Statistische Methoden der VWL und BWL

Theorie und Praxis

5., aktualisierte Auflage

Josef Schira

EXTRAS ONLINE

ALWAYS LEARNING

PEARSON

Statistische Methoden der VWL und BWL

Statistische Methoden der VWL und BWL

Theorie und Praxis

5., aktualisierte Auflage

Josef Schira

ALWAYS LEARNING

PEARSON

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

10 9 8 7 6 5 4 3 2 1

20 19 18 17 16

ISBN 978-3-86894-299-6 (Buch)
ISBN 978-3-86326-789-6 (E-Book)

Programmleitung: Martin Milbradt, mmilbradt@pearson.de
Korrektorat: Dunja Reulein, München
Titelbild: © arekmalang, fotolia
Druck und Verarbeitung: DZS-Grafik, d.o.o., Lubljana

Printed in Slovenia

Meinem Lehrer Hans Schneeweiß

Inhaltsverzeichnis

Teil III Schließende Statistik 423

Vorwort

Statistik ist die Wissenschaft vom Sammeln, Aufbereiten, Darstellen, Analysieren und Interpretieren von Fakten und Zahlen. Staatliches Interesse an Informationen über demographische, soziale und ökonomische Sachverhalte war seit Jahrhunderten die Triebfeder für die Entwicklung der Statistik und gab ihr auch den Namen. Heute stimuliert die Notwendigkeit, große Mengen von verfügbaren Daten in vielen Anwendungsgebieten in nützliche Information zu verwandeln, die theoretische und praktische Weiterentwicklung der Statistik. In allen „empirischen Wissenschaften", wie der Medizin, Biologie, Geologie, Physik, Psychologie, Soziologie und der Wirtschaftswissenschaft, um nur die bekanntesten zu nennen, ist sie zu einer der wichtigsten Methoden der Erkenntnisgewinnung geworden. Zum Studium dieser Wissenschaften gehört deshalb auch eine intensive Beschäftigung mit Statistik.

Das vorliegende Buch wendet sich vor allem an Studierende der Volks- und Betriebswirtschaftslehre im Grund- und Hauptstudium. Es kann als Textbook oder als Begleitlektüre zu Vorlesungen und Seminaren in den herkömmlichen Diplomstudiengängen an Universitäten und Fachhochschulen und in den neuen Bachelor- und Masterstudiengängen dienen. Gleichzeitig ist das Buch als allgemeines Nachschlagewerk zu den grundlegenden statistischen Fragestellungen und Methoden angelegt und kann den Studenten während seines ganzen fachwissenschaftlichen Studiums und später in der Berufspraxis begleiten.

Der Leser benötigt die üblichen Grundkenntnisse der Elementarmathematik, wie sie in den Oberschulen und Gymnasien unterrichtet wird. Die darüber hinausgehenden Anforderungen beschränken sich auf das Rechnen mit dem Summenzeichen, die Lösung linearer Gleichungssysteme und auf etwas Matrizenrechnung. Die wenigen Elemente aus der höheren Mathematik, die zur Darstellung der statistischen Theorie hilfreich sind, werden an Ort und Stelle auf verständliche Weise erläutert. Statistische Sätze und Theoreme der Wahrscheinlichkeitsrechnung werden nur dann durch mathematische Beweise ergänzt, wenn diese kurz sind, mit einfachen Mitteln bewerkstelligt werden können und dadurch das Verständnis der statistischen Theorie befördert wird. Denn die statistische Grundausbildung hat auch das Ziel, auf das Studium der fortgeschrittenen Statistik und Ökonometrie vorzubereiten.

Der Aufbau des Buches ist einerseits in dem Sinne traditionell, als er dem heute an deutschen Hochschulen für volks- und betriebswirtschaftliche Studiengänge üblichen Standardprogramm folgt. Andererseits habe ich besondere Bemühungen darauf verwandt, Theorie und Praxis einander näherzubringen. Beinahe alle behandelten Sachverhalte sind wohlbegründet und bauen aufeinander auf, der Leser muss sich nicht mit einer bloßen Rezeptesammlung zufriedengeben. Seit auf jedem Schreibtisch ein PC steht und vielfältige statistische Software angeboten wird, kann auch der Laie jede aufwendige und

anspruchsvolle statistische Auswertung per Mausklick erledigen, aus jedem Datensatz statistische Kenngrößen ermitteln und Schätzungen und Prognosen anfertigen, ob sie nun sinnvoll sind oder nicht. Deshalb wird hier besonderer Wert darauf gelegt, den theoretischen Hintergrund der statistischen Methoden klar aufzuzeigen, um so die Urteils- und Kritikfähigkeit zu fördern. Die mathematische Notation erschwert das nicht, wie manchem auf den ersten Blick erscheinen mag, sie ist vielmehr so angelegt, dass sie das Verstehen erleichtert.

Außerdem wird das Buch durch eine große Anzahl von Graphiken und Tabellen anschaulich und übersichtlich gegliedert, wodurch das Verständnis und die Orientierung zusätzlich erleichtert werden.

Beispiele

Jedes Kapitel enthält zahlreiche Beispiele, die den Text beleben. Sie dienen einerseits der didaktischen Aufbereitung und sollen abstrakte Zusammenhänge anschaulich darstellen, andererseits zeigen viele von ihnen ganz konkrete wirtschafts- und sozialwissenschaftliche Anwendungen der vorgestellten statistischen Methoden.

Kontrollfragen

Am Ende eines Kapitels finden sich Kontrollfragen. Sie helfen dem Studenten, seinen Wissensstand zu überprüfen. Wenn nötig, kann zur Beantwortung einer Kontrollfrage der Text des Kapitels noch einmal herangezogen werden. Die Fragen sind so angelegt, dass durch ihre Beantwortung der Stoff wiederholt und die Struktur und die Inhalte des jeweiligen Kapitels verdeutlicht werden.

Theorie und Praxis

Zu jedem Kapitel wird eine typische Problemstellung aus der Praxis vorgestellt. Es handelt sich dabei um interessante Fragen aus der aktuellen wirtschaftspolitischen Diskussion und um Ansätze der neueren betriebswirtschaftlichen Forschung. Der Sachverhalt wird erläutert und es wird aufgezeigt, mit welchen statistischen Methoden eine Lösung herbeigeführt werden kann und welche Schlüsse aus dem Ergebnis zu ziehen sind. Diese Praxisanwendungen sind für Studenten besonders wichtig, da sie die oft empfundene Diskrepanz zwischen statistischer Methodenlehre und substanzwissenschaftlichen Fragestellungen zu überwinden helfen.

Ergänzende Literatur

Jedem Kapitel folgt eine kurze Liste ergänzender Literatur. Die Literaturangaben sind spezifisch, es handelt sich dabei um Monographien und Aufsätze zum Gegenstand des jeweiligen Kapitels. Außerdem werden allgemeine statistische und wahrscheinlichkeitstheoretische Werke und Lehrbücher genannt, die als klassisch gelten können oder in den bezeichneten Teilen detailliertere oder weitergehende Ausführungen zu den Methoden des Kapitels anbieten.

Aufgaben und Lösungen

Eine Zusammenstellung einer Vielzahl von sorgfältig ausgearbeiteten Aufgaben schließt jedes Kapitel ab. Dabei handelt es sich zum einen Teil um einfache, auf die Inhalte des jeweiligen Kapitels zugeschnittene Übungsaufgaben, zum anderen Teil um typische Klausuraufgaben, zu deren Lösung Kenntnisse aus verschiedenen Stoffgebieten benötigt werden. Die Beschäftigung mit den Aufgaben ist für das Studium und die Prüfungsvorbereitung besonders wichtig. Die meisten Aufgaben münden in die Berechnung von numerischen Ausdrücken, wobei die im Text entwickelten Formeln zu benutzen sind. Bei manchen Aufgaben ist die Berechnung leicht von Hand oder mit dem Taschenrechner möglich, in vielen Fällen sind PC-Programme vorteilhaft. Dazu bieten sich zwar spezielle Statistik-Programme an, besonders empfohlen wird aber, die Aufgaben mit Excel-Tabellen zu lösen: Der Rechenweg und die Formeln können dort explizit umgesetzt werden, womit ein besonderer Lerneffekt verbunden ist.

In den Lösungen sind, etwas abgesetzt, die numerischen Ergebnisse genannt, was dem Benutzer die Kontrolle seiner Arbeit ermöglicht.

Danksagung

An der Entstehung dieses Buches waren viele Personen beteiligt, und ich möchte vor allem denen danken, die als Erste den Text mit kritischen Kommentaren und hilfreichen Anregungen versehen haben, namentlich Dipl.-Psych. Horst Minkmar und Dipl.-Chem. Dr. Thomas Schauer. Mein Dank gilt insbesondere meinen früheren und jetzigen Assistenten, Dipl.-Math. Dr. Wilhelm Hennerkes, Prof. Dr. rer. pol. André Kuck, Dipl.-Ök. Elsbeth Kuck, Dipl.-Ök. Dr. Nicole van de Locht und Dipl.-Volkswirt Detlef Scholz, die mich alle bei der Ausarbeitung von Beispielen und Praxisaufgaben unterstützten.

Auch an der Überarbeitung des Buches haben viele mitgewirkt. Meine Studenten und viele aufmerksame Leser haben mit ihren Fragen und Antworten dazu beigetragen, die Verständlichkeit und Lesbarkeit des Textes weiter zu verbessern.

Schließlich mussten mich meine Kollegen, meine Familie und Freunde bei der Fertigstellung des Buches öfter entbehren, ihnen sei für ihre Nachsicht gedankt. Besonderer Dank gilt auch den Lektoren bei Pearson-Studium, für die verständnisvolle Zusammenarbeit und ihren tatkräftigen Einsatz bei der Verwirklichung dieses Buchprojektes und der Pflege der Neuauflagen.

Konstanz im April 2016
J.S.

Teil I

Beschreibende Statistik

KAPITEL 1

Statistische Merkmale und Variablen

Am Anfang jeder Gewinnung von statistischer Information steht die Erhebung einer großen Zahl von Einzeldaten. Die erste Aufgabe der Statistik ist es, diese zuweilen unübersichtliche Datenmenge so darzustellen und aufzubereiten, dass danach die in der Menge der Einzeldaten verborgene Information mit statistischen Methoden herausgefiltert und analysiert werden kann. In diesem Kapitel werden die fundamentalen Konzepte der Darstellung von statistischem Datenmaterial eingeführt und gezeigt, was sie leisten und wie man mit ihnen arbeitet. Zuvor sind einige technische Begriffe zu definieren und auch ein Blick auf die Objekte zu werfen, an denen die Daten erhoben wurden.

1.1 Statistische Einheiten und Grundgesamtheiten

Die Objekte, deren Merkmale in einer gegebenen Fragestellung von Interesse sind und im Rahmen einer empirischen Untersuchung erhoben, also beobachtet, erfragt oder gemessen werden sollen, heißen ***Untersuchungseinheiten*** oder ***statistische Einheiten***.

Als statistische Einheiten können grundsätzlich alle materiellen Gegenstände oder Lebewesen sowie immateriellen Dinge auftreten: Personen, Haushalte, Unternehmungen, Waren, Länder, Ereignisse, Handlungen usw.

Beispiel [1] Statistische Einheiten können sein: Kraftfahrzeuge, Gebäude, Pferde, Studenten, Beamte, Bauernhöfe, Branchen, Äpfel, Verkäufe, Eheschließungen, Geburten, Unfälle, Girokonten.

Die statistische Einheit ist **Träger der Information**, die erhoben werden soll. Das Hauptinteresse der Statistik gilt nicht der einzelnen statistischen Einheit. In diesem Sinne interessiert sie sich nur für Massenphänomene, also dafür, was in einer *statistischen Masse*, das heißt einer bestimmten Menge von im Wesentlichen *gleichartigen Einheiten* vor sich geht. Die Abgrenzung dieser Menge muss stets sehr sorgfältig erfolgen und der jeweiligen Fragestellung der statistischen Untersuchung entsprechen. Man könnte dazu die Elemente der Menge einzeln aufzählen. Meistens wird man jedoch nicht so verfahren,

sondern zur Identifikation der gleichartigen statistischen Einheiten, die zu einer solchen statistischen Menge gehören sollen, sogenannte ***Identifikationskriterien*** angeben. In der Regel werden die statistischen Einheiten durch mindestens jeweils ein Kriterium

1. zeitlicher,
2. räumlicher und
3. sachlicher Art

identifiziert oder definiert. Diese Kriterien sollten dabei möglichst objektiv und genau sein, das heißt, es sollte nicht von subjektiven Einschätzungen abhängen, ob ein bestimmter Gegenstand diese Kriterien erfüllt oder nicht. Mit Hilfe der Identifikationskriterien wird gleichzeitig die interessierende statistische Masse abgegrenzt.

Definition: Die Menge

$$\Omega := \{ \omega \mid \omega \text{ erfüllt } IK \} \tag{1-1}$$

aller statistischen Einheiten ω, die dieselben wohldefinierten Identifikationskriterien IK erfüllen, heißt ***Grundgesamtheit***.

Häufig verwendete Synonyme für den Terminus Grundgesamtheit sind ***statistische Masse***, ***Population*** und ***Kollektiv***.

Beispiele [2] Verkehrsunfälle im Jahre 2013 in Bayern.

[3] Verkehrsunfälle mit Personenschaden im Jahre 2015 in Deutschland.

[4] Studenten in der Vorlesung am Mittwoch, den 18.11.2015 um 14.15 Uhr, im Audimax der Universität Duisburg-Essen, Campus Duisburg.

[5] Angemeldete Konkurse von Bauunternehmungen im März 2012 in Nordrhein-Westfalen.

Eine Grundgesamtheit wird damit als eine ganz gewöhnliche Menge Ω im mengentheoretischen Sinne definiert. Die Elemente ω dieser Menge sind die statistischen Einheiten, die die Identifikationskriterien erfüllen: Es sind diese Kriterien, die die Grundgesamtheit bestimmen bzw. abgrenzen, indem sie ihre Elemente definieren.

Die Identifikation von statistischen Einheiten und die Abgrenzung von Grundgesamtheiten scheint im Prinzip einfach, kann aber in der Praxis durchaus schwierig sein. Sollen für eine bestimmte Erhebung Unternehmen, Betriebe oder Arbeitsstätten erfasst werden? Soll das Einkommen erhoben werden, das von Inländern oder im Inland erzielt wird?

Die Anzahl $n(\Omega)$ ihrer Elemente heißt der ***Umfang*** einer Grundgesamtheit Ω. In der Regel hat man es in der beschreibenden Statistik mit sogenannten ***realen*** Grundgesamtheiten (Bevölkerung eines Landes, Unternehmen eines Landes etc.) zu tun. Reale Grundgesamtheiten haben stets einen endlichen Umfang n. Demgegenüber stehen hypothetische oder ***fiktive*** Grundgesamtheiten, die durchaus unendlich viele Elemente haben können – wie zum Beispiel die Menge der Würfe, die man mit einem Würfel je machen kann. Mit

derartigen Grundgesamtheiten werden wir aber erst in späteren Kapiteln Bekanntschaft machen.

1.2 Merkmale und Merkmalsausprägungen

Das Interesse der Statistik gilt nicht den statistischen Einheiten ω selbst, sondern lediglich einigen ihrer Eigenschaften, den sogenannten ***Merkmalen*** $M(\omega)$. Deshalb bezeichnet man die statistischen Einheiten auch als die ***Merkmalsträger***. Unterscheidbare Erscheinungsformen eines Merkmals heißen ***Merkmalsausprägungen*** oder ***Modalitäten***.

Beispiele [6] Das Merkmal „Geschlecht" hat die beiden Modalitäten männlich und weiblich.

[7] Das Merkmal „Familienstand" hat die vier Merkmalsausprägungen: ledig, verheiratet, geschieden, verwitwet. Oder etwas moderner: verheiratet und single.

[8] Für das Merkmal „Körpergewicht" erwachsener Menschen müssen als Ausprägungen alle Werte zwischen 30 und 300 kg zugelassen werden.

Statistische Variable

Die Begriffe ***Merkmal*** und ***Variable*** werden häufig synonym verwendet, obwohl sie streng genommen nicht ganz dasselbe bedeuten. Statistische Variablen ordnen den statistischen Einheiten ω bzw. ihren Merkmalswerten $M(\omega)$ reelle Zahlen x zu. Somit ist die ***statistische Variable*** eine reellwertige Funktion X

$$x = X(\omega) = Fkt(M(\omega))$$

der Untersuchungseinheiten ω. Man bringt deshalb gerne statistische Variablen ins Spiel, weil man mit Zahlen besser arbeiten kann. Da nun sehr häufig die Merkmalsausprägungen bereits als reelle Zahlen vorliegen, kann das Merkmal selbst als Variable benutzt werden: Die Funktion *Fkt* ist dann die *identische Funktion*.

Mit dem Symbol X bezeichnet man die Abbildung bzw. Funktion

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega) = x \ , \end{aligned}$$

aber man benutzt es auch für den Namen der statistischen Variablen und meistens eben auch für den Namen des Merkmals selbst. Man sagt einfach: „*die statistische Variable X*" oder „*das Merkmal X*".

Merkmalstypen und Messbarkeitsniveaus

Merkmale und Variablen sind nicht alle von gleicher Qualität, was die Möglichkeiten ihrer statistischen Analyse und Interpretation angeht. Es ist deshalb angebracht, sie in verschiedene Kategorien einzuteilen. Man unterscheidet zunächst qualitative und quantitative Merkmale.

1. **Qualitative Merkmale** sind solche Eigenschaften, die qualitativ, das heißt der Beschaffenheit nach, artmäßig variieren. Sie besitzen nur endlich viele Ausprägungen. Beispiele sind Geschlecht, Religionszugehörigkeit und Rechtsform von Unternehmungen.

2. **Quantitative Merkmale** sind dagegen solche Eigenschaften von Untersuchungseinheiten, die quantitativ, das heißt der Größe nach oder zahlenmäßig, variieren. Ihre Merkmalsausprägungen sind von vornherein *Zahlen*, mit oder ohne Maßeinheit. Quantitativ sind Merkmale wie Alter, Kinderzahl, Einkommen.

Auch ursprünglich qualitative Merkmale werden oft in Zahlen ausgedrückt. Drückt man das Ausbildungsniveau einer Person durch die zu seiner Erreichung mindestens erforderliche Anzahl von Jahren an Ausbildungszeit aus, spricht man von **Quantifizierung** und hat damit eine echt quantitative Variable. Ordnet man aber etwa den Ausprägungen des Merkmals „Familienstand“ die Zahlen 1 für ledig, 2 für verheiratet und 3 für verwitwet zu, spricht man von *Signierung* und hat nur scheinbar quantitative Größen.

Die quantitativen Variablen werden in stetige und diskrete unterteilt:

1. **Diskrete Merkmale** können nur ganz bestimmte (endlich viele oder schlimmstenfalls abzählbar unendlich viele) abgestufte Werte als Merkmalsausprägung haben. Diskret sind alle Merkmale, deren Ausprägungen man durch Zählen erhält, auch wenn keine Obergrenze vorhanden ist.

2. **Stetige oder kontinuierliche Merkmale** können in einem Intervall jeden reellen Wert als Ausprägung annehmen (überabzählbar unendlich viele verschiedene mögliche Merkmalsausprägungen innerhalb eines Intervalls). Stetig sind alle Merkmale, deren Ausprägungen gemessen werden. Hierzu gehören beispielsweise alle Messungen in Zeit-, Längen- oder Gewichtseinheiten.

Besonders fein abgestufte diskrete Variablen werden in der statistischen Praxis wie stetige behandelt; man spricht von **quasi-stetigen** Merkmalen. Andererseits werden im Prinzip stetige Variablen durch den Mess- oder Erhebungsvorgang zu quasi-stetigen oder gar diskreten. Denn jede Messung kann aus technischen Gründen nur mit einer bestimmten Genauigkeit durchgeführt werden, so dass dadurch das ursprünglich stetige Intervall in **diskrete Größenklassen** aufgeteilt wird. Obwohl beispielsweise die Körpergröße ein stetiges Merkmal ist, wird es in der Praxis stets nur in Abstufungen erhoben. Eine Größe von 180 cm bedeutet, dass die Person zwischen 179.5 cm und 180.5 cm misst.

Eine andere sehr wichtige Einteilung der Typen von statistischen Variablen ist die nach dem Niveau der Messbarkeit, also danach, mit welcher **Skala** oder welchem **Maßstab** sie gemessen werden können. Das Niveau der Messbarkeit bestimmt dabei, wie wir noch sehen werden, die Möglichkeiten und Grenzen der statistischen Auswertungen, die man sinnvoll mit den erhobenen Daten vornehmen kann. In der Reihenfolge aufsteigender Messbarkeit unterscheiden wir:

1. **Nominal messbare Variablen.** Ein Merkmal oder eine Variable ist *nominal skaliert*, wenn lediglich die Gleichheit oder Andersartigkeit verschiedener Ausprägungen festgestellt werden kann. Beispiele für nominal skalierte Merkmale sind Religion, Nationalität, Beruf, Rechtsform eines Unternehmens. Ein Merkmal ist immer dann nominal, wenn mit ihm keinerlei Bewertung oder Quantifizierung intendiert werden soll. Nominale Merkmale sind stets qualitativ.

2. **Ordinal messbare Variablen.** Ein Merkmal oder eine Variable ist *ordinal skaliert*, wenn die möglichen Merkmalsausprägungen unterscheidbar sind und zusätzlich in eine natürliche oder sinnvoll festzulegende Rangordnung gebracht werden können. Als Beispiele wären hier Intelligenzquotient, sozialer Status, Schulnoten oder aber Tabellenplätze der Fußball-Bundesliga zu nennen.

3. **Kardinal messbare Variablen.** Schließlich spricht man von einem *kardinal* oder *metrisch skalierten* Merkmal, wenn die verschiedenen Ausprägungen nicht nur eine Rangfolge ausdrücken, sondern außerdem der quantitative Unterschied zwischen ihnen bestimmt ist. Die Ausprägungen müssen numerisch, das heißt in Zahlen, angegeben werden. Die meisten in den Wirtschaftswissenschaften interessierenden Merkmale wie zum Beispiel BIP, Investitionen und Inflation oder aber Kosten, Umsatz und Gewinn sind kardinal skaliert.

Man unterscheidet bei kardinal skalierten Merkmalen noch, ob ihr Maßstab einen sachlogisch begründeten absoluten Nullpunkt hat oder nicht. Ist ein solcher vorhanden, lassen sich sinnvoll Quotienten aus Merkmalsausprägungen bilden, und man spricht von einem **verhältnisskalierten Merkmal.** Zum Beispiel haben die Merkmale „Gewicht“, „Einkommen“ oder „Preis“ einen absoluten Nullpunkt, und man kann sagen, der Merkmalsträger ω_1 habe ein Einkommen, das doppelt so groß ist wie das von ω_2, wenn $X(\omega_1) = 2 \cdot X(\omega_2)$.

Hat die Skala hingegen keinen absoluten Nullpunkt, liegt ein **intervallskaliertes Merkmal** vor, und nur die Differenzen zwischen den Merkmalsausprägungen können sinnvoll interpretiert werden. Ein Beispiel für eine Intervallskala ist die Messung der Temperatur in Celsius-Graden. 40º warmes Wasser ist eben nicht „doppelt so warm“ wie Wasser mit 20ºC. Aber der Temperaturunterschied zwischen 50ºC und 60ºC und der zwischen 70ºC und 80ºC wird als gleich erachtet, denn man benötigt etwa die gleiche Energiemenge, um einen Temperaturanstieg um 10º zu erzeugen. Nur die Kelvin-Skala verfügt über einen absoluten Nullpunkt bei –273.15ºC = 0 K.

1.3 Teilgesamtheiten, Stichproben

Werden die Merkmalsausprägungen des interessierenden Merkmals aller statistischen Einheiten einer Grundgesamtheit festgestellt oder ***erhoben***, spricht man von einer ***Vollerhebung*** oder ***Totalerhebung.*** Technisch erfolgt eine Erhebung – je nach Merkmalsträger und untersuchtem Merkmal – meist in Form von

Beobachtungen,
Messungen
oder Befragungen.

Oftmals ist es jedoch unpraktisch oder zu teuer, eine Vollerhebung durchzuführen, z. B. *alle* Bürger der Bundesrepublik zu ihren täglichen Ausgaben für Brot zu befragen, die Körpergröße *aller* Bundesbürger zu messen oder die Zahl der Autos, die eine bestimmte Straße befahren, an *jedem* Tag zu beobachten. Dies wird besonders deutlich, wenn man bedenkt, dass allein die Vorbereitung einer Volkszählung oder der Arbeitsstättenzählung mehrere Jahre in Anspruch nimmt. Aus diesem Grund werden häufig nur Teilgesamtheiten oder Stichproben erhoben und untersucht.

Ist Ω^* eine Auswahl oder Teilmenge von der Grundgesamtheit Ω, so erfüllt jedes Element von Ω^* die Kriterien *IK*. Wenn Ω endlich ist, gilt $n(\Omega^*) \leq n(\Omega)$.

Definition: Jede echte Teilmenge Ω^* von Ω heißt ***Teilgesamtheit*** der Grundgesamtheit. Teilgesamtheiten heißen ***Stichproben***, wenn bei der Auswahl der Elemente der Zufall wesentlich beteiligt war.

Der Zweck einer Teilerhebung besteht meist darin, die interessierenden Merkmale nur von einer Teilgesamtheit erheben zu müssen, aber auf Basis dieser Ergebnisse Aussagen über die Merkmale in der Grundgesamtheit machen zu können.

Reine Zufallsstichprobe

Bei der reinen Zufallsauswahl soll jedes Element der Grundgesamtheit die gleiche „Chance“ haben, in die Stichprobe mit aufgenommen zu werden. Auf diesem Wege wird versucht, sicherzustellen, dass kein Merkmalsträger oder keine Gruppe von Merkmalsträgern bevorzugt ausgewählt und somit die Struktur der Grundgesamtheit systematisch verfälscht wird. Es scheint paradox, dass die *Zufälligkeit* der Auswahl durch eine sorgfältige Planung der Vorgehensweise bei der Bestimmung der Merkmalsträger sichergestellt werden muss.

Repräsentative Stichprobe

Wünschenswert wäre es, eine Teilgesamtheit auszuwählen, die *repräsentativ* für die Grundgesamtheit ist, also eine Struktur bezüglich der interessierenden Merkmale

aufweist, die der Grundgesamtheit möglichst ähnlich ist. Da man diese Struktur aber vor der Erhebung noch gar nicht kennen kann, versucht man, die Repräsentanz bezüglich *anderer bekannter* Merkmale zu gewährleisten. Denn man nimmt an, dass das zu untersuchende Merkmal in einem gewissen „statistischen Zusammenhang" mit diesen anderen Merkmalen steht. Es gibt verschiedene **Auswahlverfahren**, um zu erreichen, dass die gewonnene Teilgesamtheit repräsentativ ist. Man spricht von **eingeschränkter Zufallsauswahl**.

Beispiel [9] Ein Meinungsforschungsinstitut will eine Wahlprognose erstellen. Dazu wird 3000 Wahlberechtigten die sogenannte Sonntagsfrage gestellt: „Welche Partei würden Sie wählen, wenn am nächsten Sonntag Wahl wäre?" Um verlässlichere Ergebnisse zu bekommen, wird die Stichprobe repräsentativ gestaltet: Dazu überlegt man, welche anderen Merkmale die Parteienpräferenz „statistisch beeinflussen". In der Stichprobe soll der Anteil der Frauen dem in der Grundgesamtheit aller Wahlberechtigten entsprechen. Die Altersstruktur soll mit der der Grundgesamtheit übereinstimmen. Damit ist die Stichprobe für diesen Zweck schon recht repräsentativ. Wichtig wäre sicherlich noch, die geographische Verteilung zu berücksichtigen, damit es nicht vorkommen kann, dass zu viele Befragte zufällig in Baden-Württemberg wohnen. Weiterhin wäre es gut, wenn die Berufsstruktur, wenigstens in den Ausprägungen Arbeiter, Angestellte, Beamte, Selbständige, analog wäre. Ja, und natürlich müssen Studenten in der Stichprobe sein, sonst wären die Wähler der Grünen eventuell „unterrepräsentiert".

1.4 Statistische Verteilung

Eine Grundgesamtheit, Teilgesamtheit oder Stichprobe vom Umfang n und mit den Elementen ω_i sei bezüglich eines Merkmals X untersucht worden. Von jedem Element ω_i sei sein „individueller" Merkmalswert x_i festgestellt und in der ***Urliste*** notiert worden:

Urliste						
Elemente	ω_1	ω_2	$\cdots$	ω_i	$\cdots$	ω_n
Merkmalswerte	x_1	x_2	$\cdots$	x_i	$\cdots$	x_n

Das Hauptinteresse der beschreibenden Statistik gilt aber nicht den Merkmalsträgern, sondern den Merkmalswerten.

Definition: Die Folge der n Werte

$$x_1,\ x_2,\ \cdots,\ x_i,\ \cdots,\ x_n \tag{1-2}$$

mit $x_i = X(\omega_i)$, für $i = 1, \cdots, n$, heißt ***Beobachtungsreihe der Variablen*** X oder einfach ***statistische Reihe*** X.

Spielt dabei die Reihenfolge, in der die Beobachtungen gemacht wurden, keine Rolle, ist auch die Anordnung der Werte in der statistischen Reihe ohne Bedeutung und sie könnten beliebig umgestellt werden. Die Nummerierung (Indizierung) dient nur der Unterscheidung der einzelnen Werte; eine Umnummerierung wäre zulässig und würde den Informationsgehalt der statistischen Reihe nicht verändern. Nur bei den sogenannten ***Zeitreihen*** ist das anders, diese werden aber erst in Kapitel **5** behandelt.

Häufig ist es sinnvoll, die Merkmalswerte der Urliste der Größe nach zu sortieren und umzunummerieren, so dass dann

$$x_1 \le x_2 \le x_3 \le \cdots \le x_i \le \cdots \le x_n \tag{1-3}$$

geschrieben werden kann. In der Praxis wird es oft vorkommen, dass in dieser Abfolge gleich große Werte nebeneinanderstehen, weil einzelne Ausprägungen in der statistischen Reihe mehrfach auftauchen, beispielsweise

$$\begin{array}{cccccccccccccc} 1.6 & 1.6 & 3.0 & 3.0 & 3.0 & 3.0 & 4.1 & 4.1 & 4.1 & 4.1 & 4.1 & 4.1 & 4.1 \\ 4.1 & 5.0 & 5.0 & 5.0 & 5.0 & 5.0 & 5.0\,, & & & & & & & \end{array} \tag{1-4}$$

weshalb in (1-3) ja die ≤-Zeichen stehen. Dann ordnet man die k vorkommenden, aber unterschiedlichen Variablenwerte der Größe nach zu

$$x_1 < x_2 < \cdots < x_k\,, \quad \text{mit} \quad k \le n$$

und gibt zu jedem Variablenwert x_i die ***absolute Häufigkeit***

$$n_i \ := \ \text{absH}(X = x_i) \tag{1-5}$$

an, das heißt, man gibt an, wie oft die statistische Variable X den Wert x_i in der statistischen Reihe X annimmt. Man beachte, dass k, die Anzahl der vorkommenden Merkmalsausprägungen, nicht größer als n sein kann, in der Praxis aber meist viel kleiner ist. Auf diese Weise erhalten wir eine Tabelle, die den vorkommenden Variablenwerten die zugehörigen Häufigkeiten zuordnet. Diese kann noch übersichtlicher werden, wenn statt der absoluten die ***relativen Häufigkeiten***

$$h_i \ := \ \text{relH}(X = x_i) \ = \ n_i / n\,, \quad 0 < h_i \le 1 \tag{1-6}$$

verwendet werden.

Definition: Die Tabellen

x_1	x_2	$\cdots$	x_k
n_1	n_2	$\cdots$	n_k

$$\sum n_i = n$$

und

x_1	x_2	$\cdots$	x_k
h_1	h_2	$\cdots$	h_k

$$\sum h_i = 1 \qquad (1\text{-}7)$$

heißen absolute bzw. relative ***Häufigkeitsverteilung*** der statistischen Variablen X.

Häufigkeitsverteilungen lassen sich auf sehr einfache Weise anschaulich graphisch darstellen. Man braucht nur die Häufigkeiten als Ordinate über der statistischen Variablen als Abszisse in ein Koordinatensystem einzuzeichnen. Zur Erhöhung der Anschaulichkeit verbindet man die Punkte durch senkrechte Linien mit der Abszisse: Die Längen der einzelnen Linien sind somit proportional zu den Häufigkeiten.

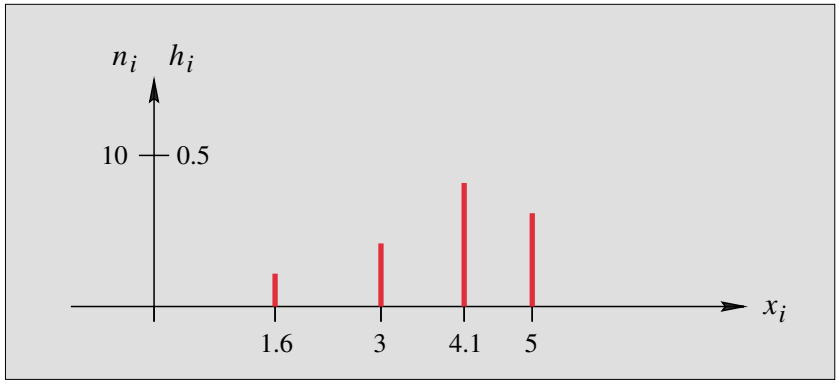


BILD 1.1 Häufigkeitsverteilung

1.5 Häufigkeitsfunktion und Verteilungsfunktion

Der einfachste Weg, zur Häufigkeitsfunktion zu gelangen, ist, ausgehend von der relativen Häufigkeitsverteilung (1-7), alle reellen Zahlen x, die nicht in der statistischen Reihe X vorkommen, mit aufzunehmen, ihnen aber die relative Häufigkeit Null zuzuweisen.

Definition: Die Funktion

$$h(x) = \begin{cases} h_i & \text{falls } x = x_i \\ 0 & \text{sonst} \end{cases} \tag{1-8}$$

heißt ***Häufigkeitsfunktion*** der statistischen Variablen X.

Diese Funktion gibt für jede reelle Zahl und damit auch für jeden möglichen Variablenwert x an, ob und mit welcher relativen Häufigkeit er in der statistischen Reihe vorkommt. Der Definitionsbereich der Häufigkeitsfunktion ist somit die ganze reelle Achse, während der Wertebereich der Funktion sich auf die rationalen Zahlen im Intervall [0,1] beschränkt. Ihre graphische Darstellung entspricht derjenigen der Häufigkeitsverteilung.

Definition: Die Funktion

$$H(x) = \sum_{x_i \le x} h(x_i) \tag{1-9}$$

heißt ***empirische Verteilungsfunktion*** der statistischen Variablen X.

Die empirische Verteilungsfunktion gibt für jedes $x \in \mathbb{R}$ die relative Häufigkeit aller Beobachtungen an, die gleich groß oder kleiner als x sind. Ihre Definitions- und Wertebereiche sind identisch mit denen der Häufigkeitsfunktion.

Der Graph von $H(x)$ hat die typische Gestalt einer ***Treppenfunktion***. Die ***Sprungstellen*** finden sich an den x-Werten mit positiver relativer Häufigkeit; an diesen Stellen springt der Funktionswert um den Betrag der relativen Häufigkeit h_i bzw. um den Wert der Häufigkeitsfunktion $h(x_i)$ nach oben. Zwischen zwei benachbarten Sprungstellen verharrt die Funktion auf konstantem Niveau.

Beispiel [10] Die Häufigkeitsfunktion $h(x)$ und die Verteilungsfunktion $H(x)$ zur statistischen Reihe (1-4) bzw. zur Verteilung

x_i	1.6	3.0	4.1	5.0
h_i	0.1	0.2	0.4	0.3

sind in BILD 1.2 dargestellt.

Es ist darauf zu achten, dass die Funktion $H(x)$ stets auf der *ganzen reellen* Achse $-\infty < x < +\infty$ erklärt ist. Sie hat im Beispiel [10] für $-\infty < x < 1.6$ den Wert $H(x) = 0$ und für $5 \le x < \infty$ den Wert $H(x) = 1$. An den Sprungstellen selbst hat die Verteilungsfunktion grundsätzlich den oberen Wert. Die empirische Verteilungsfunktion in der Definition (1-9) hat die folgenden **Eigenschaften**:

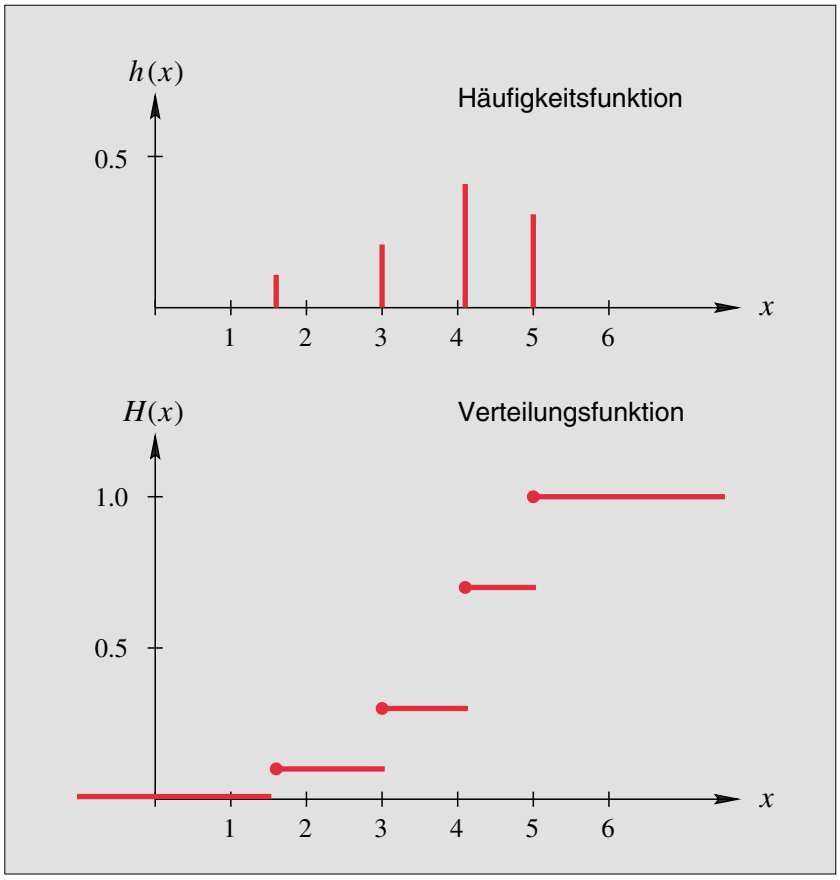


BILD 1.2 Häufigkeitsfunktion und Verteilungsfunktion

1. Die Funktion $H(x)$ ist *überall wenigstens rechtsseitig stetig*, das heißt es gilt für jedes $x \in \mathbb{R}$ (mit $\Delta x > 0$)

$$\lim_{\Delta x \to 0} H(x + \Delta x) = H(x) \ . \tag{1-10}$$

An den Sprungstellen ist sie jedoch *nur* rechtsseitig stetig; dort gilt

$$\lim_{\Delta x \to 0} H(x - \Delta x) \neq H(x) \ . \tag{1-11}$$

2. Die Funktion H ist *monoton steigend*, das heißt für jedes a und $b \in \mathbb{R}$ gilt

$$H(a) \leq H(b) \ , \quad \text{falls} \quad a < b \ . \tag{1-12}$$

3. Der *untere Grenzwert* der Verteilungsfunktion ist Null, der *obere Grenzwert* ist Eins, das heißt

$$\lim_{x\to-\infty} H(x) = 0\,, \qquad \lim_{x\to\infty} H(x) = 1\,. \tag{1-13}$$

Weiter ist anzumerken:

1. Die *Differenz*

$$H(b)-H(a) = \mathrm{relH}(a<X\le b) \tag{1-14}$$

gibt für $a < b$ die relative Häufigkeit der Beobachtungswerte der Variablen X an, die größer als a, aber nicht größer als b sind.

2. Der *Funktionswert* an jeder Stelle x gibt die relative Häufigkeit an, mit welcher Werte, die *kleiner oder gleich x* sind, in der statistischen Reihe vorkommen:

$$H(x) = \mathrm{relH}(X\le x) \tag{1-15}$$

3. An jeder Stelle $x\in\mathbb{R}$ erhält man aus der empirischen Verteilungsfunktion *die Werte der Häufigkeitsfunktion* als Differenz

$$h(x) = H(x) - \lim_{\Delta x\to 0} H(x-\Delta x) \tag{1-16}$$

zwischen dem Funktionswert und dem linksseitigen Grenzwert.

Wir beachten, dass mit der Formel (1-16) *nur an den Sprungstellen* der Verteilungsfunktion positive Differenzen herauskommen können: An allen anderen Stellen von H ist der linksseitige Grenzwert gleich dem Funktionswert, so dass die Häufigkeitsfunktion Null bleibt.

Die hier definierte empirische Verteilungsfunktion H mag aus der Sicht der beschreibenden Statistik wenig Anschaulichkeit besitzen und es scheint auch, dass man eigentlich nicht sehr viel damit anfangen kann, jedenfalls nicht viel mehr als mit der anschaulicheren Häufigkeitsfunktion h selbst. Aber die für die Anwendung sehr wichtigen Instrumente *Histogramm* und *Häufigkeitsdichte*, die im nächsten Abschnitt eingeführt werden, lassen sich am besten auf der Grundlage der Verteilungsfunktion verstehen.

Darüber hinaus dient die Beschäftigung mit H nicht zuletzt der didaktischen Hinführung zu ihrem Analogon, der *stochastischen* Verteilungsfunktion F, die in Kapitel **9** eingeführt werden wird. Diese betrifft nicht statistische Variablen, sondern sogenannte *stochastische Variablen*. Das sind Variablen, deren Werte nicht aus Beobachtungen stammen, sondern *vom Zufall abhängig* sind.

1.6 Häufigkeitsdichte und Histogramm

In der Praxis kommt es häufig vor, dass große Gesamtheiten mit einer Vielzahl verschiedener Merkmalsausprägungen untersucht werden müssen. Aus messtechnischen Gründen, aber auch aus erhebungs- oder aufbereitungstechnischen Gründen kann dabei selbst bei stetigen oder quasi-stetigen Merkmalen und vielen Einzelbeobachtungen oft nur eine endliche und verhältnismäßig kleine Zahl unterschiedlicher Merkmalsausprägungen Berücksichtigung finden, so dass für eine Variable X ***Größenklassen*** oder ***Schichten*** gebildet werden müssen. Dazu wird das von möglichen Merkmalsausprägungen belegte reelle Intervall durch geeignet gewählte ***Klassengrenzen***

$$\xi_0,\ \xi_1,\ \xi_2,\ \cdots,\ \xi_m$$

in m Abschnitte unterteilt, wie in BILD 1.3 dargestellt.

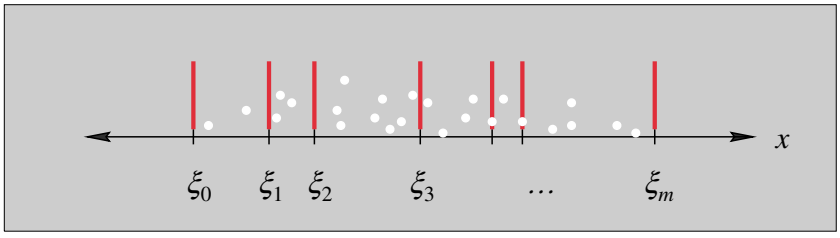


BILD 1.3 Bildung von Größenklassen

Diese m Abschnitte haben die ***Klassenbreiten***

$$\Delta_i := \xi_i - \xi_{i-1}\,, \qquad i = 1, \cdots, m \tag{1-17}$$

und die relative Häufigkeit der Werte in jeder Größenklasse sei mit

$$h_i := \mathrm{relH}(\xi_{i-1} < X \le \xi_i)\,, \qquad i = 1, \cdots, m \tag{1-18}$$

angegeben. Die weißen Punkte in BILD 1.3 sollen Beobachtungswerte darstellen, die in die einzelnen Größenklassen fallen. Fällt ein Wert genau auf die Klassengrenze, so ist er der kleineren Größenklasse zuzuordnen. Ordnet man nun diese ***Klassenhäufigkeiten*** den Klassenobergrenzen zu (eine alternative Möglichkeit wäre, die Klassenhäufigkeiten den Klassenmitten zuzuordnen), so kann aus den Werten der folgenden Häufigkeitstabelle

ξ_1	ξ_2	$\cdots$	ξ_m
h_1	h_2	$\cdots$	h_m

$$\sum h_i = 1 \tag{1-19}$$

die ***Verteilungsfunktion der Klassen*** $H_K(x)$ gezeichnet werden.

Durch diese Erhebungs- bzw. Aufbereitungstechnik ist natürlich die Information der Häufigkeitsverteilung innerhalb der Klassen verloren gegangen bzw. gar nicht erst erhoben worden. Es bieten sich zwei Möglichkeiten an, die verlorene Information annäherungsweise zu ersetzen, um die „wahre" Verteilungsfunktion $H(x)$ wenigstens ungefähr zu bestimmen.

Approximierender Polygonzug

Im oberen Teil von BILD 1.4 verbinden wir die Funktionswerte von H_K an den Sprungstellen durch gerade Linien und erhalten so eine approximierende Verteilungsfunktion $\overline{H}(x)$ als Polygonzug. Die Sprungstellen von H_K werden zu Knickstellen von $\overline{H}$, an denen sich die Steigung von $\overline{H}$ abrupt ändert, während sie dazwischen konstant ist und

$$\frac{H_K(\xi_i) - H_K(\xi_{i-1})}{\xi_i - \xi_{i-1}} = \frac{h_i}{\Delta_i} \quad , \qquad i = 1, \cdots, m$$

beträgt. Diese Vorgehensweise zur Gewinnung einer Approximation unterstellt eine „gleichmäßige Verteilung" innerhalb jeder einzelnen Größenklasse.

Definition: Ist $H_K(x)$ die Verteilungsfunktion eines nach Größenklassen erhobenen Merkmals mit den Klassenobergrenzen $\xi_1, \xi_2, \cdots, \xi_m$ und $\overline{H}(x)$ die durch einen Polygonzug approximierte Verteilungsfunktion, so heißt der Quotient

$$\boxed{\frac{H_K(\xi_i) - H_K(\xi_{i-1})}{\xi_i - \xi_{i-1}} = \frac{h_i}{\Delta_i}} \tag{1-20}$$

die (durchschnittliche) ***Häufigkeitsdichte*** der i-ten Größenklasse ($i = 1, \cdots, m$). Die erste Ableitung

$$\boxed{\overline{h}(x) := \frac{\mathrm{d}\,\overline{H}(x)}{\mathrm{d}\,x}} \tag{1-21}$$

in den Intervallen $\xi_{i-1} < x < \xi_i$ heißt ***Häufigkeitsdichtefunktion*** und ihr Graph ***Histogramm***.

Diese gleichmäßige Verteilung der Merkmalsausprägungen innerhalb einer jeden Größenklasse wird in den meisten Fällen zwar nicht mit der Realität übereinstimmen, gleichwohl stellt das Histogramm eine gute Visualisierung der Verteilung H_K dar. Nur wenn die Besetzungszahlen einzelner Größenklassen allzu gering sind, kann durch das Histogramm ein falscher Eindruck vermittelt werden.

Wie im Bild angedeutet, müssen die einzelnen „Säulen" des Histogramms, die jeweils eine Größenklasse repräsentieren, durchaus nicht die gleiche Breite Δ_i haben. Im

Gegensatz zum Graphen der Häufigkeitsfunktion gibt *nicht die Höhe der Säule, sondern die Fläche*

$$\frac{h_i}{\Delta_i} \cdot \Delta_i$$

die relative Häufigkeit in der Größenklasse an.

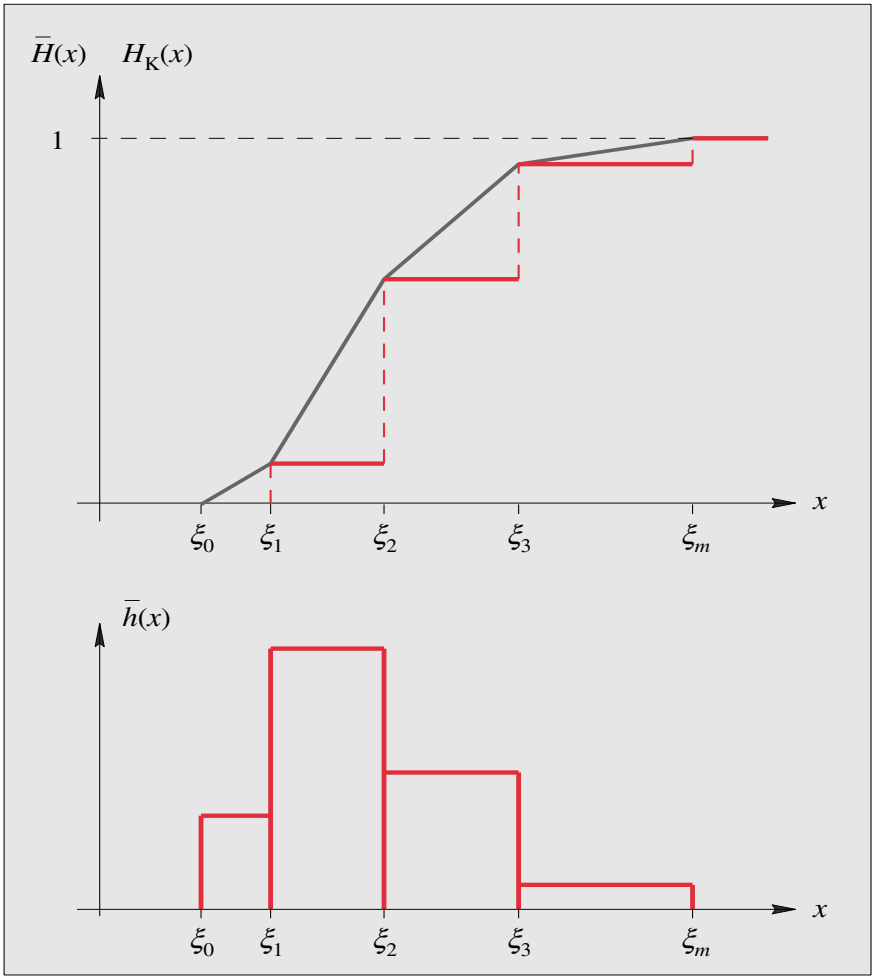


BILD 1.4 Approximierender Polygonzug und Histogramm

Die Gesamtfläche der Säulen des Histogramms ergibt somit

$$\sum_{j=1}^{m} \Delta_j \frac{h_j}{\Delta_j} = 1 .$$

Beispiel [11] Im untenstehenden Histogramm sind alle Klassenbreiten mit $\Delta_i = 10\,000$ Euro gleich. Nur die unterste und die oberste Einkommensklasse haben eine andere Breite. Deshalb entspricht bei den anderen nicht nur die Fläche sondern auch die Höhe der Säulen den Klassenhäufigkeiten, die hier in Prozent angegeben sind

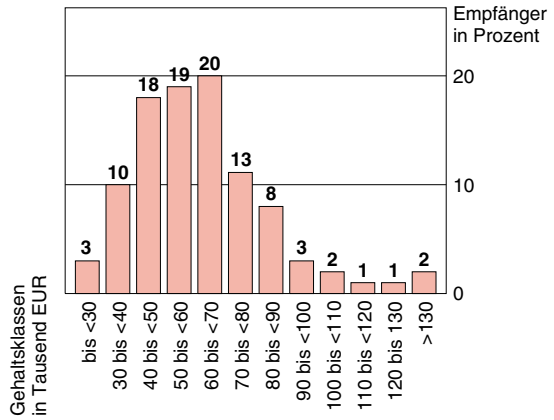


BILD 1.5 Verteilung der jährlichen Gesamtbezüge von Führungs- und Fachkräften des Außendienstes

Man beachte, dass die Approximation nur bei stetigen (oder quasi-stetigen) Merkmalen sinnvoll sein kann. Außerdem verlassen wir dadurch eigentlich den gesicherten Boden der auf Beobachtungen gründenden beschreibenden Statistik. Zwar geben wir nicht an, wie eine Verteilungsfunktion aussehen müsste, wenn in feinerer Klasseneinteilung oder ohne eine solche erhoben worden wäre, sondern es soll nur eine Annäherung an die „wahren" Verhältnisse sein. Dabei können wir uns irren, und wir wissen zunächst auch gar nicht, wie groß die Fehler sein mögen. Wir wissen auch nichts über die Fehlerwahrscheinlichkeiten. Die Unterstellung, dass die Häufigkeitsdichte über die ganze Klassenbreite hinweg gleich groß ist, erscheint in Ermangelung besserer Information sinnvoll, bedeutet aber gleichzeitig, dass sie sich an den willkürlich gewählten Klassengrenzen abrupt ändert. Dieses ist aber eher unrealistisch.

Beispiel [12] **Bevölkerungspyramiden sind Histogramme.** Die senkrechte Achse ist hier die Achse der Merkmalswerte. Die Bevölkerungspyramiden für Deutschland, Frankreich, Italien und Ungarn, aber auch die für die USA zeigen alle den für moderne Gesellschaften typischen „Bauch". Die hier und auf der folgenden Seite dargestellten Graphiken demonstrieren, dass der Begriff „Pyramide" die Form des Histogramms der Altersverteilung auch für China und Brasilien nicht mehr adäquat beschreibt. Nur die Altersstruktur in Entwicklungsländern mit hohem Bevölkerungswachstum, wie z. B. Indien, erzeugt noch das früher für die meisten Länder typische pyramidenförmige

Histogramm. Interessant ist in diesem Zusammenhang, dass sich die Auswirkungen einer Änderung des generativen Verhaltens der Bevölkerungen zuerst in Deutschland und Frankreich, dann in Ungarn und den USA, relativ spät in Italien und China und erst jüngst in Indien bemerkbar machten.

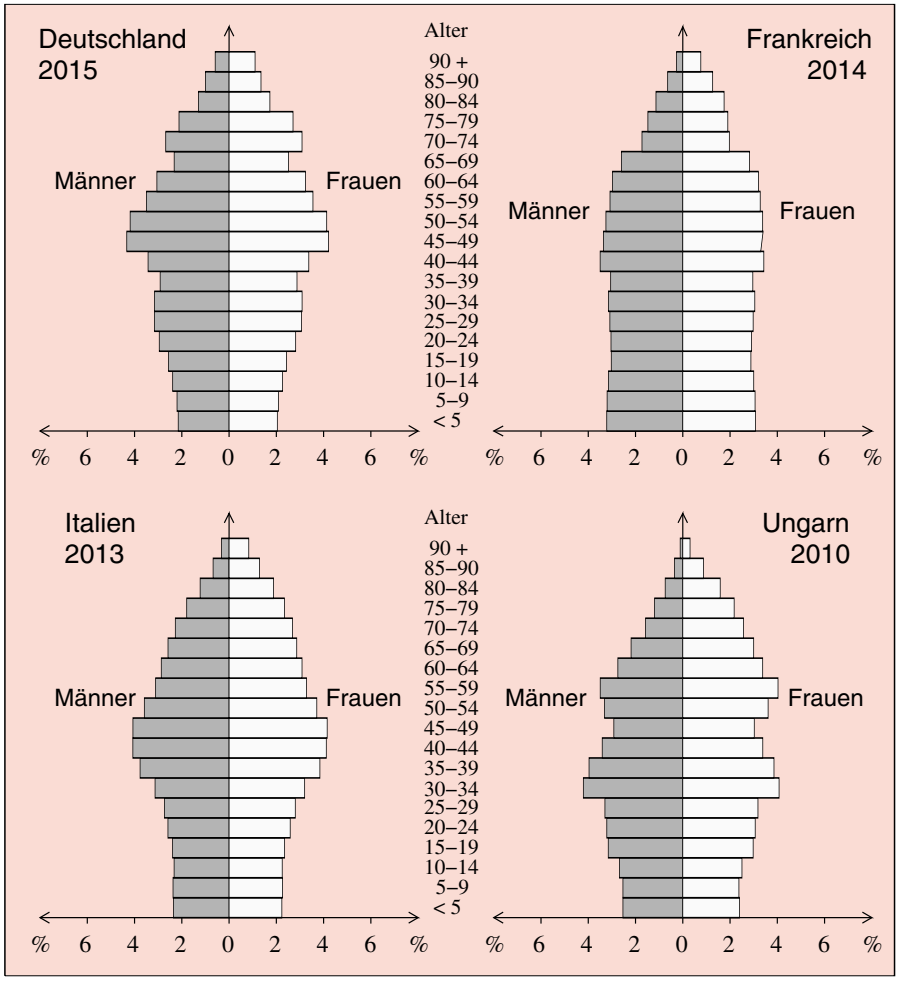


BILD 1.6 Bevölkerungspyramiden alter Länder: Europa

Die Ursachen für diese Änderungen können dabei recht unterschiedlicher Natur sein, und es lassen sich Vermutungen über die Auswirkungen des 2. Weltkriegs in Deutschland und Frankreich, der 68er-Bewegung („Pillenknick“) in Deutschland, Frankreich, Italien und den USA, des sowjetischen Einmarschs in Ungarn 1956, der Kulturrevolution und der späteren 1-Kind-Politik in China anstellen.

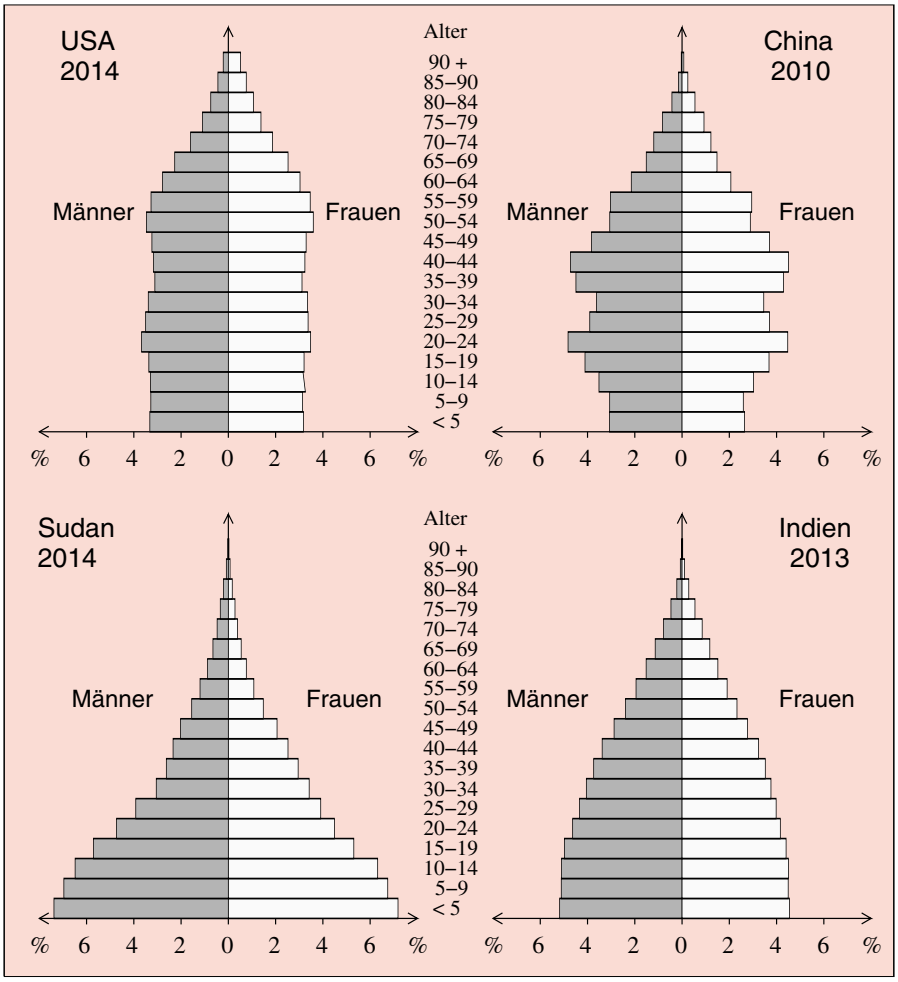


BILD 1.7 Bevölkerungspyramiden anderer Länder

Approximierende glatte Kurve

Verbindet man hingegen die Funktionswerte $H_{\mathrm{K}}(x_i)$ durch eine glatte Kurve ohne Knickstellen, so gibt man dadurch die Annahme der gleichmäßigen Verteilung innerhalb der einzelnen Größenklassen auf. Meistens ist diese Annahme auch nicht realistisch, denn sie bedeutet, dass sich die Häufigkeitsdichte an den oft willkürlich gewählten Grenzen der Größenklassen abrupt ändert. Wählt man deshalb als approximierende Verteilungsfunktion eine stetige und differenzierbare Funktion $\widetilde{H}(x)$, hat die Dichtefunktion $\widetilde{h}(x) := \mathrm{d}\,\widetilde{H}(x)/\mathrm{d}\,x$ auch keine Sprungstellen, und es gilt

$$\int_{-\infty}^{x} \tilde{h}(u)\,\mathrm{d}u = \tilde{H}(x)$$

und

$$\int_{-\infty}^{+\infty} \tilde{h}(x)\,\mathrm{d}x = \int_{\xi_0}^{\xi_m} \tilde{h}(x)\,\mathrm{d}x = \tilde{H}(\xi_m) = H(\xi_m) = 1\,.$$

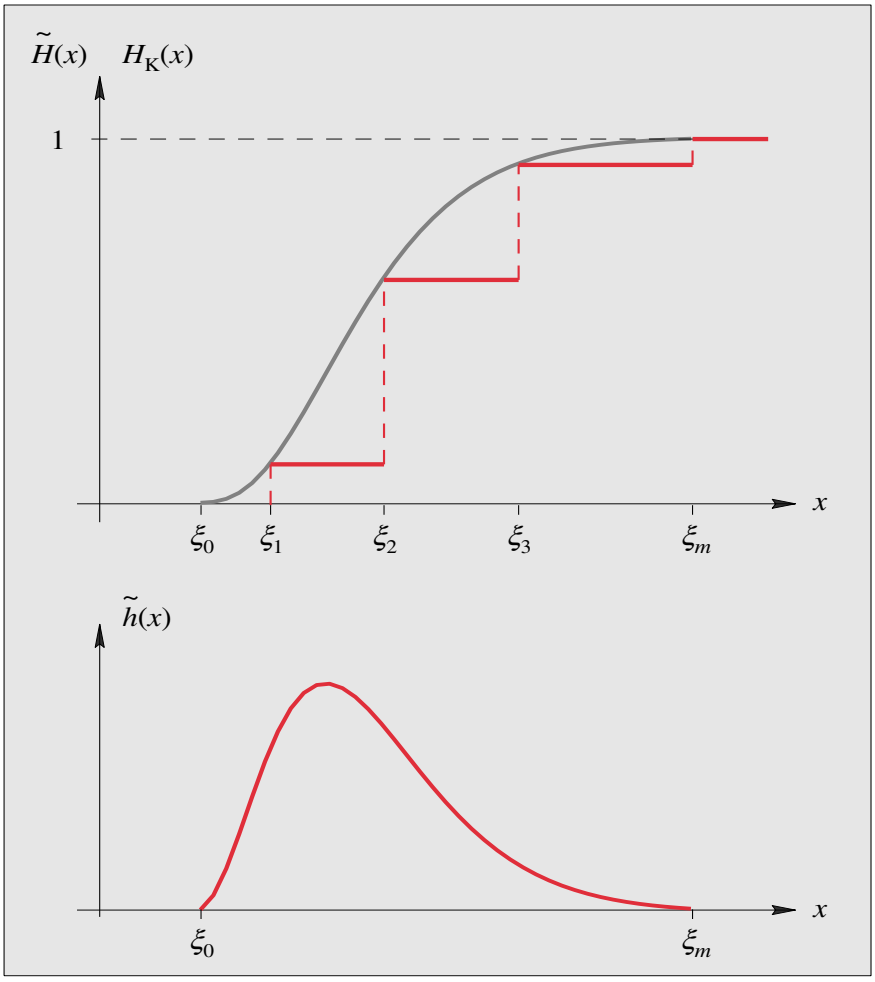


BILD 1.8 Approximierende glatte Kurven

PRAXIS

Sterben die Deutschen aus?

Die künftige demographische Entwicklung Deutschlands bereitet Sorgen. Der Vergleich der beiden Bevölkerungspyramiden in Bild 1.9 macht dies deutlich. Die rechte Pyramide ist eine Projektionsrechnung. Sie zeigt den Altersaufbau unter der Voraussetzung, dass die Geburtenrate wie seit einem Vierteljahrhundert weiterhin auf dem Niveau von 1.3 bis 1.4 Kindern pro Frau bleibt und der Einwanderungsüberschuss wie im langjährigen Durchschnitt auch künftig rund 170 000 Personen pro Jahr beträgt. Zusätzlich wird noch die absehbare Zunahme der Lebenserwartung um rund sechs Jahre berücksichtigt.

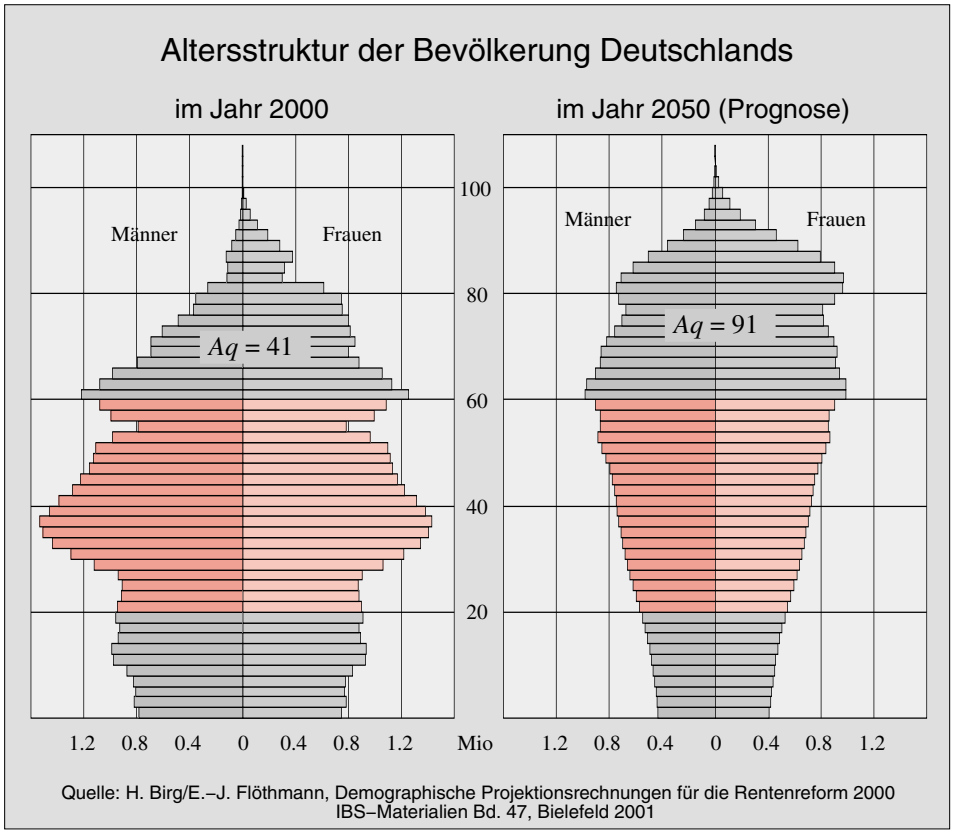


BILD 1.9 Bevölkerungspyramiden für Deutschland

So standen 100 Menschen der ökonomisch aktiven Altersgruppe 20 bis 60 im Jahre 2000 rund 41 über Sechzigjährige gegenüber, im Jahre 2014 bereits 50. Nach der Prognose würde dieser **Altenquotient** Aq im Jahre 2050 auf 91 ansteigen. Dies hätte enorme sozialpolitische Konsequenzen.

Kontrollfragen

1 Was ist der Unterschied zwischen Merkmal und Variable?

2 Welche verschiedenen Skalenarten kennen Sie? Überlegen Sie sich eigene Beispiele!

3 Warum werden in der Praxis zumeist repräsentative Stichproben erhoben?

4 Welche Eigenschaften hat die Treppenfunktion? Welchen Aussagegehalt besitzt sie?

5 Warum ist die Bildung von Größenklassen oft notwendig? Überlegen Sie sich ein Beispiel!

6 Welche Annahme liegt der approximierenden Verteilungsfunktion $\overline{H}(x)$ implizit zugrunde?

7 Was ist der Unterschied zwischen Säulendiagramm und Histogramm? Unter welcher Bedingung sehen beide gleich aus?

Ergänzende Literatur

Bohley, Peter: *Statistik*, 7. Aufl., München, Wien: Oldenbourg 2000, Kapitel III

Bourier, Günther: *Beschreibende Statistik. Praxisorientierte Einführung*, 12. Aufl., Heidelberg: Springer-Gabler 2014, Kapitel 1 und 2

Krämer, Walter: *So lügt man mit Statistik*, Frankfurt, New York: Campus-Verlag 2015

Schlittgen, Rainer: *Einführung in die Statistik: Analyse und Modellierung von Daten*, 12. Aufl., München, Wien: Oldenbourg 2012, Kapitel 1 und 2

Schwarze, Jochen: *Grundlagen der Statistik I*, 12. Aufl., Herne: NWB-Verlag 2014

Aufgaben

1.1 **Zuckerpakete.** Bei einer Nachwiegung von 20 verpackten Pfundpaketen Zucker ergaben sich folgende Werte (in g):

492	497	478	482	499	512	503
511	499	504	508	496	502	500
499	500	507	502	500	499.	

Zeichnen Sie ein Histogramm mit der

a) Klassenbreite 1 g

b) Klassenbreite 2 g .

1.2 **Merkmale.** Geben Sie zu den folgenden Merkmalen Beispiele für statistische Einheiten und Merkmalsausprägungen an. Nennen Sie Merkmalstyp und Skalierung.

Haarfarbe	Körpergröße
Verdienst	Gewicht
Abiturnote in Deutsch	Religionsbekenntnis
Geschlecht	Zugehörigkeit zu einer sozialen Schicht
Beruf	Vermögen
Kontobewegungen/Monat	

1.3 **FAZ.** Ein Kioskbesitzer notiert 200 Tage lang die Zahl der verkauften Exemplare der FAZ.

a) Geben Sie Merkmalsträger und mögliche Merkmalsausprägungen an. Um welche Merkmalstypen handelt es sich?

b) Zeichnen Sie die Verteilungsfunktion.

Verkaufte Zeitungen	Anzahl der Tage
0	21
1	46
2	54
3	40
4	24
5	10
6	5

1.4 **Statistikklausur.** Bei der letzten Statistikklausur machte sich der Prüfer die nebenstehenden Aufzeichnungen über die erreichten Punktezahlen.

a) Skizzieren Sie die Verteilungsfunktion.

Punkte von … bis unter …	Anzahl
0 – 25	50
25 – 50	90
50 – 75	170
75 – 100	90

b) Wie viele Klausurteilnehmer erzielten weniger als 90 Punkte? Erläutern Sie Ihre Antwort.

1.5 **Polygonzug und glatte Kurve.** Ein Merkmal X wurde nach Größenklassen erhoben:

Größenklassen	relative Häufigkeiten
0 – 5	0.1
5 – 8	0.7
8 – 10	0.2

a) Zeichnen Sie $H_{\mathrm{K}}(x)$ und $\overline{H}(x)$.

b) Zeichnen Sie das Histogramm.

c) Zeichnen Sie eine approximierende glatte Verteilungsfunktion nach der Freihandmethode.

1.6 **Einkommensverteilung.** Im „Statistischen Taschenbuch" 2007 des BUNDESMINISTERIUMS FÜR ARBEIT UND SOZIALES (BMAS) findet sich als Ergebnis der Einkommensteuerstatistik folgende Tabelle für 2002:

Jahreseinkünfte in Euro von . . . bis unter . . .	Steuer-pflichtige %	Gesamtbetrag der Einkünfte %
unter 2 500	3.1	0.1
2 500 – 5 000	3.7	0.4
5 000 – 7 500	4.3	0.8
7 500 – 10 000	4.4	1.1
10 000 – 12 500	4.3	1.4
12 500 – 25 000	24.2	12.8
25 000 – 37 500	23.0	19.7
37 500 – 50 000	13.6	16.3
50 000 – 125 000	17.5	33.9
125 000 – 250 000	1.4	6.5
250 000 – 500 000	0.3	2.8
500 000 und mehr	0.1	4.2
	100	100

a) Zeichnen Sie aus diesen Angaben ein Histogramm und eine Verteilungsfunktion.

b) An welcher Stelle hätte die approximierende glatte Kurve der Verteilungsfunktion – nach der Freihandmethode gezeichnet – ihre größte Steigung? Eine näherungsweise Angabe genügt.

1.7 **Examensnoten.** Ein frischgebackener Master of Arts in Economics bewirbt sich bei einem großen Stuttgarter Unternehmen und erhält postwendend eine formlose Absage. Eher empört über diese Art der Benachrichtigung ruft er den Personalchef an und befragt ihn nach den Gründen für die Ablehnung. Dieser erklärt dem Absolventen, dass das Unternehmen eine Vorauswahl nach Notendurchschnitten vornehme und er ja leider nur eine befriedigende Gesamtnote vorzuweisen habe, daher also nicht in Frage käme.

Der Bewerber erklärt dem Personalchef daraufhin, dass das arithmetische Mittel bei Noten keine Aussagekraft habe, da Zensuren ordinal skaliert seien. Zudem könne man schon gar nicht Examensnoten aus verschiedenen Fachbereichen oder gar von verschiedenen Unis miteinander vergleichen. Die Gesamtnote sei also ein denkbar schlechtes Auswahlkriterium. Zum Schluss des Gesprächs empfiehlt der Exstudent dem Personalchef die Lektüre einschlägiger Statistikliteratur.

Hat der Bewerber recht? Diskutieren Sie die Unterschiede zwischen Nominal-, Ordinal- und Kardinalskala.

1.8 **ZDF-Politbarometer.** Auf den Websites des ZDF lesen wir als Antwort auf die FAQ „Wann sind Umfragen repräsentativ?“:

„Umfragen sind repräsentativ, wenn jeder Wahlberechtigte die gleiche Chance hat, befragt zu werden. Dies wird durch eine zufällige Auswahl gewährleistet. Für jede Umfrage wählt die Forschungsgruppe Wahlen zufällig Telefonnummern aus den eingetragenen privaten Telefonanschlüssen aus. Damit auch Haushalte, die nicht im Telefonbuch stehen, in die Auswahl gelangen können, werden die drei letzten Ziffern der Telefonnummern durch Zufallszahlen ersetzt. Innerhalb eines Haushalts wird für die Befragung ausgewählt, wer zuletzt Geburtstag hatte. Dadurch sind die Ergebnisse repräsentativ für alle Wahlberechtigten.

Für jedes Politbarometer befragen Mitarbeiter der Forschungsgruppe Wahlen telefonisch etwa 1.250 Wahlberechtigte in ganz Deutschland. Die Umfragen dauern von Dienstag bis Donnerstag vor der Sendung, die am Freitag ausgestrahlt wird.“

a) Was halten Sie davon?

b) Vergleichen sie dazu Beispiel [9] dieses Kapitels.

Quelle:
http://www.zdf.de/politbarometer/faq-politbarometer-antworten-auf-haeufige-fragen-33835372.html

LÖSUNGEN

1.2

Merkmal	statistische Einheiten	Merkmals-ausprägung	Merkmals-typ	Skalierung
Haarfarbe	Männer im Alter zwischen 60 und 65	schwarz, braun, blond, grau	qualitativ	nominal
Verdienst	Studentische Hilfskräfte	8 – 12 €/Stunde	quantitativ diskret	kardinal
Abiturnote in Deutsch	Jahrgang 2000	0 – 15 Punkte	quantitativ diskret	ordinal
Beruf	Mitglieder der FDP	Arbeiter, Angest., Selbständiger	qualitativ	nominal
Kontobewe-gungen pro Monat	Girokonten der Sparkasse Duisburg	0 – 1000 Stück	quantitativ diskret	kardinal
Körpergröße	Mitglieder der dt. Basketball-Nationalmannschaft	1.60 m – 2.3 m	quantitativ stetig	kardinal
⋮				

1.3 Tage; 0, 1, 2, ... ; quantitativ, diskret

1.4 ca. 364

1.6 b) ca. 35 000

KAPITEL 2

Maßzahlen zur Beschreibung statistischer Verteilungen

Zuweilen – besonders bei statistischen Beobachtungsreihen mit sehr vielen verschiedenen Merkmalsausprägungen – verstellt die Zahlenfülle den Blick auf das Wesentliche. Selbst wenn die Verteilung nach Größenklassen gegliedert in Form einer Tabelle oder eines Histogramms aufbereitet vorliegt, entsteht doch der Wunsch, durch die Angabe einiger weniger Zahlen die gesamte Verteilung des Merkmals zu charakterisieren. Selbstverständlich bedeutet dies einen Verzicht auf Information. Solche Zahlen heißen ***Maßzahlen*** oder ***Parameter*** einer Verteilung. Zuerst möchte man ein Maß für die Lage der Verteilung, das heißt für die durchschnittliche Größenordnung der Variablenwerte einer Beobachtungsreihe haben. Als Nächstes interessiert, wie eng die einzelnen Werte beieinanderliegen oder wie weit sie streuen. Die Maßzahlen dienen gewissermaßen zur „Vermessung“ einer Verteilung. Dazu hat die deskriptive Statistik für verschiedene Fragestellungen und Anwendungssituationen eine Reihe von Konzepten entwickelt.

2.1 Arithmetisches Mittel als Lagemaß

Das für quantitative Merkmale am häufigsten verwendete Lokalisationsmaß ist das arithmetische Mittel.

Definition: Die Größe

$$\bar{x} := \frac{1}{n}\sum_{j=1}^{n} x_j \tag{2-1}$$

heißt ***arithmetisches Mittel*** oder ***Mittelwert*** einer statistischen Verteilung.

In anderer Schreibweise – unter Verwendung der absoluten Häufigkeiten – erhalten wir

$$\bar{x} := \frac{1}{n}\sum_{j=1}^{k} n_j x_j \tag{2-2}$$

beziehungsweise – unter Verwendung der relativen Häufigkeiten $h_i = n_i/n$ –

$$\bar{x} := \sum_{j=1}^{k} h_j x_j \; . \tag{2-3}$$

Diese beiden letzten Schreibweisen sehen so aus, als ob es sich um ein *gewichtetes* arithmetisches *Mittel* handeln würde. In der Tat werden die vorkommenden k Merkmalsausprägungen mit ihrer relativen Häufigkeit gewichtet.

Beispiel [1] Nehmen wir die Verteilung aus Beispiel [11] des vorigen Kapitels:

x_i	1.6	3.0	4.1	5.0
h_i	0.1	0.2	0.4	0.3

Das arithmetische Mittel ist gemäß Formel (2-3) einfach

$$\bar{x} \; = \; 0.1 \cdot 1.6 + 0.2 \cdot 3.0 + 0.4 \cdot 4.1 + 0.3 \cdot 5.0 \; = \; 3.9 \; .$$

Eigenschaften des arithmetischen Mittels

1. Zentraleigenschaft. Die Summe aller Abweichungen der Merkmalswerte einer statistischen Reihe von ihrem eigenen arithmetischen Mittel ist stets Null:

$$\sum_{j=1}^{n} (x_j - \bar{x}) \; = \; 0 \; . \tag{2-4}$$

2. Verschiebung aller Werte einer statistischen Reihe um den konstanten Wert a verschiebt das arithmetische Mittel um eben diesen Wert:

$$\begin{aligned} & y_i := x_i + a \qquad (i = 1, \cdots, n) \\ \Rightarrow \qquad & \bar{y} \; = \; \bar{x} + a \; . \end{aligned} \tag{2-5}$$

3. Homogenität. Multiplikation aller Werte einer statistischen Reihe X mit dem konstanten Faktor $b \neq 0$ multipliziert das arithmetische Mittel mit diesem Wert:

$$\begin{aligned} & z_i := b \cdot x_i \qquad (i = 1, \cdots, n) \\ \Rightarrow \qquad & \bar{z} \; = \; b \cdot \bar{x} \; . \end{aligned} \tag{2-6}$$

4. Berechnung von $\bar{x}_{\text{ges}}$ aus den Gruppenmittelwerten $\bar{x}_i$: Die statistische Reihe X mit n Elementen sei in $m < n$ disjunkte statistische

Teilreihen (Gruppen)	$X_1,\ X_2,\ \cdots,\ X_m$	
mit jeweils	$n_1,\ n_2,\ \cdots,\ n_m$	Elementen
und den Mittelwerten	$\bar{x}_1,\ \bar{x}_2,\ \cdots,\ \bar{x}_m$	

zerlegt worden. Es gilt dann

$$\bar{x}_{\text{ges}} = \frac{1}{n}\sum_{j=1}^{m} n_j \bar{x}_j \ , \tag{2-7}$$

das heißt der Mittelwert *der gesamten statistischen Reihe* kann als gewichtetes arithmetisches Mittel der Gruppenmittelwerte berechnet werden, wobei die Gewichte durch die Besetzungszahlen der einzelnen Gruppen n_j/n gegeben sind.

2.2 Median und Modus

Zur Kennzeichnung der Lage einer Verteilung findet zuweilen der Median Anwendung. Um ihn zu bestimmen, müssen die Merkmalsausprägungen in der statistischen Reihe der Größe nach geordnet sein:

$$x_1 \le x_2 \le x_3 \le \cdots \le x_i \le \cdots \le x_n \, . \tag{1-3}$$

Definition: Eine Zahl x_{Med} mit

$$\begin{array}{|c|l|}\hline x_{\text{Med}} = x_{\frac{n+1}{2}} & \text{falls } n \text{ ungerade} \\ \hline x_{\frac{n}{2}} \le x_{\text{Med}} \le x_{\frac{n}{2}+1} & \text{falls } n \text{ gerade} \\ \hline \end{array} \tag{2-8}$$

heißt ***Median*** oder ***Zentralwert*** der empirischen statistischen Reihe X.

Man sieht gleich, dass in dieser Definition x_{Med} nicht immer eindeutig bestimmt ist. Ist n gerade und sind die beiden benachbarten Werte in der Mitte nicht gleich groß, dann ist jede der beiden Zahlen und jede Zahl dazwischen ein Median. Man wählt in diesem Fall häufig

$$x_{\text{Med}} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$$

als Median.

Beispiel [2] Die statistische Reihe 4, 7, 7, 7, 12, 12, 13, 16, 19, 23, 23, 97 hat das arithmetische Mittel $\bar{x} = 20$ und den Median $x_{\text{Med}} = 12.5$. Auch 12 und 12.2 wären Mediane.

Beispiel [3] Das Durchschnittsvermögen der deutschen Haushalte betrug 2013 123 000 €, das Median-Vermögen nur 55 500 €. Wie kommt das?

Das Lagemaß Median hat den Nachteil, dass es – ganz im Gegensatz zum arithmetischen Mittel – unter Umständen nur die Zahlenwerte von einer oder zwei Ausprägungen berücksichtigt. Dafür hat dieses Lagemaß den Vorteil, dass seine Berechnung – zumindest für ungerades n – nur ordinale Messbarkeit voraussetzt, das heißt die Beobachtungswerte x_i brauchen nur auf einer Ordinalskala angesiedelt zu sein. Wir werden den Median noch als einfachen Spezialfall der sogenannten Quantile in Abschnitt **2.7** dieses Kapitels kennenlernen.

Der Modus als Lokalisationsmaß wird in jedem statistischen Lehrbuch erwähnt, jedoch wird er eher selten in der praktischen statistischen Analyse verwendet. Der Modus ist derjenige Wert in einer Stichprobe oder einer Grundgesamtheit, der am häufigsten vorkommt.

Definition: Die Zahl x_{Mod}, mit

$$h(x_{\text{Mod}}) \geq h(x_i) \qquad \text{für alle } i \tag{2-9}$$

heißt ***Modus*** oder ***Modalwert*** einer empirischen statistischen Reihe.

Somit wird mit Modus der häufigste Merkmalswert einer statistischen Reihe bezeichnet. Auch der Modus ist ein Lagemaß, das vorzugsweise bei nicht metrisch skalierten Merkmalen Verwendung findet. Im Gegensatz zum Median und arithmetischen Mittel behält es sogar seinen Sinn bei rein qualitativen Merkmalen. Obwohl Konzept und Definition sehr einfach erscheinen, birgt die sinnvolle Anwendung des Modus zuweilen doch einige Schwierigkeiten, wie die folgenden Beispiele zeigen.

Beispiele [4] Die statistische Reihe 2, 3, 3, 4, 4, 4, 5, 6 hat den Modus 4.

[5] Zwei „häufigste“ Werte gibt es in der statistischen Reihe 1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, nämlich 3 und 6. Die Werte liegen getrennt und kommen jeweils häufiger vor als ihre beiden Nachbarwerte.

Der Modus muss nicht für jede statistische Reihe in eindeutiger Weise existieren. Verteilungen, die genau einen Modus besitzen, heißen ***unimodal***.

2.3 Geometrisches Mittel

Beim geometrischen Mittel

$$G_X := \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \quad , \qquad x_i > 0 \tag{2-10}$$

werden die einzelnen Merkmalswerte multipliziert und aus dem Produkt die *n*-te Wurzel gezogen. Es ist nur definiert, wenn sämtliche Werte der statistischen Reihe *X* positiv sind.

Beispiel [6] Das geometrische Mittel aus der statistischen Reihe *X* mit den Werten 2, 6, 12, 9 ist $G_X = 6$, während das arithmetische daraus $\bar{x} = 7.25$ ist.

In der Tat ist das geometrische Mittel für jede Reihe mit nur positiven Werten stets *kleiner als das arithmetische*, es sei denn, alle Werte der Reihe sind gleich. Der Logarithmus des geometrischen Mittel entspricht dem *arithmetischen Mittel der Logarithmen*, genauer

$$\log G_X = \frac{1}{n} \sum \log x_j \ .$$

Daraus folgt wegen der Zentraleigenschaft des arithmetischen Mittels, dass die Summe der logarithmierten **Quotienten** der Werte mit ihrem geometrischen Mittel

$$\sum \left(\log x_j - \log G_X \right) = 0$$

Null ist. Wenn wir (2-10) hoch *n* nehmen

$$\begin{aligned} G_X^n &= x_1 \cdot x_2 \cdot \ldots \cdot x_n \\ 1 &= \frac{x_1}{G_X} \cdot \frac{x_2}{G_X} \cdot \ldots \cdot \frac{x_n}{G_X} \ , \end{aligned}$$

sehen wir, dass das geometrische Mittel in einer besonderen Weise „zentral“ ist: Die *n* Quotienten x_i / G_X, von denen einige kleiner als Eins, andere größer als Eins sind, multiplizieren sich gerade zu Eins auf.

Wegen dieser Eigenschaft ist das geometrische Mittel eher geeignet, *Quotienten*, *Prozente* und *Wachstumsraten* zu mitteln, wenn das arithmetische Mittel versagen würde. Deshalb ist das geometrische Mittel besonders bei der Beschreibung von Zeitreihen (siehe Kapitel **5**) ein guter Lageparameter.

Beispiel [7] In fünf aufeinanderfolgenden Jahren haben sich die Umsätze Y (in Tausend €) einer bestimmten Firma wie folgt entwickelt:

Jahr t	2011	2012	2013	2014	2015
Umsatz y_t in 1000 €	1 200	1 440	1 224	1 714	2 142
y_t / y_{t-1}		1.20	0.85	1.40	1.25
Rate $r_t := y_t / y_{t-1} - 1$		0.20	–0.15	0.40	0.25
Veränderung in % gegenüber dem Vorjahr		20%	–15%	40%	25%

(1) Würde man zur Berechnung des „durchschnittlichen Wachstums“ das *arithmetische Mittel* des prozentualen Wachstums

$$\overline{1+r} = (1.20+0.85+1.40+1.25)/4 = 4.70/4 = 1.175$$

oder $(20\%-15\%+40\%+25\%)/4 = 70\%/4 = 17.5\%$

nehmen, ergäbe diese durchschnittliche Umsatzsteigerung, von 1 200 im Jahre 2011 ausgehend, im Jahre 2015 einen Umsatz von 2 287 anstatt des tatsächlichen $Y_{2012} = 2\ 142$:

Jahr t	2011	2012	2013	2014	**2015**
Umsatz y_t in 1000 €	1 200	1 440	1 224	1 714	**2 142**
mit ∅ 17.5% pro Jahr		*1 410*	*1 657*	*1 947*	*2 287*
mit ∅ 15.59% pro Jahr		*1 387*	*1 603*	*1 853*	**2 142**

(2) Berechnen wir jedoch das *geometrische Mittel*

$$G_{1+r} = \sqrt[4]{(1+r_1)(1+r_2)(1+r_3)(1+r_4)}$$
$$= \sqrt[4]{1.20 \cdot 0.85 \cdot 1.40 \cdot 1.25} = 1.1559\ ,$$

erhalten wir eine durchschnittliche Umsatzsteigerung von 15.59%, die, wäre sie Jahr für Jahr eingetreten, 2015 gerade zum aktuellen Wert führen würde.

Die Berechnung einer *durchschnittlichen* Wachstumsrate basiert, wie im obigen Beispiel, auf der Vorstellung einer konstanten Wachstumsrate oder geometrischen Progression. Kennt man den Anfangswert und den Endwert einer Zeitreihe, sagen wir der Zeitreihe des Bruttoinlandsproduktes

$$BIP_0,\ BIP_1,\ \cdots,\ BIP_T,$$

kann man daraus natürlich die durchschnittliche Wachstumsrate r nach der Formel

$$\boxed{1+r\ =\ \sqrt[T]{\frac{BIP_T}{BIP_0}}} \tag{2-11}$$

finden, was der Berechnung eines geometrischen Mittels entspricht.

❑ *Beweis:* Würde man eine konstant bleibende Wachstumsrate r unterstellen, wäre ja gerade entsprechend der Zinseszinsformel

$$\begin{aligned} BIP_1 &= BIP_0(1+r) \\ BIP_2 &= BIP_1(1+r) = BIP_0(1+r)(1+r) \\ &\ \ \vdots \\ BIP_T &= BIP_0(1+r)^T\ . \end{aligned}$$

Eingesetzt in (2-11) ergibt das sogleich

$$\sqrt[T]{\frac{BIP_T}{BIP_0}}\ =\ \sqrt[T]{\frac{BIP_0(1+r)^T}{BIP_0}}\ =\ \sqrt[T]{(1+r)^T}\ =\ 1+r\,,$$

was zu beweisen war. ❑

2.4 Harmonisches Mittel

Bildet man von den Werten x_i einer statistischen Reihe die Kehrwerte $1/x_i$, kann man natürlich auch von diesen das arithmetische Mittel

$$\frac{1}{n}\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)$$

berechnen. Nimmt man dann vom Ergebnis wieder den Kehrwert, erhält man das sogenannte ***harmonische Mittel***

$$H_X\ :=\ \frac{n}{\sum(1/x_j)}\,. \tag{2-12}$$

Beispiel [8] Das harmonische Mittel aus der statistischen Reihe X mit den Werten 2, 6, 12, 9 ist

$$H_X = 4/(1/2 + 1/6 + 1/12 + 1/9) = 4.645 \ldots ,$$

während das arithmetische daraus $\bar{x} = 7.25$ und das geometrische $G_X = 6$ war.

Natürlich darf die statistische Reihe keine Nullen und keine negativen Werte enthalten. Für jede statistische Reihe mit (verschiedenen) positiven Werten ist

$$H_X < G_X < \bar{x}$$

das harmonische Mittel kleiner als das geometrische und das arithmetische.

Auch hier stellt sich die Frage, wann das harmonische Mittel benutzt werden sollte, das heißt wann es einen mit der Fragestellung konsistenten Mittelwert liefert. Jedenfalls wird oft das arithmetische Mittel genommen, obwohl eigentlich das harmonische verwendet werden sollte, ja sogar verwendet werden müsste.

Beispiel [9] Die durchschnittliche Geschwindigkeit von Lastzügen auf deutschen Autobahnen, als arithmetisches Mittel gerechnet, beträgt $\bar{v} = 71$ km/h. Dividiert man eine bestimmte Distanz $d = 412$ km, etwa von Hamburg nach Duisburg, durch diesen Durchschnittswert, um die Transportzeiten t, Transportkapazitäten und Transportkosten abzuschätzen, erhält man mit

$$\bar{t} = d / \bar{v} = 412 / 71 = 5.80$$

Stunden ein falsches Ergebnis. Ein richtiges Ergebnis kommt heraus, wenn das harmonische Mittel $H_V = 68$ km/h als Divisor verwendet wird:

$$\bar{t} = d / H_V = 412 / 68 = 6.06$$

Geschwindigkeiten berechnet man als Quotient aus dem zurückgelegten Weg und der dafür gebrauchten Zeit. Legt dieses Beispiel nun nahe, dass man etwa Geschwindigkeiten oder jede Messreihe von Quotienten, wie Arbeitslosenquoten, stets harmonisch mitteln sollte? Das kommt darauf an, wozu man den Mittelwert verwenden will. Im Beispiel [9] will man für eine feste Distanz d eine durchschnittliche Transportzeit berechnen. Dazu bräuchte man die einzelnen Geschwindigkeiten v_i im Nenner – oder eben das harmonische Mittel:

$$\begin{aligned} \bar{t} &= \frac{1}{n}(t_1 + \cdots + t_n) = \frac{1}{n}(d/v_1 + \cdots + d/v_n) \\ &= d \cdot \frac{1}{n}(1/v_1 + \cdots + 1/v_n) = d \cdot 1/H_V. \end{aligned}$$

Würden wir hingegen wissen wollen, wie weit die Lastzüge im arithmetischen Durchschnitt nach einer festgelegten Zeit t gekommen sind, wäre sicher die Rechnung

$$\bar{d} = \frac{1}{n}(d_1 + \cdots + d_n) = \frac{1}{n}(t v_1 + \cdots + t v_n)$$
$$= t \cdot \frac{1}{n}(v_1 + \cdots + v_n) = t \cdot \bar{v}$$

die richtige.

2.5 Streuungsmaße

Die Lageparameter wie Modus, Median, die arithmetischen Mittel und andere Lagemaße geben jeweils nur eine *zentrale Tendenz* einer Verteilung an. Nun soll aber auch noch das Ausmaß der *Streuung* oder *Variation* oder *Dispersion* einer Verteilung bzw. der Werte einer statistischen Reihe in einer Maßzahl ausgedrückt werden. Die beschreibende Statistik stellt dazu einige Maßzahlen bereit.

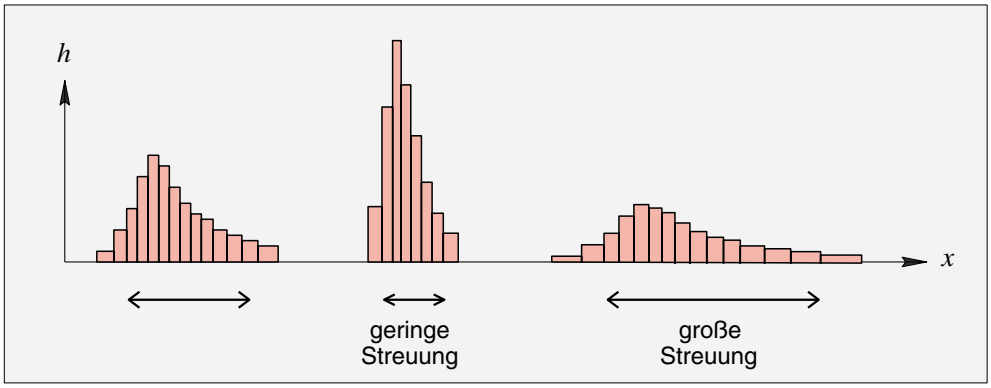


BILD 2.1 Drei Histogramme: Verteilungen mit verschiedenen Streuungen

Spannweite

Sie ist die Differenz zwischen der größten und der kleinsten Merkmalsausprägung in der statistischen Reihe:

$$\text{Spannweite} := x_{\max} - x_{\min}$$

Die Spannweite ist natürlich besonders einfach zu berechnen, denn es werden nur der größte und der kleinste Wert einer statistischen Reihe verwendet. Dies ist aber natürlich gleichzeitig die Schwäche dieses Streuungsmaßes, denn das Verhalten der übrigen Werte wird nicht beachtet.

Mittlere absolute Abweichung

Um aber alle Werte einer statistischen Reihe in ein Streuungsmaß eingehen zu lassen, könnte man versucht sein, die durchschnittliche Abweichung der Beobachtungswerte von ihrem Mittelwert zu berechnen. Man wird jedoch sogleich feststellen, dass wegen der Zentraleigenschaft des arithmetischen Mittels sich die positiven und negativen Abweichungen gerade ausgleichen, so dass als Ergebnis stets Null herauskäme. Um dem zu entgehen und ein Maß für die Streuung der Einzelwerte zu erhalten, könnte man die Absolutbeträge der Abweichungen betrachten. Die sogenannte ***mittlere absolute Abweichung***

$$\mathrm{MAA} := \frac{1}{n}\sum_{j=1}^{n}\left|x_j - \bar{x}\right| \tag{2-13}$$

wird als arithmetisches Mittel der Beträge der Abweichungen der Merkmalswerte von ihrem Mittelwert berechnet.

Mittlerer Quartilsabstand

Man ordnet die Merkmalswerte der statistischen Reihe der Größe nach

$$x_1 \le x_2 \le \cdots \le x_n$$

und teilt sie in vier Segmente mit möglichst gleich großer Anzahl von Werten. Exakt geht das natürlich nur, wenn ihre Gesamtzahl n durch 4 teilbar ist, sonst behilft man sich näherungsweise – oder verwendet die Formel aus dem Abschnitt **2.7** über die Quantile. Die drei Werte

$$Q_1 \le Q_2 = x_{\mathrm{Med}} \le Q_3,$$

die sogenannten ***Quartile***, müssen so beschaffen sein, dass sie in der gleichen Weise wie der Median x_{Med} zwischen den Segmenten liegen. Somit befinden sich zwischen Q_1 und Q_3 stets 50% der Beobachtungswerte. Dieser Bereich

$$\mathrm{IQA} := Q_3 - Q_1$$

heißt der ***Interquartilsabstand***, und das arithmetische Mittel der beiden Quartilsabstände vom Median ist nun der ***mittlere Quartilsabstand***

$$\mathrm{MQA} := \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{\mathrm{IQA}}{2}\ . \tag{2-14}$$

Einerseits wird der mittlere Quartilsabstand als Streuungsmaß vor allem dann gern verwendet, wenn es sich um ordinal skalierte Daten handelt; der zu diesem Streuungsmaß passende Lageparameter wäre dann der Median. Die Quartilsabstände verschiedener

Verteilungen sollten dann aber auch nur ordinal verglichen werden. Andererseits sind gerade bei metrischen Merkmalen die Quartile (und die weiter unten erwähnten Quantile) sehr hilfreiche *Markierungen zum Vermessen einer Verteilungsfunktion*.

Beispiel [10] Von einer statistischen Reihe mit $n = 14$ Werten suchen wir den mittleren Quartilsabstand.

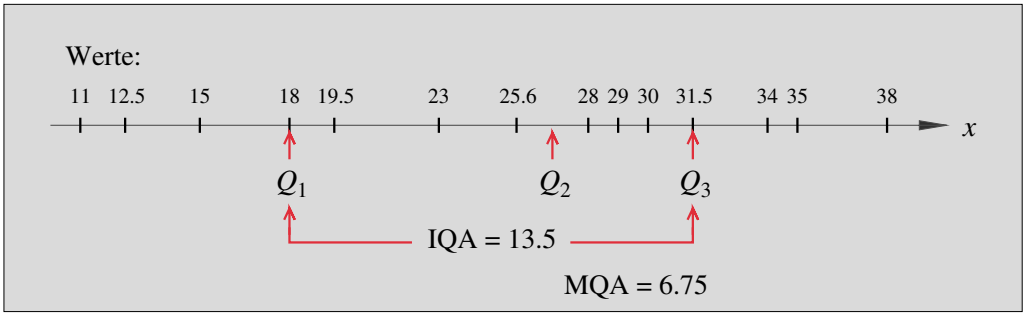


BILD 2.2 Mittlerer Quartilsabstand

Als Median nehmen wir das arithmetische Mittel der beiden Nachbarn und erhalten $Q_2 = 26.8$.

2.6 Varianz und Standardabweichung

Das am häufigsten verwendete Streuungsmaß ist die Varianz. Die Varianz hat außerdem zentrale Bedeutung in der theoretischen Statistik.

Definition: Die mittlere quadratische Abweichung vom arithmetischen Mittel

$$s_X^2 := \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 \tag{2-15}$$

heißt ***empirische Varianz*** oder kurz ***Varianz*** einer beobachteten statistischen Reihe X.

Liegt eine relative Häufigkeitsverteilung der statistischen Reihe vor, so kann die Varianz auch damit berechnet werden:

$$s_X^2 := \sum_{j=1}^{k} h_j (x_j - \bar{x})^2 \tag{2-15a}$$

Definition: Die positive Wurzel aus der Varianz

$$s_X := +\sqrt{s_X^2} \tag{2-16}$$

heißt ***Standardabweichung***.

Beispiel [11] Für die diskrete statistische Variable X sei folgende Verteilung erhoben worden:

x_i :	4	5	6
$h(x_i)$:	1/4	1/2	1/4

Sie hat den Mittelwert $\bar{x} = 5$. Ihre Varianz ist

$$\begin{aligned} s_X^2 &= (4-5)^2 \cdot (1/4) + (5-5)^2 \cdot (1/2) + (6-5)^2 \cdot (1/4) \\ &= 1/4 + 1/4 = 1/2 \end{aligned}$$

und ihre Standardabweichung

$$s_X = 1/\sqrt{2} = 0.7071 \ .$$

Beispiel [12] Bei der statistischen Reihe

3 5 9 9 6 6 3 7 7 6
7 6 5 7 6 9 6 5 3 5

rechnen wir mit folgender Arbeitstabelle:

i	x_i	n_i	h_i	$H(x_i)$	$h_i \cdot x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$h_i \cdot (x_i - \bar{x})^2$
1	3	3	0.15	0.15	0.45	–3	9	1.35
2	5	4	0.20	0.35	1.00	–1	1	0.20
3	6	6	0.30	0.65	1.80	0	0	0
4	7	4	0.20	0.85	1.40	1	1	0.20
5	9	3	0.15	1.00	1.35	3	9	1.35
		$20 = n$	1		$6.00 = \bar{x}$	0		**3.10**

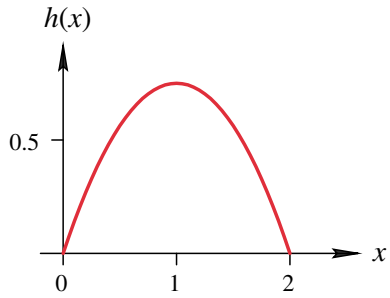
Der Mittelwert der statistischen Reihe beträgt 6, die Varianz $s_X^2 = 3.1$ und die Standardabweichung 1.761.

Beispiel [13] Die Verteilung einer statistischen Variablen sei im Intervall $0 < x < 2$ durch das Parabelstück als stetige Häufigkeitsdichtefunktion $h(x)$ approximiert.

Die Dichtefunktion lässt sich durch

$$h(x) = \frac{3}{2}(x - 0.5x^2)$$

beschreiben. Wie die Skizze zeigt, ist der Mittelwert $\bar{x}$ aus Symmetriegründen gleich Eins.



Die Varianz berechnen wir als das bestimmte Integral wie in (9-16b), wobei die Integration an die Stelle der Summation in (2-15a) tritt:

$$\begin{aligned} s_X^2 &= \int_0^2 h(x)\cdot(x-1)^2\,\mathrm{d}x = \int_0^2 \frac{3}{2}(x - \frac{1}{2}x^2)(x-1)^2\,\mathrm{d}x \\ &= \frac{3}{2}\int (x - \frac{5}{2}x^2 + 2x^3 - \frac{1}{2}x^4)\,\mathrm{d}x \\ &= \frac{3}{2}\left[\frac{1}{2}x^2 - \frac{5}{6}x^3 + \frac{1}{2}x^4 - \frac{1}{10}x^5\right]_0^2 \\ &= \frac{3}{2}\left[\frac{4}{2} - \frac{40}{6} + \frac{16}{2} - \frac{32}{10}\right] = 3 - 10 + 12 - \frac{24}{5} = \frac{1}{5} \end{aligned}$$

und die Standardabweichung als Wurzel daraus mit $s_X = 0.4472$.

Eigenschaften der Varianz

Die folgenden Eigenschaften tragen zum Verständnis des Varianzbegriffs bei und sind auch als Rechenregeln recht nützlich.

1. Die Varianz ist stets größer oder gleich Null:

$$s_X^2 \geq 0 \ .$$

2. Translation der statistischen Reihe um a = const lässt die Varianz unverändert:

$$y_i := x_i + a \qquad (i = 1, \cdots, n)$$

$$\Rightarrow \qquad s_Y^2 = s_X^2 \ .$$

3. Streckung der statistischen Reihe mit dem Faktor b = const:

$$z_i := b \cdot x_i \qquad (i = 1, \cdots, n)$$

$$\Rightarrow \qquad s_Z^2 = b^2 \cdot s_X^2 \ . \tag{2-17}$$

Was die Standardabweichung betrifft, so gilt entsprechend:

$$s_Z = |b| \cdot s_X \ . \tag{2-18}$$

4. Zu ihrer **vereinfachten Berechnung** dient die folgende Eigenschaft der Varianz. Es gilt:

$$\frac{1}{n}\sum (x_j - \bar{x})^2 = \frac{1}{n}\sum x_j^2 - \bar{x}^2 \tag{2-19}$$

❑ *Beweis:* Zum Beweis braucht man die linke Klammer nur auszumultiplizieren

$$\frac{1}{n}\sum (x_j - \bar{x})^2 = \frac{1}{n}\sum (x_j^2 - 2\bar{x}\,x_j + \bar{x}^2)$$

und dann die Mittelungsoperation auf die drei Summanden in der Klammer einzeln anzuwenden

$$= \frac{1}{n}\sum x_j^2 - 2\bar{x} \cdot \frac{1}{n}\sum x_j + \frac{1}{n}\sum \bar{x}^2$$

$$= \frac{1}{n}\sum x_j^2 - 2\bar{x} \cdot \bar{x} + \frac{1}{n} n\bar{x}^2$$

$$= \frac{1}{n}\sum x_j^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum x_j^2 - \bar{x}^2 . \qquad ❑$$

Man muss also nicht von jedem einzelnen Wert x_j das arithmetische Mittel abziehen, sondern man kann direkt die quadrierten Werte mitteln und anschließend das Quadrat des arithmetischen Mittels abziehen. In Kurzschreibweise ausgedrückt:

$$s_X^2 = \overline{x^2} - \bar{x}^2 \ . \tag{2-19a}$$

5. Obiger Sachverhalt ist nur ein Spezialfall (mit $d = 0$) des folgenden STEINERschen **Verschiebungssatzes**. Für jedes konstante d gilt:

$$\frac{1}{n}\sum (x_j - \bar{x})^2 = \frac{1}{n}\sum (x_j - d)^2 - (\bar{x} - d)^2 \tag{2-20}$$

Dabei ist $(\bar{x} - d)$ die „Verschiebung“ vom Mittelwert weg.

6. Berechnung der Gesamtvarianz s^2_{ges} aus den Gruppenvarianzen s_i^2: Die statistische Reihe X mit n Elementen sei in $m < n$ disjunkte statistische Teilreihen oder

Gruppen	$X_1,\ X_2,\ \cdots,\ X_m$	
mit jeweils	$n_1,\ n_2,\ \cdots,\ n_m$	Elementen
und den Mittelwerten	$\bar{x}_1,\ \bar{x}_2,\ \cdots,\ \bar{x}_m$	
und den Varianzen	$s_1^2,\ s_2^2,\ \cdots,\ s_m^2$	

zerlegt worden. Daraus errechnet sich die Gesamtvarianz als

$$s^2_{\text{ges}} = \frac{1}{n}\sum_{j=1}^{m} n_j s_j^2 + \frac{1}{n}\sum_{j=1}^{m} n_j (\bar{x}_j - \bar{x})^2 . \tag{2-21}$$

Der 1. Summand heißt ***innere Varianz*** oder interne Varianz. Sie ist das gewichtete Mittel aus den Varianzen *innerhalb* der m Gruppen. Der 2. Summand heißt die ***äußere Varianz*** oder externe Varianz. Sie ist die Varianz der Gruppenmittelwerte $\bar{x}_i$ ($i = 1, \cdots, m$), also die Varianz *zwischen* den Gruppen.

Minimaleigenschaft der Varianz

Da im Verschiebungssatz

$$\frac{1}{n}\sum (x_j - \bar{x})^2 = \frac{1}{n}\sum (x_j - d)^2 - (\bar{x} - d)^2$$

der Term $(\bar{x} - d)^2$ nie negativ sein kann, gilt für jedes $d \neq \bar{x}$ stets

$$\frac{1}{n}\sum (x_j - \bar{x})^2 < \frac{1}{n}\sum (x_j - d)^2 ,$$

das heißt, die mittlere quadratische Abweichung vom arithmetischen Mittel $\bar{x}$ ist kleiner als die mittlere quadratische Abweichung von irgendeinem anderen Wert (Minimaleigenschaft).

Mit n multipliziert, erhält man für die **Summe der quadrierten Abweichungen** von irgendeinem d:

$$\text{SQA}(d) := \sum (x_j - d)^2 \geq \sum (x_j - \bar{x})^2 .$$

Somit könnten wir das arithmetische Mittel auch anders definieren, nämlich als diejenige Zahl d, die SQA minimiert. Das ist das ***Prinzip der kleinsten Quadrate***:

$$\text{SQA}(d) \xrightarrow[d]{} \text{Minimum} \tag{2-22}$$

Das Minimum von SQA finden wir durch Differentiation nach d und Nullsetzen:

$$\frac{\mathrm{d\,SQA}}{\mathrm{d}\,d} = \sum_{j=1}^{n} 2(x_j - d)(-1) = 0 \tag{2-23}$$

$$\sum (x_j - d) = 0$$

$$\sum x_j - \sum d = \sum x_j - nd = 0$$

$$d = \frac{1}{n}\sum x_j$$

und erhalten als Lösung

$$d = \bar{x}\ .$$

Variationskoeffizient

Definition: Der Quotient aus Standardabweichung und Absolutbetrag des Mittelwertes einer statistischen Reihe mit $\bar{x} \neq 0$

$$VK_X := \frac{s_X}{|\bar{x}|} \tag{2-24}$$

heißt ***Variationskoeffizient***.

Der Variationskoeffizient ist ein relatives Maß. Er misst die Streuung *relativ zum Niveau bzw. zur absoluten Größenordnung* der statistischen Reihe.

Beispiel [14] Der Aktienkurs der Volkswagen-Aktie wies in einem Zeitraum von 250 Handelstagen bei einem Mittelwert von $\bar{x} = 174.56$ Euro eine Standardabweichung von $s_X = 10.28$ Euro auf. Für den gleichen Zeitraum ermittelt man für die Aktie der BMW AG eine Standardabweichung von $s_Y = 7.02$ Euro bei einem Mittelwert von $\bar{y} = 55.44$ Euro. Die beiden Variationskoeffizienten betragen

$VK_X = 10.28/174.56 = 0.0589$ für VW und
$VK_Y = 7.02/55.44 = 0.1266$ für BMW.

Somit streute die BMW-Aktie trotz geringerer Standardabweichung relativ stärker.

Man verwendet den Variationskoeffizienten auch gerne als Maß für die **Volatilität eines Aktienkurses**. Die Volatilität gilt als Einschätzung des zukünftigen Risikos einer Anlage in dieser Aktie.

2.7 Quantile

Eigentlich sind die Quartile nur ein Spezialfall der allgemeiner definierten Quantile.

Definition: Eine Zahl $x_{[q]}$ mit $0 < q < 1$ heißt ***q*-Quantil**, wenn sie die statistische Reihe X so aufteilt, dass mindestens $100 \cdot q\,\%$ ihrer Beobachtungswerte kleiner oder gleich $x_{[q]}$ sind und gleichzeitig mindestens $100 \cdot (1-q)\,\%$ größer oder gleich $x_{[q]}$ sind, also

$$\text{relH}(X \le x_{[q]}) \ge q \quad \text{und} \quad \text{relH}(X \ge x_{[q]}) \ge 1-q \tag{2-25}$$

Somit wäre das untere Quartil Q_1 ein 25%-Quantil, das obere Quartil Q_3 ein 75%-Quantil und der Median ein 50%-Quantil

$$\begin{aligned} Q_1 &= x_{[0.25]} && \text{unteres Quartil} \\ Q_2 &= x_{[0.50]} = x_{\text{Med}} && \text{Median} \\ Q_3 &= x_{[0.75]} && \text{oberes Quartil .} \end{aligned}$$

Neben dem Median und den Quartilen sind noch die ***Dezile***

$$x_{[0.1]},\ x_{[0.2]},\ x_{[0.3]},\ \cdots,\ x_{[0.9]}$$

und die ***Perzentile***

$$x_{[0.01]},\ x_{[0.02]},\ x_{[0.03]},\ \cdots,\ x_{[0.99]}$$

beliebte Quantile. Da bei *stetig approximierten Verteilungsfunktionen* für die q-Quantile stets gilt, dass

$$H(x_{[q]}) = q\,,$$

findet man die Quantilswerte auch sehr leicht aus der Verteilungsfunktion, indem man nachschaut, bei welchem Argumentwert sie den Funktionswert q hat, also wenn man die Verteilungsfunktion *umkehrt*, das heißt

$$x_{[q]} = H^{-1}(q)\,.$$

Das geht auch bei *treppenförmigen Verteilungsfunktionen* meistens gut und man findet einen wohldefinierten in der statistischen Reihe vorkommenden Wert an einer Sprungstelle. Nur wenn man dabei genau auf einer Treppenstufe von H landet, ist die Umkehrfunktion nicht eindeutig bestimmt: Man hätte gewissermaßen alle Werte zwischen den fraglichen benachbarten Sprungstellen x_i und x_{i+1} zur Auswahl. In der Tat ist in diesem Fall *jeder Wert*

$$x_i \le x_{[q]} \le x_{i+1}$$

zwischen den beiden ein q-Quantil! Um einen *eindeutigen Wert* zu erhalten, wählt man dann in der Regel das arithmetische Mittel aus beiden. Zwar haben diskrete Verteilungsfunktionen im streng mathematischen Sinne keine Umkehrfunktion, aber man kann graphisch vorgehen. BILD 2.3 veranschaulicht dies: Beim Median hat man hier Glück und trifft auf eine Sprungstelle, bei den beiden Quartilswerten muss gemittelt werden.

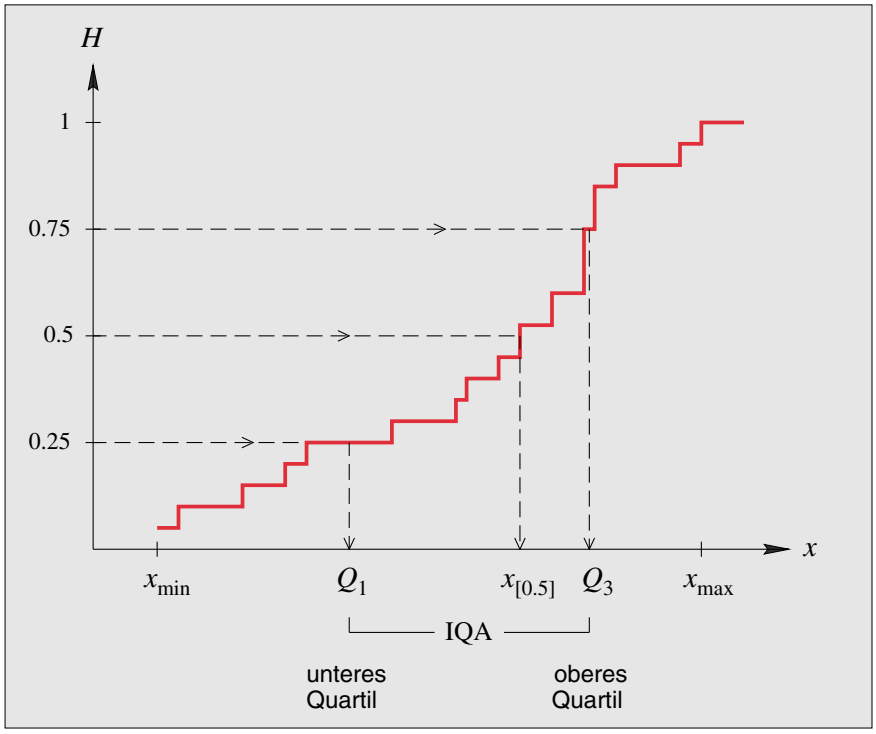


BILD 2.3 Quantile und wie man sie aus der Verteilungsfunktion ableitet

Natürlich findet man das q-Quantil auch ohne den Umweg über den Graphen der Verteilungsfunktion. Ist n der Umfang der statistischen Reihe mit den der Größe nach geordneten und nummerierten Werten und erweist sich das Produkt nq als eine ganze Zahl, so ist einfach

$$x_{[q]} = \frac{1}{2}(x_{nq} + x_{nq+1}) .$$

Ist jedoch nq keine ganze Zahl, nimmt man die *nächstgrößere ganze* Zahl $<nq>$. Im Beobachtungswert mit dieser Indexnummer hat man dann

$$x_{[q]} = x_{<nq>}$$

direkt das q-Quantil.

Beispiel [15] In der nachfolgenden Tabelle sind die größten Industrieunternehmen in Deutschland mit dem jeweiligen Umsatz (2014, in Mio. Euro) aufgeführt.

Unternehmen	Umsatz	Unternehmen	Umsatz
20 Volkswagen	202 458	10 ThyssenKrupp	41 304
19 Daimler AG	192 872	9 Continental	34 506
18 E.ON	111 556	8 Shell Deutschland	25 565
17 BMW	80 400	7 Hochtief	22 099
16 BASF	74 326	6 EnBW	21 003
15 Siemens	71 920	5 ZF Friedrichshafen	18 415
14 Bosch	48 951	4 Ford GmbH	17 400
13 RWE	48 468	3 Linde AG	17 047
12 Deutsche BP	47 952	.2 Henkel	16 428
11 Bayer	42 239	1 Vattenfall	14 654

Quelle: Wikipedia

Weil $nq = 20 \cdot 0.75 = 15$ eine ganze Zahl ist, wäre

$$x_{[0.75]} = \frac{1}{2}(71\,920 + 74\,326) = 73123$$

ein 75%-Quantil. 75% der genannten Industrieunternehmen liegen mit ihrem Umsatz nicht darüber. Der Median $x_{[0.5]} = 41\,771$ ist kleiner als der Mittelwert $\bar{x} = 57\,428$. Ein 80%-Quantil finden wir in

$$x_{[0.80]} = \frac{1}{2}(74\,326 + 80\,400) = 77\,363\,.$$

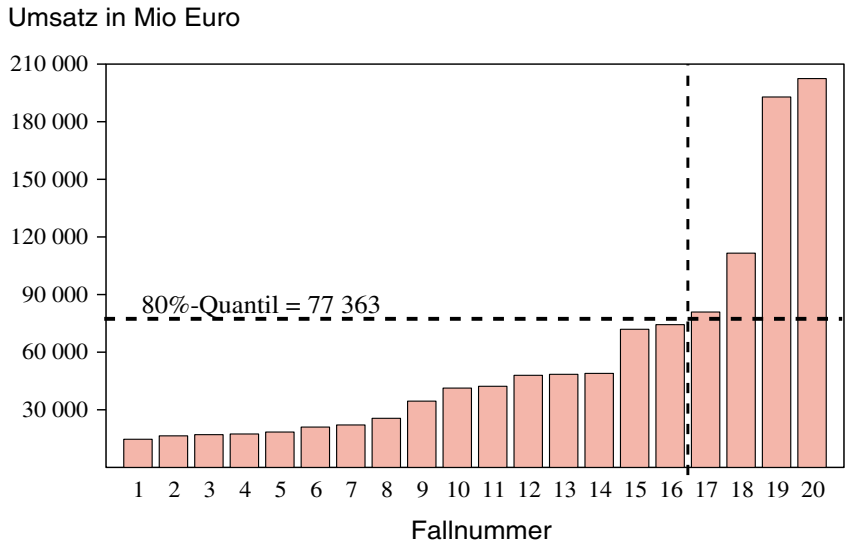


BILD 2.4 Umsatz von 20 Industrieunternehmen und 80%-Quantil

Fünf-Punkte-Zusammenfassung und Box-Plot

In der Tat kann man durch die Angabe von wenigen Werten die Gestalt einer Häufigkeitsverteilung recht gut analysieren. In der Praxis beliebt ist die sogenannte ***Fünf-Punkte-Zusammenfassung***

$$(x_{\min},\ x_{[0.25]},\ x_{\mathrm{Med}},\ x_{[0.75]},\ x_{\max})\,.$$

Sie teilt den Datensatz in vier Teile, so dass in jedem Teil etwa ein Viertel der Beobachtungswerte liegt. Sie enthält den Median als Lagemaß, die Spannweite und den Interquartilsabstand IQA als Streuungsmaße. Besonders in der graphischen Darstellung als ***Box-Plot***

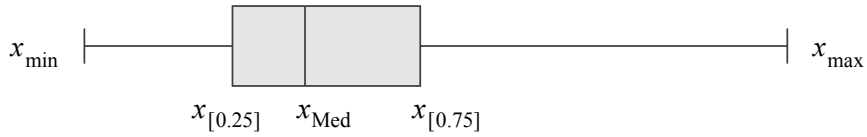


BILD 2.5 Box-Plot einer linkssteilen Verteilung

vermitteln diese fünf Werte einen guten Gesamteindruck von der Verteilung. Die Box wird von den Quartilen begrenzt und vom Median geteilt. Ihre Länge entspricht dem Interquartilsabstand IQA und sie enthält die Hälfte der Beobachtungswerte. Auf den ersten Blick kann man erkennen, ob die Verteilung symmetrisch ist. Je länger die beiden sogenannten ***Whiskers*** („Antennen", „Fühler") im Vergleich zur Box sind, desto geringer ist die *Wölbung*[1] der Verteilung.

Zäune und Ausreißer

Zuweilen verwendet man Box-Plots in etwas modifizierter Form. Liegt das Minimum oder das Maximum zu weit von der Box entfernt, könnten die Whiskers recht lang werden. Man begrenzt ihre Länge dann durch die ***Zäune***

$$z_{\mathrm{unten}} = x_{[0.25]} - 1.5 \cdot \mathrm{IQA}$$

$$z_{\mathrm{oben}} = x_{[0.75]} + 1.5 \cdot \mathrm{IQA}$$

und erklärt die weiter außen liegenden Beobachtungswerte zu ***Ausreißern***. Die Ausreißer werden einzeln eingezeichnet.

---

1 Zur Definition der Wölbung vgl. Abschnitt **9.8**

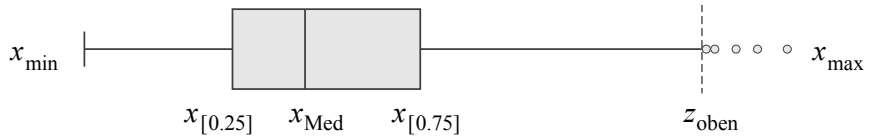


BILD 2.6 Box-Plot einer linkssteilen Verteilung mit oberem Zaun

Beobachtungswerte, die weiter als das Anderthalbfache des IQA abseits vom 50%-Pulk der Werte liegen, als Ausreißer zu definieren, ist natürlich nicht unproblematisch. Will man damit singuläre Werte kennzeichnen, die gar nicht zum Gesamtbild passen, untypisch weit von der Masse der Daten entfernt sind oder eventuell auf Messfehlern beruhen? Manchmal bezeichnet man erst diejenigen Werte, die weiter als 2.5 Interquartilsabstände entfernt liegen, als *strenge* Ausreißer, die anderen als *milde*.

Boxplots ermöglichen einen raschen und übersichtlichen Vergleich mehrerer Verteilungen; sie lassen sich leicht unter- oder nebeneinander auf derselben Skala darstellen, wie das Beispiel [16] zeigt.

Beispiel [16] Die Verteilungen der Arbeitskosten in den einzelnen EU-Beitrittsländern unterscheiden sich sehr stark, sowohl in der durchschnittlichen Lohnhöhe als auch in der Streuung.

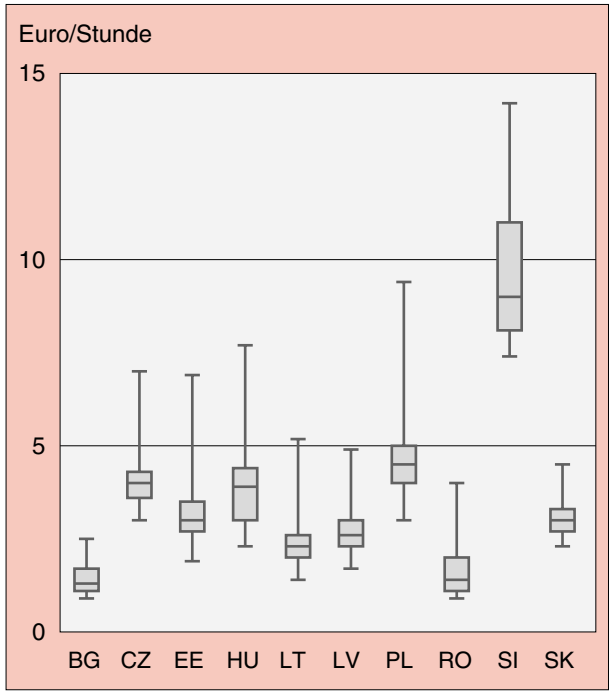


BILD 2.7
Arbeitskosten bei EU-Beitrittskandidaten im Jahre 2000.

Quelle: Eurostat

2.8 Konzentrationsmaße

Die Streuungsmaße geben Auskunft darüber, wie stark oder wie weit die einzelnen Merkmalsausprägungen einer statistischen Reihe voneinander oder von einem Zentralwert abweichen. Auch das Konzept der Konzentration will eine ähnliche Struktureigenschaft empirischer Verteilungen beschreiben, allerdings unter einem anderen Gesichtspunkt. Die Frage nach der Konzentration richtet sich auf den Anteil, den einzelne statistische Einheiten ω_i an der gesamten Summe der Merkmalswerte in einer statistischen Reihe

$$S := x_1 + x_2 + \cdots + x_n$$

haben. Es könnte ja sein, dass sich die Merkmalssumme S zu einem großen Teil oder fast vollständig auf nur wenige Merkmalsträger „konzentriert". Wenn in einer Branche 5% der Firmen 80% des Branchenumsatzes machen, liegt eine hohe Umsatzkonzentration vor. Wenn in einem Land 3% der Einwohner über 50% des Gesamtvermögens verfügen, wird man das für eine hohe Vermögenskonzentration oder sehr ungleiche Vermögensverteilung halten. Unter Konzentration versteht man also eine *Ungleichheit in der Verteilung der Merkmalssumme auf die Merkmalsträger*.

Absolute Konzentration

Ein hoher Anteil der Merkmalssumme S entfällt auf eine kleine *absolute Anzahl* von Merkmalsträgern. Beispiel: Vier Firmen machen 62% des Gesamtumsatzes des Lebensmitteleinzelhandels.

Relative Konzentration

Ein hoher Anteil der Merkmalssumme S entfällt auf einen *kleinen Anteil* der Merkmalsträger. Oder auch umgekehrt ausgedrückt: Ein geringer Anteil der Merkmalssumme entfällt auf einen hohen Anteil der Merkmalsträger. Beispiel: Nur knapp 4% des Einkommensteueraufkommens in Deutschland wurden 2007 vom unteren Drittel der Steuerpflichtigen aufgebracht.

Von Konzentration in diesem Sinne zu sprechen, geht natürlich nur dann, wenn Summen von Merkmalsausprägungen und auch die Gesamtsumme S sachlich sinnvoll interpretiert werden können (extensives Merkmal).

Die statistischen Phänomene, die man mit der Fragestellung der Konzentration analysieren möchte, können sowohl betriebswirtschaftlicher als auch wirtschaftspolitischer Natur sein. Oft gilt Konzentration als unerwünscht oder wird gar als ungerecht oder unmoralisch empfunden. Man denke nur an die Verteilung des Einkommens, der