050\,, & \text{für } x \in [50,60) \\ 0.0250\,, & \text{für } x \in [60,70) \\ 0.0600\,, & \text{für } x \in [70,75) \\ 0.0500\,, & \text{für } x \in [75,80) \\ 0.0100\,, & \text{für } x \in [80,90) \\ 0.0025\,, & \text{für } x \in [90,110) \end{cases}$$

Das zugehörige Histogramm hat dann das folgende Aussehen:



Da über die Verteilung der Beobachtungen innerhalb der Klassen keine Information vorliegt, trifft man die Annahme, daß diese Beobachtungen sich innerhalb der Klassen gleichmäßig verteilen. Das bedeutet, daß man für die kumulierten Häufigkeiten die relativen Häufigkeiten für die Klassenobergrenzen aufaddiert und die so erhaltenen Punkte durch Geradenstücke verbindet. Die Steigung eines solchen Geradenstücks ist dann genau der Wert der Dichtefunktion in dieser entsprechenden Klasse.
Die zu Beispiel 2.2.1 gehörende Verteilungsfunktion hat dann folgendes Aussehen:



Mit dieser Funktion lassen sich die Anteilswerte unmittelbar angeben:
Es bezeichnet nämlich F(x) den Anteil derjenigen Männer, deren Körpergewicht höchstens x kg beträgt.
Beispielsweise ist der Anteil der Männer, deren Gewicht höchstens 70 kg beträgt 30 %, der Anteil, deren Gewicht höchstens 80 kg beträgt 85 %, usw.

Fällt der Wert x nicht auf eine Klassengrenze, sondern ins Innere einer Klasse, so läßt sich der entsprechende Anteil aufgrund der Annahme, daß sich alle Beobachtungen innerhalb einer Klasse gleichmäßig verteilen, durch eine einfache Dreisatzrechnung bestimmen:
Allgemein gilt:

$$(1) \qquad F(x) = F(x_i^u) + \frac{x-x_i^u}{\Delta K_i}\frac{n_i}{n} = F(x_i^u) + (x-x_i^u)f(x) \quad \text{für } x \in K_i .$$

Dabei bedeutet x_i^u die Untergrenze der Klasse K_i.

2.3 Quantile

2.3.1 Definition

Unter einem **α-Quantil (oder α-Fraktil)** versteht man denjenigen Wert x, für den gilt: $F(x) = \alpha$ $(0 \le \alpha \le 1)$.

Bei stetigen Merkmalen kann dieses x eindeutig bestimmt werden, bei diskreten Merkmalen wird man im allgemeinen kein solches x finden. Man bezeichnet dann als α-Quantil diejenige beobachtete Ausprägung x_i, für die gilt:

$$F(x) \ge \alpha \text{ und } 1-F(x)+f(x) \ge \alpha.$$

In Worten bedeutet dies, daß als α-Quantil diejenige Beobachtung genommen wird, bei der die kumulierten relativen Häufigkeiten erstmalig den Wert α übersteigen.
Für spezielle Quantile gibt es entsprechende Bezeichnungen, etwa:
Das **1. Quartil**: Es ist der Bereich von Beobachtungen zwischen der ersten Beobachtung und der kleinsten Beobachtung $x_{1/4}$, für die $F(x_{1/4}) \ge 0.25$.

Das **2. Quartil** ist der Bereich von Beobachtungen zwischen der Beobachtung $x_{1/4}$ und der kleinsten Beobachtung $x_{2/4}$, für die $F(x_{2/4}) \ge 0.50$.

Das **3. Quartil** ist der Bereich von Beobachtungen zwischen der Beobachtung $x_{2/4}$ und der kleinsten Beobachtung $x_{3/4}$, für die $F(x_{3/4}) \ge 0.75$

Das **4. Quartil** ist der Bereich von Beobachtungen zwischen der Beobachtung $x_{3/4}$ und der kleinsten Beobachtung $x_{4/4}$, für die $F(x_{4/4}) = 1.00$ (d.h. der letzten Beobachtung).

Für Beispiel 2.2.1 erhält man etwa als 0.30-Quantil den Wert x = 70.
Das erste Quartil, also der Punkt x mit $F(x) = 0.25$ bestimmt sich aus Gleichung (1) durch Umformung nach x:

$$x = x_i^u + \frac{F(x)-F(x_i^u)}{\frac{n_i}{n}}\Delta K_i .$$

In Beispiel 2.2.1 ist offensichtlich $x = 60 + \frac{0.25\text{-}0.05}{\frac{25}{100}} 10 = 68$.

Bei einer Unterteilung in 10% Schritte (statt wie bei Quantilen in 25% Schritte) heißen die Quantile **Dezile**.

2.4 Aufgaben

Aufgabe 2.4.1
Bei einer Erhebung zu dem Merkmal X: *Eintrittspreise der Fahrgeschäfte auf dem Hamburger Dom* (das ist der Hamburger Jahrmarkt) erhielt man folgende Beobachtungen:

x_k (in DM)	2.00	2.50	3.00	3.50	4.00	5.00	6.00
n_k	3	7	5	5	15	10	5

a) Bestimmen Sie die empirische Häufigkeitsfunktion und die empirische Verteilungsfunktion des Merkmals X.
b) Zeichnen Sie die Graphen der empirischen Häufigkeitsfunktion und der empirischen Verteilungsfunktion des Merkmals X.
c) Welcher Anteil der Geschäfte verlangt einen Eintrittspreis von
- höchstens DM 4.00
- mehr als DM 3.50
- mindestens DM 3.50
- mindestens DM 2.80 und höchstens DM 4.90?

Drücken Sie diese Anteile mit Hilfe der Verteilungsfunktion und der Häufigkeitsfunktion aus.

Aufgabe 2.4.2
Bei einer Erhebung zu dem Merkmal X: *Fahrtdauer einer Fahrt mit den jeweiligen Fahrgeschäften auf dem Hamburger Dom* (in Sekunden) erhielt man folgende Beobachtungen:

	120 bis 150	5
über	150 bis 180	8
über	180 bis 200	8
über	200 bis 220	10
über	220 bis 250	15
über	250 bis 300	4

a) Bestimmen Sie die empirische Dichtefunktion und die empirische Verteilungsfunktion des Merkmals X.
b) Zeichnen Sie die Graphen der empirischen Dichtefunktion und der empirischen Verteilungsfunktion des Merkmals X.
c) Welcher Anteil der Geschäfte bietet eine Fahrtzeit von
- höchstens 220 sec
- mehr als 220 sec
- mindestens 220 sec
- höchstens 185 sec
- mindestens 175 sec und höchstens 240 sec ?

Drücken Sie diese Anteile mit Hilfe der Verteilungsfunktion aus.
d) Wie lange muß eine Fahrt mindestens dauern, damit sie zu den 30 % der am längsten dauernden Fahrten gehört?
e) Bestimmen Sie das 1. bis 4. Quartil und interpretieren Sie die Werte.

Aufgabe 2.4.3
In einer Klinik wird an 50 Tagen die Anzahl der Geburten pro Tag ermittelt:

Anzahl x der täglichen Geburten	Anzahl der Tage mit x Geburten
0	6
1	10
2	29
3	3
4	2

a) Bestimmen Sie die relativen Häufigkeiten und den Anteil der Tage mithöchstens x Geburten, und geben Sie die empirische Häufigkeitsfunktionf und die empirische Verteilungsfunktion F explizit an.
b) Stellen Sie die Funktionen f und F graphisch dar.
c) Wie hoch ist der Anteil der Tage mit 3 oder mehr Geburten täglich ?
d) Wie hoch ist der Anteil der Tage, an denen kein Kind geboren wird ?
e) Berechnen Sie den Wert der empirischen Verteilungsfunktion an der Stelle x = 3.6 und interpretieren Sie das Ergebnis.

3 Verdichtung der Daten

3.1 Lageparameter

Um die Charakteristiken von vorliegendem Datenmaterial überblicken zu können, verdichtet man die Daten mit Hilfe von Maßzahlen.

3.1.1 Definition
Die Ausprägung, die am häufigsten beobachtet wurde, für die also gilt

$$x_{mod} := x_i \text{ mit } f(x_i) = \max \{f(x_k) \mid k=1,\ldots,K\}$$

heißt **Modus** oder **Modalwert**.

Offensichtlich findet der Modalwert innerhalb der deskriptiven Statistik sinnvolle Anwendung nur bei diskreten Merkmalen. Bei theoretischen Verteilungen (s. u. Kap. 8) kann der Modus entsprechend auch auf stetige Verteilungen angewandt werden.

3.1.2 Definition
Der Wert x_{med}, für den gilt

$$F(x_{med}) \geq 0.5 \text{ und } 1-F(x_{med})+f(x_{med}) \geq 0.5,$$

also die Obergrenze des 2. Quartils, heißt **Median**.

3.1.3 Definition
Ist X ein diskretes Merkmal, dann ist

$$(1) \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} b_i = \sum_{k=1}^{K} x_k \, f(x_k)$$

das **arithmetische Mittel** oder der **Mittelwert** von X.

Ist X ein stetiges Merkmal und bezeichnet x_k^* die Klassenmitte der k-ten Klasse, dann ist

$$(1)' \qquad \bar{x} = \sum_{k=1}^{K} x_k^* \frac{n_k}{n}$$

das **arithmetische Mittel** oder der **Mittelwert** von X.

Zu Beispiel 2.1.1 ist x_{mod} nicht eindeutig zu bestimmen, denn die Ausprägungen 2, 3 und 6 wurden jeweils 20 mal beobachtet. Dagegen ergibt sich für den Median und das arithmetische Mittel $x_{med} = 3$ und $\bar{x} = 3.5$.
Zu Beispiel 2.2.1 kann der Modalwert nicht bestimmt werden, da das Merkmal stetig ist; für den Median gilt $x_{med} = 73.333$ und es ist $\bar{x} = 73.625$.

3.2 Streuungsparameter

3.2.1 Definition

Die Differenz zwischen der größten und der kleinsten beobachteten Ausprägung wird als **Spannweite** bezeichnet.

3.2.2. Definition

Ist X ein diskretes Merkmal, dann ist

$$(2) \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - \bar{x})^2 = \sum_{k=1}^{K}(x_k - \bar{x})^2 f(x_k)$$

die **mittlere quadratische Abweichung** oder die **empirische Varianz** des Merkmals X.

Ist X ein stetiges Merkmal und bezeichnet x_k^* die Klassenmitte der k-ten Klasse, dann ist

$$(2)' \qquad s^2 = \sum_{k=1}^{K}(x_k^* - \bar{x})^2 \frac{n_k}{n}$$

die **mittlere quadratische Abweichung** oder die **empirische Varianz** des Merkmals X.

Die positive Wurzel von s^2 wird als (empirische) **Standardabweichung** bezeichnet.

Aus später deutlich werdendem Grund wird auch die folgende *Modifikation* von empirischer Varianz und Standardabweichung benutzt:

$$s^{*2} = s^2 \frac{n}{n-1} \qquad \text{und} \qquad s^* = \sqrt{s^2 \frac{n}{n-1}}$$

d.h. anstatt durch n wird durch (n-1) in (2) bzw. (2)' geteilt.

Zu Beispiel 2.1.1 erhält man $s^2 = 3.05$ und $s = 1.7464$ und zu Beispiel 2.2.1 erhält man $s^2 = 87.7969$ und $s = 9.37$.

3.3 Die Streuungszerlegung

Häufig setzen sich Stichproben aus gesondert erhobenen Teilstichproben zusammen, beispielsweise für die gesamte BRD aus den einzelnen Bundesländern. In solchen Fällen ist der folgende Satz von Bedeutung.

3.3.1 Satz

Sei eine Stichprobe von n Beobachtungen in I Teilstichproben jeweils vom Umfang n_i (i=1,2,...,I) gegeben, d.h.

$$\{x_1,...,x_n\} = \{x_{11},x_{12},...,x_{1n_1};x_{21},x_{22},...,x_{2n_2};...;x_{I1},x_{I2},...,x_{In_I}\};\ n = \sum_{i=1}^{I} n_i ,$$

dann gilt für die Mittelwerte

$$\bar{x} = \sum_{i=1}^{I} \frac{n_i}{n} \bar{x}_i \tag{3}$$

und für die empirischen Varianzen

$$s^2 = \sum_{i=1}^{I} \frac{n_i}{n} s_i^2 + \sum_{i=1}^{I} \frac{n_i}{n} (\bar{x} - \bar{x}_i)^2 \tag{4}$$

Beweis:

Für (3) ergibt sich:

$$\bar{x} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \frac{x_{ij}}{n} = \sum_{i=1}^{I} \frac{n_i}{n} \sum_{j=1}^{n_i} \frac{x_{ij}}{n_i} = \sum_{i=1}^{I} \frac{n_i}{n} \bar{x}_i$$

Für (4) ergibt sich für den Fall I = 2 (Die Verallgemeinerung auf I ≥ 2 ist offensichtlich.):

$$s^2 = \frac{1}{n} \left(\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2 + \left(\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2 \right) \right) =$$

$$= \frac{1}{n} \left(\sum_{j=1}^{n_1} ([x_{1j} - \bar{x}] + [\bar{x} - \bar{x}_1])^2 + \left(\sum_{j=1}^{n_2} ([x_{2j} - \bar{x}] + [\bar{x} - \bar{x}_2])^2 \right) \right) =$$

$$= \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} ([x_{ij} - \bar{x}_i]^2 + [\bar{x} - \bar{x}_i]^2 - 2[x_{ij} - \bar{x}_i][\bar{x} - \bar{x}_i]) =$$

$$= \sum_{i=1}^{2} \frac{n_i}{n} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \frac{1}{n_i} \sum_{j=1}^{n_i} (\bar{x} - \bar{x}_i)^2 \right) -$$

$$- \frac{2}{n} \sum_{i=1}^{2} (\bar{x} - \bar{x}_i) \left(\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) \right).$$

Wegen $\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = 0$ ergibt sich daraus

$$s^2 = \frac{n_1}{n} s_1^2 + \frac{n_2}{n} s_2^2 + \frac{n_1}{n} (\bar{x} - \bar{x}_1)^2 + \frac{n_2}{n} (\bar{x} - \bar{x}_2)^2 .$$

■

3.4 Konzentrationsmaße

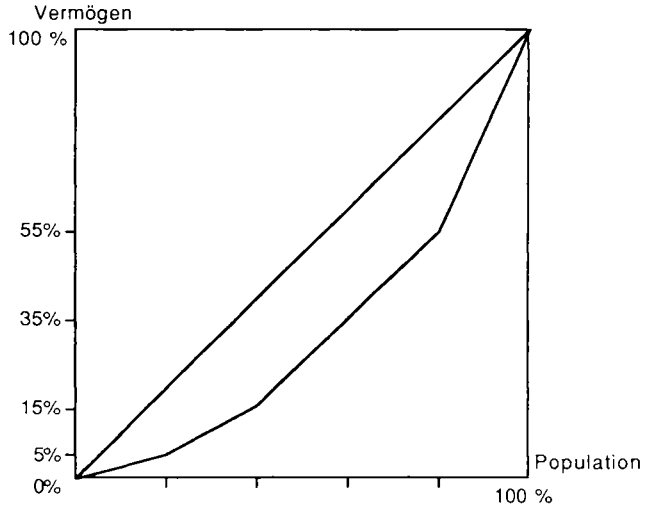
Konzentrationsmaße werden benutzt, um Vermögens-, Einkommens- oder ähnliche Verteilungen auf einer Population darzustellen. Dabei ordnet man vorgegebenen Anteilen der Population die entsprechenden Anteile des betrachteten Merkmals folgendermaßen zu:
Man ordnet die Anteile der Population in der Weise, daß der Anteil des betrachteten Merkmals wächst, wie im folgenden Beispiel dargestellt:

3.4.1 Beispiel

Seien 5 Eigentümer von Vermögen verglichen, deren Vermögen

$\{z_1\ z_2\ z_3\ z_4\ z_5\} = \{20\ 10\ 40\ 90\ 40\}$ bzw. sortiert
$\{y_1\ y_2\ y_3\ y_4\ y_5\} = \{10\ 20\ 40\ 40\ 90\}$, bzw. skaliert
$\{w_1\ w_2\ w_3\ w_4\ w_5\} = \{5\%\ 10\%\ 20\%\ 20\%\ 45\%\}$ bzw. kumuliert
$\{F_1\ F_2\ F_3\ F_4\ F_5\} = \{5\%\ 15\%\ 35\%\ 55\%\ 100\%\}$ betrage.

Das bedeutet, daß die *ärmste* der fünf betrachteten Personen ein Vermögen von 10 Einheiten, die nächste ein Vermögen von 20 Einheiten besitzt usw. bis zur *reichsten*, die ein Vermögen von 90 Einheiten besitzt.
Graphisch läßt sich dieser Sachverhalt folgendermaßen darstellen:



Bei Gleichheit der Verteilung stimmt die untere Kurve mit der Diagonalen überein.