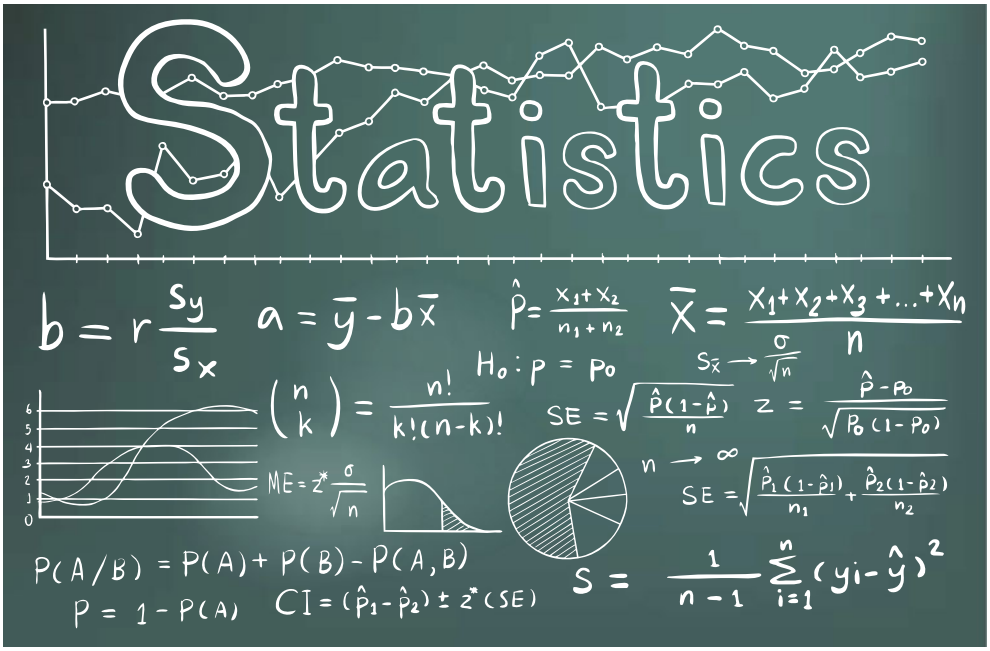
Statistik

IV



Eindimensionale deskriptive Statistik – ein Merkmal aussagekräftig beschreiben

15

Wie kann man große Datenmengen aussagekräftig beschreiben?

Welche grafischen Darstellungen gibt es für Daten?

Welche Kenngrößen gibt es, und was bedeuten sie?

Teil IV

L. Göllmann et al., *Mathematik für Ingenieure: Verstehen – Rechnen – Anwenden*, DOI 10.1007/978-3-662-53867-8_15

In diesem Kapitel lernen wir die wichtigsten Elemente der eindimensionalen deskriptiven Statistik kennen, also des Teilgebiets der Statistik, das sich mit der Beschreibung von eindimensionalen Daten beschäftigt. Im Wesentlichen haben wir es immer mit der gleichen Fragestellung zu tun: Wir stellen uns vor, wir haben eine große Menge an Daten erhoben und möchten nun gerne die wesentlichen Eigenschaften dieser Datenmenge darstellen. Die Rohdaten können wir sicherlich nicht präsentieren, denn sie sind zu wenig aussagekräftig.

Die einfachste Möglichkeit besteht darin, die Daten grafisch darzustellen. Welche Arten der grafischen Aufbereitung es gibt, werden wir nach den unvermeidlichen Fachbegriffen in Abschn. 15.3 kennenlernen.

Eine weitere Möglichkeit der Beschreibung von Daten stellt die empirische Verteilungsfunktion dar. Auf diese werden wir auch in abgewandelter Form wieder in der Wahrscheinlichkeitstheorie treffen, denn das Konzept lässt sich auch dort anwenden.

Darüber hinaus haben sich Statistiker vor uns aber auch schon damit beschäftigt, welche Kenngrößen besonders geeignet sind, um Daten zu beschreiben. Diese Kenngrößen werden in Abschn. 15.6 vorgestellt.

15.1 Was soll das eigentlich mit der Statistik?

Ingenieure brauchen immer mehr statistische Kenntnisse. Qualitätskontrolle, statistische Prozesskontrolle und viele Themen mehr sind ohne Statistik heute nicht mehr denkbar. Trotzdem gibt es noch immer Hochschulen, in denen Statistik kein Pflichtmodul in der Ingenieursausbildung ist. Die Autoren des vorliegenden Werkes sind aber übereinstimmend der Ansicht, dass in ein Mathebuch für Ingenieure auch ein Statistikteil gehört. Beginnen wir mit der folgenden Übersicht über die wichtigsten Teilgebiete der Statistik bzw. Stochastik. Denn genau so ist der vorliegende Teil des Buches aufgebaut:

Die Stochastik besteht im Wesentlichen aus drei Teilgebieten: der deskriptiven Statistik, der Wahrscheinlichkeitsrechnung und der induktiven Statistik. Jeder einzelne Teilbereich ist wichtig und schön, aber ich finde nach wie vor, dass die induktive Statistik die Königsdisziplin der Statistik ist. Aber wie immer im Leben ist es ein langer und steiniger Weg in die Königsklasse. Wir beginnen in diesem Buch in Kap. 15 mit der deskriptiven Statistik. Die deskriptive Statistik hat zur Aufgabe, große Datenmengen für Menschen, insb. für Menschen, die sich nicht so gut mit Statistik auskennen, fassbar zu machen. Man stelle sich dazu (zumindest in der Praxis) Excel-Tabellen mit einigen Zehntausend Einträgen vor, da sollte klar sein, dass man da noch so lange rauf- und runterscrollen können, ohne dass man auch nur einen blassen Schimmer davon hat, was mit diesen Datensätzen los ist. Hier setzt die deskriptive Statistik mit ihren Instrumenten an: Wir werden hier grafische Darstellungen kennenlernen, denn ein Bild sagt mehr als 1000 Formeln, und wir werden einige Kenngrößen besprechen, die von den statistischen Experten als charakteristisch für Datenmengen angesehen werden. Übrigens machen wir es am Anfang nicht zu kompliziert: Wir interessieren uns erstmal nur für einen Sachverhalt, z. B. Gewichte, Noten oder Farben, und analysieren die zugehörigen Datensätze. Im folgenden Kap. 16 wird die Sache dann doppelt (oder quadratisch?) so kompliziert, denn nun schauen wir uns nicht nur einen Sachverhalt an, sondern zwei Sachverhalte gleichzeitig. Nun steht die Frage nach Zusammenhängen und Beeinflussungen im Vordergrund: Wie kriegt man raus, ob Werbeaktivitäten tatsächlich einen Einfluss auf den Umsatz haben? Hängen Haar- und Augenfarbe irgendwie voneinander ab? Haben Noten in Mathematik und Informatik einen Zusammenhang? All das sind Fragestellungen, die wir nach Durcharbeiten dieses Kapitels beantworten können. Denn je nachdem, mit was für Daten wir es zu tun haben, müssen wir vollkommen unterschiedliche Methoden anwenden. Das nun folgende Kap. 17 fällt aus dem Rahmen. Bis dahin haben wir es mit „real existierenden“Daten zu tun gehabt. Nun geht es um theoretische Hirngespinste, nämlich um Wahrscheinlichkeiten. Die können wir nicht anfassen, und es gibt auch Menschen, die das Konzept der Wahrscheinlichkeiten mit dem Argument ablehnen, dass sie auch nicht an den Zufall glauben. Nun, wie man selbst das künftig handhaben will, ist jedem selbst überlassen, aber man sollte das Konzept zumindest verstanden haben, sonst kann man weit verbreitete Argumentationsformen nicht verstehen.

Über Wahrscheinlichkeiten werden wir hier zweierlei lernen: einerseits die Modellierung und den Umgang mit Wahrscheinlichkeiten, ein absolutes Highlight hier der Umgang mit bedingten Wahrscheinlichkeiten. Andererseits lernen wir die wichtigsten Zufallsvariablen und ihre Verteilungen kennen. Ein Thema von hoher praktischer Relevanz, da zufallsbehaftete Vorgänge in freier Wildbahn nur selten ein Namensschild mit der ihnen zugehörigen Verteilung tragen, also müssen wir sie wohl oder übel selbst erkennen und zuordnen können. Und egal ob wir im wirtschaftlichen, im technischen oder im naturwissenschaftlichen Bereich arbeiten, überall stoßen wir auf (mehr oder weniger) zufällige Sachverhalte, die mit ein bisschen Sachverstand viel besser untersucht und kontrolliert werden können.

Naja, und für uns hier ganz pragmatisch sind Wahrscheinlichkeiten ein notwendiges Instrument, um im letzten Kap. 18 endlich zur Krönung der Statistik zu kommen, der induktiven Statistik. Hier lehnen wir uns mal richtig weit aus dem Fenster und gehen im Vergleich zur deskriptiven Statistik einen Schritt weiter: Anstatt uns darauf zu beschränken, die Datenmengen zu beschreiben, die uns vorliegen, erlauben uns die Methoden der induktiven Statistik, ausgehend von den vorliegenden Daten, also einer Stichprobe, Rückschlüsse auf die Allgemeinheit zu ziehen. Und das geht nachvollziehbar für Dritte nur, wenn man das Konzept der Wahrscheinlichkeiten akzeptiert (oder zumindest benutzt). Hierbei sind zwei Problemstellungen von herausragender Wichtigkeit.

Die erste Problemstellung besteht darin, dass man oft bestimmte Größen in einer Situation (vorurteilsfrei) schätzen möchte. Wie groß ist der zu erwartende Inhalt einer Lieferung von Ravioli-Dosen, wenn man nicht jede Dose öffnen kann? Wie viele Silvester-Raketen einer Lieferung funktionieren ordnungsgemäß? Man kann wohl kaum jede abfeuern, wenn man noch welche verkaufen möchte … Zwei Möglichkeiten gibt es hier, die Werte zu schätzen: Entweder wir wollen eine Zahl als Ergebnis erhalten, dann benötigen wir einen Punktschätzer. Oder uns

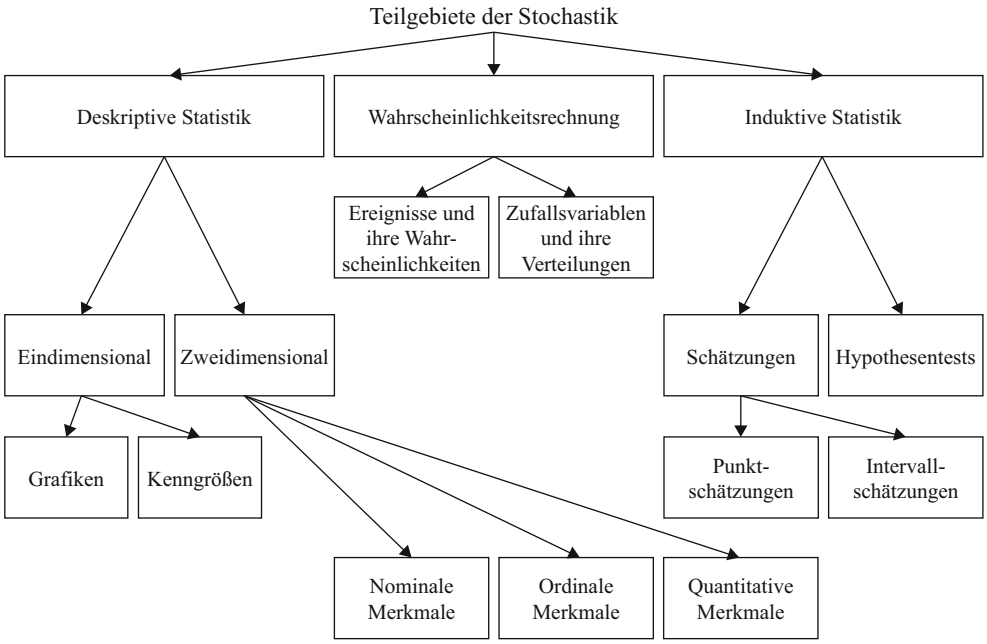


Abb. 15.1 Teilgebiete der Stochastik

genügt es, wenn wir mit einer hinreichenden Sicherheit sagen können, dass der gesuchte Wert zwischen a und b liegt. Dann suchen wir einen Intervallschätzer.

Die andere Problemstellung betrifft die Situation, wenn wir bereits eine Vermutung haben, die die uns unbekannte Größe betrifft. Wenn man als Leiter der Wareneingangskontrolle einer Fabrik vermutet, dass der Hersteller der Zuliefererteile diese zu klein gefertigt hat, dann ist er nicht mehr vorurteilsfrei, sondern er hat eine Hypothese bezüglich der Größe, die er überprüfen bzw. testen möchte. Auch das ist eine induktive Fragestellung, weil er nur eine Stichprobe untersucht, aber Schlussfolgerungen auf die gesamte Lieferung ziehen will.

15.2 Grundbegriffe

Wenn man sich mit deskriptiver Statistik beschäftigt, muss man einige Fachtermini kennen, die im allgemeinen deutschen Sprachgebrauch oftmals nicht sinngemäß verwendet werden.

Definition: Merkmal, Merkmalsausprägung und Merkmalsträger

Unter einem **Merkmal** versteht man einen Sachverhalt, den man mit statistischen Methoden untersucht.

Unter einer **Merkmalsausprägung** versteht man eine Konkretisierung des Sachverhalts, den man mit statistischen Methoden untersucht.

Unter einem **Merkmalsträger** versteht man ein zu untersuchendes Objekt, das das Merkmal aufweist, das man untersucht.

Betrachten wir hierzu ein Beispiel:

Beispiel

Nehmen wir an, wir untersuchen die Länge der innerhalb eines Tages bei einer Produktion anfallenden Verschnittreste. Dann untersuchen wir das Merkmal „Länge“. Mögliche Merkmalsausprägungen sind z. B. 17.2 mm, 18.6 mm etc. Merkmalsträger sind die Verschnittreste des Produktionsprozesses. ◄

Bleiben wir direkt bei dem Beispiel. Wie beschrieben wurde nur die Länge der innerhalb eines Tages angefallenen Verschnittreste bei unserer Produktion untersucht, nicht die Länge jedes jemals existierenden Verschnittrestes auf der Welt. Das hat den Vorteil, dass wir eine klar abgegrenzte Menge untersuchen.

Definition: Grundgesamtheit

Die Menge aller zu betrachtenden Merkmalsträger heißt **Grundgesamtheit**.

Das klingt zwar sehr logisch, der Teufel steckt aber im Detail, hier in der genauen Abgrenzung. Denn nicht immer ist klar, was

nun zur Grundgesamtheit gehören soll und was nicht. Nehmen wir das weibliche Kaufverhalten. Dann stellt sich die Frage, sollen alle Frauen dieser Welt in der Grundgesamtheit sein oder nur die deutschen oder nur die erwachsenen Frauen? Und es kommt ein zweites Problem in der Praxis hinzu: Im vorliegenden Beispiel kann man sicherlich noch jeden einzelnen Verschnittrest vermessen, die meisten Grundgesamtheiten sind allerdings so groß, dass man nicht mehr jeden Merkmalsträger untersuchen kann. Man beschränkt sich auf einige und versucht dann Rückschlüsse auf die Grundgesamtheit zu ziehen. (Wie man mit vernünftigen Methoden Rückschlüsse zieht, ist Forschungsgebiet der induktiven Statistik. Das lernen wir in Kap. 18.)

Definition: Stichprobe

Unter einer **Stichprobe** versteht man eine zufällig ausgewählte Teilmenge der Grundgesamtheit.

Die Anzahl der Elemente der Stichprobe heißt **Stichprobenumfang** und wird in der Regel mit n bezeichnet.

Auch hier stecken, wie oft, die praktischen Schwierigkeiten in scheinbaren Kleinigkeiten, hier im Wort „zufällig“. Das bedeutet nämlich, dass jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit besitzt, für die Stichprobe ausgewählt zu werden. Und das ist in der Praxis oft ein Problem. Bei Schrauben, die man in einen Eimer wirft, umrührt und mit verschlossenen Augen einige auswählt, mag das ja noch funktionieren, aber wie zieht man zufällig Menschen? Früher gab es immer den Tipp, man sollte sich das Telefonbuch nehmen, irgendeine Seite aufschlagen und mit dem Finger auf einen Namen tippen. Das geht aber nicht, da immer mehr Menschen gar nicht im Telefonbuch stehen und diese logischerweise gar nicht ausgewählt werden können.

Wie geht die Praxis mit dem Problem um? Praktiker denken sich komplizierte Kriterien aus, die Interviewer bei der Auswahl ihrer Befragungsteilnehmer beachten müssen ($x\,\%$ Frauen, $y\,\%$ aus jeder Altersklasse, $z\,\%$ Beamte etc.) und nennen das Ganze dann repräsentative Stichprobe. Dies ist erstens natürlich alles nicht zufällig, also immer noch keine Stichprobe, und zweitens gibt es noch ein großes logisches Problem. Nehmen wir an, wir wollen das Wahlverhalten der deutschen Wahlberechtigten bei der nächsten Bundestagswahl prognostizieren. Logischerweise können wir nicht alle Wahlbeteiligten befragen, also müssen wir eine Stichprobe ziehen. Was machen die Meinungsforschungsinstitute? Sie befragen Männer, Frauen, Alte, Junge, Städter und Landbevölkerung etc. Hunderte von Kriterien müssen erfüllt sein, und dann jubeln sie: Die Stichprobe ist repräsentativ. Man fragt sich aber: Wofür ist diese Stichprobe repräsentativ? Für alle Wahlberechtigten? Stimmen die Menschen in der Stichprobe also genauso ab wie alle Wahlberechtigten? Woher wissen die Meinungsforscher das? Die Umfrage wird doch gerade durchgeführt, weil man das Wahlverhalten nicht kennt! Klingt irgendwie, als ob sich die Katze in den Schwanz beißt, oder?

Soviel zu den absoluten Grundlagen. Es ist nicht erstaunlich, dass nicht alle Merkmale gleich sind und daher auch unterschiedlich behandelt werden müssen. Erfahrungsgemäß fällt es Studierenden etwas schwer, die einzelnen Merkmalsarten auseinanderzuhalten. Dabei ist es ganz einfach, wenn man sich immer die richtigen Fragen stellt. Schauen wir uns das im Einzelnen an:

Definition: Merkmalsarten

Die erste Frage, die man sich stellen muss, ist die Frage, ob es bei dem vorliegenden Merkmal Sinn macht, einen Durchschnittswert zu bilden. Wenn man sich diese Frage mit Ja beantwortet, hat man es mit einem **quantitativen Merkmal** zu tun. Wenn die Durchschnittsbildung nicht sinnvoll ist, handelt es sich bei dem Merkmal um ein **qualitatives Merkmal**.

Quantitative Merkmale können noch genauer untergliedert werden: Hierzu muss man sich fragen, ob zwischen zwei Ausprägungen immer noch mindestens eine weitere Ausprägung liegen kann. Wichtig ist hierbei das Wörtchen „immer“! Wenn die Antwort Ja ist, handelt es sich um ein **stetiges (oder auch kontinuierliches) Merkmal**, sonst um ein **diskretes Merkmal**.

Auch qualitative Merkmale kann man genauer klassifizieren. Hierzu lautet die Frage, ob eine Reihenfolge im Sinne von „ist besser als“ Sinn macht. Wenn es eine solche Reihenfolge gibt, haben wir ein **ordinales Merkmal**, ansonsten ist das **Merkmal nominal**.

Abb. 15.2 stellt die Fragen und die Folgerungen noch einmal übersichtlich dar.

Beispiel

- Die Veränderung einer Oberfläche durch einen Fertigungsprozess (z. B. durch Verfärbung oder Aufrauhung etc.) ist ein qualitativ nominales Merkmal, da man weder die durchschnittliche Oberflächenfarbe berechnen, noch eine Reihenfolge im Sinne von „ist besser als“ angeben kann.
- Die Länge von Verschnittresten ist definitiv ein quantitatives Merkmal, da man durchaus an der durchschnittlichen Länge interessiert ist. Ob man es als stetiges oder diskretes Merkmal betrachtet, hängt von der Messgenauigkeit ab. Wenn man „nur“ auf den Millimeter genau misst, also z. B. 17 mm, 18 mm etc. handelt es sich streng genommen um ein diskretes Merkmal, denn zwischen 17 mm und 18 mm gibt es keine weitere Merkmalsausprägung mehr. Falls man aber eine (im Prinzip) unendliche Messgenauigkeit annimmt, kann man das Merkmal auch als stetig auffassen. ◄

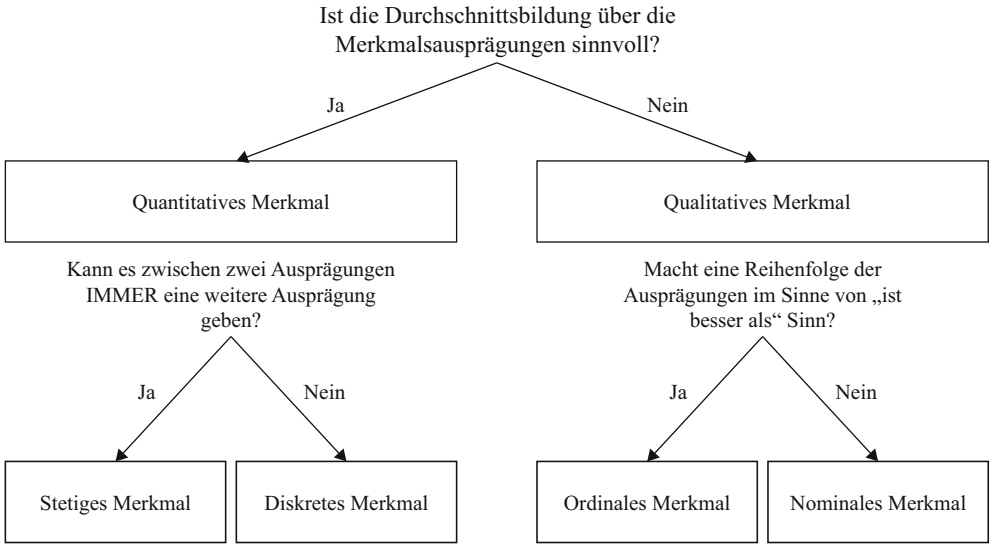


Abb. 15.2 Klassifizierungen von Merkmalen

15.3 Grafische Darstellungen von Daten

Oft versucht man, sich einen Überblick über die Datenlage zu verschaffen, indem man sich die Datenverteilung, also die Häufigkeiten der einzelnen Merkmalsausprägungen, bildlich darstellt. Zum Beispiel gibt es hierzu Balken-, Säulen-, Stab- und Kreissektordiagramme, aber auch die in Abschn. 15.5 behandelten Histogramme gehören streng genommen zu den grafischen Darstellungen. Zuerst brauchen wir aber noch einige Bezeichnungen.

Definition: Absolute und relative Häufigkeiten, Urliste und Häufigkeitstabelle

- Die Anzahl, wie oft eine Merkmalsausprägung vorkommt, nennt man **absolute Häufigkeit** der Merkmalsausprägung bzw. n_i.
- Der Anteil, wie oft eine Merkmalsausprägung vorkommt, nennt man **relative Häufigkeit** der Merkmalsausprägung bzw. h_i, wobei gilt:

$$h_i = \frac{n_i}{n}.$$

- In einer **Urliste** trägt man die erhobenen Merkmalsausprägungen in der Reihenfolge der Erhebung ein. Manchmal werden Urlisten in Strichlisten umgeformt.
- In einer **Häufigkeitstabelle** trägt man in tabellarischer Form folgende Informationen ein:
  - Merkmalsausprägung
  - Absolute Häufigkeit
  - Relative Häufigkeit (prozentuale Anteile)
  - Gradzahl für das Kreissektordiagramm (eher veraltet)

Es gibt ein paar Eigenschaften, die Häufigkeiten immer aufweisen. Manchmal kann man das ausnutzen, um zu überprüfen, ob man sich verrechnet hat:

Häufigkeitsregeln

- Absolute und relative Häufigkeiten sind immer positiv oder 0:

$$n_i \geq 0, h_i \geq 0.$$

- Wenn man die absoluten Häufigkeiten aller Merkmalsausprägungen addiert, erhält man den Stichprobenumfang:

$$\sum_{i=1}^{k} n_i = n$$

für die Merkmalsausprägungen x_i, $i = 1, 2, \ldots, k$.

- Wenn man die relativen Häufigkeiten aller Merkmalsausprägungen addiert, ist das Ergebnis 1:

$$\sum_{i=1}^{k} h_i = 1$$

für die Merkmalsausprägungen x_i, $i = 1, 2, \ldots, k$.

Beispiel

Nehmen wir an, bei der Erhebung der Länge von 10 Verschnittresten ergab sich folgende Urliste: 17.2; 18.6; 19.2; 15.6; 15.3; 16.8; 17.2; 16.7; 16.7; 15.6.

Dann lautet die (geordnete) Häufigkeitstabelle:

Merkm.auspräg.	Abs. Hfgk.	Rel. Hfgk.	Prozent
15.3	1	0.1	10 %
15.6	2	0.2	20 %
16.7	2	0.2	20 %
16.8	1	0.1	10 %
17.2	2	0.2	20 %
18.6	1	0.1	10 %
19.2	1	0.1	10 %

◀

Balkendiagramm

In einem **Balkendiagramm** wird für jede Merkmalsausprägung ein waagerechter Balken einer einheitlichen Dicke gezeichnet. Die Länge des Balkens entspricht der absoluten bzw. relativen Häufigkeit der Merkmalsausprägung.

Beispiel

Für die obige Erhebung sieht das Balkendiagramm wie in Abb. 15.3 aus.

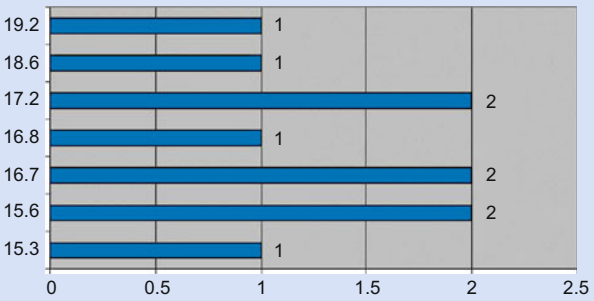


Abb. 15.3 Beispiel eines Balkendiagramms ◀

Säulendiagramm

In einem **Säulendiagramm** wird für jede Merkmalsausprägung eine senkrechte Säule einer einheitlichen Breite gezeichnet. Die Länge der Säule entspricht der absoluten bzw. relativen Häufigkeit der Merkmalsausprägung.

Beispiel

Für die obige Erhebung sieht das Säulendiagramm wie in Abb. 15.4 aus.

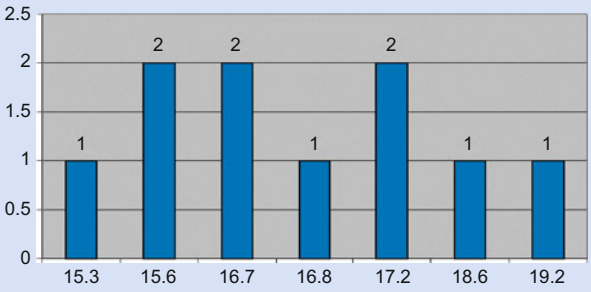


Abb. 15.4 Beispiel eines Säulendiagramms ◀

Stabdiagramm

In einem **Stabdiagramm** wird für jede Merkmalsausprägung eine senkrechte Linie gezeichnet. Meistens schließt die Linie mit einem etwas dickeren Punkt ab. Die Länge der Linie entspricht der absoluten bzw. relativen Häufigkeit der Merkmalsausprägung.

Beispiel

Für die obige Erhebung sieht das Stabdiagramm wie in Abb. 15.5 aus.

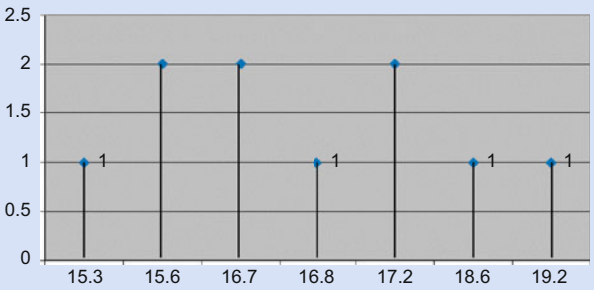


Abb. 15.5 Beispiel eines Stabdiagramms ◀

Liniendiagramm

Ein **Liniendiagramm** darf man nicht in jedem Fall anwenden. Es eignet sich nur für quantitative Merkmale, und zwar streng genommen auch nur für stetige. Zusätzlich muss man eine sinnvolle Interpretation für die Verbindungslinien haben. Daher macht diese Diagrammart für die Länge der Verschnittreste aus unserem Beispiel wenig Sinn.

Beispiel

Am häufigsten werden Liniendiagramme zur Darstellung von Veränderungen eines quantitativen Merkmals

im Zeitablauf verwendet. Hier kann man die Linien als Annäherung der Veränderung zwischen den gemessenen Zeitpunkten interpretieren. Betrachten wir daher die folgende Abbildung, in der der Mietpreis je Quadratmeter einer Lagerfläche in einem Jahr dargestellt wird. Gemessen wurde aber nur einmal pro Monat:

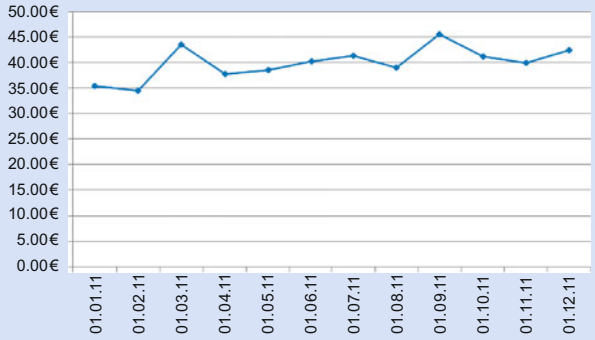


Abb. 15.6 Beispiel eines Liniendiagramms ◀

Streng genommen haben wir es beim Liniendiagramm mit einem Diagramm zu tun, das eigentlich in das folgende Kapitel gehört, denn wir betrachten hier zwei Merkmale gleichzeitig (Zeit und Umsatz).

Kreissektordiagramm

In einem **Kreissektordiagramm** repräsentiert die Dicke eines „Tortenstückes" die absolute bzw. relative Häufigkeit der zugehörigen Merkmalsausprägung. Um ein Kreissektordiagramm zu zeichnen, zeichnet man zuerst einen Kreis der gewünschten Größe und einen senkrechten Radius ein. Anschließend müssen die Winkel der einzelnen Tortenstücke berechnet werden. Dies geschieht mithilfe der Formel:

$$\alpha_i = h_i \cdot 360^\circ \text{ bzw. } \alpha_i = \frac{n_i}{n} \cdot 360^\circ.$$

An den entsprechenden Stellen werden Radien gezogen.

Beispiel

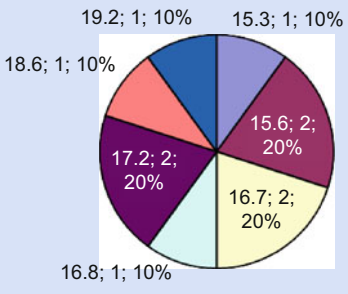


Abb. 15.7 Beispiel eines Kreissektordiagramms

Für die obige Erhebung sieht das Kreissektordiagramm wie in Abb. 15.7 aus. ◀

15.4 Empirische Verteilungsfunktionen

Empirische Verteilungsfunktionen antworten (auf unterschiedliche Weise) auf die Frage: Wie groß ist der Anteil der Merkmalsträger, die höchstens die Merkmalsausprägung x aufweisen? Es gibt daher eine empirische Verteilungsfunktion nur für quantitative Merkmale. Empirische Verteilungsfunktionen sind nicht zuletzt deswegen so wichtig, weil wir mit ihnen üben können, uns für den Anteil ... „höchstens" ... zu interessieren. Das wird vor allem im Zusammenhang mit Wahrscheinlichkeiten wichtig, wo wir auf die (wirklich wahre) Verteilungsfunktion treffen werden. Dort hat man oft keine weiteren Informationen.

Die empirische Verteilungsfunktion beschreibt die Häufigkeitsverteilung eines quantitativen Merkmals

Bislang haben wir uns nur für die Häufigkeiten einzelner Merkmalsausprägungen interessiert. Viel häufiger interessiert man sich in der Praxis aber dafür, wie groß der Anteil der Messwerte ist, der höchstens einen bestimmten Wert betragen.

Beispiel

Man könnte sich fragen, wie hoch im obigen Beispiel der Anteil der Verschnittreste ist, die höchstens 17.2 mm lang sind. Dafür muss man alle relativen Häufigkeiten der Merkmalsausprägungen addieren, die höchstens 17.2 mm betragen. Das sind in diesem Fall:

$$0.1 + 0.2 + 0.2 + 0.1 + 0.2 = 0.8,$$

also beträgt der Anteil 80 %. ◀

Diesen Anteil kann man nun in Abhängigkeit von der betrachteten Höchstgrenze ausdrücken:

Definition: Empirische Verteilungsfunktion

Seien $x_1, x_2, \ldots, x_n$ Ausprägungen eines quantitativen Merkmals. Dann ist

$$F_n(x) = \sum_{i=1}^{k} h_i(x_i)$$

Teil IV

der Anteil der Merkmalsausprägungen mit $x_k \leq x$ (x ist hierbei eine Variable). $F_n(x)$ heißt **empirische Verteilungsfunktion**.

Schauen wir uns das an unserem alten Beispiel einmal an:

Beispiel

Zur Erinnerung: Die gemessenen Längen betrugen (geordnet): 15.3, 15.6, 15.6, 16.7, 16.7, 16.8, 17.2, 17.2, 18.6, 19.2.

Um die empirische Verteilungsfunktion anzugeben, muss man sich für jeden möglichen Wert x (von $-\infty$ bis $+\infty$) überlegen, wie hoch der Anteil der Messwerte ist, die höchstens diesen Wert x betragen.

Dabei kann für x zwar prinzipiell jede Zahl eingesetzt werden, auch Zahlen, die keine Messwerte sind. Es tritt aber nur bei den Messwerten eine Veränderung der empirischen Verteilungsfunktion ein. Denn nur dort ändern sich ja die relativen Häufigkeiten.

Für alle Werte x, die kleiner als der erste Messwert sind, muss $F_n(x) = 0$ sein. Wir fragen uns immer: „Wie groß ist der Anteil der Messwerte, die z. B. höchstens 1.40 sind?“ Die Antwort lautet: 0.

Daher lautet die empirische Verteilungsfunktion für das Verschnittbeispiel wie folgt:

$$F_n(x) = \begin{cases} 0, & x < 15.3 \\ 0.1, & 15.3 \leq x < 15.6 \\ 0.3, & 15.6 \leq x < 16.7 \\ 0.5, & 16.7 \leq x < 16.8 \\ 0.6, & 16.8 \leq x < 17.2 \\ 0.8, & 17.2 \leq x < 18.6 \\ 0.9, & 18.6 \leq x < 19.2 \\ 1, & 19.2 \leq x \end{cases}$$

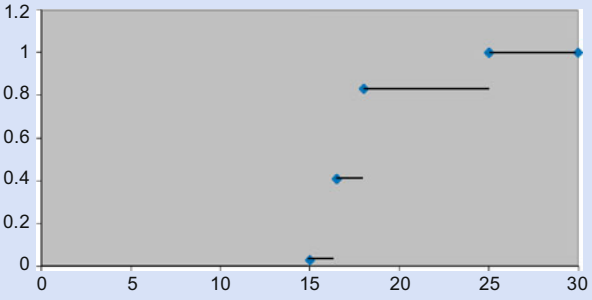


Abb. 15.8 Empirische Verteilungsfunktion am Beispiel der Verschnittreste

Oft bereitet der Unterschied zwischen den Zeichen $\leq$ und $<$ am Anfang Schwierigkeiten. Dabei ist es ganz einfach, wenn man sich an den Messwerten die Frage stellt: „Wie groß ist der Anteil der Messwerte, die z. B. höchstens 1.679 sind?“. Die Antwort lautet fünf Messwerte, also 0.5. Aber wie groß ist der Anteil der Messwerte, die z. B. höchstens 1.68 sind? Hier sind es sechs Messwerte, also 0.6. An dieser Stelle muss sich $F_n(x)$ also verändern.

Grafisch sieht das Ganze dann wie in Abb. 15.8 aus. ◀

Empirische Verteilungsfunktionen besitzen einige Eigenschaften, die nützlich sind, wenn man überprüfen will, ob das, was man gezeichnet hat, tatsächlich eine Verteilungsfunktion ist.

Satz (Eigenschaften einer empirischen Verteilungsfunktion)

- F_n ist eine Treppenfunktion, d. h., sie besteht aus waagerechten Geradenstücken. Die Sprünge passieren an den Messwerten, die Höhe des Sprunges ist immer die relative Häufigkeit h_i des Messwertes.
- F_n ist monoton wachsend.
- F_n geht für kleine x-Werte gegen 0, für große x-Werte gegen 1, d. h.

$$\lim_{n \to -\infty} F_n(x) = 0$$

und

$$\lim_{n \to \infty} F_n(x) = 1.$$

Es gilt sogar: $F_n(x) = 0$ für alle Werte x, die kleiner als der kleinste Messwert sind, und $F_n(x) = 1$ für alle Werte x, die größer oder gleich dem größten Messwert sind.
- F_n ist rechtsseitig stetig, d. h., wenn man von rechts nach links auf der Funktion entlangwandert, gehört der „linkeste“ Punkt immer mit zur Gerade. ◀

Wir haben jetzt übrigens schon mehrfach über geordnete Daten gesprochen. Auch hierfür gibt es eine Schreibweise, die Missverständnisse erspart.

Definition: geordnete Daten

Gegeben seien n Beobachtungswerte $x_1, x_2, \ldots, x_n$ eines ordinalen oder quantitativen Merkmals. Dann heißen $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ die geordneten Daten, in dem Sinne, dass $x_{(i)} \leq x_{(j)}$, falls $i \leq j$, dass also $x_{(1)}$ der kleinste Messwert ist, $x_{(n)}$ der größte usw.

Die empirische Verteilungsfunktion für klassierte Daten beschreibt die Häufigkeitsverteilung eines klassierten Merkmals

Wenn man eine empirische Verteilungsfunktion für große Datensätze erstellt, kann man oftmals nichts mehr erkennen. Daher betrachtet man in diesem Fall nicht mehr jeden Datensatz einzeln, sondern fasst die Datensätze zu sinnvollen Klassen zusammen.

Wenn man aber nur noch weiß, wie viele Datensätze in den einzelnen Klassen sind, aber nicht mehr, wo sie genau in den Klassen liegen, kann man die empirische Verteilungsfunktion nicht mehr genau bestimmen. Daher hat man sich in der Statistik dazu entschieden, auf Nummer sicher zu gehen. Das bedeutet, dass man bei der Beantwortung der alten Frage „Wie groß ist der Anteil der Messwerte, die höchstens x sind?“ nur die Datensätze berücksichtigt, bei denen man sich sicher sein kann, dass sie tatsächlich kleiner oder gleich x sind. Diejenigen, bei denen man es nicht genau weiß, ignoriert man. Das sind immer genau die Datensätze, die sich innerhalb der aktuellen Klasse befinden, denn bei denen weiß man nicht, ob sie sich nicht vielleicht alle am rechten Rand befinden.

So erklärt es sich, dass die empirische Verteilungsfunktion für klassierte Daten wie folgt definiert ist:

Definition: Empirische Verteilungsfunktion für klassierte Daten

Gegeben seien n Beobachtungswerte $x_1, x_2, \ldots, x_n$, die in k Klassen $(A_1, A_2, \ldots, A_k)$ eingeteilt sind. Die Klassen lauten dabei wie folgt:

$$A_1 = [e_0, e_1],\ A_2 = (e_1, e_2],\ \ldots,\ A_k = (e_{k-1}, e_k].$$

h_i sei die relative Häufigkeit der Beobachtungen in der Klasse A_i. Dann heißt die Funktion

$$\hat{F}_n(x) = \begin{cases} 0 & , \quad x < e_1 \\ \sum_{j=1}^{i-1} h_j, & e_{i-1} \le x < e_i, 2 \le i \le k \\ 1 & , \quad e_k \le x \end{cases}$$

die **empirische Verteilungsfunktion der klassierten Daten**.

Das sehen wir uns an einem Beispiel an:

Beispiel

Bei der Messung der Länge von 100 Verschnittresten wurde nur festgehalten, ob sich die Länge in einer der folgenden Klassen befindet:

$$A_1 = [0; 15.0], \quad A_2 = (15.0; 16.5],$$
$$A_3 = (16.5; 18.0], \quad A_4 = (18.0; 25.0]$$

Es ergaben sich folgende relative Häufigkeiten:

$$h_1 = 0.03,\ h_2 = 0.38,\ h_3 = 0.42,\ h_4 = 0.17$$

Daher lautet die empirische Verteilungsfunktion für die klassierten Daten:

$$\hat{F}_{100}(x) = \begin{cases} 0 & , \quad x < 15.0 \\ 0.03, & 15.0 \le x < 16.5 \\ 0.41, & 16.5 \le x < 18.0 \\ 0.83, & 18.0 \le x < 25.0 \\ 1 & , \quad 25.0 \le x \end{cases}$$

Grafisch sieht das Ganze dann wie in Abb. 15.9 aus.

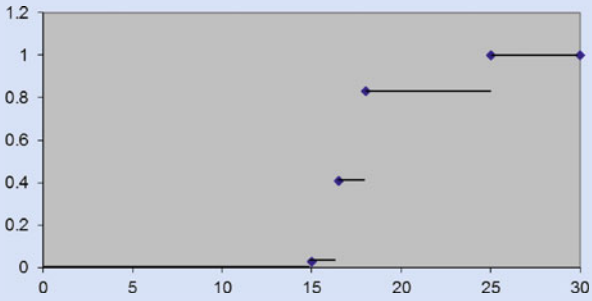


Abb. 15.9 Empirische Verteilungsfunktion für klassierte Daten am Beispiel der Verschnittreste ◀

Die linear interpolierte empirische Verteilungsfunktion für klassierte Daten beschreibt die Häufigkeitsverteilung eines klassierten Merkmals näherungsweise

Mit der empirischen Verteilungsfunktion für klassierte Daten macht man sicherlich keinen Fehler bei der Beantwortung der schon mehrfach aufgeführten Frage, aber befriedigend ist die Antwort auch nicht. Schließlich ist es extrem unwahrscheinlich, dass sich alle Messwerte ganz rechts an der Intervallgrenze befinden. Da möchte man gerne ein wenig mutiger und damit realistischer sein.

Auch hierzu haben Statistiker eine Idee entwickelt. Wenn man unterstellt, dass die Daten in den einzelnen Klassen gleichverteilt sind, verändert die empirische Verteilungsfunktion ihre Gestalt: Statt einer Treppenfunktion entsteht eine stückweise lineare Funktion, die aus Geradenstücken zusammengesetzt ist. Auch diese kann man mit wenig Aufwand berechnen:

Definition: Linear interpolierte empirische Verteilungsfunktion für klassierte Daten

Gegeben seien n Beobachtungswerte $x_1, x_2, \ldots, x_n$, die in k Klassen $A_1, A_2, \ldots, A_k$ eingeteilt sind. Die Klassen lauten dabei wie folgt:

$A_1 = [e_0, e_1], A_2 = (e_1, e_2], \ldots, A_k = (e_{k-1}, e_k]$ mit den zugehörigen Klassenbreiten $d_i = e_i - e_{i-1}$.

h_i sei die relative Häufigkeit der Beobachtungen in der Klasse A_i. Dann heißt die Funktion

$$F_n^*(x) = \begin{cases} 0 & , x < e_0 \\ \sum_{j=1}^{i-1} h_j + \frac{h_i}{d_i}(x - e_{i-1}), & e_{i-1} \le x < e_i, 1 \le i \le k \\ 1 & , e_k \le x \end{cases}$$

die **linear interpolierte empirische Verteilungsfunktion für klassierte Daten**.

Auch das sehen wir uns an dem Beispiel von oben an:

Beispiel

Bei der Messung der Länge von 100 Verschnittresten wurde nur festgehalten, ob sich die Länge in einer der folgenden Klassen befindet:

$$A_1 = [0; 15.0], \quad A_2 = (15.0; 16.5],$$
$$A_3 = (16.5; 18.0], \quad A_4 = (18.0; 25.0]$$

mit den Klassenbreiten

$$d_1 = 15.0, d_2 = 1.5, d_3 = 1.5, d_4 = 7.0.$$

Es ergaben sich folgende relative Häufigkeiten:

$$h_1 = 0.03, h_2 = 0.38, h_3 = 0.42, h_4 = 0.17$$

Daher lautet die linear interpolierte empirische Verteilungsfunktion für die klassierten Daten:

$$F_{100}^*(x) = \begin{cases} 0 & , x < 0 \\ 0 + \frac{0.03}{15}(x - 0) & , 0 \le x < 15 \\ 0.03 + \frac{0.38}{1.5}(x - 15) & , 15 \le x < 16.5 \\ 0.41 + \frac{0.42}{1.5}(x - 16.5), & 16.5 \le x < 18 \\ 0.83 + \frac{0.17}{7}(x - 18) & , 18 \le x < 25 \\ 1 & , 25 \le x \end{cases}$$

Vereinfacht ergibt sich

$$F_{100}^*(x) = \begin{cases} 0 & , x < 0 \\ 0.002 \cdot x & , 0 \le x < 15 \\ 0.2533 \cdot x - 3.77 & , 15 \le x < 16.5 \\ 0.28 \cdot x - 4.21 & , 16.5 \le x < 18 \\ 0.0243 \cdot x + 0.393, & 18 \le x < 25 \\ 1 & , 25 \le x \end{cases}$$

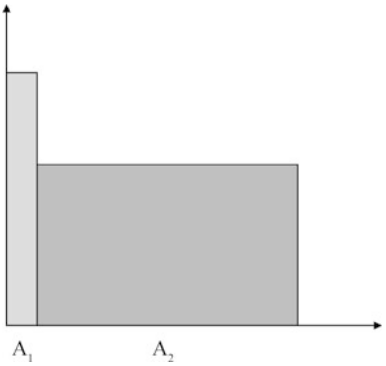


Abb. 15.11 Kontraintuitives Säulendiagramm

In Abb. 15.10 ist diese Funktion grafisch dargestellt.

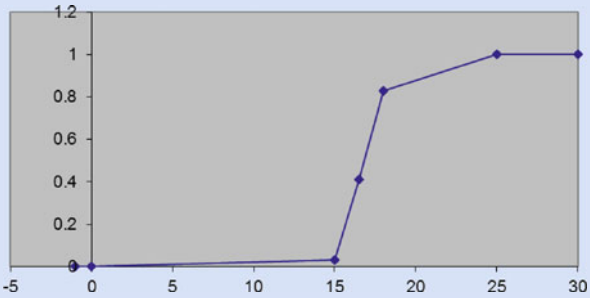


Abb. 15.10 Linear interpolierte empirische Verteilungsfunktion für klassierte Daten am Beispiel der Verschnittreste ◀

15.5 Histogramme

Die Diagramme, die wir bisher betrachtet haben, haben die Eigenschaft, dass immer die Länge des Balkens/der Säule/des Stabes die relative Häufigkeit repräsentiert. Untersuchungen haben aber ergeben, dass das menschliche Auge streng genommen nicht auf die Länge eines Balkens reagiert, sondern auf seinen Flächeninhalt. Dazu betrachten wir Abb. 15.11.

In welcher Klasse, würde man rein intuitiv sagen, liegen mehr Datensätze? Die meisten Menschen antworten auf diese Frage, dass in der zweiten Klasse mehr Datensätze liegen, der Flächeninhalt des Balkens ist nämlich größer, obwohl die Höhe niedriger ist.

In der Situation, dass alle Daten bekannt sind, spielt dieser Sachverhalt keine Rolle, weil alle Säulen gleich breit sind und daher Länge und Flächeninhalt eines Balkens proportional sind. Zum Problem wird es erst dann, wenn man es mit klassierten Daten zu tun hat, denn es ist nicht gefordert, dass alle Klassen gleich breit sind, und dann kann es zu Missinterpretationen kommen.

Der Ausweg lautet: Man zeichnet kein Säulendiagramm, sondern ein **Histogramm**. Das Histogramm berücksichtigt den

obigen Sachverhalt und setzt den Flächeninhalt der Säule einer Klasse gleich der relativen Häufigkeit dieser Klasse h_i. Leider kann man bei Kenntnis des Flächeninhalts allein noch kein Rechteck zeichnen, man braucht Höhe und Breite. Die Breite ist die Klassenbreite d_i, und als Höhe des Rechtecks nimmt man k_i $k_i = \frac{h_i}{d_i}$. Sehen wir uns das Histogramm zum Verschnittbeispiel an:

Beispiel

Klasse A_i	Breite d_i	Rel. Hfgk. h_i	Höhe k_i
$A_1 = [0; 15.0]$	15.0	0.03	0.002
$A_2 = [15.0; 16.5]$	1.5	0.38	0.2533
$A_3 = [16.5; 18.0]$	1.5	0.42	0.28
$A_4 = [18.0; 25.0]$	7.0	0.17	0.0243

In Abb. 15.12 ist das Histogramm grafisch dargestellt.

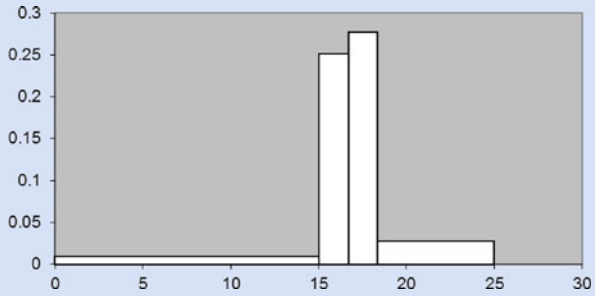


Abb. 15.12 Histogramm am Beispiel der Verschnittreste ◀

15.6 Kenngrößen von Daten

Erinnern wir uns kurz, dass die Aufgabenstellung der deskriptiven Statistik darin besteht, dass man die wichtigsten Eigenschaften großer Datenmengen erfasst. Dann erscheint es naheliegend, dass man z. B. einen Wert angibt, der nach irgendwelchen Kriterien in der Mitte der Datensätze liegt. Solche Parameter nennt man **Lageparameter**. Zusätzlich will man aber auch wissen, wie sich die Datensätze um diesen Lageparameter „tummeln", also ob sie eng neben dem Lageparameter liegen oder ob sie weit verstreut um ihn verteilt sind. Parameter, die das messen, nennt man **Streuungsparameter**. Für beide Arten von Kenngrößen lernen wir nun einige Beispiele kennen.

Lageparameter beschreiben die Mitte der Datensätze

Der bekannteste Lageparameter ist sicherlich der „Durchschnitt", auch **arithmetisches Mittel** genannt.

Definition: Arithmetisches Mittel

Gegeben seien n Messwerte $x_1, x_2, \ldots, x_n$ eines quantitativen Merkmals.

- Dann heißt $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ das arithmetische Mittel der x_i.
- Wenn $\alpha_1, \alpha_2, \ldots, \alpha_n$ Gewichte sind mit $\alpha_i \geq 0$ und $\sum_{i=1}^{n} \alpha_i = 1$, dann heißt $\tilde{x} = \sum_{i=1}^{n} \alpha_i \cdot x_i$ gewichtetes arithmetisches Mittel der x_i.

Leider hat das arithmetische Mittel einige Nachteile: Es kann nur bei quantitativen Merkmalen berechnet werden. Hinzu kommt, dass es sehr empfindlich auf Ausreißer reagiert. Wer schon einmal einen Ausrutscher in den Noten hatte, weiß, wovon die Rede ist. Aber während man bei Noten noch argumentieren kann, dass es gerechtfertigt ist, dass der Durchschnitt absackt, wenn man eine Note verhaut, wird das Ganze kritisch, wenn man es mit Ausreißern aufgrund von Messfehlern zu tun hat.

Aber es gibt Alternativen: Da die Durchschnittsbildung nur bei quantitativen Merkmalen Sinn macht, brauchen wir auch Lageparameter, die sich (neben den quantitativen Merkmalen) auch für qualitative Merkmale eignen. Für (mindestens) ordinale Merkmale kann man auch den **Median** berechnen. Dieser Wert liegt in der Mitte, da er den Wert angibt, für den 50 % aller Messwerte kleiner oder gleich und 50 % der Messwerte größer oder gleich sind.

Definition: Empirischer Median

Gegeben seien n Messwerte $x_1, x_2, \ldots, x_n$ eines quantitativen oder ordinalen Merkmals. Dann ist der empirische Median der Messwerte definiert als

$$x_{\text{med}} = \hat{Q}_{0,5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , n \text{ ungerade} \\ \left[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}\right] & , n \text{ gerade} \end{cases}$$

Bei quantitativen Daten mit geradem Stichprobenumfang wählt man häufig aus dem Intervall des Medians den Mittelpunkt aus und bezeichnet diesen als Median. In diesem Fall (quantitatives Merkmal und gerader Stichprobenumfang) gilt:

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

Der Median muss also nach dieser Definition nicht unbedingt eindeutig sein. Im zweiten Fall ist jeder Wert aus dem angegebenen Intervall ein Median, da jeder Wert des Intervalls die oben genannte Forderung erfüllt.

Achtung Ein beliebter Anfängerfehler besteht darin, die Daten nicht der Größe nach zu ordnen, dann kann das Ergebnis

gar nicht mehr richtig werden. Also, unbedingt Daten der Größe nach ordnen! Das bedeutet die runde Klammer um den Index, s. Definition: geordnete Daten. ◀

Die Idee des Medians, dass 50 % aller Messwerte kleiner oder gleich und 50 % der Messwerte größer oder gleich dem Median sind, lässt sich mit wenig Aufwand verallgemeinern. Die so entstandenen Kenngrößen sind die **Quantile**.

Definition: Empirisches p-Quantil

Gegeben seien n Messwerte $x_1, x_2, \ldots, x_n$ eines quantitativen oder ordinalen Merkmals. Dann ist das empirische p-Quantil der Messwerte definiert als

$$\hat{Q}_p = \begin{cases} x_{(\lfloor np \rfloor + 1)} & , np \text{ nicht ganzzahlig} \\ \left[x_{(np)}, x_{(np+1)}\right] & , np \text{ ganzzahlig} \end{cases}$$

Keine Angst vor $\lfloor np \rfloor$. Es heißt untere Gaußklammer und bedeutet, dass man die größte ganze Zahl unterhalb der Kommazahl nehmen soll, z. B. $\lfloor 7.9 \rfloor = 7$, $\lfloor 15 \rfloor = 15$.

Das bedeutet für das p-Quantil, dass $p \cdot 100\,\%$ aller Messwerte kleiner oder gleich und $(1-p)\cdot 100\,\%$ der Messwerte größer oder gleich dem p-Quantil sind. Auch für das p-Quantil gilt, dass man bei quantitativen Daten oft die Mitte des Intervalls als Wert angibt.

Einige Quantile sind so wichtig, dass sie einen eigenen Namen bekommen haben:

Definition: Spezielle Quantile

- $\hat{Q}_{0.5}$ heißt Median.
- $\hat{Q}_{0.25}$ heißt unteres Quartil.
- $\hat{Q}_{0.75}$ heißt oberes Quartil.
- $\hat{Q}_{0.125}$ heißt erstes Oktil.
- $\hat{Q}_{0.1}$ heißt erstes Dezentil.

Der **Boxplot** ist eine sehr ansprechende grafische Darstellung der Lageparameter Median, Quartile und kleinster und größter Messwert. Man kann einen Boxplot nur für quantitative Messwerte zeichnen, denn man braucht eine x-Achse für die Merkmalsausprägungen. Oberhalb dieser x-Achse macht man beim kleinsten und beim größten Messwert einen dicken Punkt, bei den Quartilen und dem Median einen senkrechten Strich. Dann verbindet man die drei senkrechten Linien mit zwei waagerechten Linien zu einem Kasten und die beiden dicken Punkte durch waagerechte Linien mit dem Kasten. Fertig ist der Boxplot.

Anbei sehen wir den Boxplot für die Maße der zehn Verschnitte, die wir schon die ganze Zeit untersuchen. Die Berechnung der Kenngrößen können wir sofort im nächsten Beispiel nachlesen.

In jedem der vier Bereiche $[x_{(1)}; \hat{Q}_{0.25}]$, $[\hat{Q}_{0.25}; x_{\text{med}}]$, $[x_{\text{med}}; \hat{Q}_{0.75}]$, $[\hat{Q}_{0.75}; x_{(n)}]$ liegt dann jeweils ein Viertel der Datensätze. Und das bedeutet, man kann schon an der Zeichnung

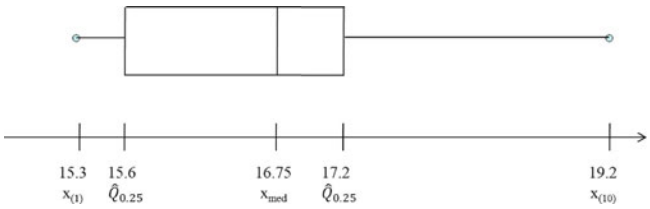


Abb. 15.13 Boxplot am Beispiel der Verschnittreste

sehen, wo sich Messwerte häufen und wo nur wenige liegen. Sie häufen sich nämlich dort, wo der Bereich kurz ist (im obigen Beispiel im ersten und im dritten Bereich), und wenige liegen, wo der Bereich lang ist (im obigen Beispiel betrifft das den zweiten und den vierten Bereich). Lange „tails“ (Das ist der Bereich zwischen den Quartilen und den Extremwerten.) sind übrigens ein Indiz dafür, dass evtl. Ausreißer in den Daten vorhanden sind. Das muss nicht der Fall sein, aber man sollte sich die Daten noch einmal ansehen.

Ein weiterer Kennwert zur Lage einer Datenreihe besteht z. B. in dem Wert, der am häufigsten vorkommt. Denn ihm misst man aufgrund seiner Häufigkeit eine besondere Aussagekraft bei. Ein weiterer Vorteil des **Modalwertes** liegt darin, dass er auch problemlos für nominale Daten verwendet werden kann.

Definition: Modalwert

Die Merkmalsausprägung, die bei den Messwerten am häufigsten auftritt, heißt Modalwert x_{mod}.

Jetzt kennen wir die wichtigsten Lageparameter. Schauen wir sie uns noch einmal an einem Beispiel an:

Beispiel

Bei der Erhebung der Verschnittreste ergaben sich folgende geordnete Messwerte: 15.3, 15.6, 15.6, 16.7, 16.7, 16.8. 17.2, 17.2, 18.6, 19.2.

Dann beträgt

$$\bar{x} = \frac{1}{10}(15.3 + 15.6 + 15.6 + 16.7 + 16.7 + 16.8 + 17.2 + 17.2 + 18.6 + 19.2) = 16.89.$$

Dies kann man auch durchaus als gewichtetes arithmetisches Mittel auffassen, wenn man die relative Häufigkeit der Messwerte als Gewichte versteht:

$$\tilde{x} = 0.1 \cdot 15.3 + 0.2 \cdot 15.6 + 0.2 \cdot 16.7 + 0.1 \cdot 16.8 + 0.2 \cdot 17.2 + 0.1 \cdot 18.6 + 0.1 \cdot 19.2 = 16.89$$

Der Median ist nicht eindeutig ($n = 10$ gerade). Prinzipiell ist das gesamte Intervall $[x_{(5)}, x_{(6)}]$ Median, also gilt:

$$x_{\text{med}} = \hat{Q}_{0.5} = [16.7; 16.8]$$

Da es sich bei der Länge um ein quantitatives Merkmal handelt, kann man den Median auch festlegen:

$$x_{\text{med}} = \frac{1}{2} \cdot (16.7 + 16.8) = 16.75$$

Auch der Modalwert ist nicht eindeutig festgelegt: 15.6, 16,7 und 17.2 kommen jeweils am häufigsten (nämlich doppelt) vor und sind somit alle Modalwerte. ◀

Streuungsparameter beschreiben die Lage der Daten im Vergleich zum Lageparameter

Die Streuungsparameter gehören zu jeweils einem Lageparameter, in dem Sinne, dass es immer einen Lageparameter gibt, der das entsprechende Abstandsmaß für einen gegebenen Datensatz minimiert.

Die **empirische Varianz** und die **empirische Standardabweichung** gehören zum arithmetischen Mittel. Die Idee der beiden Streuungsparameter ist auch schnell erzählt und eigentlich sehr einleuchtend:

Man möchte wissen, wie weit die Datensätze im Durchschnitt vom arithmetischen Mittel entfernt sind, also muss man von jedem Datensatz x_i seinen Abstand zum arithmetischen Mittel $\overline{x}$ berechnen. Wenn man hierüber den Durchschnitt bildet, stellt man fest, dass die Abstände teilweise positiv und teilweise negativ sind, was immer dazu führt, dass sie sich komplett bei der Addition aufheben und das Ergebnis somit immer 0 wird.

Was tut man also? Man quadriert alle Abstände, dann werden sie definitiv alle positiv, und nichts hebt sich auf. Darüber bildet man noch den Durchschnitt und fertig ist die empirische Varianz. Wäre da nicht noch das Problem der Interpretation. Was bedeutet es, wenn z. B. 25 herauskommt?

Streng genommen heißt das, dass der durchschnittliche quadratische Abstand der Messwerte vom arithmetischen Mittel 25 beträgt, aber dieser Wert besitzt keine anschauliche Bedeutung. Daher geht man zum Schluss noch einmal hin und zieht aus der Varianz die Wurzel. Dieser Wert heißt dann empirische Standardabweichung und kann als durchschnittlicher Abstand der Messwerte vom arithmetischen Mittel angesehen werden.

Definition: Varianz und Standardabweichung

- Die empirische Varianz s^2 berechnet sich wie folgt:

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

- Die empirische Standardabweichung berechnet sich wie folgt:

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

Die Definition der empirischen Varianz ist nicht nützlich für praktische Berechnungen. Viel einfacher ist die Berechnung der empirischen Varianz mithilfe der folgenden Formel:

Es gilt: $s^2 = \overline{x^2} - (\overline{x})^2$.

Man darf nicht glauben, dass dabei 0 rauskommt. Man muss die Formel lesen können.

$\overline{x^2}$ bedeutet, dass man jeden einzelnen Messwert quadrieren und über die quadrierten Werte den Durchschnitt bilden muss.

$(\overline{x})^2$ bedeutet, dass man das arithmetische Mittel über die Messwerte bilden und die Zahl zum Schluss quadrieren muss.

Daher ist das Ergebnis nur dann gleich 0, wenn alle Messwerte identisch sind.

Achtung Die empirische Varianz kann nicht negativ werden! Wenn man ein negatives Ergebnis berechnet, kann man sich darauf verlassen, dass man sich irgendwo verrechnet hat! ◀

Beispiel

Berechnen wir einmal die empirische Varianz und die empirische Standardabweichung für das Verschnittrestebeispiel.

Es ergibt sich Folgendes:

$$\overline{x^2} = 286.711,$$
$$\overline{x}^2 = 16.89^2 = 285.2721.$$

Also gilt für die empirische Varianz:

$$s^2 = \overline{x^2} - (\overline{x})^2 = 286.711 - 285.2721 = 1.4389.$$

Die empirische Standardabweichung ergibt sich somit zu

$$\sqrt{s^2} = \sqrt{1.4389} = 1.19954157.$$

Das bedeutet also, dass die Längen der Verschnittreste durchschnittlich um ca. 1.2 mm von der Durchschnittslänge 16.89 mm nach oben bzw. nach unten abweichen. ◀

Auch wenn die Standardabweichung eine nachvollziehbare inhaltliche Bedeutung hat, fällt es schwer, Zahlenwerte der Standardabweichung einzuschätzen. Ist eine Standardabweichung von $s = 10\,000$ groß oder klein? Prinzipiell klingt die Zahl sehr groß, wenn aber das arithmetische Mittel $\overline{x} = 100\,000\,000$ beträgt, erscheint eine Standardabweichung von 10 000 eher gering. Man würde schlussfolgern, dass die Daten sehr eng am Mittelwert „kleben“.

Um einen besseren Eindruck zu gewinnen, betrachtet man das Verhältnis von Standardabweichung und Mittelwert.

Definition: Variationskoeffizient

Der **Variationskoeffizient** bildet das Verhältnis von empirischer Standardabweichung zu arithmetischem Mittel:

$$V = \frac{s}{\overline{x}}.$$

Beispiel

Im vorliegenden Beispiel ergibt sich demnach:

$$V = \frac{s}{\overline{x}} = \frac{1.19954157}{16.89} = 0.071020815.$$

s beträgt also ca. 7 % von $\overline{x}$. ◀

Ein anderer Weg, den durchschnittlichen Abstand von einem Lageparameter zu ermitteln, besteht darin, dass man die einzelnen Abstände nicht quadriert, sondern den Betrag bildet, bevor man den Durchschnitt berechnet. Dieses Abstandsmaß gehört zum Median, daher berechnet er sich wie folgt:

Definition: Durchschnittliche Abweichung vom Median

Die **durchschnittliche Abweichung vom Median** lautet:

$$d_m = \frac{1}{n} \sum_{i=1}^{n} |x_i - x_{\text{med}}|$$

Beispiel

Im Beispiel der Verschnittreste ergibt sich also:

$$\begin{aligned} d_m &= \frac{1}{10} \cdot (|15.3 - 16.75| + |15.6 - 16.75| \\ &\quad + |15.6 - 16.75| + |16.7 - 16.75| \\ &\quad + |16.7 - 16.75| + |16.8 - 16.75| \\ &\quad + |17.2 - 16.75| + |17.2 - 16.75| \\ &\quad + |18.6 - 16.75| + |19.2 - 16.75|) \\ &= \frac{1}{10} \cdot (1.45 + 1.15 + 1.15 + 0.05 + 0.05 \\ &\quad + 0.15 + 0.45 + 0.45 + 1.85 + 2.45) \\ &= 0.91 \end{aligned}$$

◀

Aufgaben

15.1 Einem Unternehmen können unter anderem folgende Merkmale zugeordnet werden:

Mitarbeiterzahl, Familienstand des Chefs, hergestellte Produkte, Qualität der Produkte, Umsatz eines Jahres, Rechtsform, Betriebsklima, Kundenzufriedenheit, Sitz der Zentrale, Wert der Immobilien, Gründungsjahr.

1. Geben Sie zu den einzelnen Merkmalen jeweils eine mögliche Menge an Ausprägungen an.
2. Bestimmen Sie die jeweils zugrunde liegende Art der Merkmale.

15.2 In der nachfolgenden Liste sind die Massen der wöchentlich abgeholten Abfallcontainer einer Werkstatt angegeben (in kg):

800, 850, 1000, 1500, 850, 900, 1000, 1200, 850, 1400, 800, 1800, 900, 2000, 1200, 1000, 850, 2500, 1500, 800, 600, 1400, 1000, 2700, 850, 1000, 1200, 1400, 850, 1000, 1300, 1700, 1500, 850, 2500, 2000, 900, 1400, 1000, 2000

1. Berechnen Sie die absolute und die relative Häufigkeitsverteilung.
2. Zeichnen Sie die empirische Verteilungsfunktion.
3. Berechnen Sie den Anteil der Container, die eine Masse von
   (a) weniger als 1000 kg,
   (b) mehr als 1500 kg,
   (c) zwischen 1100 und 2600 kg
   erreicht haben.
4. Bestimmen Sie den Median, den Modalwert und das arithmetische Mittel.
5. Fertigen Sie ein Histogramm für die Klassenbildung [600, 1000], [1000,1500], [1500,2000], [2000,2700] an.
6. Zeichnen Sie für die obige Klasseneinteilung die empirische Verteilungsfunktion für klassierte Daten und die linear interpolierte empirische Verteilungsfunktion der klassierten Daten.

15.3 In der nachfolgenden Tabelle sind die Arbeitskosten je geleisteter Arbeitsstunde im Jahr 2000 in verschiedenen Ländern der EU aufgelistet (Eurostat, Pressemitteilung 23/2003).

Land	Kosten in €
Dän.	27.1
BRD	26.54
Gr.	10.4
Sp.	14.22
Fr.	24.39
Irl.	17.34
Lux.	24.33
Niederl.	22.99
Ö.	23.6
Por.	8.13
Fin.	22.13
Sw.	28.56
GB	23.85

Berechnen Sie das arithmetische Mittel, die Standardabweichung und das untere sowie das obere Quartil und den Variationskoeffizienten.