wi
wirtschaft

Josef Schira

Statistische Methoden der VWL und BWL

Theorie und Praxis

3., aktualisierte Auflage

PEARSON
Studium

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 2008 in Bayern.

[3] Verkehrsunfälle mit Personenschaden im Jahre 1999 in Deutschland.

[4] Studenten in der Vorlesung am Mittwoch, den 23.04.2008 um 14.15 Uhr, im Audimax der Universität Duisburg-Essen, Campus Duisburg.

[5] Angemeldete Konkurse von Bauunternehmungen im April 2009 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

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

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 meist 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 hat 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* Merkmale zu gewährleisten. Denn man nimmt an, dass das zu untersuchende Merkmal in einem gewissen „statistischen Zusammenhang" mit diesen anderen Merkmalen steht. Es gibt unterschiedliche *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

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

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

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

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***

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

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

$\sum h_i = 1$ (1-7)

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.

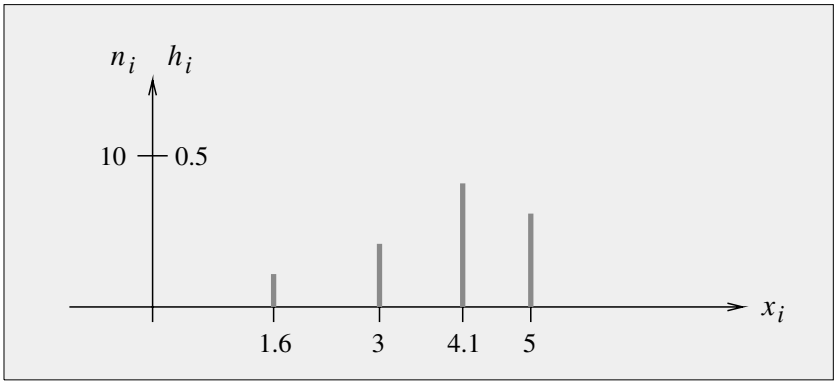


BILD 1.1 Häufigkeitsverteilung

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**:

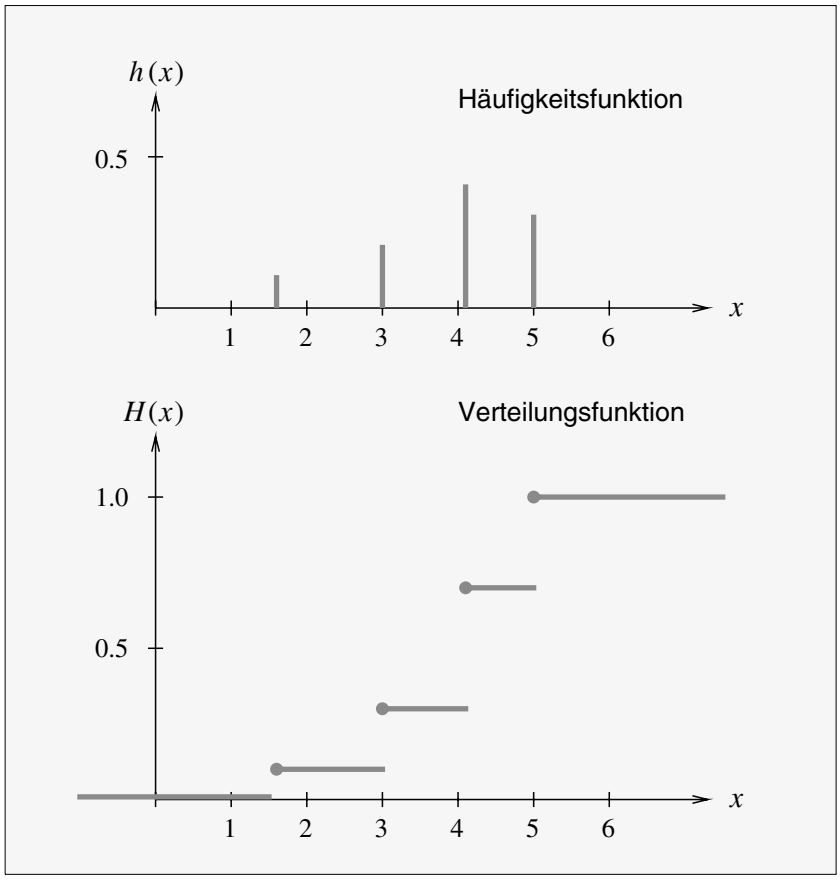


BILD 1.2 Häufigkeitsfunktion und Verteilungsfunktion

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.

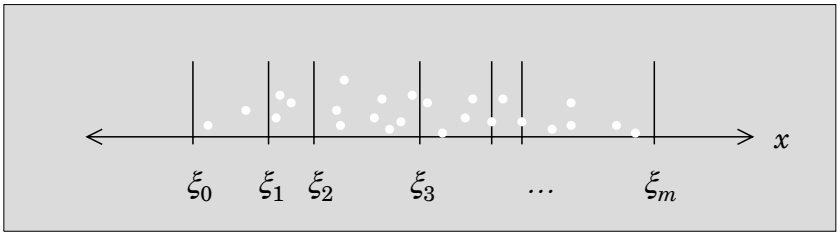


BILD 1.3 Bildung von Größenklassen

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

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

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

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

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.

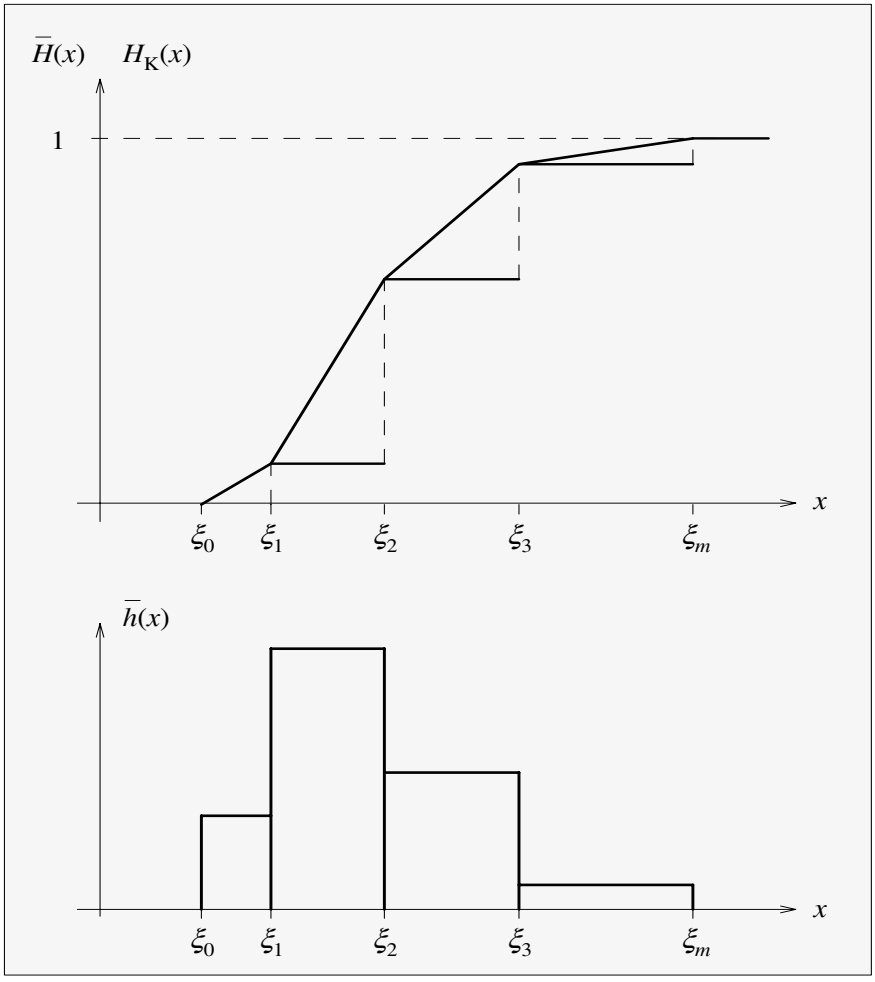


BILD 1.4 Approximierender Polygonzug und Histogramm

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

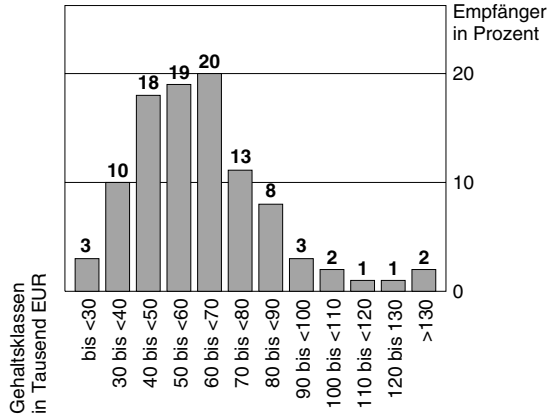


BILD 1.5 Verteilung der jährlichen Gesamtbezüge von Führungs- und Fachkräften des Außendienstes

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 Brasilien bemerkbar machten.

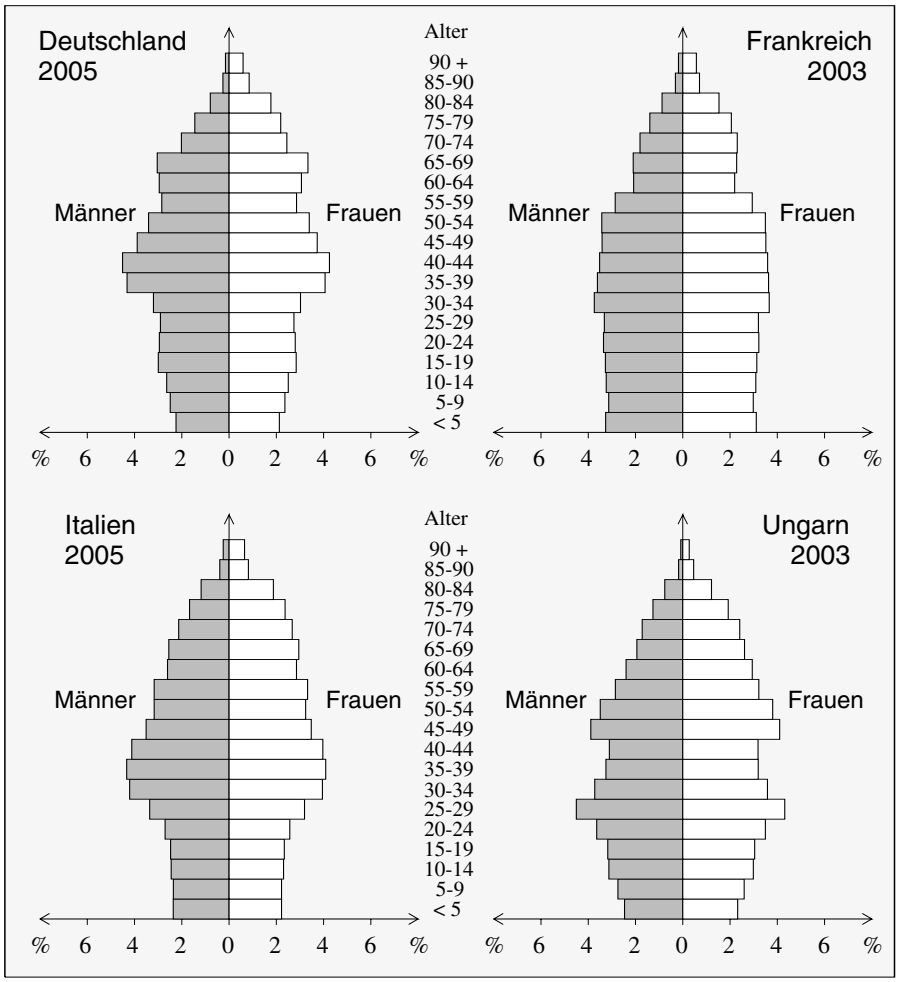


BILD 1.6 Bevölkerungspyramiden alter Länder: Europa

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.

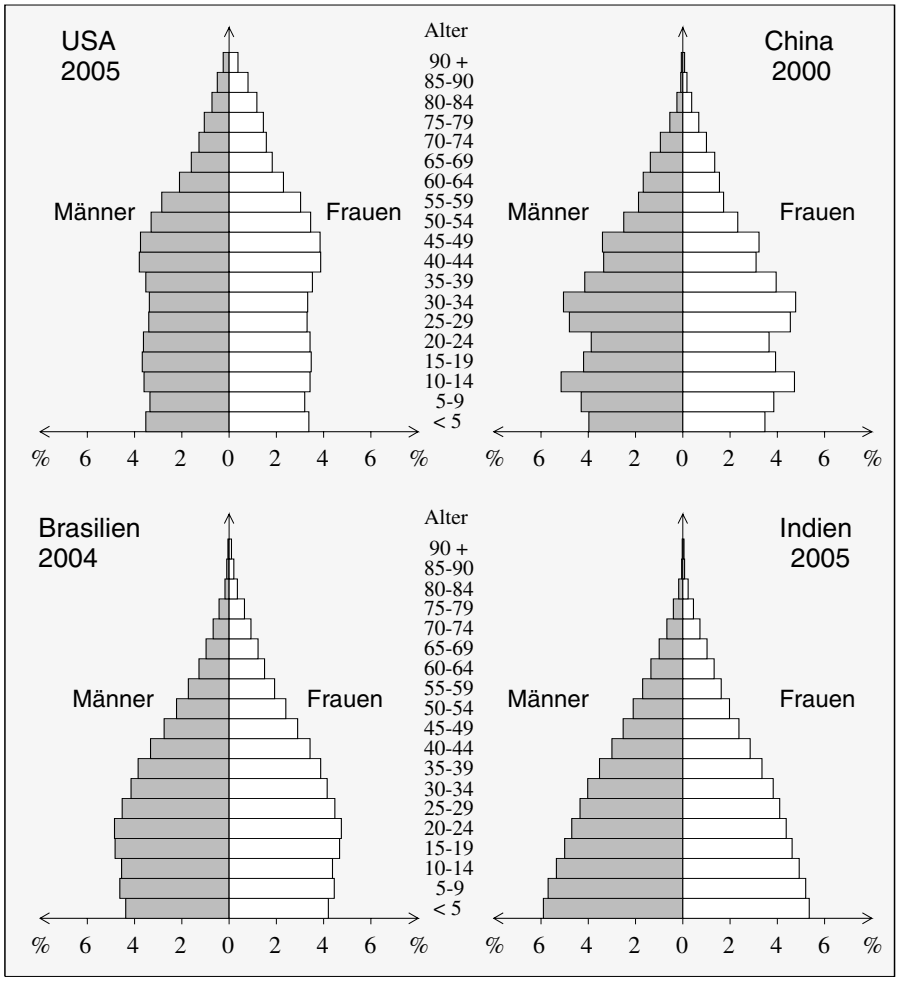


BILD 1.7 Bevölkerungspyramiden anderer Länder

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

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

und

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

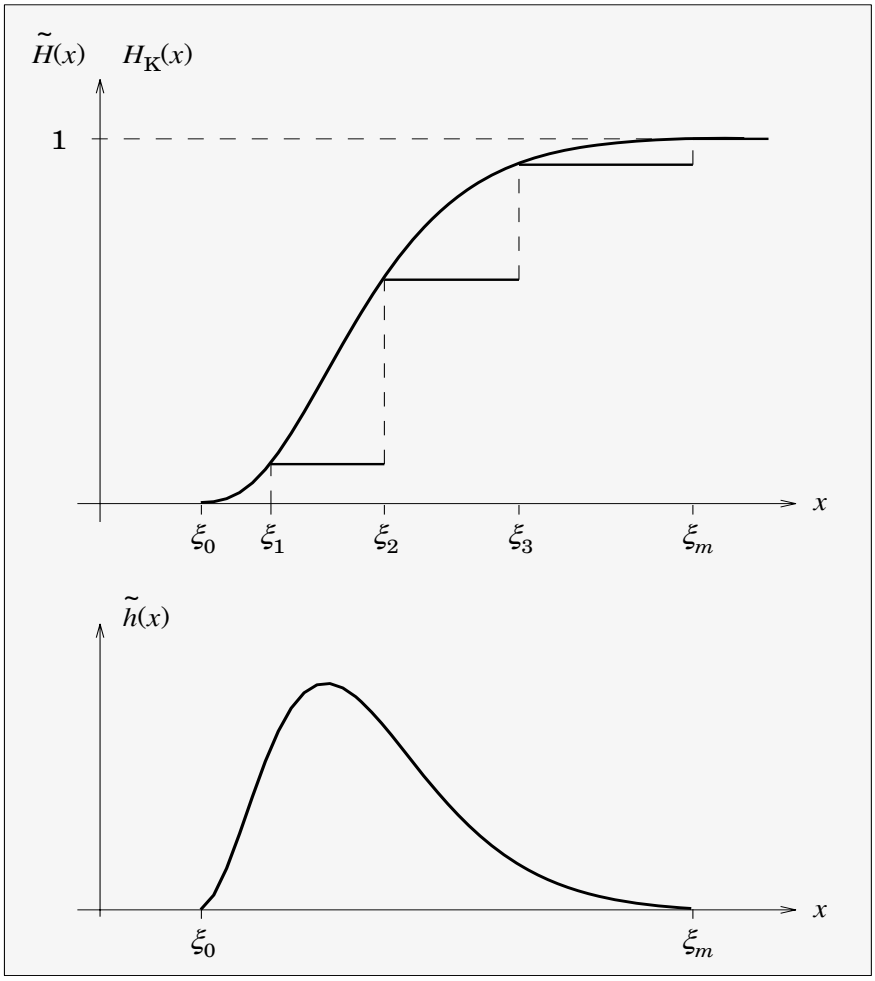


BILD 1.8 Approximierende glatte Kurven

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.

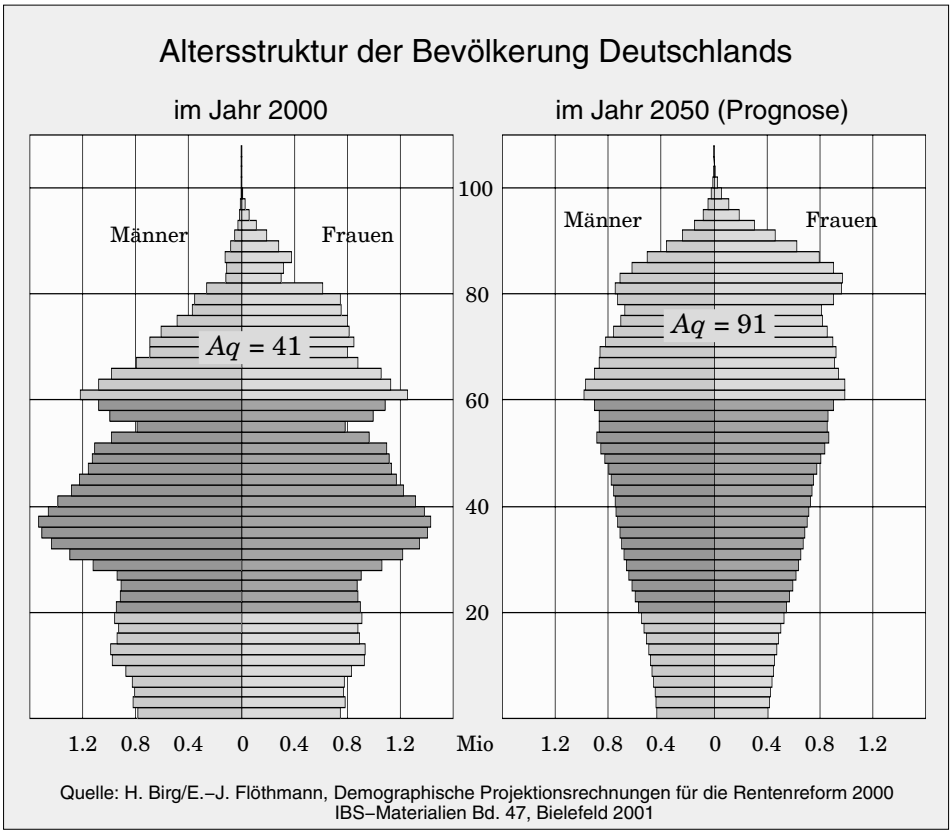


BILD 1.9 Bevölkerungspyramiden für Deutschland

So standen 100 Menschen der ökonomisch aktiven Altersgruppe 20 bis 60 im Jahre 2000 rund 41 über Sechzigjährige gegenüber. 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

Hochstädter, Dieter: *Statistische Methodenlehre*, 8. Aufl., Frankfurt am Main: Harri Deutsch, 1996

Krämer, Walter: *So lügt man mit Statistik*, 4. Aufl., München: Piper, 2003

Schlittgen, Rainer: *Einführung in die Statistik: Analyse und Modellierung von Daten*, 9. Aufl., München, Wien: Oldenbourg, 2003, Kapitel 1 und 2

Schwarze, Jochen: *Grundlagen der Statistik I*, 10. Aufl., Herne: Neue Wirtschaftsbriefe, 2005

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.

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

a) Skizzieren Sie die Verteilungsfunktion.

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 die approximierende Verteilungsfunktion als ein Polynom 3. Grades

$$\widetilde{H}(x) = ax^3 + bx^2 + cx$$

im Intervall [0,10]. Berechnen Sie dazu die Koeffizienten a, b und c.

d) Wie lautet die approximierende Dichtefunktion $\tilde{h}(x)$?
Zeichnen Sie sie in das Histogramm ein.

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

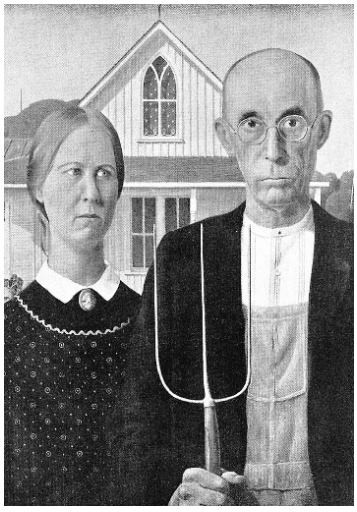
Jahreseinkünfte in Euro von ... bis unter ...	Steuerpflichtige %	Gesamtbetrag der Einkünfte %
unter 2 500	3.1	0.1
2 500 – 5 000	3.7	0.4
5 000 – 7 500	4.3	0.8
7 500 – 10 000	4.4	1.1
10 000 – 12 500	4.3	1.4
12 500 – 25 000	24.2	12.8
25 000 – 37 500	23.0	19.7
37 500 – 50 000	13.6	16.3
50 000 – 125 000	17.5	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 **Diplomnoten.** Ein frischgebackener Master of Arts in Ökonomie 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 Diplomnoten 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 Amerikaner und Deutsche in Durchschnittswerten



	USA	Deutschland
BIP pro Kopf	47 025 $	46 498 $
Arbeitseinkommen	47 688 $	38 626 $
Arbeitsstunden/Jahr	1 804	1 436
Alter	36.7	43.4
Lebenserwartung	78.1	79.3
Kinder pro Frau	2.1	1.4
TV-Konsum pro Tag	3	2
Body-Mass-Index	35.1	25.5
Alkohol Liter/Jahr	8.6	12.0

Quelle: FRANKFURTER ALLGEMEINE SONNTAGSZEITUNG 02.11.2008

a) Sind sie wirklich so viel dicker als wir oder
b) rechnen die Amerikaner das Merkmal Body-Mass-Index in Pounds und Inches? Rechnen Sie um!

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

1.5 c) $a = -0.005333$; $b = 0.096$
$c = -0.3267$

d) $\tilde{h}(x) = -0.016\,x^2 + 0.192\,x - 0.327$

1.6 b) ca. 35 000

1.8 a) nein b) ja