Statistik

für Sozial- und Wirtschaftswissenschaften

Lehrbuch mit Übungsaufgaben

Von
Universitätsprofessor
Dr. Peter Hackl
und
Universitätsdozent
Dr. Walter Katzenbeisser

11., durchgesehene Auflage

R. Oldenbourg Verlag München Wien

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Hackl, Peter:
Statistik für Sozial- und Wirtschaftswissenschaften : Lehrbuch mit Übungsaufgaben / von Peter Hackl ; Walter Katzenbeisser. - 11., durchges. Aufl. – München ; Wien : Oldenbourg, 2000
ISBN 3-486-25468-5
NE: Katzenbeisser, Walter:

Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
www.oldenbourg-verlag.de

Gedruckt auf säure- und chlorfreiem Papier
Druck: R. Oldenbourg Graphische Betriebe Druckerei GmbH

ISBN 3-486-25468-5

Inhaltsverzeichnis

Vorwort zur neunten Auflage

Diese neunte Auflage ist eine wesentliche Überarbeitung der früheren Auflagen unseres Buches. Zielsetzung des Buches, das vielfach als Textbuch zu Vorlesungen und Übungen der statistischen Grundausbildung eingesetzt wird, ist es, das für die Sozial- und Wirtschaftswissenschaften notwendige, statistische Instrumentarium zu vermitteln. Wir haben uns bemüht, eine anwendungsbezogene Darstellung der statistischen Verfahren sowie deren Grundlagen zu geben. Wir verzichten weitgehend auf mathematische Beweise der behandelten Verfahren; für den interessierten Leser geben wir Hinweise auf die weiterführende Literatur. Entsprechend unseren Erfahrungen, die wir in wiederholt abgehaltenen Lehrveranstaltungen gewonnen haben, scheinen uns Erläuterungen an Hand illustrierender Beispiele für den angesprochenen Leserkreis dem Verständnis förderlicher und motivierender zu sein.

Bei der Überarbeitung haben wir vor allem darauf Wert gelegt, die Lesbarkeit des Buches zu verbessern. Im Sinn der immer besseren Verfügbarkeit von Statistik-Software haben wir an vielen Stellen des Buches das Statistik-Programmpaket MINITAB angesprochen: Wir haben uns bemüht, an allen in Frage kommenden Stellen MINITAB-Prozeduren zu besprechen und zur Lösung von Beispielen einzusetzen; die Übersicht im Anhang soll das aktive Anwenden von MINITAB zum Lösen von Aufgaben erleichtern.

Jedes Kapitel des Buches gliedert sich in drei Teile: Im ersten Teil werden die jeweils relevanten Definitionen, Sätze und Techniken eingeführt und an Hand geeigneter Beispiele erläutert. Im zweiten Teil werden typische Anwendungen der besprochenen Methoden an Hand gelöster und kommentierter Beispiele in exemplarischer Weise dargestellt. Am Ende der Kapitel sind eine Reihe von Übungsaufgaben und deren Lösungen angegeben. Die durch die Vorgabe der Lösungen unterstützte Bearbeitung der Übungsaufgaben soll dem Studenten helfen, sein Verständnis zu vertiefen und Übung bei der selbständigen Handhabung der statistischen Methoden zu erlangen.

Seit Herbst 1989 leitet unser Mitautor der bisherigen Auflagen, Univ.Prof. Dr. Wolfgang Panny, das Extraordinariat "Angewandte Informatik" an der

hiesigen Wirtschaftsuniversität. Als Folge der damit verbundenen Veränderung seines Wirkungsschwerpunktes steht er als Mitautor dieser und zukünftiger Auflagen nicht mehr zur Verfügung. Wir sind ihm für seinen langjährigen Beitrag zum Entstehen dieses Buches und zu seiner laufenden Verbesserung verbunden.

Zu Dank verpflichtet sind wir Herrn Stefan Katzenbeisser für das Anfertigen von Graphiken und Frau Doris Müller für die Arbeit, das Manuskript in eine mit TEX verarbeitbare Form zu bringen. Herrn Dipl.Vw. Martin Weigert vom Oldenbourg-Verlag danken wir für die nun schon jahrelange, angenehme Zusammenarbeit und für seine Geduld.

Wir hoffen, daß die Neuauflage bei Kollegen und Studenten wohlwollen aufgenommen wird.

Peter Hackl
Walter Katzenbeisser

Vorwort zur elften Auflage

Die Vorauflage war derart rasch vergriffen, daß wir uns auf eine Durchsicht des Textes beschränken konnten.
Wir hoffen, daß auch die 11. Auflage bei Kollegen und Studenten freundlich aufgenommen wird.

Peter Hackl
Walter Katzenbeisser

Verzeichnis der Tabellen

Kapitel 1

Statistik: Begriff und Probleme

1.1 Was ist Statistik?

Für den Begriff Statistik gibt es eine große Zahl von verschiedenen Definitionen. Die wohl treffendste Charakterisierung ist nach unserer Meinung, wenn Statistik als **die Methoden des Lernens aus der Empirie** bezeichnet wird. Aus der Erfahrung zu lernen bedeutet, daß wir Gesetzmäßigkeiten in den um uns ablaufenden Prozessen erkennen. Natürlich bedarf nicht jeder Lernprozeß ausgeklügelter Methoden. Das Kleinkind, das sich – wie jeder von uns – die Finger an einem Bügeleisen verbrennt, lernt die Lektion gründlich und ohne komplizierte Methodik. Der Techniker, der die Bedingungen kennen möchte, unter denen sein chemischer Reaktor einen maximalen Output liefert, ist gut beraten, beim Identifizieren der Kontrollvariablen und beim Festlegen der anzuwendenden Prozeßbedingungen statistische Methoden zu benützen; die Zahl der notwendigen Experimente und damit die Kosten sind umso geringer, je mehr Informationsgehalt die in diesen Experimenten gewonnenen Daten haben. Die Ökonom verwendet ökonometrische Methoden, wenn er Abhängigkeiten und Wirkungsweisen in seinem ökonomischen System studieren oder ein Modell zum Zweck der Prognose erstellen möchte. Ein Medikament wird von der Zulassungsbehörde nicht zum Gebrauch freigegeben, wenn nicht durch nach strengen, statistischen Grundsätzen geplante und ausgewertete Studien die Wirksamkeit des Medikaments und das Fehlen von Nebenwirkungen nachgewiesen worden ist.

Die Statistik ist eine eigenständige wissenschaftliche Disziplin, deren Wurzeln in eine Reihe von anderen Gebieten reichen. So wird man von einem professionellen Statistiker erwarten, daß er eine solide Ausbildung in Mathematik hat, und daß er mit Daten umgehen kann und es insbesondere versteht,

Computer zur Analyse und graphischen Aufbereitung der Daten einzusetzen. Darüberhinaus muß er als Anwender auch ein Grundwissen und Verständnis für die Sachprobleme haben, auf die er statistische Methoden anwenden möchte.

Umgekehrt ist jeder Umgang mit Daten eine dem Wesen nach statistische Tätigkeit, sodaß – insbesondere im Zeitalter der Alltäglichkeit der elektronischen Datenverarbeitung – für fast jeden Beruf und jede Tätigkeit ein statistisches Grundwissen von Vorteil ist. Insbesondere sollte der Konsument von statistischen Ergebnissen in der Lage sein, eine kritische Distanz zu diesen Ergebnissen einzunehmen.

Die meisten wirtschaftlichen Berufe sind in diesem Sinn besonders herausgefordert. Die Ergebnisse betrieblichen und ökonomischen Handelns werden in besonderem Maße mittels elektronischer Datenverarbeitung aufbereitet und zahlenmäßig dargestellt. Kennzahlen des betrieblichen Geschehens, Optimierung der betrieblichen Prozesse, Qualitätsstatistik, Marktforschung, betriebsökonometrische Modellierung und Prognosen über die betriebliche oder ökonomische Entwicklung sind Beispiele für mehr oder weniger komplizierte statistische Aufgabenstellungen für den Ökonomen. Bei methodisch schwierigen Problemen wird wohl ein Wirtschaftsstatistiker beigezogen oder mit der Aufgabe befaßt werden. In jedem Fall muß der Betriebswirt oder Ökonom in der Lage sein, den statistischen Sachverhalt gemeinsam mit dem Statistiker herauszuarbeiten, die Methodenauswahl zu verstehen und an der Interpretation der Ergebnisse mitzuwirken.

Die Anwendung von statistischen Methoden besteht entsprechend dem Gesagten aus folgenden Tätigkeiten:

- dem Erarbeiten des statistischen Sachverhaltes der Problemstellung
- der Auswahl der geeigneten, statistischen Methode
- der Erhebung der notwendigen Daten
- der Ausführung der entsprechenden Datenanalyse
- der Interpretation der Ergebnisse

Der Gegenstand dieses Buches sind die statistischen Methoden. Die Komplikation der Anwendung dieser Methoden kommt daher, daß sie Verständnis des sachlichen Gehalts der Problemstellung und Erfahrung in der Anwendung statistischer Methoden erfordert. Dieses Wissen kann natürlich in einer Einführung in die Statistik kaum vermittelt werden.

1.2 Problemstellungen der Statistik

Statistik beschäftigt sich damit,

- Untersuchungen zu planen,
- die erhobenen Daten zu beschreiben (**beschreibende Statistik**) und
- Entscheidungen über interessierende Phänomene, die durch die Daten beschreiben werden, zu treffen (**schließende Statistik**).

Im Rahmen dieses Buches werden wir uns mit den beiden letzteren Fragestellungen befassen. Dies primär deswegen, da für geplante Experimente in der Ökonomie – im Gegensatz zu den Naturwissenschaften – nur wenig Spielraum besteht. Dies heißt aber nicht, daß dem Design der Datenerhebung keine große Bedeutung zukommt: Ist eine Studie schlecht geplant, so ist der Informationsgehalt der Daten gering, und es kann auch kein entsprechend aussagekräftiges Ergebnis der Studie erwartet werden.

Nach dem Gesagten teilen wir statistische Prozeduren grundsätzlich ein in

- deskriptive und
- schließende Prozeduren.

Deskriptive Statistik. Deskriptive, statistische Methoden sind graphische und numerische Verfahren und Techniken, um Daten übersichtlich darzustellen. Ziel ist die Reduktion der Daten auf einige wenige charakteristische Größen, sodaß die Beobachtungen leichter interpretiert werden können, ohne daß es durch die Datenreduktion zu einem Verlust an relevanter Information kommt.

Der zweite Aspekt der Statistik ist der **schließende** Aspekt. Um den Unterschied zwischen der beschreibenden und der schließenden Fragestellung zu illustrieren, muß man die beiden Begriffspaare **Gesamtheit(=Population), Stichprobe** und **Parameter, Statistik** unterscheiden:

- Eine **Population** ist die Menge aller Untersuchungseinheiten in einer Studie.
- Eine **Stichprobe** ist eine Teilmenge einer Population, an der die im Sinn der Problemstellung interessierenden Merkmale beobachtet (gemessen) werden.
- Ein **Parameter** ist eine charakteristische Größe einer Gesamtheit, über die auf der Basis einer Stichprobe eine Aussage getroffen werden soll.

- Eine **Statistik** ist eine entsprechende Größe einer Stichprobe, die zur Aussage über einen Parameter herangezogen wird.

Obwohl man natürlich Aussagen über die Gesamtheit machen will, ist es oft unumgänglich, sich auf eine Stichprobe zu beschränken, da z.B.

- die Gesamtheit zu groß ist, als daß alle Einheiten untersucht werden könnten, oder

- durch die Messung die Untersuchungseinheit zerstört würde.

Aber auch in Fällen, in denen eine Vollerhebung möglich wäre, beschränkt man sich fast stets – aus ökonomischen Erwägungen – auf die Analyse einer Stichprobe.

Schließende Statistik. Verfahren der schließenden Statistik erlauben es, über Parameter einer Population Aussagen zu treffen, wenn nur Informationen aus einer Stichprobe, d.h. Statistiken, zur Verfügung stehen.

Natürlich interessieren primär die Parameter der Gesamtheit und nicht die Statistiken der Stichprobe. Diese sind nur insoweit von Interesse, als sie Informationen über die unbekannten Parameter enthalten. Wir verwenden die aus der Stichprobe erhaltene Statistik, um zu Aussagen über die Gesamtheit oder über einen unbekannten Parameter der Gesamtheit zu gelangen. Wichtig ist dabei, die Gesamtheit exakt zu beschreiben und abzugrenzen.

Basis der Verfahren der schließenden Statistik sind wahrscheinlichkeitstheoretische Überlegungen. Man modelliert den Mechanismus, der die Generierung der Daten wahrscheinlichkeitstheoretisch beschreibt, um auf Grund dieses Modells zu Aussagen über die interessierenden Parameter zu gelangen.

Es ist natürlich naheliegend, daß wir einen Fehler begehen, wenn wir nur auf der Basis einer Stichproben zu Aussagen über unbekannte Parameter gelangen. Wir werden sehen, daß man durch Randomisierung, d.i. die zufällige Ausswahl der Untersuchungseinheiten, den Stichprobenfehler kontrollieren kann.

Die Einteilung in beschreibende und schließende Statistik bestimmt auch die Gliederung des Buches. Deskriptive und explorative, datenanalytische Verfahren werden in den Kapiteln 2 bis 5 behandelt und an Hand des am Ende von Kapitel 2 vorgestellten Datensatzes illustriert. Nach einer Zusammenstellung wichtiger wahrscheinlichkeitstheoretischer Werkzeuge in den Kapiteln 6 bis 8 werden Verfahren der schließenden Statistik in den Kapiteln 9 bis 13 gebracht.

1.3 Datentypen, Messniveaus und Skalen

Charakteristika von Personen oder Dingen, den Untersuchungseinheiten, die zahlenmäßig ausgedrückt werden können, nennen wir **Merkmale** oder **Variable**. Die verschiedenen Werte, die ein Merkmal annehmen kann, heißen seine **Merkmalsausprägungen**. Der **Wert** eines Merkmals ist die Merkmalsausprägung, die wir in einer konkreten Situation beobachten oder messen. Die Menge aller möglichen Merkmalsausprägungen bildet den Wertebereich des Merkmals.

Um die Merkmalsausprägung zu erhalten oder zu messen, brauchen wir einen "Maßstab" oder eine **Skala**. Je nach Art und Qualität der Skala sprechen wir von verschiedenen Meßniveaus, die eine Hierarchie zunehmender Meßqualität bilden. Wir unterscheiden

- Nominalskala
- Ordinalskala
- metrische Skala

Mit dieser Klassifizierung weitgehend synonym ist die Einteilung in qualitative, Rang- und quantitative Merkmale: Nominalskalierte Merkmale werden auch als **qualitative**, ordinalskalierte Merkmale als Rangmerkmale bezeichnet. Als **quantitative** Merkmale werden metrische Merkmale, von manchen Autoren jedoch ordinale *und* metrische Merkmale angesprochen.

Eine weitere Einteilung ist die nach dem Wertebereich des Merkmals in

- stetige Merkmale und
- diskrete Merkmale

1.3.1 Merkmalstypen nach dem Meßniveau

Die Unterscheidung der Meßniveaus ist für den Statistiker wichtig, da zur Analyse unterschiedlich skalierter Merkmale auch unterschiedliche statistische Verfahren zu verwenden sind. Noch wichtiger ist die Unterscheidung für den Anwender der Statistik, da der Informationsgehalt von Daten umso größer ist, je höher das Meßniveau ist. Ein wesentlicher Teil der wissenschaftlichen Arbeit besonders in den Wirtschaftswissenschaften zielt darauf ab, die Meßqualität zu verbessern. Die Entwicklung des **Intelligenzquotienten** und die Messung von **Konsumentenzufriedenheit** sind Beispiele für Bestrebungen, komplexe Sachverhalte meßbar zu machen bzw. das Meßniveau zu erhöhen.

Nominalskalierte Merkmale

Die Nominalskala ist das unterste Meßniveau. Ein Merkmal heißt **nominal** oder nominalskaliert, wenn seine verschiedenen Ausprägungen nur durch ihren Namen unterschieden sind. Die auf einer Nominalskala "gemessenen" Werte sind nur Substitute oder Kodierungen für die Merkmalsausprägungen. Durch das Kodieren werden die Beobachtungen in Klassen eingeteilt, von denen jede alle solche Untersuchungseinheiten umfaßt, denen ein gemeinsamer Wert auf der Nominalskala zugeordnet wird. Voraussetzung ist, daß jede Untersuchungseinheit genau einer Klasse entspricht.

Beispiel 1.1 Das einfachste Beispiel nominalskalierter Merkmale sind die **dichotomen** Variablen. Sie können nur zwei Ausprägungen annehmen, etwa die Zahlen 0 und 1 oder die Zahlen 1 und 2. Mit den Zahlen 0 und 1 kann eine Population in die beiden Klassen der Männer und Frauen eingeteilt werden, wenn die Merkmalsausprägungen der Variablen "Geschlecht" als *weiblich* = 0 und *männlich* = 1 kodiert werden. Mit der Kodierung *blau* = 1, *grün* = 2, *braun* = 3, *grau* = 4 und *schwarz* = 5 für die Merkmalsausprägungen der – nominalskalierten – Variablen "Augenfarbe" kann die Population in fünf andere Klassen zerlegt werden.

Es ist zu beachten, daß die Zahlen, die den Merkmalsausprägungen zugeordnet werden, willkürliche sind. Sie repräsentieren nicht etwa eine Ordnung, die zwischen den einzelnen Ausprägungen besteht. Daher dürfen die Zahlen auch nicht in arithmetischen oder anderen Rechenoperationen verarbeitet werden. Man kann sich leicht vorstellen, daß viele statistische Operationen, etwa das Berechnen des Mittelwertes, für nominalskalierte Merkmale keinen Sinn geben.

Ordinalskalierte Merkmale

Ein Merkmal heißt **ordinal** oder ordinalskaliert, wenn seine Ausprägungen zueinander in einer Ordnungsbeziehung wie "größer", "kleiner", "besser", etc. stehen.

Beispiel 1.2 Das klassische Beispiel einer Ordinalskala ist die Notenskala mit den Werten 1 bis 5 für die Ausprägungen *sehr gut* bis *nicht genügend.* Die Skala zerlegt die Menge der Studenten in fünf Klassen von Studenten, die mit *sehr gut* etc. benotet wurden, zwischen denen darüber hinaus auch eine Rangordnung besteht: Die mit *sehr gut* benoteten Studenten haben eine bessere Leistung erbracht als jene, die mit *gut* benotet wurden, usw. Allerdings haben Abstände oder Quotienten zwischen den Ausprägungen keine Bedeutung. Man kann etwa nicht sagen, daß ein Student, der mit *sehr gut* benotet wurde, doppelt so gut ist wie einer mit einem *gut.*

Andere Beispiele von Ordinalskalen sind

(a) Beliebtheitsskalen von Personen, sonstige Präferenzskalen,

(b) Hierarchie in einer Organisation

(d) Güteklassen von Obst oder Lebensmitteln

(e) ATP Punkteliste des Internationalen Tennisverbandes

Wie in dem Beispiel angedeutet, geben die Skalenwerte keine Information über die Abstände zwischen den Merkmalsausprägungen. Dementsprechend dürfen die Zahlen nur in solchen arithmetischen oder anderen Rechenoperationen verarbeitet werden, die die Ordnungsrelation der Werte unverändert läßt, wie das Sortieren nach der Größe.

Metrisch skalierte Merkmale

Ein Merkmal heißt **metrisch** oder metrisch skaliert, wenn es in Vielfachen bestimmter Einheiten gemessen wird. Metrische Merkmale sind gleichzeitig ordinalskaliert. Man unterscheidet **verhältnisskalierte Merkmale** und **intervallskalierte Merkmale** je nachdem, ob die Skala einen natürlichen Bezugspunkt besitzt oder nicht.

Beispiel 1.3 Das Gewicht einer Person in Kilogramm oder der Umsatz eines Unternehmens in öS sind Beispiele für metrische Merkmale. Beide Skalen sind Verhältnisskalen: der Wert Null ist ein natürlicher Bezugspunkt. Ein klassisches Beispiel für ein intervallskaliertes Merkmale ist die Temperaturmessung in Celsius Graden. Diese Temperaturskala ergibt sich durch die willkürliche Einteilung des Intervalls zwischen Gefrier- und Siedepunkt von Wasser in 100 Teile. Eine alternative Temperaturskala ist die nach Fahrenheit, die ebenfalls willkürlich ist. Ein natürlicher Bezugspunkt wäre die Temperatur, bei der die Moleküle keine thermische Bewegung mehr ausführen; die Kelvin-Skala basiert auf diesem Bezugspunkt.

1.3.2 Merkmalstypen nach dem Wertebereich

Nach dem Wertevorrat des Merkmals unterscheiden wir **diskrete** und **stetige** Merkmale. In ersterem Fall ist die Zahl der möglichen Merkmalsausprägungen endlich oder abzählbar; sie sind typisches Ergebnis eines Zählvorganges. Die Merkmalsausprägung eines stetigen Merkmals kann jeder Wert eines Intervalls der reellen Zahlengeraden sein.

Beispiel 1.4 Die Zahl der defekten Produkte in der Tagesproduktion, die Zahl der Tage eines Jahres mit Frost, die Zahl derer unter 2000 Befragten, die mit "ja" antworten, sind Beispiele für diskrete Merkmale.

Das Gewicht einer Person, die Verspätung eines Zuges, die Temperatur des Katalysators in einem chemischen Prozeß sind Beispiele für stetige Merkmale.

Obwohl auch für die Einteilung in diskrete und stetige Merkmale gilt, daß zur Analyse unterschiedlicher Merkmale unterschiedliche statistische Verfahren zu verwenden sind, ist diese Einteilung von geringerer praktischer Bedeutung. Der Grund liegt darin, daß die Abgrenzung nicht sehr streng ist. Diskrete Merkmale werden oft wie stetige behandelt, wenn die Schrittweite der Maßeinheit in Bezug auf die beobachtete Größe kein ist. Beispielsweise werden monetäre Größen (der Umsatz eines Unternehmens in öS) meist als stetig betrachtet, obwohl sie, bedingt durch die Nichtteilbarkeit der kleinsten Währungseinheit, diskrete Größen sind. Umgekehrt ist Messung jedes stetigen Merkmals, bedingt durch eine endliche Meßgenauigkeit, eine diskrete Größe.

Kapitel 2

Deskriptive und explorative Datenanalyse: ein Merkmal

Wenn statistische Verfahren angewendet werden sollen, um Aussagen über einen bestimmten Realitätsaspekt zu machen, so benötigen wir numerische Daten, die diese Realität beschreiben. Wir haben in Kapitel 1 einiges darüber erfahren, wie wir uns die notwendigen Daten besorgen können. In diesem Kapitel werden wir elementare Analyseverfahren kennenlernen, mit denen diese Daten oder bestimmte Charakteristika dieser Daten graphisch oder zahlenmäßig dargestellt werden können. Diese Darstellungen sind essentiell für jede statistische Analyse, da oft schon das Beschreiben der Daten die im Sinn der Aufgabenstellung wesentlichen Aspekte sichtbar macht, und dieses "Erforschen" der Daten oft die weiteren Analyseschritte bestimmt. Dieses Kapitel ist diesen ersten Schritte der statistischen Analyse in empirischen Studien, nämlich

- der graphischen Darstellung und
- dem Bestimmen der – für eine Entscheidungssituation – relevanten Charakteristika

der gesammelten Daten gewidmet, wobei nur jeweils ein Merkmal betrachtet wird. Im folgenden Kapitel werden wir uns mit der Analyse von mehr als einer Variablen und insbesondere der Charakterisierung von Beziehungen zwischen den Merkmalen befassen. Wir beginnen mit der verdichteten Darstellung der Daten als Häufigkeitsverteilung der beobachteten Merkmalsausprägungen in Abschnitt 2.1; die Parameter, die eine Häufigkeitsverteilung charakterisieren, werden in Abschnitt 2.2 vorgestellt. In diesen beiden Abschnitten werden auch Methoden der explorativen Datenanalyse (EDA) behandelt; dabei geht es um spezielle Methoden der deskriptiven Statistik, die in den letzten zwanzig Jahren unter Nutzung der Möglichkeiten der EDV und

vieler neuer Ideen entwickelt wurden. Schließlich stellen wir in Abschnitt 2.3 weitere Methoden der graphischen Darstellung von Daten vor.

2.1 Die Häufigkeitsverteilung

Ziel jeder Datenanalyse ist die **Datenreduktion**, d.i. die Zusammenfassung der Daten so, daß die zugrundeliegenden Strukturen deutlich hervortreten. Das Ergebnis des Sammelns von Daten ist eine Liste oder eine Computerdatei, die für jede Beobachtung (neben einer Identifikation) eine entsprechende Merkmalsausprägung enthält. Aus dieser Zahlenmenge die für eine Situation relevante Information abzulesen, erfordert es, diese Information sichtbar zu machen. Eine wirksame Methode dazu ist das Zusammenfassen der Daten zu einer Verteilung der Häufigkeiten, mit denen die Ausprägungen des interessierenden Merkmals beobachtet wurden. Diese Häufigkeitsverteilung kann in Tabellenform oder – für manche Zwecke besser noch – als graphische Darstellung wiedergegeben werden.

Wir behandeln zunächst den Fall eines qualitativen oder diskret-quantitativen Merkmals X mit k Ausprägungen. Sollen n Beobachtungen eines solchen Merkmals dargestellt werden, so kommen dafür ein **Balkendiagramm** (im Englischen *bar chart*) oder ein **Histogramm** in Frage. Die Ausprägungen $x_1, \ldots, x_k$ des Merkmals X seien mit den **absoluten Häufigkeiten** $H_1, \ldots, H_k$ beobachtet worden:

$$H_j = \text{Anzahl der Beobachtungen mit } X = x_j$$

für $j = 1, \ldots, k$. Natürlich gilt $\sum_j H_j = n$. Dividiert man die H_j durch n, so erhält man die **relativen Häufigkeiten**

$$h_j = \frac{H_j}{n}.$$

Relativen Häufigkeiten werden oft in Prozenten angegeben.

Beispiel 2.1 Für die dichotome Variable "Geschlecht" ergibt sich für eine Stichprobe vom Umfang 50 die Häufigkeitsverteilung

		H_j	h_j
weiblich	1	20	0.4
männlich	2	30	0.6
gesamt		50	1.0

Eine analoge Tabelle für die qualitative Variable "Augenfarbe" erhält man zu

		H_j	h_j
braun	1	19	0.388
grün	2	12	0.245
blau	3	15	0.306
grau	4	2	0.041
schwarz	5	1	0.020
gesamt		49	1.0

wobei für eine Beobachtung die Merkmalsausprägung der Augenfarbe fehlt (sie ist "missing"). Aus diesen Tabellen können wir wichtige Charakteristika, etwa die am häufigsten vorkommenden Merkmalsausprägungen, ablesen.

In graphischer Form kann die Häufigkeitsverteilung als Balkendiagramm dargestellt werden: Die Balken über den einzelnen Merkmalsausprägungen haben eine Länge, die proportional seiner Häufigkeit ist. Die Balken können senkrecht oder waagrecht angeordent sein. Die *Abbildung 2.1* zeigt Balkendiagramme zu den Variablen "Geschlecht" und "Augenfarbe" unserer 50 Studenten, die mittels der MINITAB-Prozedur `histogram` erzeugt wurden. Als Merkmalsausprägungen gibt MINITAB die numerischen Codes an, die in den Tabellen von Beispiel 2.1 auch ausgewiesen sind.

Abbildung 2.1: Balkendiagramme zu den Variablen (a) "Geschlecht" und (b) "Augenfarbe" von 50 Studenten.

```
MTB > hist 'AFA'

Histogram of AFA   N = 49   N* = 1

Midpoint   Count
       1      19  ******************
       2      12  ************
       3      15  ***************
       4       2  **
       5       1  *

MTB > hist 'SEX'

Histogram of SEX   N = 50

Midpoint   Count
       0      20  ********************
       1      30  ******************************
```

Die obigen Tabellen und Balkendiagramme zeigen sogenannte eindimensionale Häufigkeitsverteilungen: Sie betreffen jeweils nur ein Merkmal. Betrach-

ten wir zwei (oder mehrere) Merkmale simultan, so erhalten wir **zweidimensionale** (oder höherdimensionale) Häufigkeitsverteilungen. Alle möglichen Kombinationen von Merkmalsausprägungen der beiden (oder mehreren) Merkmale bilden dann eine sogenannte **Kreuzklassifikation**. Die entstehende Tabelle heißt **Kontingenztafel**.

Beispiel 2.2 Für die beiden Merkmale "Geschlecht" und "Augenfarbe" gibt es 2×5 oder 10 Merkmalskombinationen. So gibt es etwa weibliche Personen mit blauen Augen, männliche mit schwarzen Augen, etc. Die Häufigkeitsverteilung kann in einer zweidimensionalen Kontingenztafel mit zwei Zeilen und fünf Spalten, auch 2×5-Tafel genannt, dargestellt werden. Die absoluten Häufigkeiten zeigt die folgende Tafel:

	blau	*grün*	*braun*	*grau*	*schwarz*	gesamt
weiblich	6	7	7	0	0	20
männlich	9	5	12	2	1	29
gesamt	15	12	19	2	1	49

Beachte! Die beiden (eindimensionalen) Häufigkeitsverteilungen der Merkmale "Geschlecht" und "Augenfarbe" können als Zeilen- bzw. Spaltensummen der 2×5-Tafel abgelesen werden.

Aus einer zweidimensionalen Kontingenztafel können zwei verschiedene Arten von relativen Häufigkeiten abgeleitet werden. Dividiert man alle Eintragungen der Tafel durch den Stichprobenumfang, so erhält man die relativen Häufigkeiten der entsprechenden Merkmalskombinationen. Eine andere Art von Tafel ergibt sich, wenn man die Eintragungen der Zeilen auf die jeweilige Zeilensumme bezieht: Dann erhält man sogenannte **bedingte relative Häufigkeiten** des einen Merkmals, wobei die Bedingung darin besteht, daß das andere Merkmal einen bestimmten Wert hat.

Beispiel 2.3 Dividiert man alle Eintragungen der 2×5-Tafel von Beispiel 2.2 durch den Stichprobenumfang 49, so erhält man die relativen Häufigkeiten der verschiedenen Merkmalskombinationen: so hat die relative Häufigkeit der Merkmalskombination *weiblich und blau* den Wert 0.122. Eine bedingte Verteilung des Merkmals der "Augenfarbe" gibt es für die beiden Geschlechter: Bezieht man die Häufigkeiten der Eintragungen einer Zeile auf die jeweilige Zeilensumme, so hat z.B. die Augenfarbe *braun* einen Anteil von 0.35 bei den weiblichen und einen von 0.41 bei den männlichen Personen. Solche bedingte relative Häufigkeiten erlauben die Beantwortung der Frage, ob die Verteilung der Augenfarben über die beiden Geschlechter gleich ist.

Eine analoge Darstellung ist auch für stetige Merkmale möglich. Allerdings ist es dazu notwendig, die Merkmalsausprägungen in Klassen zu gruppieren und damit zu diskretisieren. Dann kann man die sogenannten **Klassenhäufigkeiten** in einem Histogramm darstellen, einer Variante des Balkendiagramms, bei dem die Fläche jedes Balkens proportional der jeweiligen Klassenhäufigkeit ist. Graphische Darstellungen von stetigen Merkmalen, die den Verzicht auf Information ganz oder teilweise vermeiden, der mit dem Diskretisieren verbunden ist, sind das Punkt- und das Stem & Leaf Diagramm.

Beispiel 2.4 In der Elektroabteilung eines Warenhauses wurden bei 50 Kunden die folgenden Beträge in Rechnung gestellt:

10390	2950	12730	6260	6965	13610	8030
7530	21200	9345	4820	8580	16042	9050
10250	8512	9827	7360	8550	9240	7490
4785	7000	6000	6983	8500	17140	9340
10395	10150	12725	7290	25650	9340	15937
11410	7824	12260	10865	11860	9748	8640
12334	9620	7710	24020	6210	11644	4520
3470						

Natürlich ist die Tafel der Beobachtungen nicht sehr informativ; dieser Mangel wird mit zunehmender Zahl der Beobachtungen immer gravierender. Mehr Einsicht geben graphische Darstellungen der Häufigkeitsverteilung.

Klasse (j)	Rechnungsbeträge	H_j	h_j
1	weniger als 2000.-	0	0.00
2	2000.- bis 4000.-	2	0.04
3	4000.- bis 6000.-	3	0.06
4	6000.- bis 8000.-	12	0.24
5	8000.- bis 10000.-	14	0.28
6	10000.- bis 12000.-	8	0.16
7	12000.- bis 14000.-	5	0.10
8	14000.- bis 16000.-	1	0.02
9	16000.- bis 18000.-	2	0.04
10	18000.- bis 20000.-	0	0.00
11	20000.- bis 22000.-	1	0.02
12	22000.- bis 24000.-	0	0.00
13	24000.- bis 26000.-	2	0.04

Die *Abbildung 2.2* zeigt (a) ein Punktdiagramm, (b) ein Stem & Leaf Diagramm und (c) ein Histogramm. Zum manuellen Erstellen des Histo-

gramms wurden 13 Klassen gebildet und die Tabelle der **Klassenhäufigkeiten** erstellt, wie sie in der folgenden Tabelle gezeigt ist. Dazu wurde für jede Klasse die Zahl der Beobachtungen ausgezählt, die in dem der Klasse entsprechenden Intervall enthalten sind.

In der letzten Spalte sind die relativen Häufigkeiten als Anteile der Rechnungsbeträge in der jeweiligen Klasse angegeben.

Abbildung 2.2: Darstellungen der Rechnungsbeträge von 50 Kunden der Elektroabteilung eines Warenhauses aus Beispiel 2.4; (a) Punktdiagramm; (b) Stem & Leaf Diagramm; (c) Histogramm.

(a) Punktdiagramm

```
MTB > dotp 'RBETR'

                    . . .
                   .: : : .   :
          .. .: :.:::::::::.:.: .    : .       .     . .
     +---------+---------+---------+---------+---------+---RBETR
     0       5000      10000     15000     20000     25000
```

(b) Stem & Leaf Diagramm

```
MTB > Stem-and-Leaf 'RBETR'

Stem-and-leaf of RBETR     N  = 50
Leaf Unit = 1000

    2     0 23
    5     0 444
   17     0 666667777777
  (14)    0 88888899999999
   19     1 00000111
   11     1 22223
    6     1 5
    5     1 67
    3     1
    3     2 1
    2     2
    2     2 45
```

(c) Histogramm

```
MTB > hist 'RBETR';
SUBC> start 1000;
SUBC> incr 2000.
```

```
Histogram of RBETR   N = 50

Midpoint   Count
    1000       0
    3000       2  **
    5000       3  ***
    7000      12  ************
    9000      14  **************
   11000       8  ********
   13000       5  *****
   15000       1  *
   17000       2  **
   19000       0
   21000       1  *
   23000       0
   25000       2  **
```

Das **Punktdiagramm** erhält man durch das Einzeichnen eines Punktes für jede Beobachtung an der Stelle der Achse der Merkmalsskala, die der Merkmalsausprägung der Beobachtung entspricht.

Das Stem & Leaf-Diagramm (im Deutschen auch Stamm & Blatt Diagramm) ist eine einfache Möglichkeit, die Form einer Häufigkeitsverteilung graphisch darzustellen, ohne auf die numerischen Werte der Daten zu verzichten. Seine Verwendung zum Visualisieren der erhobenen Daten ist bei nicht zu vielen (≤ 50) Beobachtungen zu empfehlen.

Definition 2.1 *Konstruktion eines* **Stem & Leaf Diagramms***:*

1. *Zerlege den Wert jeder Beobachtung in das Blatt (letzte Stelle) und den Stamm (die übrigen Stellen).*
2. *Ordne den Stamm von oben nach unten mit wachsenden Werten an, zeichne eine Linie rechts neben den Stamm.*
3. *Füge die Blätter in der Folge steigender Werte rechts neben den Stamm.*

Das Stem & Leaf Diagramm erlaubt eine einfache und schnelle visuelle Inspektion der Daten. Außerdem können bestimmte Maßzahlen der Verteilung einfach abgelesen oder berechnet werden.

Das Histogramm ist ähnlich dem Stem & Leaf Diagramm, wobei die Flächen der einzelnen Säulen proportional der Zahl der Beobachtung im entsprechenden Intervall sind.

Definition 2.2 *Konstruktion eines* **Histogramms**:

1. *Ordne die n Beobachtungen nach steigender Größe, und bestimme die Spannweite der Häufigkeitsverteilung, d.i. der Abstand von der kleinsten zur größten Beobachtung.*

2. *Zur Festlegung der Klassen unterteile die Spannweite in Intervalle gleicher Länge; die Zahl k der Klassen soll etwa $\sqrt{n}$ betragen und zwischen fünf und 20 liegen. Die Klassenmitten sollen "einfache" Zahlen sein.*

3. *Bestimme die Zahl der Beobachtungen jeder Klasse, d.s. die absoluten Klassenhäufigkeiten; die relativen Häufigkeiten erhält man durch Dividieren der absoluten Häufigkeiten durch die Zahl der Beobachtungen.*

4. *Zeichne das Histogramm. Bei gleichen Klassenbreiten sind die Höhen der Felder proportional den Häufigkeiten; bei ungleichen Klassenbreiten sind die Höhen proportional den Quotienten aus Häufigkeit und Klassenbreite.*

Wie sich aus der Definition 2.2 ergibt, können Häufigkeitstabelle und Histogramm für **absolute Häufigkeiten** oder für **relative Häufigkeiten** konstruiert werden.

Durch die Faustregel, daß die Zahl der Klassen etwa $\sqrt{n}$ betragen soll, wird vermieden, daß das Histogramm "zu unregelmäßig" (zu viele Klassen) oder "zu glatt" (zu wenige Klassen) ist.

Beispiel 2.5 Ein Histogramm für das Merkmal "Körpergröße" ergibt sich folgendermaßen: Die kleinste Beobachtung ist 153 cm, die größte Beobachtung ist 195 cm. Die Spannweite der n=50 Beobachtungen ist daher 42 cm. Teilt man diese in $\sqrt{50} \approx 7$ Klassen, so kann z.B. eine Klasseneinteilung gewählt werden, die das Intervall $(150, 199]$ in Klassen zu je 7 cm einteilt. Die absoluten und relativen Klassenhäufigkeiten für alle Befragten und die absoluten Häufigkeiten für männliche (m) und weibliche (w) Personen zeigt die folgende Tabelle.

		Häufigkeit			
Klasse(j)	Körpergröße	H_j	h_j	m	w
1	(150-157]	1	0.02	1	0
2	(157-164]	3	0.06	3	0
3	(164-171]	13	0.26	9	4
4	(171-178]	12	0.24	6	6
5	(178-185]	7	0.14	1	6
6	(185-192]	12	0.24	0	12
7	(192-199]	2	0.04	0	2
	gesamt	50	1.00	20	30

Das Histogramm für die "Körpergröße" zeigt die *Abbildung 2.3.* Sie zeigt eine zweigipfelige Verteilung; die beiden Gipfel entsprechen den männlichen (m) und weiblichen (w) Teilpopulationen.

Abbildung 2.3: Histogramm der "Körpergröße" von 50 Personen aus Beispiel 2.5.

```
MTB > histogram 'GRO';
SUBC> start 154;
SUBC> increment 7.

Histogram of GRO   N = 50

Midpoint   Count
  154.00       1  *
  161.00       3  ***
  168.00      13  *************
  175.00      12  ************
  182.00       7  *******
  189.00      12  ************
  196.00       2  **
```

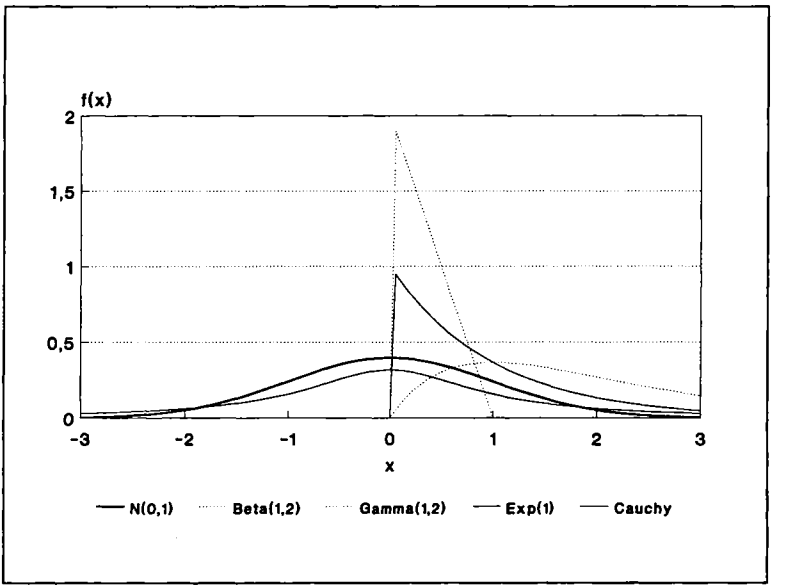
Charakteristika des Stem & Leaf Diagramms bzw. des Histogramms, die für die Analyse wichtig sein können, sind:

- das Niveau oder Zentrum der Verteilung; die Beobachtung, die in der Folge der steigenden Werte in der Mitte liegt – der Median –, wird gerne als Maß für die Lage des Zentrums genommen
- ein Maß der Variabilität der Verteilung; es sagt uns, wie sehr sich die Beobachtungen unterscheiden
- die Symmetrie oder Schiefe der Verteilung und andere Aspekte der Form der Verteilung
- "Ausreißer", d.s. Beobachtungen, die abseits der Masse der übrigen Daten liegen; Lücken in der Verteilung; sonstige Besonderheiten

Dem Beschreiben einer Häufigkeitsverteilung durch Charakteristika wie dem Zentrum der Verteilung wird der folgende Abschnitt gewidmet. Dort wird auch eine weitere Art der graphischen Darstellung der Häufigkeitsverteilung gebracht, das sogenannte Box- oder Box & Whisker-Plot, das einen raschen Eindruck von einigen dieser Charakteristika gibt und Details der Form außer Acht läßt. Daneben interessieren manchmal Charakteristika der Form: Die in *Abbildung 2.2* gezeigte Häufigkeitsverteilung nennt man unimodal oder eingipfelig. Sie ist unsymmetrisch und rechtsschief, womit angedeutet werden

soll, daß die Verteilung nach großen Werten mehr auslädt als nach kleinen Werten. Andere Formen der Verteilung werden mit symmetrisch, linksschief, zweigipfelig, mehrgipfelig, etc. bezeichnet; es ist leicht vorstellbar, was mit diesen Eigenschaften gemeint ist. Die *Abbildung 2.4* zeigt einige typische Verteilungen.

Abbildung 2.4: Verschiedene Verteilungstypen



Es empfiehlt sich stets, vor einer weiteren Analyse der Daten die graphische Darstellung der Daten sorgfältig in Augenschein zu nehmen. Abweichungen vom zu erwartenden Muster der Daten, etwa Ausreißer, sollen vor der numerischen Analyse geklärt werden.

2.2 Charakteristika einer Verteilung

Das Histogramm komprimiert eine Datenmenge zu einer graphischen Darstellung, aus der wesentliche Charakteristika der Häufigkeitsverteilung visuell rasch erfaßt werden können. Unter diesen Charakteristika sind solche der Lage und der Streuung für viele Fragestellungen die entscheidenden. Diese Charakteristika können – wie auch andere – durch Maßzahlen zahlenmäßig dargestellt werden.

Einige Charakteristika einer Häufigkeitsverteilung basieren auf den sogenannten **Quantilen**. Die Beobachtungen eines zumindest ordinalskalierten Merkmals X seien $x_1, \ldots, x_n$. Die nach ihrer Größe aufsteigend sortierten

x_i wollen wir mit $x_{(1)}, \ldots, x_{(n)}$ bezeichnen; dabei gilt $x_{(1)} \leq \ldots \leq x_{(n)}$; $x_{(1)}$ ist die kleinste und $x_{(n)}$ die größte Beobachtung. Das p-Quantil einer Datenmenge ist, grob gesprochen, jene Beobachtung, die größer als $100p\%$ und kleiner als $100(1-p)\%$ der Daten ist. Mit Hilfe der oben eingeführten Notation können wir das p-Quantil folgendermaßen definieren.

Definition 2.3 *Das* **p-Quantil** $\tilde{x}_p$ *einer Datenmenge* $\{x_1, \ldots, x_n\}$ *ist durch*

$$\tilde{x}_p = \begin{cases} x_{(\lceil np \rceil)}, & \text{wenn } np \text{ nicht ganzzahlig,} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}), & \text{wenn } np \text{ ganzzahlig} \end{cases}$$

gegeben, wobei $\lceil x \rceil$ *die nächstgrößere ganze Zahl zu* x *bedeutet.*

Das 0.25-Quantil, auch **1. Quartil** oder **unteres Quartil** genannt und mit Q_u bezeichnet, kommt in der Folge der sortierten Beobachtungen an jener Stelle, die rechts von dem Viertel der noch kleineren und links von den drei Viertel der größeren Beobachtungen liegt. Analog ist das 0.75-Quantil oder **3. Quartil** oder **obere Quartil** Q_o zu verstehen. Das 0.5-Quantil hat auch den Namen Median.

Wie schon im Abschnitt 1.2 über Datentypen gesagt, ist die Zulässigkeit von Rechenoperationen vom Skalierungsniveau der Merkmale abhängig. Dementsprechend müssen für verschiedene Skalierungsniveaus unterschiedliche Maßzahlen eingeführt werden.

2.2.1 Lagemaße

Lagemaße sind Maße, die Information über die Lage der Verteilung auf der (rellen) Achse der Merkmalsausprägungen, d.h. über die "Größe" der Beobachtungen, geben. Die am häufigsten verwendeten Maße der Lage einer Verteilung sind der Mittelwert, der Median und der Modus. Von anderen Möglichkeiten, die Lage der Verteilung zu charakterisieren, sollen die sogenannten *robusten* Lagemaße erwähnt werden.

Arithmetisches Mittel

Das wohl wichtigste Lagemaß ist das arithmetische Mittel $\bar{x}$, das allerdings metrisch skalierte Merkmale voraussetzt.

Definition 2.4 *Das* **arithmetische Mittel** *(der* **Mittelwert***) von* n *Beobachtungen* $x_1, \ldots, x_n$ *ist durch*

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

gegeben.

Da das Addieren der Werte x_i nur für metrisch skalierte, nicht aber für ordinal- oder nominalskalierte Merkmale zulässig ist, darf das arithmetische Mittel sinnvollerweise auch nur für metrisch skalierte Merkmale verwendet werden.

Eine gerne benützte Veranschaulichung des Mittelwertes in Begriffen des täglichen Lebens basiert auf der Mechanik: Stellt man sich die Verteilung der Daten als Masseverteilung vor, bei der jede Beobachtung die Masse 1 hat, so ist der Mittelwert als Schwerpunkt der Verteilung interpretierbar. Die reelle Achse als Waagebalken würde sich horizontal stellen, wenn sich der Drehpunkt der Waage im Mittelwert befindet.

Einige Eigenschaften des Mittelwerts sind in folgendem Satz zusammengefaßt.

Satz 2.1 Eigenschaften des Mittelwertes:

(a) Aus $y_i = ax_i + b$, $i = 1, \ldots, n$, *folgt für beliebige Konstante* a *und* b

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

(b) Die Summe der Abweichungen der Beobachtungen vom Mittelwert ist Null:

$$\sum_i (x_i - \bar{x}) = 0.$$

Bei der Berechnung von Mittelwerten sind die folgenden Sonderfälle zu beachten:

(a) Sind die Merkmalsausprägungen diskret, und können sie nur die k Werte $x_1, \ldots, x_k$ annehmen, so ergibt sich der Mittelwert zu

$$\bar{x} = \frac{1}{n} \sum_i x_i H_i = \sum_i x_i h_i,$$

wenn die Werte x_i mit den Häufigkeiten H_i $(i = 1, \ldots, n)$ beobachtet wurden.

(b) Interpretation des Mittelwertes von dichotomen Beobachtungen: Seien die $x_1, \ldots, x_n$ die Beobachtungen eines dichotomen Merkmals, das nur die beiden Werte 0 und 1 annehmen kann; dann ist

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_i x_i \\ &= \frac{1}{n} [\text{Anzahl 1 unter den Beobachtungen}] \\ &= [\text{relative Häufigkeit der "1"}]. \end{aligned}$$

(c) **Gewogenes Mittel:** Es seien k Teilmengen von Daten gegeben: $x_{11}, \ldots, x_{1n_1}; \ldots; x_{k1}, \ldots, x_{kn_k}$; wobei $n = n_1 + \ldots + n_k$. Dann ergibt sich der Mittelwert $\bar{x}$ zu

$$\begin{aligned} \bar{x} &= \frac{1}{n}(x_{11} + \ldots + x_{1n_1} + \ldots + x_{k1} + \ldots + x_{kn_k}) \\ &= \frac{1}{n}(n_1\bar{x}_1 + \ldots + n_k\bar{x}_k) = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{x}_i \end{aligned}$$

Beachte! Das Mittel ist der gewogene Mittelwert (nicht der einfache Mittelwert) der Teilmengen-Mittelwerte. Ausnahme ist der Fall, daß alle Teilmengen den gleichen Umfang haben, d.h. $n_1 = \ldots = n_k$.

Siehe dazu auch Beispiel 2.12

Modus und Median

Als Lagemaß für nominalskalierte Merkmale bietet sich der **Modus** der Häufigkeitsverteilung an:

Modus = häufigster Wert der Verteilung.

Beispiel 2.6 In der Stichprobe, die der Häufigkeitsverteilung in Beispiel (2.1) zugrundeliegt, ist der Modus des Merkmals "Geschlecht" die Merkmalsausprägung *männlich*; der des Merkmals "Augenfarbe" ist *braun*. Da die beiden Verteilung eingipfelig sind, sind die Modi eindeutig.

Für ordinalskalierte Merkmale kann man als Lagemaß natürlich ebenfalls den Modus verwenden. Da aber eine Ordnung zwischen den Merkmalsausprägungen definiert und daher ihr Sortieren möglich ist, kann man darüber hinaus solche Maße einführen, die auf dem Sortieren der Beobachtungen beruhen. Das wichtigste entsprechende Lagemaß ist der **Median** $\tilde{x}$, das ist das 0.5-Quantil oder jene Merkmalsausprägung, die am "mittleren Platz" in der Folge der sortierten Beobachtung steht. In Analogie zur *Definition 2.3* der Quantile definiert man den Median wie folgt.

Definition 2.5 *Der* **Median** $\tilde{x}$ *von* n *Beobachtungen* $x_1, \ldots, x_n$ *ist durch*

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & \textit{wenn } n \textit{ ungerade} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \textit{wenn } n \textit{ gerade} \end{cases}$$

gegeben.

Der Median ist somit jener Wert, der die Häufigkeitsverteilung "halbiert", d.h. es liegen (höchstens) 50% der Beobachtungen links bzw. rechts des Medians. Der Median kann nun unmittelbar durch abzählen aus dem Stem & Leaf Diagramm gefunden werden.

Beispiel 2.7 Den Median der Variablen "Schuhgröße" erhält man aus dem Stem & Leaf Diagramm durch abzählen zu

$$\tilde{x} = \frac{1}{2}(x_{(25)} + x_{(26)}) = 41.5$$

Dabei wurde berücksichtigt, daß der Stichprobenumfang $n = 50$ eine gerade Zahl ist.

Beachte! Median und Mittelwert fallen zusammen, wenn die Häufigkeitsverteilung symmetrisch ist. Der Mittelwert bzw. der Median ist dann Symmetriezentrum.

Robuste Lagemaße

Maßzahlen, deren Wert von Ausreißern nicht oder nur geringfügig verzerrt werden, nennen wir robuste Maßzahlen. Solche **Ausreißer** können sich etwa als Folge der Vermengung von Daten aus verschiedenen Populationen oder durch Übertragungsfehlern in einer Stichprobe befinden und sind Werte, die nach oben oder unten von der Masse der übrigen Beobachtungen abweichen. Robuste Lagemaße sollen demnach auch dann die Lage der Häufigkeitsverteilung einigermaßen richtig angeben, wenn die Daten durch Ausreißer verfälscht sind.

Vergleicht man den Median mit dem Mittelwert so zeigt sich, daß der Median robuster als der Mittelwert ist: Da der Median von den meisten Beobachtungen nur ihre Sortierfolge berücksichtigt, haben einige extreme Beobachtungen auf seinen Wert keinen Effekt; demgegenüber gehen in die Berechnung des Mittelwertes alle Beobachtungen mit dem gleichen Gewicht ein, und er reagiert deshalb empfindlich auf extreme Beobachtungen.

Neben dem Median werden zu den robusten Lagemaßen einige Maßzahlen gerechnet, die sich durch "Robustifizieren" des Mittelwertes ergeben, nämlich der getrimmte und der winsorisierte Mittelwert gezählt.

(a) Der α-**getrimmte Mittelwert** ergibt sich als Mittelwert der verbleibenden Beobachtungen, wenn die größten und die kleinsten 100α-% der Beobachtungen weggelassen werden.

(b) Der **winsorisierte Mittelwert** ist der Mittelwert der modifizierten Stichprobe, die sich ergibt, wenn die Beobachtungen, die größer als das dritte Quartil (kleiner als das erste Quartil) sind, durch das dritte (erste) Quartil ersetzt werden.

Sowohl getrimmter als auch winsorisierter Mittelwert werden durch einen kleinen Anteil von extremen Beobachtungen oder Ausreißern nicht verfälscht.

2.2.2 Streuungsmaße

Streuungsmaße beschreiben die Variabilität der Daten. Streuungsmaße basieren auf Differenzen der Merkmalsausprägungen zwischen einzelnen Beobachtungen oder – was im wesentlichen dasselbe ist – auf Abweichungen der Merkmalsausprägungen vom Mittelwert. Beispielsweise definierte C. Gini eine Streuungsmaßzahl Δ_G als durchschnittliche absolute Differenz aller Paare von Beobachtungen

$$\Delta_G = \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j| .$$

Allerdings ist dieses Streuungsmaß wenig gebräuchlich. Wiederum müssen wir je nach Datentyp bzw. Messniveau verschiedene Maßzahlen unterscheiden.

Die Varianz

Das wichtigste Streuungsmaß ist das Varianz s^2, die wie das arithmetische Mittel metrisch skalierte Merkmale voraussetzt. Die Varianz ist die mittlere quadratische Abweichung der Beobachtungen vom Mittelwert.

Definition 2.6 *Die* **Varianz** *von n Beobachtungen $x_1, \ldots, x_n$ ist durch*

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

gegeben. Die (positive) Wurzel aus der Varianz heißt die **Standardabweichung**

$$s = \sqrt{s^2} .$$

Der Quotient

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

wird **Variationskoeffizient** *genannt.*

Ein Vorteil der Standardabweichung gegenüber der Varianz liegt darin, daß die Standardabweichung die gleiche Maßeinheit wie die Beobachtungen hat, was für die Varianz nicht zutrifft. Demnach wird man zur Interpretation von Ergebnisse die Standardabweichung der Varianz vorziehen.

Vielfach wird die Varianz als

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

definiert. Die Argumente für den Nenner "n" oder "$n-1$" beruhen auf schätztheoretischen Überlegungen; wir werden im Kapitel 9 "Konzepte der statistischen Inferenz" darauf eingehen.

Am Rande sei erwähnt, daß man die Varianz als Trägheitsmoment interpretieren kann, wenn man die Häufigkeitsverteilung wiederum als Masseverteilung auffaßt.

Sollen Streuungsmaße verschiedener Merkmale verglichen werden, so müssen Unterschiede im Niveau der Merkmale berücksichtigt werden. Eine Standardabweichung von 10 Einheiten ist bei einem Mittelwert von 10 anders zu sehen als bei einem Mittelwert von 1000. Der Variationskoeffizient v ist ein normiertes Streuungsmaß. Er eignet sich besonders zum Vergleich der Streuung von Merkmalen unabhängig von deren Mittelwerten. Es ist nur sinnvoll, wenn das Merkmal nichtnegativ ist.

Den Aufwand zur praktischen Berechnung der Varianz kann man stark reduzieren, wenn man vom Verschiebungssatz Gebrauch macht: Anstelle von n Subtraktionen braucht man nur eine auszuführen.

Satz 2.2 Verschiebungssatz:

$$s^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2.$$

Der Verschiebungssatz ergibt sich aus der Definition der Varianz durch Ausquadrieren. Einige weitere Eigenschaften der Varianz sind in folgendem Satz zusammengefaßt.

Satz 2.3 *Eigenschaften der Varianz:*

(a) Aus $y_i = ax_i + b$ *folgt für alle reellen* a *und* b*:*

$$s_y^2 = a^2 s_x^2.$$

(b) Aus $\sum_i (x_i - z)^2 = ns^2 + (\bar{x} - z)^2$ *folgt für alle reellen* z

$$\frac{1}{n}\sum_i (x_i - z)^2 \to \min \quad \textit{für} \quad z = \bar{x}.$$

Ähnlich dem gewogenen arithmetischen Mittel kann man die Gesamtvarianz als gewogene Varianz von k Teilmengen von Beobachtungen errechnen:

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

wobei $\bar{x}$ der (gewogene) Mittelwert ist.

Berechnung von Mittelwert und Varianz aus einer Häufigkeitsverteilung

Bei der praktische Berechnung von Mittelwert und Varianz kann man manchmal nicht mehr auf die einzelnen Beobachtungen, die Rohdaten x_i, i =

$1, \ldots, n$, zurückgreifen, sondern nur mehr auf die Verteilung der Klassenhäufigkeiten, die durch k Zahlenpaare a_j, H_j $j = 1, \ldots, k$, gegeben ist. Dabei bedeutet a_j die Klassenmitte der j-ten Klasse und H_j die absolute Häufigkeit, mit der die j-te Klasse besetzt ist. Aus diesen Daten kann man mit stark reduzierten Rechenaufwand Näherungswerte für Mittelwert und Varianz bestimmen:

$$\bar{x} \doteq \frac{1}{n}\sum_{j=1}^{k} a_j H_j$$

$$s^2 \doteq \frac{1}{n}\sum_{j=1}^{k} a_j^2 H_j - \bar{x}^2.$$

Aus den Formeln erkennt man, worin die Näherung besteht. Es wird unterstellt, daß alle Beobachtungen der j-ten Klasse den Wert der Klassenmitte a_j haben. Demnach ist der Fehler der Näherung umso größer, je stärker die einzelnen Beobachtungen von den Klassenmitten abweichen. Wie man sich überlegen kann, ist der Fehler beim Mittelwert nicht systematisch, während die Varianz überschätzt wird. Eine Korrektur des systematischen Fehlers bietet die **Sheppard Korrektur**:

$$s^2_{korr} = s^2 - \frac{b^2}{12};$$

hier ist b die (für alle Klassen als gleich vorausgesetzte) Breite der Klassen.

Spannweite, Interquartilsabstand, MAD

Für metrisch skalierte Daten können Streuungsmaße auch als Differenzen zweier geordneter Werte definiert werden. So ist die **Spannweite** $R = x_{(n)} - x_{(1)}$ als Differenz zwischen dem größten und dem kleinsten Wert definiert. Ein offensichtlicher Nachteil dieses Maßes ist, daß es sehr empfindlich auf extreme Beobachtungen oder Ausreißer reagiert; es ist nicht robust. Diesen Nachteil hat der Interquartilsabstand I nicht, der nur die inneren 50% der Häufigkeitsverteilung und somit nicht die extremen Beobachtungen berücksichtigt.

Definition 2.7 *Der* **Interquartilsabstand** I *ist definiert als*

$$I = \tilde{x}_{0.75} - \tilde{x}_{0.25} = Q_o - Q_u.$$

Ein wenig verwendetes Streuungsmaß ist die **mittlere absolute Abweichung** oder MAD: Sie ist die durchschnittliche absolute Abweichung der Beobachtungen vom Median

$$d_{\tilde{x}} = \frac{1}{n}\sum_i \mid x_i - \tilde{x} \mid .$$

2.2.3 Weitere Maße

Neben den Lage- und Streuungsmaßen interessieren je nach Problemstellung auch andere Maße, die spezielle Aspekte der Form der Verteilung beschreiben. Eine besondere Bedeutung haben Schiefe- und Wölbungsmaße.

Schiefe

Die **Schiefe** einer Häufigkeitsverteilung ist ein Maß für ihre Asymmetrie. Rechts- bzw. linksschiefe Verteilungen sind durch lange rechte bzw. linke Schwänze der Verteilung charakterisiert. Maße, die diese Asymmetrie messen, werden so definiert, daß rechtsschiefe Verteilungen positive und linksschiefe Verteilungen negative Schiefemaße aufweisen. Für eingipfelige Verteilungen gelten die folgenden Relationen zwischen den Lagemaßen Modus, Median und Mittelwert:

rechtsschief	linksschief
$Modus < Median < \bar{x}$	$Modus > Median > \bar{x}$

Die Pearson'schen Schiefekoeffizienten SK spiegeln diese Relationen wieder:

$$\mathrm{SK} = \begin{cases} \frac{\bar{x} - Modus}{s}, \\ 3\frac{(\bar{x} - Median)}{s}. \end{cases}$$

Ein weiteres Schiefemaß ist der von Fisher eingeführte Momentkoeffizient

$$g_1 = \frac{m_3}{s^3},$$

wobei $m_3 = \frac{1}{n}\sum_i (x_i - \bar{x})^3$ und s^3 die dritte Potenz der Standardabweichung ist. Für symmetrische Verteilungen sind alle Schiefemaße gleich Null. Beachte, daß diese Schiefemaße nur dann sinnvoll sind, wenn die Verteilung eingipfelig ist.

Wölbung

Ein Maß für die **Wölbung** bzw. Steilheit, Exzeß oder Kurtosis der Verteilung ist durch

$$g_2 = \frac{m_4}{s^4} - 3,$$

gegeben, wobei $m_4 = \frac{1}{n}\sum_i (x_i - \bar{x})^4$ und s^4 die vierte Potenz der Standardabweichung ist. Dieses Maß hat für die Normalverteilung bzw. die Gaußsche Glockenkurve den Wert 0. Diese Verteilung hat große praktische und theoretische Bedeutung. Der Wert $g_2 < 0$ entspricht einer gegenüber der Gaußschen Glockenkurve **abgeplatteten** (platykurtischen, im Englischen auch *heavy tail*) Verteilung, ein $g_2 > 0$ einer verglichen mit der Gaußschen Glockenkurve **spitzen** (leptokutischen, im Englischen auch *light tail*) Verteilung.

Beispiel 2.8 Betrachten wir wieder die Variable "Körpergröße" für die weiblichen und männlichen Personen getrennt. Die wichtigsten Parameter der beiden Häufigkeitsverteilungen sind in der folgenden Tabelle zusammengefaßt:

```
MTB > desc 'GRO';
SUBC> by 'SEX'.

        SEX       N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
GRO       0      20   168.75   170.00   168.89     5.97     1.34
          1      30   182.13   184.50   182.23     7.82     1.43

        SEX     MIN      MAX       Q1       Q3
GRO       0  153.00   182.00   166.25   172.75
          1  168.00   195.00   176.00   187.25
```

Beachte! Die von MINITAB berechnete Varianz ist das $n/(n-1)$-fache der in *Definition 2.8* definierten Stichprobenvarianz s^2. Auch zur Berechnung der Quartile verwendet MINITAB eine modifizierte Formel.

Konzentrationsmaße

Ein weiteres Charakteristikum einer Verteilung wird durch **Konzentrationsmaße** beschrieben. Voraussetzung ist, daß das Merkmal nur nichtnegative Werte annimmt. Das Konzentrationsmaß gibt an, in welchem Ausmaß die Summe der beobachteten Merkmalsausprägungen auf die Untersuchungseinheiten verteilt ist: Ist der gesamte Einkommen in einer Population gleichmäßig auf alle Personen verteilt oder sind es einige wenige Personen, die über fast das ganze Einkomen verfügen. Das ist die klassische Problemstellung der Konzentrationsmessung. Andere Anwendungen betreffen die Verteilung des Vermögens auf die Angehörigen einer Population, die Verteilung der Beschäftigten auf die Betriebe, die Verteilung der Marktanteile, etc. Ein Konzentrationsmaß soll einen Wert nahe bei 0 haben, wenn keine Konzentration vorliegt; es soll einen Wert nahe bei 1 haben, wenn die Beobachtungen konzentriert sind.

Die Überlegungen, wie die Ungleichheit in der Einkommensverteilung gemessen werden kann, haben zum Konzept der Lorenzkurve geführt: Die nach der Größe sortierten Einkommen von n Personen seien $0 \leq x_1 \leq \ldots \leq x_n$. Die Lorenzkurve ist der Polygonzug, der die Punkte $(i/n, v_i)$, $i = 0, \ldots, n$, verbindet, wobei $v_0 = 0$ und $v_i = \sum_{j \leq i} x_j / \sum_{i=1}^{n} x_n$ der Anteil der i ärmsten Einkommensbezieher am gesamten Einkommen ist (siehe *Abbildung 2.5*). Von einem Konzentrationsmaß verlangen wir, daß es bei Gleichverteilung des Einkommens den Wert 0, bei hoher Konzentration einen Wert nahe bei 1 hat. Gleichverteilung des Einkommens bedeutet, daß für jedes p $100p\%$ der Personen auch $100p\%$ des Einkommens beziehen; in diesem Fall ist die

Lorenzkurve die Diagonale D von $(0,0)$ nach $(1,1)$ in *Abbildung 2.5*. Es ist einsichtig, daß alle möglichen Lorenzkurven im Bereich unterhalb der Diagonale D liegen müssen.

Als Konzentrationsmaß definiert man

$$\text{KM} = \frac{\text{Fläche zwischen } D \text{ und Lorenzkurve}}{\text{maximale Fläche zwischen } D \text{ und Lorenzkurve}}.$$

Gini hat als Konzentrationsmaß den Koeffizienten

$$G = \frac{\Delta_G}{2\overline{x}}$$

definiert. Er berücksichtigt (vergl. die Definition des Gini-Koeffizienten Δ_G in Abschnitt2.2.2) die Einkommensunterschiede zwischen allen Paaren von Einkommensbeziehern und ist normiert durch das Dividieren durch $\overline{x}$. Im Fall, daß alle Einkommen gleich sind, gilt $G = 0$. Es läßt sich zeigen, daß $G = 1$, wenn alle außer einer Person das gleiche Einkommen haben (der Fall größter Konzentration); in allen anderen Fällen gilt $0 < G < 1$. Gini konnte auch zeigen, daß $G = 2F$, wobei F die Fläche zwischen Diagonale D und Lorenzkurve ist. Somit kann G geometrisch interpretiert werden als Anteil der Fläche F an der größtmöglichen Fläche.

Zur praktischen Berechnung verwendet man die Formel

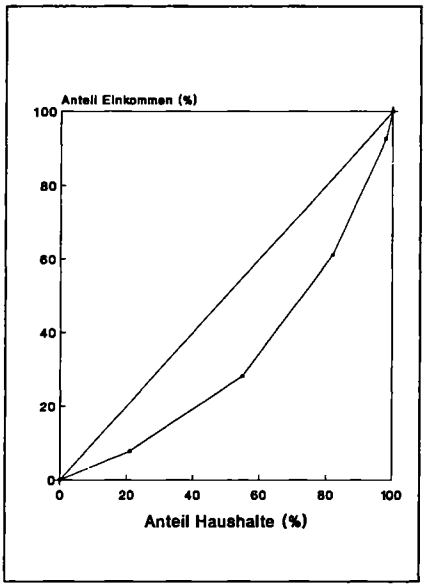
$$G = \frac{2\sum_{i=1}^{n} i x_i - (n+1)\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i}.$$

Beispiel 2.9 Die folgende Tabelle gibt die Einkommen der Haushalte einer Gemeinde im Jahr 1991 an.

Einkommen (in 1000.-)	Anteil der Haushalte in %	Σ	Anteil des Einkommens in %	Σ
unter 100	21.2	21.2	7.8	7.8
100–200	23.9	55.1	20.4	28.2
200–300	27.1	82.2	33.0	61.2
300–500	16.0	98.2	31.5	92.7
über 500	1.8	100.0	7.3	100.0

Die *Abbildung 2.5* zeigt die entsprechende Lorenzkurve. Da die Daten nur in gruppierter Form zur Verfügung stehen, ergibt sich ein Polygonzug, der den tatsächlichen Verlauf annähert und im Extremfall nur in den Knickpunkten mit der tatsächlichen Lorenzkurve übereinstimmt. Das KM der in *Abbildung 2.5* gezeigten Kurve beträgt 0.355 oder 35.5%.

Abbildung 2.5: Lorenzkurve für die Konzentration der Haushaltseinkommen, d.i. der Verlauf des kumulierten Anteils der Haushalte über dem kumulierten Anteil am Einkommen.



2.3 Weitere graphische Verfahren

In den vorangegangenen Abschnitten dieses Kapitels haben wir mehrere Möglichkeiten kennengelernt, mittels graphischer Darstellungen wesentliche Charakteristika einer Datenmenge wiederzugeben. Solche graphische Darstellungen sind sehr hilfreich für das Verständnis der Daten (explorativer Aspekt), aber auch beim Beschreiben der Daten (deskriptiver Aspekt), etwa zum Zweck der Kommunikation über Analyseergebnisse. Es läßt sich denken, daß die Möglichkeiten für graphische Darstellungen von statistischem Material nur durch die menschliche Phantasie beschränkt sind. In diesem Abschnitt wird eine besonders informative Form der Darstellung einer Datenmenge, das Box- oder Box & Whiskers-Plot, vorgestellt. Daneben wird eine Reihe weiterer graphischer Verfahren der explorativen und deskriptiven Statistik behandelt, die in der angewandten Statistik eine größere Verbreitung haben.

2.3.1 Andere Darstellungen der Häufigkeitsverteilung

Das Box-Plot

Das Box-Plot, auch Box & Whiskers-Plot genannt, ist eine Darstellung der wesentlichen Charakteristika einer Häufigkeitsverteilung mit einem sehr hohen Informationsgehalt. Das Box-Plot zeigt Lage- und Streuungsmaße sowie Symmetrie der Häufigkeitsverteilung an und visualisiert extreme Beobachtungen. Die *Abbildung 2.6* zeigt das Beispiel eines Box-Plots.

Definition 2.8 *Konstruktion eines* **Box-Plots**:

1. *Zeichne ein Rechteck ("box") für die mittleren 50% der Verteilung mit* $\tilde{x}_{0.25} = Q_u$ *als untere und* $\tilde{x}_{0.75} = Q_o$ *als obere Begrenzung; zeichne in der Höhe des Medians eine Mittellinie und für den Mittelwert ein "+" ein.*

2. *Bestimme die inneren Grenzen ("inner fences")* $Q_u - 1.5I$ *und* $Q_o + 1.5I$, *wobei* $I = Q_o - Q_u$ *der Interquartilsabstand ist; die Datenmenge zwischen den inneren Grenzen nennt man "the main body of the data".*

3. *Bestimme die äußeren Grenzen ("outer fences")* $Q_u - 3I$ *und* $Q_o + 3I$.

4. *Verbinde die Datenpunkte außerhalb der Box und innerhabl der inneren Grenzen durch zwei Gerade; die Bezeichnung Box & Whiskers-Plot kommt von der Ähnlichkeit dieser Geraden mit den Schnurrbarthaaren ("whiskers") von Katzen.*

5. *Trage die Beobachtungen zwischen den inneren und äußeren Grenzen als '+' ein; diese Beobachtungen heißen Ausreißer ("outliers").*

6. *Trage die Beobachtungen, die außerhalb der äußeren Grenzen liegen, als Punkte ein; diese Beobachtungen heißen extreme Ausreißer ("far outliers").*

Beispiel 2.10 Betrachten wir die beiden Box-Plots der Variablen "Körpergröße" für die weiblichen und männlichen Personen getrennt: die dazu notwendigen Größen kann man direkt aus den beiden stem and leaf plots entnehmen.

weiblich	männlich
$n_0 = 20$	$n_1 = 30$
$x_{(1)} = 153$	$x_{(1)} = 182$
$x_{(n)} = 182$	$x_{(n)} = 195$
$\tilde{x} = 170$	$\tilde{x} = 184.5$
$Q_u = 165.5$	$Q_u = 176$
$Q_o = 172.5$	$Q_o = 187$
$I = 6$	$I = 11$
innere Grenzen [156.5; 181.5]	innere Grenzen [159.5; 203.5]
äußere Grenzen [147.5; 190.5]	äußere Grenzen [143.0; 220.0]

Abbildung 2.6 zeigt die beiden Box-Plots und erlaubt einen graphischen Vergleich der beiden Populationen.

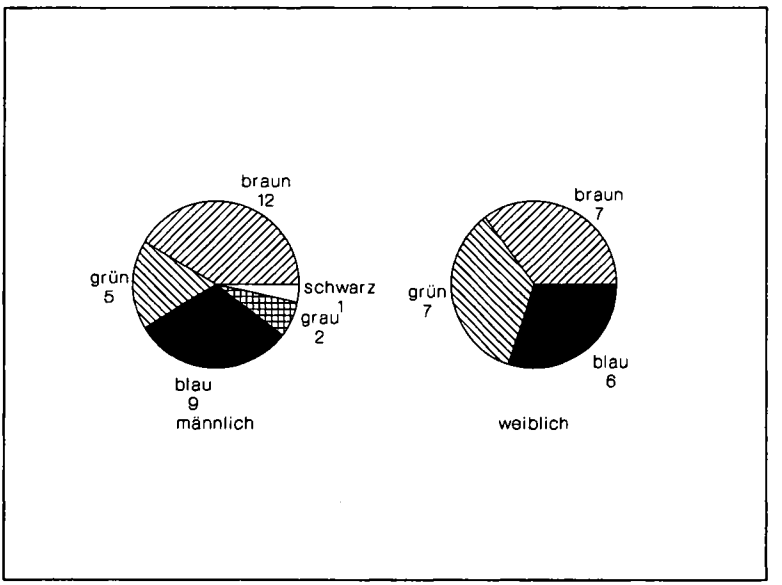
Abbildung 2.6: Box-Plots der Variablen "Körpergröße" für weibliche und männliche Personen.

```
MTB > boxplot 'GRO';
SUBC> by SEX.

SEX

                        ---------
0     *         --------I   +   I          *
                        ---------

                                    ---------------
1                         ----------I           +  I----------
                                    ---------------
  --+---------+---------+---------+---------+---------+----GRO
152.0     160.0     168.0     176.0     184.0     192.0
```

Das Kreisdiagramm

Eine oft benützte Darstellung der (relativen) Häufigkeiten der Ausprägungen von ordinalskalierten Merkmalen ist das **Kreisdiagramm**, im Englischen *pie chart* (Tortendiagramm) genannt (siehe *Abbildung 2.7*). Die Konstruktion erklärt sich von selbst.

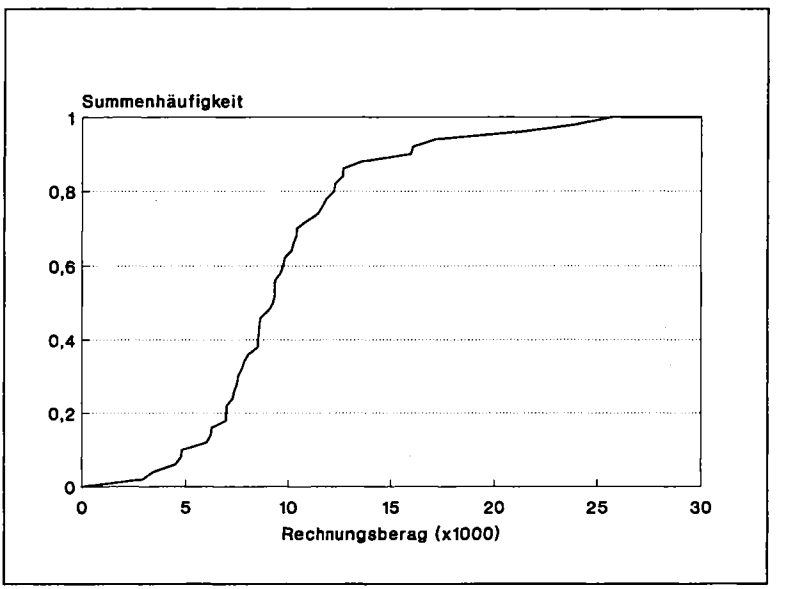
Abbildung 2.7: Kreisdiagramme.



Die Summenhäufigkeitskurve

Kumulative (relative) Häufigkeiten können über den entsprechenden Merkmalsausprägungen durch einen Polygonzug graphisch dargestellt werden (siehe *Abbildung 2.8*). Eine solche Darstellung nennt man **Summenhäufigkeitskurve** oder **empirische Verteilungsfunktion**, im Englischen *ogive*. Daraus – wie aus der entsprechenden Summenhäufigkeitsfunktion – können Anteile der Beobachtungen abgelesen werden, die weniger als ein bestimmter Wert sind. Aus *Abbildung 2.8* erkennt man, daß 60% (80%) der Rechnungsbeträge weniger als öS 9.750 (12.260) betragen. Zur Konstruktion ergänzt man die Tabelle der Häufigkeitsverteilung um die Spalte der kumulierten Häufigkeiten und zeichnet danach die Summenhäufigkeitskurve. Sie ist eine nichtfallende Kurve.

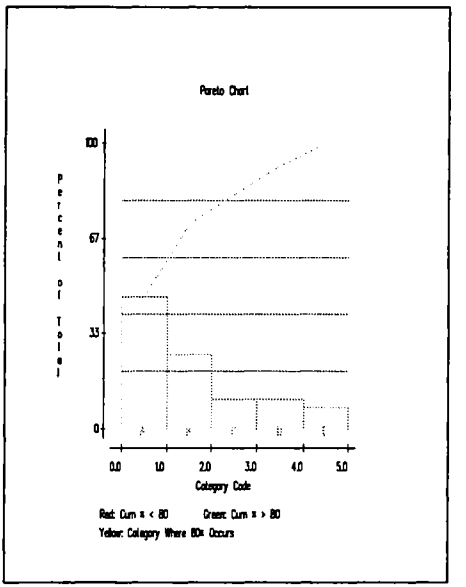
Abbildung 2.8: Summenhäufigkeitsfunktion.



Das Pareto Diagramm

Eine besondere Form von Häufigkeitsverteilung ist das **Pareto Diagramm**. Es zeigt die Häufigkeiten, mit denen verschiedene Typen von Defekten beobachtet wurden, wobei die Merkmalsausprägungen nach fallenden Häufigkeiten geordnet sind (siehe *Abbildung 2.9*). Das Pareto Diagramm ist ein effektvolles Instrument, die wichtigste Fehlerursache zu identifizieren, und ist eine wertvolle Methoden des modernen Qualitätsmanagements. In der *Abbildung 2.9* sind die Häufigkeiten angegeben, mit denen verschiedenen Ursachen (A,B, ..., E) von Materialdefekten beobachtet wurden. Man sieht, daß die häufigste Ursache (A) für mehr als 50% der Defekte verantwortlich ist.

Abbildung 2.9: Pareto Diagramm.



2.3.2 Bilderdiagramme

Unter dem Begriff Bilderdiagramme verstehen wir alle Darstellungen von Daten, die eine bildliche Auflösung des Sachverhaltes benützen, wie sie in Zeitungen und Zeitschriften alltäglich zu finden sind. Derartigen graphische Auflösungen sind einprägsam und transportieren, wenn sie gut gemacht sind, die zu vermittelnde Aussage besser als viele Worte. Es gibt keine Rezepte für das Anfertigen von derartigen Graphiken; es bleibt der Geschicklichkeit und Phantasie des Statistikers oder des ihm helfenden Graphikers überlassen, wie das darzustellende Datenmaterial am besten umgesetzt werden kann. Mehr als mit anderen Formen der graphischen Darstellung kann man mit Bilderdiagrammen irreführen.

2.3.3 Zeitreihendiagramme

Eine Zeitreihe ist eine Menge von Beobachtungen, die sich durch in der Zeit wiederholte Beobachtung derselben Variablen ergibt. Je nach Beobachtungsintervall unterscheiden wir monatliche, jährliche, tägliche, stündliche, etc. Daten. Ein Zeitreihendiagramm ist ein Polygonzug, der die Beobachtungen über der (horizontalen) Zeitachse darstellt. Im Kapitel 4 ber die Analyse von Zeitreihen machen ausgiebig von Zeitreihendiagrammen Gebrauch. Besser als die Zahlen laßen diese Diagramme den Verlauf und seine Charakteristika wie Trend und Saisonalität einer Zeitreihe erkennen.

Eine für die Prozeßkontrolle wichtige Klasse von Zeitreihendiagrammen sind die Shewhart Kontrollkarten, benannt nach dem Amerikaner W.A. Shewhart, der sie in den 20er Jahren vorgeschlagen hat. Die Shewhart Kontrollkarte zeigt den Polygonzug der Beobachtungen, sodaß man gut erkennet, wie sehr die Realisationen des Prozesses um seinen Mittelwert streuen. Die Variation eines solchen Prozesses ist typischerweise von der Prozeßumgebung bestimmt, der Mittelwert wird eingestellt. Neben dem Polygonzug der Beobachtungen enthält das Diagramm drei Linien: die Mittellinie, die den Mittelwert des kontrollierten Prozesses bestimmt, und zwei Kontrollgrenzen; diese Grenzen werden bei der Konstruktion der Kontrollkarte so gelegt, daß Beobachtungen außerhalb sehr unwahrscheinlich sind, solange der Prozeß stabil ist, d.h., wie geplant läuft. Wird eine Merkmalsausprägung beobachtet, die die obere Kontrollgrenze über- oder die untere Kontrollgrenze unterschreitet, so wird der Prozeß gestoppt und nach Ursachen für eine so extreme Beobachtung gesucht. Damit soll geklärt werden, ob der Prozeß den Zustand der Stabilität verlassen hat, oder ob der so extreme Wert zufällig realisiert wurde, obwohl der Prozeß stabil läuft (Fehlalarm). Die Shewhart Kontrollkarte und andere derartige Kontrollkarten sind ein mächtiges Instrument in der Prozeßkontrolle, einem wichtigen Anwendungsbereich der Statistik im technischen Bereich.

Der Datensatz "WU-Studenten"

In einer Erhebung unter ca. 1000 Studenten der Wirtschaftsuniversität Wien, die im Sommersemester 1992 die Proseminare aus Statistik besuchten, wurden Daten zu den folgenden zehn persönlichen Merkmalen erhoben. In der ersten Spalte der Tabelle ist das Kürzel angegeben, mit dem das Merkmal angesprochen wird.

Kürzel	Merkmal
SEX	Geschlecht (0: weiblich, 1: männlich)
AGE	Alter
STUD	Studienrichtung (1: BWL, 2: HW, 3: Sonstige)
NOTE	Note im Proseminar aus Mathematik I
GROESSE	Körpergröße (in cm)
GEWICHT	Gewicht (in kg)
SCHUH	Schuhgröße
FARBE	Augenfarbe

Das Ergebnis der Erhebung ist in der folgenden Tabelle enthalten.

Es soll an dieser Stelle auf zwei Punkte hingewiesen werden, die uns im Zusammenhang mit echten Daten stets beschäftigen werden und für die Qualität der Ergebnisse bedeutsam sind, aber in der statistischen Literatur und in den Publikationen statistischer Analysen kaum Beachtung finden.

- Der erste Schritt einer Datenanalyse sollte stets eine **Plausibilitätskontrolle** der erhobenen Beobachtungen sein. Als Methoden kommen dabei numerische Überprüfungen und graphische Darstellungen der Daten zur Anwendung. Als Illustration aus dem vorliegenden Datenmaterial können wir die Beobachtung 21 heranziehen: Eine Schuhgröße 74 ist wohl nur durch einen Datenfehler erklärlich. Tatsächlich handelt es sich, wie der Vergleich mit dem Originalbeleg zeigt, um einen Übertragungsfehler: Der dort angegebene Wert ist 47.

- Es kommt immer wieder vor, daß einzelne Werte nicht erhoben wurden und im Datensatz fehlen (im Englischen *missing observations*), sei es, weil der Befragte die Antwort nicht weiß, sie verweigert oder ein Fehler bei der Datenerfassung oder -übertragung passiert ist. Solche **fehlende Beobachtungen** müssen speziell gekennzeichnet werden, damit sie in der statistischen Analyse entsprechend berücksichtigt werden können. In der Tabelle unserer Daten wurden fehlende Beobachtungen der Variablen "Note" mit -9 kodiert. Für die Variable "Augenfarbe" fehlt die Beobachtung 32. Fehlen einzelne Daten einer Beobachtung, so ist in Abhängigkeit vom statistischen Analyseverfahren

	SEX	ALT	STR	NOT	GRO	GEW	SCH	AFA
1	0	21	1	−9	173	48	40	3
2	0	21	1	−9	173	70	40	3
3	1	20	2	−9	180	75	43	1
4	1	20	1	−9	187	80	44	2
5	1	19	1	5	177	68	42	1
6	1	22	1	3	192	82	44	2
7	1	22	1	4	183	88	43	1
8	1	28	3	4	184	85	44	3
9	1	24	1	−9	186	75	42	3
10	1	21	1	1	185	70	44	2
11	1	21	1	−9	169	80	43	1
12	1	23	1	−9	187	75	43	2
13	0	24	3	5	170	60	39	2
14	0	23	1	3	168	57	39	1
15	1	19	1	3	189	80	44	3
16	0	20	1	2	172	60	39	1
17	0	19	1	2	167	53	37	1
18	1	23	1	4	189	65	43	4
19	1	21	1	−9	169	66	40	3
20	1	25	1	−9	179	63	42	1
21	1	29	1	−9	187	67	44	5
22	1	23	1	−9	177	75	41	3
23	0	20	1	2	169	63	37	2
24	0	24	1	4	172	51	38	3
25	1	22	1	−9	170	54	39	3
26	0	22	1	−9	166	60	39	3
27	0	20	1	−9	163	52	38	2
28	0	20	1	−9	163	54	36	1
29	1	21	1	−9	172	75	42	3
30	1	24	1	−9	176	84	43	1
31	1	26	3	3	194	86	47	4
32	1	33	1	−9	176	76	40	−9
33	1	23	1	−9	191	82	45	3
34	1	27	1	2	186	83	44	1
35	1	21	1	1	186	78	43	1
36	1	22	1	2	183	74	44	1
37	0	21	1	1	170	56	38	2
38	1	24	1	3	188	81	44	3
39	0	24	1	1	173	63	39	1
40	1	23	1	−9	173	73	41	1
41	1	20	1	−9	195	83	48	1
42	0	25	1	3	182	68	40	2
43	1	21	1	−9	186	70	41	2
44	0	19	1	−9	170	63	39	2
45	0	28	1	−9	160	52	38	1
46	0	19	1	−9	153	52	37	3
47	0	20	1	−9	170	62	37	1
48	0	21	1	3	173	79	40	2
49	1	24	1	3	168	57	41	1
50	0	26	3	3	168	82	39	3