

1 Beschreibende Statistik

Umgangssprachlich und auch inhaltlich sind mit dem Wort Statistik zwei Bedeutungen verbunden. Es geht einmal um die (mathematische) Disziplin Statistik, also um die *Wissenschaft* Statistik. Dann geht es auch um Statistiken, z. B. um Statistiken von Kosten, um Statistiken in der Bevölkerungspolitik, also um Zahlenmaterial aus Erhebungen. Hier stellt sich die Statistik als *Ergebnis einer Tätigkeit* dar. In der beschreibenden Statistik geht es um eine Beschreibung des Ist-Zustandes. Nach Ferschel ist Statistik das Bestreben, die Dinge so zu sehen, wie sie wirklich sind ([53]). Doch das ist offenbar recht schwer, wie Fehldeutungen und Manipulationen in Wirtschaft, Wissenschaft und Politik belegen (Kütting [102], [104], Krämer [85]). Es verwundert nicht, dass das Ansehen der Statistik in der Öffentlichkeit gering ist:

- Es gibt die gewöhnliche Lüge, die Notlüge und die Statistik.
- Trauen Sie keiner Statistik, die Sie nicht selbst gefälscht haben.
- Zahlen können lügen, Lügner können zählen.
- Mit Zahlen kann man alles, und daher nichts beweisen.
- Statistik ist die Kunst, mit richtigen Zahlen etwas Falsches zu beweisen.
- Statistik ist das Umgraben von Datenfriedhöfen.

Diese unrühmlichen Einschätzungen sollten positive Impulse für eine sachgerechte Information freisetzen. Denn nicht die Sprache der Statistik – das sind die Zahlen und Graphiken – lügt, sondern es „lügen“ allenfalls die Menschen, die mit den Zahlen und Graphiken umgehen.

1.1 Die historische Entwicklung der Statistik – ein kurzer Abriss

Die Ursprünge der Statistik liegen weit zurück. Für die beschreibende (deskriptive) Statistik gibt man im Allgemeinen drei Quellen an: die **Amtliche Statistik**, die **Politische Arithmetik** und die **Universitätsstatistik**.

1.1.1 Die Amtliche Statistik

Der Sinn einer **Amtlichen Statistik** liegt darin, Informationen darüber zu gewinnen, wie die organisierte Gesellschaft am besten „verwaltet“ werden kann. Man möchte Kenntnisse haben z. B. über die Anzahl der Bewohner, über den Landbesitz, über die Bodenschätze, über den Viehbestand usw. Einfache statistische Erhebungen wurden bereits vor Jahrtausenden durchgeführt, z. B. in Ägypten, China, Indien, Griechenland, Rom. Schon 3000 v. Chr. nahmen die Pharaonen in Ägypten Volkszählungen, Landvermessungen und Viehbestandszählungen vor. Neben steuerlichen Zwecken dienten solche Volkszählungen auch zur Erstellung von Verzeichnissen für den Frondienst und „Militärdienst“.

Auch in der Bibel werden Volkszählungen erwähnt.

Vgl. für die folgenden Ausführungen:

- a) Die Heilige Schrift des Alten und Neuen Testaments (Übersetzung von Hamp, V./Stengel, M./Kürzinger, J.). Aschaffenburg 1975²⁵.
- b) Schürer, E.: Geschichte des jüdischen Volkes im Zeitalter Jesu Christi. Band 1. Hildesheim - New York 1970.
- c) Schneider, G.: Das Evangelium nach Lukas, Kapitel 1 – 10. Würzburg 1977.
- d) Scharbert, J.: Numeri. Würzburg 1992.
- e) Cornfeld, G./Botterweck, G. J. (Hrsg.): Die Bibel und ihre Welt (2 Bände). Herrsching 1991.

Im Alten Testament verweisen wir auf die Stellen Exodus (2. Buch Moses) Kap. 30, Verse 11–16; Kap. 38, Vers 25f.; Numeri (4. Buch Moses) Kap. 1, Verse 1–54; Kap. 26, Verse 1–51 und Verse 57–65; 2. Buch Samuel Kap. 24, Vers 1f. (Parallelbericht im 1. Buch der Chronik, Kap. 21, Verse 1–5). Im Neuen Testament nennen wir die Stellen Apostelgeschichte Kap. 5, Vers 37 und Lukas Kap. 2, Verse 1ff. Im Folgenden gehen wir auf einige Stellen ein.

In Exodus, 30. Kap., 11. Vers und folgende heißt es: „¹¹Der Herr sprach zu Moses: ¹²Wenn du die Kopfbzahl der Israeliten bei ihrer Musterung feststellst, dann soll jeder ein Lösegeld für sein Leben anlässlich der Musterung an den Herrn entrichten, damit nicht eine Plage sie bei ihrer Musterung treffe. ¹³Dieses soll ein jeder, der zur Musterung kommt, entrichten: Ein halbes Silberstück nach heiligem Gewicht, 20 Gera das Silberstück, ein halbes Silberstück also als Weihegabe an den Herrn. ¹⁴Jeder, der zur Musterung kommt, von 20 Jahren an und darüber, soll die Weihegabe an den Herrn entrichten.“

Anmerkungen:

- Nach Überlieferung musste jeder Israelit jedes Jahr zum Unterhalt des Heiligtums ein als Sühnegeld bezeichnetes halbes Silberstück (einen halben Schekel) zahlen. Die Anzahl der halben Schekel ergab auch die Anzahl der Männer

über 20 Jahre, d. h. der Wehrdienstpflichtigen (vgl. Exodus, Kap. 38, Vers 25f.).

- Der Gedanke der Entsühnungen aus Anlass einer Volkszählung wird in Beziehung gebracht mit der ersten königlichen Volkszählung durch David (2. Buch Samuel, 24. Kap., Vers 1f.). Denn die nach Durchführung der Volkszählung über das Volk Israel hereinbrechende Pest wurde als Strafe für Davids Volkszählung angesehen (2. Buch Samuel, 24. Kap., Verse 10–15).

Das Buch Numeri (Zahlen) verdankt sogar seinen Namen einer Volkszählung. Es beginnt mit Zählungen und Musterungen der wehrfähigen Männer Israels. „¹Der Herr redete zu Moses in der Wüste am Sinai im Offenbarungszelt am ersten Tage des zweiten Monats im zweiten Jahr nach dem Auszug aus dem Lande Ägyptens folgendes: ²Nehmt die Gesamtzahl der ganzen Gemeinde der Israeliten auf, und zwar nach ihren Sippen und Familien mit Zählung der einzelnen Namen; alles, das männlich ist, nach seiner Kopfzahl! ³Von 20 Jahren und darüber sollt ihr alle Kriegstüchtigen in Israel scharenweise mustern, du und Aaron!“ (Numeri, 1. Kap., Verse 1–3). Diese erste Zählung durch Moses fand nach dieser Tradition während des Exodus am Sinai statt. Gezählt wurden 603 550 (Numeri, 2. Kap., Vers 32). Eine zweite Volkszählung fand nach dieser Tradition durch Moses vor dem Einzug in das Gelobte Land statt: „¹Nach dieser Heimsuchung sprach der Herr zu Moses und Eleasar, dem Sohn des Priesters Aaron: ²Nehmt die Gesamtzahl der Israelitengemeinde auf, von 20 Jahren an nach ihren Familien geordnet, alle die heerespflichtig sind in Israel!“ (Numeri, 26. Kap., Vers 1f.). Gezählt wurden 601 730 Gemusterte. „⁵³An diese werde das Land als Erbbesitz nach dem Verhältnis der Namenszahl verteilt.“ (Numeri, 26. Kap., Vers 53). „⁵⁵Doch soll man das Land durch das Los verteilen; ...“ (Numeri, 26. Kap., Vers 55).

Anmerkungen: (vgl. auch *Scharbert*, a.a.O., S. 18f, S. 109)

- Die Zahlen selbst sind mit Sicherheit unhistorisch. Denn wenn man Frauen, Kinder und Greise einbezieht, müssten ungefähr 2 Millionen Menschen auf der Wanderung gewesen sein. Die Probleme einer solchen Völkerwanderung (Verpflegung etc.) dürften kaum zu lösen gewesen sein. Denkbar ist, dass man bei den Zahlen an eine endzeitliche Fülle dachte.
- Bemerkenswert ist, dass schon nach Kategorien (Sippen, Familien) getrennt gezählt wurde.
- Auch das Verfahren der Landverteilung ist interessant. Es werden zwei sich gegenseitig behindernde Verfahren genannt. Der Widerspruch kann aufgelöst werden, wenn man annimmt, dass Gott (Jahwe) die Verteilung durch Los (also durch Zufall) so lenken wird, dass eine gerechte Verteilung nach der Größe der Stämme erfolgt.

Überblickt man diese Volkszählungen und Prozeduren, so wird verständlich, dass Volkszählungen nicht beliebt waren: Furcht vor Heeresdienst, Angst vor

Besteuerung, Einschränkung der Persönlichkeit, Offenlegung der Privatsphäre, Angst vor Strafen. Die Unbeliebtheit von Volkszählungen hat sich bis heute erhalten.

Die wohl bekannteste Volkszählung aus der Bibel ist die aus dem Neuen Testament bei Lukas, Kap. 2, Verse 1ff. Es handelt sich um das Weihnachtsevangeli- um nach Lukas. „¹In jenen Tagen geschah es, dass vom Kaiser Augustus der Befehl erging, das ganze Reich zu beschreiben und einzutragen. ²Diese erste Eintragung geschah, als Quirinius Statthalter von Syrien war. ³Alle gingen hin, sich eintragen zu lassen, ein jeder in seine Stadt. ⁴Auch Joseph ging von Galiläa, aus der Stadt Nazaret, hinauf nach Judäa in die Stadt Davids, die Bethlehem heißt – weil er aus dem Haus und Geschlecht Davids war – ⁵um sich eintragen zu lassen zusammen mit Maria, seiner Vermählten, die gesegneten Leibes war.“ Nach diesem Bericht des Lukas wurde Jesus dann in Bethlehem geboren.

Anmerkungen:

- Bei der Eintragung handelte es sich um eine Volkszählung (einen Zensus) hauptsächlich für steuerliche Zwecke. Aus historischer Sicht ist aber der bei Lukas erwähnte Zensus nicht unumstritten. Mit großer Sicherheit weiß man nämlich, dass Quirinius die in der Apostelgeschichte (Kap. 5, Vers 37) erwähnte Volkszählung durchgeführt hat. Das war aber im Jahre 6 oder 7 nach Christus. Andererseits ist nach heutigem Kenntnisstand ein Zensus im *gesamten* Reich (Reichszensus) unter Augustus nicht bezeugt. Wegen Einzelheiten dieser wissenschaftlichen Diskussion verweisen wir auf entsprechende Literatur (s. *Schürer*, a.a.O., S. 508 – 543; *Schneider*, a.a.O., S. 68f).
- Der römische Censur, der auf Servius Tullius (um 550 v. Chr.) zurückgeht, war eine sich regelmäßig alle 5 Jahre wiederholende Erhebung der Bevölkerung. Seine Bezeichnung hat sich in einigen Ländern wie z. B. in den USA für regelmäßig stattfindende Bestandsaufnahmen bis heute erhalten. Auch in der Bundesrepublik Deutschland findet seit 1957 jährlich ein Mikrozensus statt, bei dem 1 % aller Haushalte erfasst werden. Interessant ist nun, dass diese „kleine Zählung“ durch *Interviewer* vorgenommen wird. Dadurch können Rückfragen bei komplexen Sachverhalten sofort geklärt werden, so dass die Ergebnisse zuverlässiger erscheinen mögen und vielleicht auch sind als bei einer reinen Zählung.

Zur Amtlichen Statistik zählen auch die Inventarien, die Karl der Große anfertigen ließ, und das sogenannte Domesday Book um ca. 1084/85, das Wilhelm der Eroberer anlegen ließ. Letzteres enthält die Zählungen der Einwohner, die auch nach Ständen statistisch aufgegliedert waren, und Zählungen ihres Grund- und Viehbesitzes. Ferner erwähnen wir die sog. Populationslisten (für Geburten, Trauungen und Todesfälle) unter dem Kurfürsten Friedrich Wilhelm um 1683. Die letzten Stufen in dieser Entwicklung sind bei uns die Statistischen Jahrbücher für die Bundesrepublik Deutschland, die vom Statistischen Bun-

desamt herausgegeben werden, und die Publikationen der Statistischen Landesämter und Kommunen. Der Hintergrund solcher Erhebungen ist in seiner praktischen Bedeutung für Regierungen und Verwaltungen zu sehen.

1.1.2 Die Politische Arithmetik

Im 17. und 18. Jahrhundert traten international zwei neue Aspekte hinzu: die sog. Politische Arithmetik und die sog. Universitätsstatistik.

Als Begründer der in England aufgewachsenen **Politischen Arithmetik** gelten *John Graunt* (1620 – 1674) und *Sir William Petty* (1623 – 1687). Durch Vergleich von Geburtenhäufigkeiten und Sterbezahlen versuchte man Bevölkerungsentwicklungen zu beobachten. Nicht Einzelercheinungen waren wichtig, sondern zu (homogenen) Klassen zusammengefasste Massenerscheinungen. Man fragte nach Ursachen und Regelmäßigkeiten. (Vgl. hierzu auch Biehler [18].) J. Graunt war von Haus aus Tuchkleinhändler, später war er Kommissar für die Wasserversorgung Londons. Das Material fand Graunt in Geburts- und Todeslisten, in Tauf- und Sterberegistern. Seine grundlegende Schrift erschien 1662: *Natürliche und politische Beobachtungen über die Totenlisten der Stadt London, führnehmlich ihre Regierung, Religion, Gewerbe, Luft, Krankheiten und besondere Veränderungen betreffend* ... W. Petty war nach dem Medizinstudium Professor für Anatomie in Oxford, war aber sehr vielseitig interessiert. Sein Werk *Political Arithmetic* gab der Strömung ihren Namen.

Als weitere Vertreter der Politischen Arithmetik erwähnen wir noch E. Halley und J. P. Süßmilch. Der Astronom *Edmond Halley* (1656 – 1742) – nach ihm ist der von ihm vorausgesagte Halley-Komet benannt – verfasste aufgrund von Kirchenbüchern der Stadt Breslau die ersten Sterbetafeln mit Sterbewahrscheinlichkeiten. Vertreter der Politischen Arithmetik war in Deutschland der preußische Prediger und nachmalige Oberkonsistorialrat *Johann Peter Süßmilch* (1707 – 1767) mit seinem Buch *Die göttliche Ordnung in den Veränderungen des menschlichen Geschlechts aus der Geburt, Tod und Fortpflanzung desselben erwiesen von Johann Peter Süßmilch, Prediger beim hochlöblichen Kalksteinschen Regiment* (1741). Wie der Titel schon andeutet, betrachtete Süßmilch die Gesetzmäßigkeiten als der göttlichen Ordnung zugehörig.

1.1.3 Die Universitätsstatistik und ihre Weiterentwicklung

Mit dem Terminus **Universitätsstatistik** ist die an Universitäten vertretene bzw. etablierte wissenschaftliche Disziplin gemeint. Es geht also nicht um Statistiken an den Universitäten. Da man für die zentrale Verwaltung von Staaten Ausbildungsmöglichkeiten brauchte, entstand an den Universitäten im 17. Jahrhundert ein erweitertes Lehr- und Ausbildungsangebot. Es betraf die Lehre von den Staatsmerkwürdigkeiten. Es ging um Staatsbeschreibungen. Schon

1660 kündigte der Rechtshistoriker *Hermann Conring* (1606 – 1681), Professor an der ehemaligen Universität Helmstedt, eine Vorlesung zu „Notitia rerum publicarum“ oder „Staatenkunde“ an und behandelte Staatsbeschreibungen unter den Gesichtspunkten Bevölkerung, Staatsform, Verwaltung, Finanzen. *Gottfried Achenwall* (1719 – 1772), Professor in Marburg und später in Göttingen, führte den Namen *Statistik* für diese neue Disziplin ein. Der Name geht zurück auf das italienische Wort *statista*, was Staatsmann bedeutet, oder auf das lateinische Wort *status (rei publicae)* was Zustand (des Staates) bedeutet. Die moderne Staatswissenschaft prägte also den wissenschaftlichen Charakter der Statistik. Wenn auch der Ursprung des Namens Statistik mit Achenwall untrennbar verbunden ist, kann man aber nicht sagen, dass er der Begründer der Statistik ist. Das folgt auch schon aus unseren früheren Darlegungen. Mittelpunkt war bei ihm noch nicht die zahlenmäßige Erforschung von Massenerscheinungen. Dieser Gesichtspunkt wurde von *Karl Knies* (1821 – 1898) hervorgehoben.

Von den deutschen Universitäten breitete sich die Statistik auf andere Länder aus: Österreich, Ungarn, Italien (Venetien), Belgien, Frankreich, England, USA. Dabei ist interessant, dass die Statistik in den USA um 1845 eingeführt wurde, und zwar an der Universität Virginia im Department of Moral Philosophy. Diese Verbindung von Philosophie mit Statistik ist bemerkenswert.

Mit dem Entstehen der Wahrscheinlichkeitstheorie machte sich der *Einfluss wahrscheinlichkeitstheoretischer Überlegungen* auf die Statistik bemerkbar. Schon *Jakob Bernoulli* (1654 – 1705) hatte den Zusammenhang der mathematischen Wahrscheinlichkeit und der statistischen Wahrscheinlichkeit (Gesetz der großen Zahlen) gesehen, wie seiner *Ars conjectandi*, die acht Jahre nach seinem Tode erschien, zu entnehmen ist.

Auch für den belgischen Astronomen und Statistiker *Lambert Adolph Jacob Quetelet* (1796 – 1874) war die Wahrscheinlichkeitsrechnung ein wichtiger Bezugspunkt für seine Forschungen. Er war überzeugt, dass soziale und gesellschaftliche Erscheinungen auf Gesetzmäßigkeiten verweisen, die man durch statistische Erhebungen entdecken und erforschen könnte. 1846 führte er – vielleicht die erste – wissenschaftlich fundierte Volkszählung in Belgien durch. Bekannt geworden ist er auch durch seinen aus Erhebungen am Menschen errechneten „mittleren Menschen“ (*homme moyen*) als Idealtyp des Menschen. Diese Theorie war und ist heftig umstritten. Quetelet kann aber als Begründer der Anthropometrie angesehen werden. 1853 organisierte Quetelet den ersten Internationalen Statistikkongress in Brüssel. Das bedeutete eine Stärkung und Förderung der internationalen Zusammenarbeit, die dann 1885 zur Gründung des Internationalen Statistischen Instituts (ISI) führte.

Zu erwähnen ist in diesem Zusammenhang auch der in der Wahrscheinlichkeitsrechnung bekannte *Sir Francis Galton* (1822 – 1911), der übrigens ein Vetter von Charles Darwin war. Er entwickelte das nach ihm benannte Galtonbrett, das der Demonstration der Binomialverteilung dient. Ferner entwickelte er die Korrelationsrechnung zur Auswertung seiner Daten (vornehmlich zur Vererbungs-

lehre). Sein Schüler *Karl Pearson* (1857 – 1936) war Mitbegründer der Zeitschrift *Biometrika*, einer statistischen Zeitschrift (1900).

Im 20. Jahrhundert entwickelten sich mit den Methoden der Wahrscheinlichkeitstheorie neue Verfahren und Möglichkeiten, es ist der Beginn der **mathematischen Statistik**. War lange Zeit die Gesamterhebung das Mittel der Statistik zur Beschreibung der Umwelt, wird jetzt eine repräsentative Teilerhebung (Teiluntersuchung) durchgeführt (Stichprobenverfahren) und aus den Ergebnissen der Teilerhebung durch mathematische Verfahren auf die Gesamtheit zurückgeschlossen. So ist neben der rein deskriptiven (beschreibenden) Statistik die *induktive* (schließende) Statistik getreten. Als *mathematische Statistik* hat sie sich zu einem selbstständigen Zweig der Mathematik entwickelt. Es ist ein Verdienst von *Sir Ronald Aylmer Fisher* (1890 – 1962) die Versuchsplanung eingeführt zu haben, und damit den großen Anwendungsbereich der Statistik in Wirtschafts- und Sozialstatistik begründet zu haben. *Egon Sharpe Pearson* (1895 – 1980), Sohn von Karl Pearson, ist zusammen mit dem in Russland geborenen *Jerzy Neymann* (1894 – 1981) besonders auf dem Gebiet des Testens von Hypothesen bekannt geworden (Neymann-Pearsonsche Theorie des Prüfens von Hypothesen). Statistische Hypothesen (Annahmen) werden mit Hilfe statistischer Tests überprüft, d. h. aufgrund einer Stichprobe wird eine Entscheidung über Annahme oder Ablehnung der Hypothese herbeigeführt. Unter dem Einsatz mathematischer Methoden gibt es inzwischen eine Fülle von Verfahren zur Überprüfung von Hypothesen.

Ganz allgemein kann man sagen, dass es heute in der Statistik nicht nur um eine Beschreibung, sondern auch um eine Auswertung und kritische Beurteilung von erhobenen Daten geht. Statistik und Wahrscheinlichkeit sind heute miteinander verkettet.

Die historisch entstandene Gliederung der Statistik in *Deskriptive Statistik* und *Induktive Statistik* wird heutzutage unter dem Einfluss des Anwendungsgedankens infrage gestellt. Auch unter dem Einfluss der von *Tukey* 1977 eingeführten *Explorativen Datenanalyse* (EDA) ist man geneigt, Ideen und Methoden der induktiven (schließenden) Statistik schon in heuristischer Form frühzeitig einzusetzen (vgl. *Tukey* [172]). So unterwirft man in der EDA u. a. die Datenmenge auch systematischen und probierenden Reduktionen und Umgestaltungen – die modellärmer als in der induktiven Statistik sind – in der Erwartung, dass einfache Zusammenhänge als Muster sichtbar werden und so evtl. zu begründeten Vermutungen führen können.

1.2 Grundbegriffe der beschreibenden Statistik und Aufbereitung der Daten

1.2.1 Statistische Erhebung, Daten, Merkmale, Merkmalsausprägungen

In der beschreibenden Statistik geht es um eine Datenerfassung in Sachsituationen, um die Datenaufbereitung und um eine erste vorsichtige Dateninterpretation.

Didaktische Vorbemerkungen

Grundsätzlich gibt es verschiedene Vorgehensmöglichkeiten zur Einführung in die beschreibende Statistik:

- Man kann Datenlisten oder graphische Darstellungen von Daten vorgeben. Zeitungen, Bücher und statistische Jahrbücher liefern für alle Altersstufen interessantes Datenmaterial (siehe auch die Tabellen und Graphiken in den nachfolgenden Kapiteln). Die Aufgabe kann dann darin bestehen, dieses Datenmaterial richtig zu lesen und zu verstehen und evtl. weiter aufzubereiten. So können beispielsweise Fragen nach anderen graphischen Darstellungen oder Fragen nach Kennzahlen wie Mittelwerte und Streuungswerte gestellt werden.
- Man kann durch eine statistische Erhebung das Datenmaterial finden lassen, das dann aufbereitet wird. Wegen der damit verbundenen hohen Motivation heben Richtlinien und Lehrpläne für den Mathematikunterricht diesen Weg besonders hervor. Planung, Durchführung und anschließende Auswertung einer selbst durchgeführten Erhebung können die Schüler tatsächlich stärker motivieren als eine von außen an sie herangetragene Fragestellung durch Vorgabe von irgendwelchen Daten. Die Schüler können so eigene Erfahrungen sammeln und müssen zudem stets im Gespräch mit ihren Mitschülern bleiben. Allerdings wird man berücksichtigen müssen, dass die Planung, Durchführung und Auswertung einer eigenen Erhebung mehr Zeit beansprucht als die Auswertung stets neu vorgegebener Daten.

In Praktika mit Studenten haben wir mehrfach mit beiden Wegen unterrichtliche Erfahrungen sammeln können. Es kann bestätigt werden, dass der Motivationsschub zu Beginn einer selbst durchgeführten Erhebung sehr stark ist, es muss aber auch eingestanden werden, dass die Aufbereitung immer wieder desselben Datenmaterials unter neuen Fragestellungen schnell das Interesse der Schüler abflachen lässt: „Schon wieder diese Daten!“ Schließlich muss man auch damit rechnen, dass sich die erhobenen Daten nicht immer zur Vorbereitung neuer Fragestellungen (wie z. B. zur Motivation der Frage nach Streuungsmaßen) eignen. Geht man den anderen Weg, so kann man mit Vorteil die Möglichkeit

ausnutzen, für jede neue Fragestellung neue Datenmengen aus neuen aktuellen Sachproblemen wählen zu können. Diese Varianz der Sachgebiete bewirkt aufgrund unserer Erfahrung ebenfalls eine hohe intrinsische Motivation.

Unter Berücksichtigung dieser Überlegungen empfiehlt sich ein Mittelweg als Mischung aus beiden Wegen. In jedem Fall sollte aber zumindest eine im Umfang kleine realisierbare Erhebung etwa im Umfeld der Schule von den Schülern durchgeführt werden. Mögliche Themen wären etwa:

- Erhebung zum Fernsehverhalten der Mitschüler einer bestimmten Klassenstufe,
- Erhebung über aktiv betriebene Sportarten der Schüler einer Schule,
- Erhebung zu Berufs- und Studienwünschen der Schüler der Abgangsklassen,
- Erhebung über Mitgliedschaften von Schülern in Jugendverbänden.

Es ist wichtig, dass die Schüler das Thema der Erhebung selbst bestimmen. Dieses Vorgehen ist besonders *zu Beginn* einer unterrichtlichen Behandlung von großem Vorteil. Schüler lernen so unmittelbar die Schwierigkeiten einer Datenerhebung kennen, und sie sind motiviert, die Daten aufzubereiten und auszuwerten, da sie ja das Thema interessiert und die Daten evtl. interessante Informationen über die Fragestellung liefern.

Welchen Weg man auch beschreitet, stets müssen dabei einige Grundbegriffe der beschreibenden Statistik eingeführt werden, wie z. B. Erhebung, statistische Einheit, Merkmal, Merkmalsausprägung und Häufigkeit. Die Vermittlung einer fachspezifischen Sprache erleichtert dann später das Unterrichtsgespräch.

Grundlegende Begriffe der Statistik

- Unter einer **statistischen Masse** (empirischen Grundgesamtheit) versteht man die durch die Identifikationsmerkmale (z. B. *weibliche* und *männliche* Bevölkerung in *Nordrhein-Westfalen* im Jahre *2010* unter *18 Jahre*) ausgezeichnete und abgegrenzte Menge von Einheiten, in der eine statistische Erhebung zur Untersuchung eines oder mehrerer Merkmale (z. B. Alter, Staatsangehörigkeit) durchgeführt wird.
- Unter einer **statistischen Einheit** (Beobachtungseinheit, Merkmalsträger) versteht man das Einzelobjekt (den Informationsträger) einer statistischen Untersuchung. Jede statistische Einheit muss wie die statistische Masse eindeutig identifizierbar bzw. abgrenzbar sein. Dieses geschieht durch die Identifikationsmerkmale.
- Bei den **Identifikationsmerkmalen** unterscheidet man
 - *sachliche Identifikationsmerkmale* (z. B. weibliche Bevölkerung unter 18 Jahren),
 - *räumliche Identifikationsmerkmale* (z. B. in Nordrhein-Westfalen),
 - *zeitliche Identifikationsmerkmale* (z. B. im Jahre 2010).
- Deckt sich die Menge der untersuchten statistischen Einheiten (Merkmals-träger) mit der statistischen Masse (der empirischen Grundgesamtheit), so

spricht man von einer **Totalerhebung** (z. B. Volkszählung), wird nur ein Teil der statistischen Einheiten untersucht, spricht man von einer **Teilerhebung** oder **Stichprobe** (z. B. Mikrozensus). Eine Totalerhebung ist aufwendig, teuer und nicht immer durchführbar. Will man beispielsweise mittels einer Totalerhebung die Lebensdauer von Glühlampen (Kerzenbirne, 40 Watt, klar, bestimmtes Fabrikat, bestimmter Produktionszeitraum) feststellen, so führt das zwangsläufig zu einer Zerstörung aller Glühbirnen dieses Typs. Deshalb führt man Stichproben durch. Sie sparen Kosten und sind bei den heutigen Methoden (Repräsentativerhebung) äußerst zuverlässig. Auch Volkszählungen (Totalerhebung) werden aus den genannten Gründen deshalb in der Bundesrepublik Deutschland in der Regel nur alle 10 Jahre durchgeführt. Durch den schon erwähnten Mikrozensus, der seit 1957 eingeführt ist, wird die Zeitspanne überbrückt.

- Unter einem **Merkmal** versteht man eine bei einer statistischen Untersuchung interessierende Eigenschaft der statistischen Einheiten. Die statistischen Einheiten heißen deshalb Merkmalsträger und sind es auch.
- Die möglichen Werte (Kategorien), die ein Merkmal annehmen kann, nennt man **Merkmalsausprägungen** (Modalitäten). Beispiel: Merkmal „Geschlecht“, Modalitäten „männlich“ bzw. „weiblich“.
- Registrierte Merkmalsausprägungen werden als statistische **Daten** bezeichnet. Sie sind also beobachtete Werte eines bestimmten Merkmals in einer bestimmten Grundgesamtheit. Die „Beobachtung“ erfolgt nach einem festgelegten Verfahren. Die Daten werden in einer Liste, die als **Urliste** bezeichnet wird, angegeben.
- Ein **Merkmal** heißt **erschöpfend** bezüglich der Grundgesamtheit (bzgl. der statistischen Masse), wenn sich *jedem* Merkmalsträger aus der Grundgesamtheit eine Merkmalsausprägung des Merkmals zuordnen lässt. So ist das Merkmal „Staatsangehörigkeit“ in der Grundgesamtheit Europa mit den vier Ausprägungen „deutsch“, „französisch“, „griechisch“, „italienisch“ sicher nicht erschöpfend. Durch Hinzufügung der Modalität „sonstige“ ist es aber erschöpfend.

Anmerkung:

Durch Hinzufügen der Modalität „sonstige“ kann man ein Merkmal stets zu einem erschöpfenden Merkmal machen. Man betrachte einmal unter diesem Aspekt Fragebögen und Statistiken.

- Von Bedeutung ist die Unterscheidung verschiedener **Merkmalstypen**. Denn um statistische Methoden anwenden zu können, muss feststehen, ob und in welchem Umfang mit registrierten Merkmalsausprägungen (den Daten) gerechnet werden darf. Es geht ja stets um eine Beschreibung der Wirklichkeit. Dazu ist eine Analyse der Sachsituation erforderlich, die zu einem adäquaten mathematischen Modell führt. Damit sind dann auch mögliche Rechenoperationen festgelegt. Die Unterscheidung verschiedener Merkmalstypen liefert in dieser Hinsicht einen ersten Beitrag. Wir unterscheiden:

qualitative Merkmale (lateinisch *qualitas*: Beschaffenheit, Eigenschaft), **Rangmerkmale** und **quantitative Merkmale** (lateinisch *quantus*: wie groß).

- Die **qualitativen Merkmale** werden auch **nominalskalierte Merkmale** genannt (lateinisch *nomen*: Benennung, Wort). Bei ihnen sind die Merkmalsausprägungen nur Beschreibungen, sind also nicht messbar. Die Merkmalsausprägungen lassen sich in keine Reihenfolge bringen, sie stehen gleichberechtigt nebeneinander. Man kann nur feststellen, ob sie bei einer statistischen Einheit zutreffen oder nicht. Beispiele für nominalskalierte Merkmale sind: *Haarfarbe* (z. B. mit den Ausprägungen rot, blond, schwarz, sonstige), *Beruf* (z. B. mit den Ausprägungen Schlosser, Maurer, Elektriker, Kaufmann, Lehrer, Richter), *Staatsangehörigkeit* (z. B. mit den Merkmalsausprägungen deutsch, italienisch, spanisch, französisch), *Familienstand* (z. B. mit den Ausprägungen ledig, verheiratet, verwitwet, geschieden), *Geschlecht* (mit den Merkmalsausprägungen weiblich, männlich).

Qualitative Merkmale erlauben nur Vergleiche der Art „gleich“ bzw. „ungleich“, z. B. Staatsangehörigkeit von Person A ist gleich der Staatsangehörigkeit von Person B. Auch bei einer Codierung der Merkmalsausprägungen durch Zahlen folgt daraus nicht, dass sie sich anordnen lassen. Codiert man z. B. beim Geschlecht „weiblich“ durch „1“ und „männlich“ durch „0“, so macht die Aussage $0 < 1$ doch keinen Sinn.

- Die **Rangmerkmale** werden auch **ordinalskalierte Merkmale** genannt (lateinisch *ordo*: Reihe, Ordnung). Die Merkmalsausprägungen der ordinalskalierten Merkmale lassen sich in eine Reihenfolge bringen. Beispiel: *Leistungsnoten* mit den Ausprägungen sehr gut, gut, befriedigend, ausreichend, mangelhaft, ungenügend. Die Abstände zwischen verschiedenen Ausprägungen sind aber nicht gleich und nicht mathematisch interpretierbar (siehe didaktischen Hinweis 3, S. 13). Bei ordinalskalierten Merkmalen sind aber Vergleiche der Art „folgt vor“, „ist größer“, „ist besser“ möglich und erlaubt. Weitere Beispiele für Rangmerkmale sind die *Handelsklassen bei Obst* (1. Qualität, 2. Qualität usw. oder Handelsklasse A, Handelsklasse B usw.). Auch nach einer Codierung z. B. der Güteklassen bei Obst durch Zahlen ergeben arithmetische Operationen aber keinen Sinn. Wenn bei Obst etwa „Qualität 1“ durch „1“, „Qualität 2“ durch „2“ codiert wird, so macht eine Aussage „ $1 + 2 = 3$ “ vor diesem Hintergrund keinen Sinn.
- Die **quantitativen Merkmale** werden auch als **metrischskalierte Merkmale** bezeichnet. Bei den metrischskalierten Merkmalen sind ihre Ausprägungen angeordnet (eine Reihenfolge liegt fest) und die Abstände zwischen den Merkmalsausprägungen sind mathematisch interpretierbar. Die quantitativen Merkmale haben als Ausprägungen reelle Zahlen. Bei-

spiele: *Anzahl* der Autos mit Katalysator in den EURO-Ländern im Jahre 2010, *Körpergröße* von bestimmten Personen, *Alter* von Schülern einer bestimmten Klasse, *Gewicht* von Personen, *Einkommen* einer festgelegten Personengruppe. Erst die metrischskalierten Merkmale erlauben Vergleiche, Summen- und Differenzenbildungen, sowie die Berechnung des arithmetischen Mittels. Es ist z. B. sinnvoll zu sagen und interpretierbar: Person A hat dreimal so viel verdient wie Person B.

Zählt man bei den Merkmalsausprägungen für jede Ausprägung aus, wie oft sie auftritt, so erhält man die **absolute Häufigkeit** dieser Merkmalsausprägung.

Definition 1.1 (Absolute Häufigkeit, relative Häufigkeit)

Es sei n die Anzahl der statistischen Einheiten, und es seien x_i ($i = 1, 2, \dots, N$) mögliche Merkmalsausprägungen. Dann heißt die Anzahl der statistischen Einheiten mit der Merkmalsausprägung x_i die **absolute Häufigkeit** $H_n(x_i)$ der Merkmalsausprägung x_i . Es entsteht eine **Häufigkeitsverteilung**.

Der Anteil der statistischen Einheiten mit der Merkmalsausprägung x_i an der Gesamtzahl n der statistischen Einheiten heißt **relative Häufigkeit** (syn.: Quote) $h_n(x_i)$ der Merkmalsausprägung x_i , also: $h_n(x_i) := \frac{H_n(x_i)}{n}$. ♦

Die Summe aller relativen Häufigkeiten ist 1.

Didaktische Hinweise

1. Welche Begriffe und welche der genannten Bezeichnungen für die drei Merkmalstypen man im Unterricht verwendet, hängt von der Schulform und der Klassenstufe ab. Wichtiger als die Bezeichnungen sind allerdings die mit ihnen verbundenen Sachinhalte, die der Schüler schon durchschauen sollte. Die Bezeichnungen qualitativ und quantitativ sind griffig, doch das Wort qualitativ kann auch falsche Assoziationen hervorrufen. Denn das Wort qualitativ bezeichnet etwas hinsichtlich der Qualität, und Qualität bedeutet in der Umgangssprache nicht nur Beschaffenheit, sondern beinhaltet auch Güte und Wert. Diese mit dem Wort qualitativ evtl. verbundene Wertung kann bei Schülern dann leicht Irritationen hervorrufen, wenn Merkmalsausprägungen bei qualitativen Merkmalen genannt werden. Die Aufzählung der Berufe Schlosser, Maurer, Lehrer und Richter könnte als Wertung gesehen werden und als Diskriminierung verstanden werden. Die Bezeichnung „nominalskaliertes Merkmal“ scheint in diesem Sinne treffender zu sein. Die Terminologie „-skaliertes Merkmal“ für die drei Merkmalsarten ist für Schüler nicht einfach. Dazu folgende Anmerkung: Jedes Merkmal hat Ausprägungen. Um diese „messen“ zu können, ist eine Skala notwendig. Je nach Art des Merkmals lassen sich seine Ausprägungen durch die folgenden Skalen messen: Nominalskala, Ordinalskala, Metrische Skala. Daraus ergeben sich die genannten Bezeichnungen für die Merkmale.

Nach Abwägen der Vor- und Nachteile würden wir für den Unterricht in der Sekundarstufe I die Verwendung der Benennungen qualitatives Merkmal, Rangmerkmal und quantitatives Merkmal empfehlen.

2. Im Zusammenhang mit quantitativen Merkmalen unterscheidet man häufig zwischen *diskreten* und *stetigen* Merkmalen. Bei den *quantitativ diskreten* Merkmalen können die Ausprägungen nur isolierte Werte annehmen. Beispiel: Die an einer Kreuzung zu einer bestimmten Zeit vorbeifahrenden Autos. Bei den *quantitativ stetigen* Merkmalen können die Ausprägungen die Werte eines Intervalls (Kontinuums) annehmen. Beispiel: Körpergröße, Füllgewicht.

Eine weitere Verfeinerung der metrischskalierten Merkmale in *intervallskalierte* und *proportionalskalierte Merkmale* halten wir für den Schulunterricht für nicht erforderlich. Das klassische Beispiel für ein intervallskaliertes Merkmal ist die Temperatur, die ja z. B. in °Celsius oder °Fahrenheit oder °Réaumur oder °Kelvin gemessen werden kann. Eine Aussage „6 Grad ist doppelt so warm wie 3 Grad“ ist nur sinnvoll, wenn dieselbe Einheit zugrundegelegt wird. Längen und Gewichte sind Beispiele für proportionalskalierte Merkmale.

3. Häufig werden *Leistungsnoten* für die einzelnen Fächer wie quantitative Merkmale behandelt. Man bildet z. B. die Durchschnittsnote als arithmetisches Mittel der Noten. Das ist sicher nicht korrekt, denn die Noten sind nur Rangmerkmale. So ist der Unterschied zwischen „1 sehr gut“ und „2 gut“ sicher anders als der Unterschied zwischen „2 gut“ und „3 befriedigend“. Und wie steht zu diesen Unterschieden der Noten der Unterschied zwischen „4 ausreichend“ und „5 mangelhaft“? Die Unterschiede zwischen den Noten sind nicht gleich. Das erkennt man ganz deutlich, wenn man bedenkt, dass die Noten *verbal* festgelegt sind.

Beispielhaft sei die Bewertung von Prüfungsleistungen gemäß der Lehramtsprüfungsordnung – LPO vom 27.03.2003, zuletzt geändert durch Gesetz vom 27.06.2006 (nach dem Stand vom 01.07.2009) in NRW – angegeben.

Man erkennt in dem nachfolgenden Zitat, dass die verbale Beschreibung der Noten Fragen unbeantwortet lässt. Was bedeutet „durchschnittliche Anforderung“ (befriedigend), was bedeutet „genügt trotz ihrer Mängel noch den Anforderungen“ (ausreichend) usw.?

Die Zuordnung Note \rightarrow Zahl ist ziemlich willkürlich. Die Problematik der arithmetischen Durchschnittsbildung wird besonders deutlich, wenn man andere Zuordnungen Note \rightarrow Zahl als die oben angegebenen, weithin üblichen wählt.

- „(1) Die einzelnen Prüfungsleistungen sind mit einer der folgenden Noten zu bewerten:

| | | | | |
|---|---|--------------|---|---|
| 1 | = | sehr gut | = | eine ausgezeichnete Leistung |
| 2 | = | gut | = | eine Leistung, die erheblich über den durchschnittlichen Anforderungen liegt |
| 3 | = | befriedigend | = | eine Leistung, die durchschnittlichen Anforderungen entspricht |
| 4 | = | ausreichend | = | eine Leistung, die trotz ihrer Mängel noch den Anforderungen genügt |
| 5 | = | mangelhaft | = | eine Leistung, die wegen erheblicher Mängel den Anforderungen nicht mehr genügt |
| 6 | = | ungenügend | = | eine Leistung, die in keiner Hinsicht den Anforderungen entspricht |

- (2) Die Note der Prüfungsleistung wird aus dem arithmetischen Mittel der Einzelnoten der Prüfenden gebildet.“

1.2.2 Graphische Darstellungen der Daten

In diesem Abschnitt und in den nächsten Abschnitten geht es um eine Beschreibung und Strukturierung des Datenmaterials. Man spricht von einer *Aufbereitung* der Daten. Die Bezeichnung Aufbereitung stammt von dem preußischen Statistiker *E. Engel* (1821 – 1896), der sie aus der Bergmannssprache übernahm. Ziel einer Aufbereitung der Daten ist es, wesentliche Informationen einer Erhebung übersichtlich zu vermitteln.

Wir behandeln zunächst die graphischen Darstellungsmöglichkeiten wie *Tabelle*, *Stabdiagramm*, *Kreisdiagramm*, *Blockdiagramm*, *Histogramm*, *Stengel-Blatt-Diagramm* und die *empirische Verteilungsfunktion*. Diese Graphiken werden im Abschnitt 1.2.4 noch ergänzt durch die „Fünf-Zahlen-Zusammenfassung“ (Five-digit-Display) und das „Kastenschaubild“ (Box-Plot-Diagramm).

In den folgenden Abschnitten 1.2.3 bis 1.2.5 werden dann Lageparameter, Streuungsparameter und Lineare Regression und Korrelation besprochen.

Generell gilt: Zur sachgemäßen Interpretation der Daten muss das den Daten zugrundeliegende Begriffsfeld bekannt sein. Das kann nicht immer vorausgesetzt werden und muss gegebenenfalls erarbeitet werden (siehe spätere Beispiele „Verurteilte wegen Vergehen und Verbrechen“ und „Länge der Grenzen Deutschlands“).

Urliste, Tabelle

Bei der Aufbereitung von Daten geht man von der **Urliste** aus. Die Urliste ist eine Aufstellung aller ermittelten Daten $x_1, x_1, x_3, \dots, x_n$. Diese sind entweder in der Reihenfolge der Erhebung oder schon nach anderen Kriterien (etwa Größe oder Häufigkeit) aneinandergereiht. Im folgenden geben wir die Urliste in Form einer **Tabelle** an.

| Urliste | |
|---------------------------------|--------------------------|
| bei einem registrierten Merkmal | |
| Merkmalsträger i | Merkmalsausprägung x_i |
| 1 | x_1 |
| 2 | x_2 |
| 3 | x_3 |
| 4 | x_4 |
| 5 | x_5 |

Hat man weitere Merkmale registriert, so hat die Urliste weitere Spalten für die Merkmalsausprägungen y_i, z_i usw.

Häufig erstellt man die Urliste tabellarisch in Form einer Strichliste (z. B. bei Verkehrszählungen an einer Kreuzung). Dabei bündelt man je fünf Striche zu einer „Einheit“: |||| und gibt die absolute Häufigkeit $H_n(x_i)$ an: 31 Fahrradfahrer, 22 Personenkraftwagen, 6 Lastwagen, 15 Fußgänger.

Neben den absoluten Häufigkeiten $H_n(x_i)$ sind auch die relativen Häufigkeiten $h_n(x_i)$ von Interesse. Die relativen Häufigkeiten werden häufig als prozentualer Anteil angegeben. Durch Rundungen kann die Summe der relativen Häufigkeiten geringfügig von 100 % (bzw. 1) abweichen. Im folgenden Beispiel 1.1 (Länge der Grenzen Deutschlands mit den Nachbarländern nach dem Stand vom 31.12.2000) ergeben die gerundeten Anteile 100 %. Im Beispiel 1.4 (Personalkosten der Krankenhäuser 2007) ergibt sich 100,1 %. Im Beispiel 1.5 (Gestorbene in der Bundesrepublik Deutschland) ergibt die Summe der relativen Klassenhäufigkeiten der zu acht Klassen zusammengefassten Ausgangsdaten den Wert 1.

Beispiel 1.1
(Länge der Grenzen Deutschlands 2000)

Länge der Grenzen mit den Nachbarländern
der Bundesrepublik Deutschland (Stand: 31.12.2000)

| Gemeinsame Grenze mit (Land) | km | Anteil in % |
|------------------------------|-------------------|-------------|
| Dänemark | 67 ¹⁾ | 1,8 |
| Niederlande | 567 ²⁾ | 15,1 |
| Belgien | 156 | 4,1 |
| Luxemburg | 135 | 3,6 |
| Frankreich | 448 | 11,9 |
| Schweiz | 316 ³⁾ | 8,4 |
| Österreich | 815 ⁴⁾ | 21,7 |
| Tschechische Republik | 811 | 21,6 |
| Polen | 442 | 11,8 |
| | 3757 | 100,0 |

- 1) Landgrenze, Seegrenze nicht endgültig festgelegt.
 - 2) Festlandgrenze (ohne Dollart und Außenbereich der Ems)
 - 3) Vom Dreiländereck Deutschland - Frankreich - Schweiz bis einschließlich
Konstanzer Bucht (mit Exklave Büsingen, aber ohne Obersee des Bodensees)
 - 4) Ohne Bodensee
- (Quelle der Daten: Statistisches Jahrbuch 2009 für die Bundesrepublik Deutschland.
Wiesbaden 2009, S. 21)



Beispiel 1.2
(Verurteilte Personen) Im diesem Beispiel wird die statistische Masse nach *zwei* Merkmalen untersucht. Das Merkmal „Verurteilte Person“ hat die Merkmalsausprägungen Jugendlicher, Heranwachsender, Erwachsener, das Merkmal „Verurteilter wegen Vergehen im Straßenverkehr“ hat die Merkmalsausprägungen „ohne Trunkenheit“ und „in Trunkenheit“. Das führt zu einer erweiterten Form der Tabelle. Die Tabelle erhält zwei Eingänge: den Spalteneingang und den Zeileneingang. Die Tabelle enthält ferner eine Spalte bzw. Zeile für Zeilen- bzw. Spaltenzusammenfassungen. Sie werden *Randspalte* bzw. *Randzeile* genannt. Die Schnittstelle von Randspalte und Randzeile gibt die Summe der statistischen Einheiten an oder (bei Prozentangaben) 100 % (siehe Beispiel 1.1).

Wegen Vergehen im Straßenverkehr im Jahre 2007
Verurteilte in der Bundesrepublik Deutschland

| | Jugendliche | Heranwachsende | Erwachsene | |
|---|-------------|----------------|------------|--------|
| Verurteilte mit Vergehen ohne Trunkenheit | 5516 | 8832 | 80652 | 95000 |
| Verurteilte mit Vergehen in Trunkenheit | 1424 | 9394 | 106028 | 116846 |
| | 6940 | 18226 | 186680 | 211846 |

(Quelle der Daten: Statistisches Jahrbuch 2009 für die Bundesrepublik Deutschland. Wiesbaden 2009, S. 275.)

Anmerkungen zu diesem Beispiel:

Auch wenn die Daten an sich schon beeindruckend sind, muss man zur sachgemäßen Beurteilung der Daten zusätzlich Sachkenntnisse über die in der Tabelle genannten Begriffe aus der Rechtskunde besitzen:

- Vergehen sind von Verbrechen zu unterscheiden. Nach § 12 Verbrechen und Vergehen des Strafgesetzbuches (StGB) gilt:
„(1) Verbrechen sind rechtswidrige Taten, die im Mindestmaß mit Freiheitsstrafe von einem Jahr oder darüber bedroht sind.
(2) Vergehen sind rechtswidrige Taten, die im Mindestmaß mit einer geringeren Freiheitsstrafe oder die mit einer Geldstrafe bedroht sind.“
- Jugendlicher ist, wer zur Zeit der Tat 14, aber noch nicht 18 Jahre alt ist (Jugendgerichtsgesetz (JGG)). Heranwachsende im Sinne des Strafrechts sind Personen von 18 bis einschließlich 20 Jahre (JGG). Erwachsene sind 21 Jahre und älter.
- Erwachsene unterliegen ausschließlich den Vorschriften des allgemeinen Strafrechts, Jugendliche werden nach Jugendstrafrecht behandelt. Heranwachsende nehmen bei Anwendung des Strafrechts eine Sonderstellung ein. Bei ihnen kann allgemeines Strafrecht oder Jugendstrafrecht zur Anwendung kommen. Ein wesentliches Entscheidungskriterium ist hierfür zum Beispiel die „Reife“ des Heranwachsenden, d. h. die sittliche und geistige Entwicklung des Heranwachsenden.
- Verurteilte sind Straffällige, gegen die entweder nach allgemeinem Strafrecht eine Freiheitsstrafe, Strafarrrest und/oder Geldstrafe verhängt worden ist, oder deren Straftat nach Jugendstrafrecht mit Jugendstrafe und/oder Maßnahmen geahndet worden ist. Die Jugendstrafe beträgt mindestens 6 Monate. Maßnahmen sind Zuchtmittel (z. B. Verwarnung, Auferlegung besonde-

rer Pflichten, Freizeitarrrest) und Erziehungsmaßregeln (z. B. Schutzaufsicht, Fürsorgeerziehung).

- Im Strafrecht gibt es auch noch den Begriff Kind. Kinder sind Personen, die noch keine 14 Jahre alt sind. Sie sind strafunmündig/schuldunfähig (§ 19 StGB).
- Welcher Gruppe ein Mensch zugeordnet wird, hängt von seinem Alter zur Tatzeit ab.

■

Anmerkung:

Eine Tabelle soll eine kurze zutreffende Überschrift tragen, die den Leser über das Untersuchungsobjekt informiert. Die Eingangszeilen und Eingangsspalten sollen präzise Benennungen tragen.

Stabdiagramm

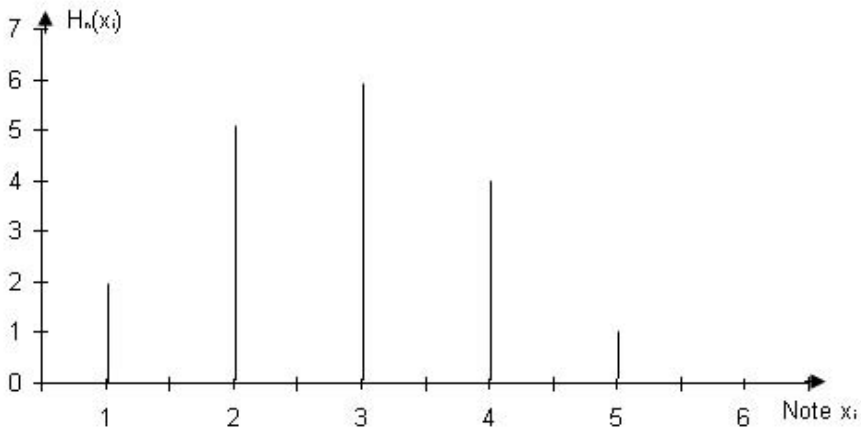
Einen hohen Grad an Anschaulichkeit gewinnt man, wenn man die absoluten und relativen Häufigkeiten graphisch darstellt. Um Häufigkeiten darzustellen, gibt es verschiedene Möglichkeiten wie Stabdiagramm, Kreisdiagramm, Blockdiagramm und Histogramm. Dabei kann es sein, dass man einen Informationsverlust in Kauf nehmen muss, insbesondere dann, wenn die Zahlen nicht gleichzeitig im Diagramm übermittelt werden. Anschaulichkeit *und* möglichst umfassende Informationen sollten aber stets im Blick bleiben. Deshalb erhalten die genannten Diagramme häufig auch die realen Zahlen.

Das **Stabdiagramm** verwendet Stäbe in einem rechtwinkligen Koordinatensystem. Auf der y -Achse werden die Häufigkeiten abgetragen, und auf der x -Achse werden die Merkmalsausprägungen notiert. Bei qualitativen Merkmalen ist die Einteilung auf der Achse für die Merkmalsausprägungen willkürlich (Nominalskala). Die Abstände zwischen den Ausprägungen können beliebig gewählt werden. Aus optischen Gründen sollten auch bei nominalskalierten Daten die Abstände zwischen den Merkmalsausprägungen gleich gewählt werden. Die Anordnung der Merkmalsausprägungen ist bei qualitativen (nominalskalierten) Merkmalen beliebig. Bei Rangmerkmalen hat die Einteilung jedoch der Anordnung der Merkmalsausprägungen zu folgen. Die Stablänge gibt die absolute bzw. relative Häufigkeit der Merkmalsausprägungen an. Wenn man Vergleiche anstellen möchte, ist die Verwendung der relativen Häufigkeiten statt der absoluten Häufigkeiten zu empfehlen. Die Summe der Längen sämtlicher Stäbe ergibt bei der Verwendung relativer Häufigkeiten Eins.

Beispiel 1.3

(Klausurnoten) Ein Schüler hat seine Klausurnoten aus den letzten Jahren im Fach Mathematik aufgeschrieben: 3, 4, 3, 2, 5, 4, 2, 3, 2, 1, 2, 1, 4, 3, 2, 3, 4, 3.

Darstellung der Daten im Stabdiagramm:

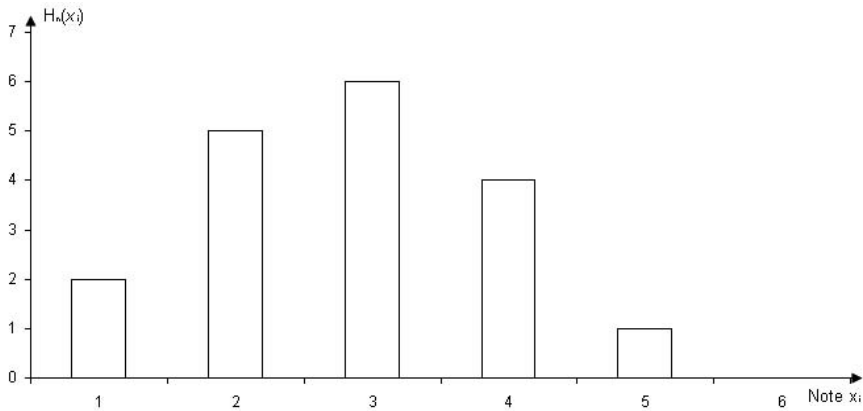


Ist der Stichprobenraum sehr groß, können große absolute Häufigkeiten auftreten. Das führt zu Schwierigkeiten, wenn man im Stabdiagramm die absoluten Häufigkeiten darstellen möchte. Man „hilft“ sich dann häufig so, dass das Stabdiagramm „durchtrennte“ Stäbe oder „abgeknickte“ Stäbe enthält.

Nicht sachgemäß ist es, die Endpunkte der Stäbe bei qualitativen und diskreten Merkmalen durch Strecken miteinander zu verbinden.

Häufig verwendet man bei Stabdiagrammen zur „optischen Aufbesserung“ Rechtecke als Stäbe. Nach wie vor soll aber die *Höhe* der Rechtecke ein Maß für die absolute bzw. für die relative Häufigkeit der Merkmalsausprägungen sein. Da das Auge aber die Größe der Fläche wahrnimmt, müssen die Rechtecke eine *gemeinsame Breite haben, wenn die Höhe der Rechtecke ein Maß für die Häufigkeit ist*. Anderenfalls sind Fehlinterpretationen nicht auszuschließen.

Stabdiagramm für Beispiel 1.3 (Klausurnoten) mit Rechtecken als Stäbe:



Bei graphischen Darstellungen in Zeitungen und Zeitschriften befinden sich die Stäbe häufig in horizontaler Lage. Solche Stabdiagramme nennt man auch **Balkendiagramme**.

Kreisdiagramm

Verwendet man Kreis- und Blockdiagramme und Histogramme, um Häufigkeiten darzustellen, so wird die *Fläche* als Mittel der Veranschaulichung herangezogen.

Beim **Kreisdiagramm** wird jeder Merkmalsausprägung ein Kreissektor zugeordnet. Bezeichnet $h_n(x_i)$ die relative Häufigkeit der Merkmalsausprägung x_i , so ist der Mittelpunktswinkel α_i des zugehörigen Kreissektors bestimmt durch

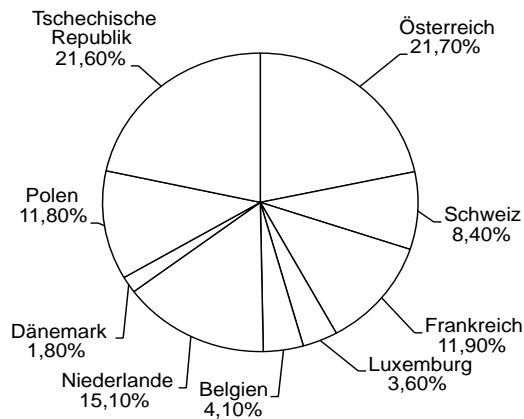
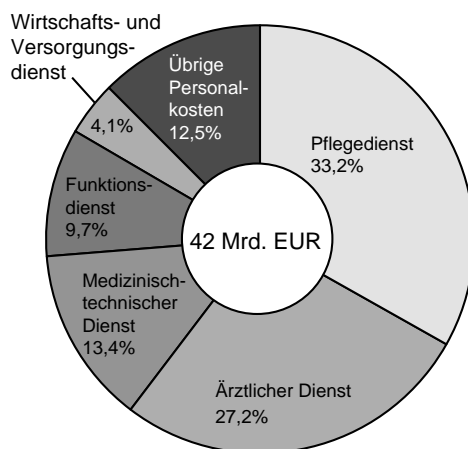
$$\alpha_i = h_n(x_i) \cdot 360^\circ.$$

Die relative Häufigkeit 1 bzw. 100 % entspricht dem Winkel 360° .

Im nachstehenden Kreisdiagramm für das Beispiel 1.1 (Länge der Grenzen Deutschlands) erleichtern die zusätzlich angegebenen Prozentzahlen das richtige Lesen. Der zugehörige Kreissektor unterstützt also das Einprägen der Zahlen. Fehlen die Anteilsangaben, so erhält man durch das Kreisdiagramm optisch nur eine Vorstellung von den Größenverhältnissen. Zur exakten rechnerischen Bestimmung der relativen Häufigkeit $h_n(x_i)$ müsste in diesem Fall zuerst der zugehörige Winkel α_i gemessen werden.

Im Kreisdiagramm findet man gelegentlich Angaben zur Datenmenge. Stichprobenumfang, Einheitenangaben, Prozentangaben oder Jahreszahlen sind häufige Angaben. (Siehe das nachfolgende Kreisdiagramm für das Beispiel 1.4. Die Summe der Prozentangaben ergeben durch Rundungen 100,1 %.)

Kreisdiagramm für das Beispiel 1.1 (Deutsche Ländergrenzen)

**Beispiel 1.4****(Personalkosten der Krankenhäuser 2007)**Personalkosten der Krankenhäuser 2007
in der Bundesrepublik Deutschland

(Abbildung entnommen: Statistisches Jahrbuch 2009 für die Bundesrepublik Deutschland. Wiesbaden 2009, Seite 243. Die noch spezifizierteren genauen Daten sind a.a.O. auf Seite 252 angegeben.)

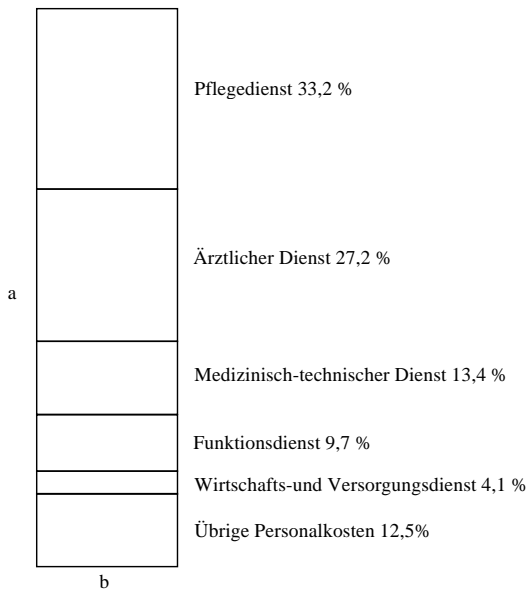


Blockdiagramm

Auch **Blockdiagramme** benutzen *Flächen* zur Darstellung der Häufigkeitsverteilungen. Man unterteilt ein Rechteck mit der Breite b und der Länge a in Teilrechtecke für die relativen Häufigkeiten der Merkmalsausprägungen x_i , $i = 1, \dots, n$. Die Wahl von a und b ist beliebig. Die Teilrechtecke für die Merkmalsausprägungen x_i haben dieselbe Breite b . Die Länge l_i des Rechtecks für das Merkmal x_i berechnet sich nach

$$l_i = h_n(x_i) \cdot a.$$

Blockdiagramm für das Beispiel 1.4 (Personalkosten der Krankenhäuser)



Hinweis:

Häufig werden durch die moderne Computergraphik auch **dreidimensionale Darstellungen** in der beschreibenden Statistik üblich. Da dann für den Betrachter das *Volumen* das bestimmende optische Element ist, kann das bei oberflächlicher Betrachtung leicht zu Fehlinterpretationen führen. Besonders häufig treten sog. quaderförmige Säulendiagramme und Tortendiagramme auf. Der optisch gefällige Eindruck kann nicht darüber hinwegtäuschen, dass die Ablesegenauigkeit evtl. erschwert ist, wenn Zahlenangaben fehlen.

Histogramm

Für die Darstellung von Häufigkeitsverteilungen quantitativer Merkmale sind grundsätzlich alle bisher genannten Graphiken geeignet.

In der Praxis hat man es bei quantitativen Merkmalen häufig mit einer großen Anzahl von Merkmalsausprägungen zu tun, so dass man sie aus Gründen der Übersichtlichkeit zu Klassen zusammenfasst. Bei stetigen quantitativen Merkmalen findet eine solche **Klassenbildung** häufig schon bei der Datenerhebung statt. Die graphische Darstellung von Klassenhäufigkeiten führt zu **Histogrammen**. Am folgenden Beispiel aus der Praxis erläutern wir das Vorgehen.

Beispiel 1.5
(Gestorbene in der Bundesrepublik Deutschland)

Gestorbene in der Bundesrepublik Deutschland im Jahr 2007
männlichen Geschlechts nach Altersgruppen
(ohne Totgeborene, nachträglich beurkundete Kriegssterbefälle und
gerichtliche Todeserklärungen; einschließlich Ausländer)

| Alter von ... bis unter ... Jahren | Gestorbene 2007 | Alter von ... bis unter ... Jahren | Gestorbene 2007 |
|---------------------------------------|--------------------|---------------------------------------|--------------------|
| 0 – 1 | 1 518 | 45 – 50 | 10 931 |
| 1 – 5 | 301 | 50 – 55 | 15 460 |
| 5 – 10 | 220 | 55 – 60 | 20 949 |
| 10 – 15 | 223 | 60 – 65 | 26 431 |
| 15 – 20 | 990 | 65 – 70 | 48 440 |
| 20 – 25 | 1 503 | 70 – 75 | 56 006 |
| 25 – 30 | 1 575 | 75 – 80 | 65 827 |
| 30 – 35 | 1 755 | 80 – 85 | 59 926 |
| 35 – 40 | 3 257 | 85 – 90 | 42 055 |
| 40 – 45 | 6 535 | 90 und mehr | 27 237 |
| | | Insgesamt | 391 139 |

(Quelle der Daten: Statistisches Jahrbuch 2009 für die Bundesrepublik Deutschland.
Wiesbaden 2009, S. 59)

Wie man erkennt, wählt man in der beschreibenden Statistik bei der Klasseneinteilung generell *halboffene* (meistens *rechtsoffene*) Intervalle (Klassen). Die erste bzw. letzte Klasse kann zudem links bzw. rechts unbeschränkt sein. Generell soll die Anzahl der Klassen aus Gründen der Übersichtlichkeit nicht zu groß sein (≤ 20).

Im Beispiel 1.5 verändern wir im Folgenden die vorgegebene kleinschrittige Klasseneinteilung und wählen eine Einteilung in acht Klassen. Ferner haben wir die letzte unbeschränkte Altersklasse „90 und mehr“ nach oben durch 105

abgeschlossen (siehe nachfolgende Tabelle). Die relativen Klassenhäufigkeiten stellen wir in einem Histogramm dar.

Gestorbene in der Bundesrepublik Deutschland im Jahr 2007
männlichen Geschlechts nach Altersgruppen

| Alter von ... bis unter ... Jahren | absolute Klassen- häufigkeit | relative Klassen- häufigkeit | Klassen- breite Δ_i | Häufigkeits- dichte f_i |
|---------------------------------------|------------------------------------|------------------------------------|-------------------------------|------------------------------|
| 0 – 15 | 2262 | 0,00578 | 15 | 0,000386 |
| 15 – 25 | 2493 | 0,00637 | 10 | 0,000637 |
| 25 – 45 | 13122 | 0,03355 | 20 | 0,001678 |
| 45 – 65 | 73771 | 0,18861 | 20 | 0,009431 |
| 65 – 75 | 104446 | 0,26703 | 10 | 0,026703 |
| 75 – 85 | 125753 | 0,32150 | 10 | 0,032150 |
| 85 – 90 | 42055 | 0,10752 | 5 | 0,021504 |
| 90 – 105 | 27237 | 0,06964 | 15 | 0,004643 |
| | 391139 | 1 | | |

Bei einem Histogramm werden in einem rechtwinkligen Koordinatensystem über den einzelnen Klassen Rechtecke gezeichnet. Als Maß für die absolute bzw. relative Klassenhäufigkeit ist die *Fläche* der Rechtecke (und nicht ihre Höhe) festgelegt. Auf der horizontalen Achse werden die Klassenbreiten dargestellt. Seien allgemein n Klassen $[x_1, x_2[$, $[x_2, x_3[$, \dots , $[x_n, x_{n+1}]$ mit $x_1 < x_2 < \dots < x_{n+1}$ gegeben, so versteht man unter der **Klassenbreite** Δ_i der i -ten Klasse die Differenz

$$\Delta_i = x_{i+1} - x_i, \quad i = 1, 2, \dots, n.$$

Die Klassenbreite legt eine Seite des Rechtecks fest. Auf der vertikalen Achse werden nicht die absoluten bzw. relativen Klassenhäufigkeiten abgetragen, sondern die sog. Häufigkeitsdichten f_i . Die **Häufigkeitsdichte** f_i (also die Rechteckshöhe) ist für relative Klassenhäufigkeiten der Quotient

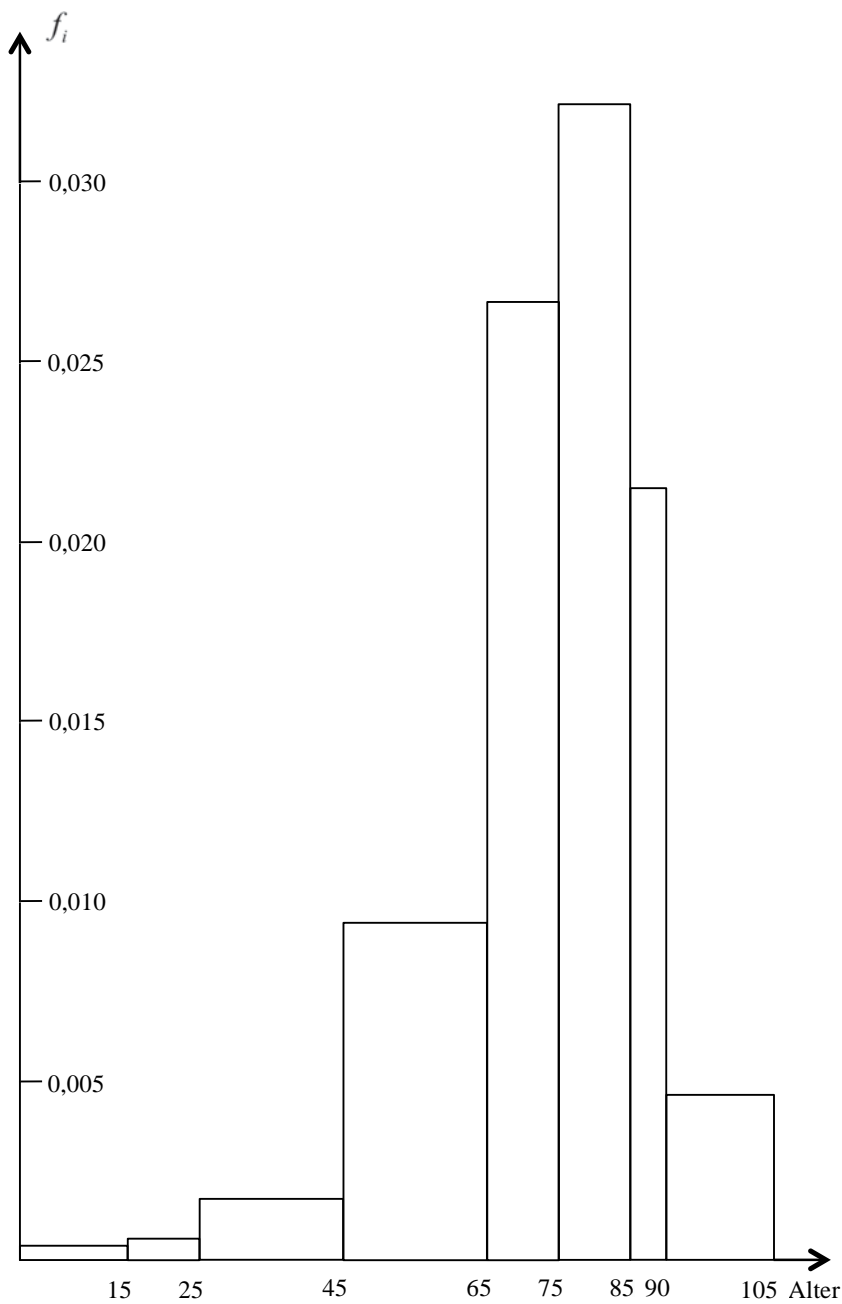
$$f_i = \frac{\text{relative Klassenhäufigkeit der Klasse } i}{\text{Klassenbreite } \Delta_i}.$$

Das Produkt „Klassenbreite Δ_i · Häufigkeitsdichte f_i “ ist also die Maßzahl des Flächeninhalts des Rechtecks und gibt damit die relative Klassenhäufigkeit der Klasse i an.

Sind alle Klassen gleich breit, können die Höhen der Rechtecke unmittelbar als Maß für die Klassenhäufigkeit angesehen werden.

Analog wird auch die Häufigkeitsdichte für die absolute Klassenhäufigkeit bestimmt.

Histogramm für Beispiel 1.5 (Gestorbene in der Bundesrepublik Deutschland)
gemäß obiger Klasseneinteilung



Didaktischer Hinweis:

Es sei darauf hingewiesen, dass durch die Festlegung der Klassenbreiten und Klassenanzahl die Gefahr fahrlässiger oder sogar gewollter Täuschung besteht.

Wir empfehlen deshalb dem Leser zur Herstellung eines Histogramms

- ausgehend von der ursprünglichen Klasseneinteilung im Statistischen Jahrbuch 2009 eine solche Klassenunterteilung vorzunehmen, die sich von unserer deutlich unterscheidet, und das dazu gehörige Histogramm zu erstellen,
- für die ursprünglich vorgegebene Klasseneinteilung im Statistischen Jahrbuch 2009 das Histogramm zu erstellen. Was kann man hier beobachten?

Ferner möchten wir in diesem Zusammenhang auch hinweisen auf Beispiel 1.7 (Gehaltsstatistik eines Betriebes). ■

Stengel-Blatt-Diagramm

Um auf elementarer Ebene Daten übersichtlich anzuordnen und gleichzeitig Klassenhäufigkeitsverteilungen deutlich zu machen, kann man das **Stengel-Blatt-Diagramm** (*stem-and-leaf-display*) einsetzen. Es gehört zu den Methoden der Explorativen Datenanalyse (abgekürzt heute als EDA), die Tukey 1977 in seinem Buch *Exploratory Data Analysis* dargestellt hat.

Beim Stengel-Blatt-Diagramm werden nur die „führenden“ Ziffern der Daten berücksichtigt und nach einem bestimmten Schema notiert. Die erste bzw. die ersten (zwei) Ziffern der Daten werden links von einem „senkrecht“ zur Heftseite gezogenen Strich, die zweite bzw. die dritte Ziffer (allgemeiner: die direkt auf sie folgende Ziffer) rechts vom Strich in der gleichen Zeile aufgeschrieben. Die anderen nachfolgenden Ziffern der Daten bleiben unberücksichtigt. Die links vom Trennstrich geschriebene Ziffernfolge bildet den Stengel (Stamm), die rechts geschriebenen Ziffern sind die Blätter. Die im Stamm untereinanderstehenden Zahlen markieren also die Klassen, die rechts vom Strich in der Zeile hinter einer „Stammzahl“ stehenden Ziffern geben die Beobachtungswerte innerhalb der Klasse an. Diese Ziffern werden der Größe nach geordnet.

Beispiel 1.6

(Körpergewicht von Kindern) Bei einer medizinischen Untersuchung einer Schulklasse wurden bei den 30 Kindern folgende Körpergewichte (in kg) notiert (Urliste):

| | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 35 | 27 | 36 | 42 | 50 | 32 | 35 | 29 | 44 | 40 |
| 36 | 38 | 45 | 40 | 42 | 34 | 38 | 43 | 45 | 42 |
| 37 | 45 | 51 | 48 | 31 | 34 | 46 | 30 | 38 | 35 |

bestimmten Stelle auf. Durch die Folge der Summen der relativen Häufigkeiten kann eine Funktion bestimmt werden, die als die (kumulative) **empirische Verteilungsfunktion** H bezeichnet wird. H wird für alle $x \in \mathbb{R}$ definiert und nimmt natürlich nur Werte aus $[0, 1]$ an. Bei der Größe nach geordneten Merkmalsausprägungen a_1, a_2, \dots, a_s definiert man $H(x)$ durch

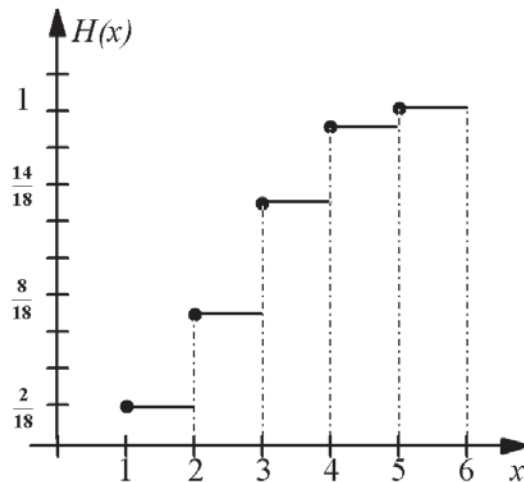
$$H(x) := \begin{cases} 0 & \text{für } x < a_1 \\ \sum_{i=1}^r h_n(a_i) & \text{für } a_r \leq x < a_{r+1}, \quad r+1 < s \\ 1 & \text{für } x \geq a_s \end{cases}.$$

Die **empirische Verteilungsfunktion** für das Beispiel 1.3 (Klausurnoten) ist gegeben durch:

$$H(x) = \begin{cases} 0 & \text{für } x < 1 \\ \frac{2}{18} & \text{für } 1 \leq x < 2 \\ \frac{7}{18} & \text{für } 2 \leq x < 3 \\ \frac{13}{18} & \text{für } 3 \leq x < 4 \\ \frac{17}{18} & \text{für } 4 \leq x < 5 \\ 1 & \text{für } x \geq 5 \end{cases}.$$

Die empirische Verteilungsfunktion ist bei diskreten Merkmalen (wie in diesem Beispiel) eine Treppenfunktion. Im folgenden Graphen bedeutet der Punkt \bullet , dass der Funktionswert bei $x = a_r$ angenommen wird. Die empirische Verteilungsfunktion ist rechtsseitig stetig.

Graph der empirischen Verteilungsfunktion für Beispiel 1.3



In der Regel hat man es bei den angesprochenen Fragen mit klassierten Daten zu tun. Hier sind entsprechend die Klassenhäufigkeiten aufzuaddieren (zu kumulieren). Hat man eine Klasseneinteilung mit den Klassen k_1, k_2, \dots, k_s mit

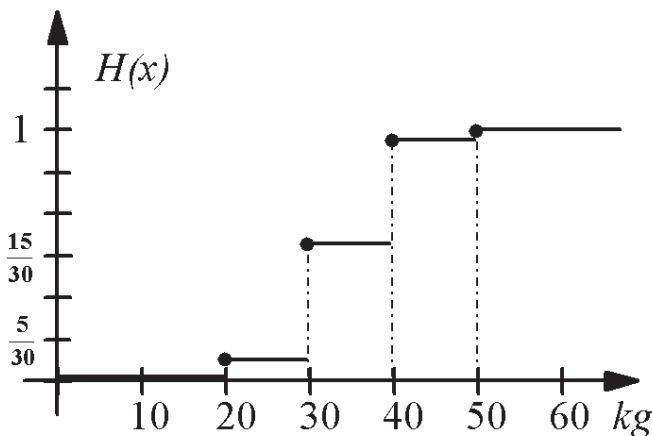
$k_i = [x_{i-1}, x_i[$, und bedeutet $h_n(k_i)$ die relative Häufigkeit der Klasse k_i , so nimmt man als Näherung der empirischen Verteilungsfunktion die Funktion $H : \mathbb{R} \rightarrow [0, 1]$ mit

$$H(x) := \begin{cases} 0 & \text{für } x < x_0 \\ \sum_{i=1}^r h_n(k_i) & \text{für } x \in k_r, \text{ also } x_{r-1} \leq x < x_r, 1 \leq r < s \\ 1 & \text{für } x \geq x_s \end{cases} .$$

Für die empirische Verteilungsfunktion des Beispiels 1.6 (Körpergewicht von Kindern) erhalten wir, wenn wir die vier Klassen

$$[20, 30[, \quad [30, 40[, \quad [40, 50[, \quad [50, 60[$$

bilden, das folgende Bild:



1.2.3 Lageparameter

Zur Beschreibung der Daten, insbesondere wenn die Daten sehr umfangreich sind, gibt man geeignet gewählte Kennziffern (statistische Maßzahlen, auch Parameter genannt) an. Sie sollen die Daten gut repräsentieren, überschaubar und mit Daten aus ähnlichen Erhebungen vergleichbar machen. Man unterscheidet zwischen Lageparametern und Streuungsparametern. Die Lageparameter wie z. B. arithmetisches Mittel, Median usw. geben Aufschluss über das Zentrum einer Verteilung. Die Streuungsparameter wie z. B. Spannweite, empirische Standardabweichung geben Aufschluss über die Streuung der Werte einer Verteilung. Lageparameter und Streuungsparameter ergänzen also einander und gehören zur genaueren Beschreibung einer Verteilung zusammen.

Wir besprechen zunächst die Lageparameter arithmetisches Mittel, geometrisches Mittel, harmonisches Mittel, Median (allgemeiner: Quantile) und den Modalwert.

Dass eine Beschäftigung mit Mittelwerten dringend geboten erscheint, ergibt sich aus Erfahrungsberichten:

- 1975 fand der National Assessment of Educational Progress, dass nur 69 % der Erwachsenen richtig einen einfachen Mittelwert berechnen konnten, und dass 45 % der Erwachsenen Schwierigkeiten hatten, eine Steuertabelle zu gebrauchen (vgl. Goodman [60]).
- Untersuchungen von Barr [8] zeigten, dass Studenten (69 % studierten Ingenieurwissenschaften, 31 % Naturwissenschaften) nur oberflächliche Vorstellungen von Median und Modalwert hatten. Aufgrund einer Analyse der verwirrten Ansichten kann man annehmen, dass die Studenten zum Teil nicht wussten, wie eine Häufigkeitstabelle konstruiert ist.
- Shahani [161] zeigt an einigen eindrucksvollen Beispielen, wie die falsche Verwendung von Mittelwerten in bestimmten Sachzusammenhängen überraschend falsche Aussagen liefern kann.
- H.-J. Schmidt [151] berichtet über einen Test, bei dem Schüler aufgrund von 4 Versuchen die Formel für Quecksilberoxid herleiten sollten. 57,7 % der Schüler benutzten bei der Lösung lediglich Einzelwerte, sie tun so, als ob nur 1 Versuch vorliegt. Wird eine Mittelwertbildung aus den 4 Versuchen vorgenommen, so ist sie in 16,3 % prinzipiell falsch.
- Die *Süddeutsche Zeitung* berichtet in ihrem Magazin vom 21.08.1998 über eine EMNID-Umfrage. Befragt wurden 1000 Deutsche: Was bedeutet 40 %? Es war eine von den drei Antwortmöglichkeiten a) ein Viertel, b) 4 von 10, c) jeder Vierzigste auszuwählen. Ein Drittel der Befragten gab eine falsche Antwort.

Arithmetisches Mittel

Das arithmetische Mittel ist der wohl bekannteste und am häufigsten gebrauchte Mittelwert.

Definition 1.2 (Arithmetisches Mittel)

Es seien x_1, x_2, \dots, x_n Daten eines quantitativen Merkmals. Dann heißt

$$\bar{x} := \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.1)$$

arithmetisches Mittel dieser Daten. ◆

Man bildet also die Summe aller Daten und dividiert die Summe durch die Anzahl der Daten.

Didaktische Hinweise und Ergänzungen

1. Das **arithmetische Mittel** kann nur bei quantitativen Merkmalen benutzt werden, denn nur diese Merkmale gestatten die Durchführung der zur Berechnung des arithmetischen Mittels notwendigen Operationen.
2. Aus der obigen Definitionsgleichung (1.1) für das arithmetische Mittel folgt durch eine elementare Umformung

$$n \cdot \bar{x} = (x_1 + x_2 + x_3 + \dots + x_n) = \sum_{i=1}^n x_i,$$

- d. h. die Summe aller n Einzelwerte kann man sich ersetzt denken durch das Produkt $n \cdot \bar{x}$, also durch die Summe von n gleich großen (errechneten) Werten \bar{x} . Das arithmetische Mittel \bar{x} nimmt also eine Ersatzfunktion wahr.
3. Wenn man die Summe von n realen Daten unterschiedlicher Größe gemäß Punkt 2. durch die Summe von n gleich großen Daten der Größe \bar{x} ersetzen kann, ergibt sich daraus durch einfache Rechnung

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Das bedeutet: **Die algebraische Summe der Abweichungen** (nach oben und nach unten) **aller Daten** x_i ($i = 1, 2, 3, \dots, n$) **von ihrem arithmetischen Mittel \bar{x} ist Null.**

Diese Eigenschaft könnte man auch als definierende Eigenschaft für die Definition des arithmetischen Mittels wählen. Das hat aber den Nachteil, dass dann die Definition nicht unmittelbar eine Berechnungsvorschrift für das arithmetische Mittel liefert.

4. Das arithmetische Mittel wird von einzelnen Daten, die extrem von den anderen Daten abweichen, stark beeinflusst. Wir betrachten ein Beispiel: Fünf Schüler wollen das restliche Geld von einer Fahrt, das jeder noch hat, unter sich aufteilen, so dass jeder gleich viel hat. A besitzt 3 Euro, B 2 Euro, C noch 5 Euro, D 1 Euro, E noch 4 Euro. Es beträgt $\bar{x} = 3$ Euro. Hat Schüler E statt 4 Euro noch 15 Euro, so ist der neue Mittelwert $\bar{x} = 26 : 5 = 5,2$ [Euro].
5. Für eine Berechnung des arithmetischen Mittels kann die unter Punkt 3. genannte Beziehung einen interessanten Weg eröffnen. Man geht von einem angenommenen Wert als arithmetisches Mittel aus und versucht durch Ausnutzen der Eigenschaft 3. das exakte arithmetische Mittel zu bestimmen. Wir erläutern das „**Verfahren zur Bestimmung des arithmetischen Mittels durch Korrektur eines geschätzten arithmetischen Mittels**“ zunächst am Beispiel unter Punkt 4 (Geldbeträge von 5 Schülern). Wir gehen aus von einem beliebigen Wert als Schätzwert für \bar{x} . Für das konkrete Beispiel wählen wir die Zahl 2,3. Jetzt bestimmen wir alle Abweichungen der realen Daten von 2,3. Wir erhalten: $3 - 2,3 = 0,7$; $2 - 2,3 =$

$-0,3$; $5 - 2,3 = 2,7$; $1 - 2,3 = -1,3$; $4 - 2,3 = 1,7$. Die Summe der Abweichungen beträgt: $0,7 - 0,3 + 2,7 - 1,3 + 1,7 = 3,5$. Jeder der fünf realen Werte weicht also im Mittel um $3,5 : 5 = 0,7$ vom geschätzten arithmetischen Mittel $2,3$ ab. Deshalb addieren wir $0,7$ zu $2,3$. Wir erhalten 3 , diese Zahl ist das arithmetische Mittel im Beispiel.

In Aufgabe 1 des Abschnitts 1.2.8 ist das Verfahren allgemein zu beschreiben und zu begründen.

Gewogenes arithmetisches Mittel

Treten Daten mehrfach auf, kann man sie als Summe gleicher Summanden zu einem Produkt zusammenfassen.

Wir formulieren den Sachverhalt allgemein:

Sind $x_1, x_2, x_3, \dots, x_n$ Daten eines quantitativen Merkmals und kommt x_i insgesamt g_i mal vor, so gilt für das arithmetische Mittel

$$\bar{x} = \frac{g_1 x_1 + g_2 x_2 + \dots + g_n x_n}{g_1 + g_2 + \dots + g_n} = \frac{\sum_{i=1}^n g_i x_i}{\sum_{i=1}^n g_i}. \quad (1.2)$$

Die Faktoren g_i in Gleichung (1.2) drücken also aus, wie oft die Daten x_i jeweils in der Liste vorkommen.

Gleichung (1.2) kann aber auch so gedeutet werden, dass einige Daten ein anderes (vielleicht ein höheres) „Gewicht“ haben als andere. In Verallgemeinerung führt das zu folgender

Definition 1.3 (Gewogenes arithmetisches Mittel)

Sind $x_1, x_2, x_3, \dots, x_n$ Daten eines quantitativen Merkmals, so heißt

$$\bar{x} := \frac{g_1 x_1 + g_2 x_2 + \dots + g_n x_n}{g_1 + g_2 + \dots + g_n} = \frac{\sum_{i=1}^n g_i x_i}{\sum_{i=1}^n g_i}$$

mit $g_i \geq 0$ für $i = 1, 2, 3, \dots, n$, und $\sum_{i=1}^n g_i > 0$ **gewogenes arithmetisches Mittel der Daten**. Die nichtnegativen Zahlen g_i heißen **Gewichtungsfaktoren** oder kurz **Gewichtsfaktoren**. ♦

Das gewogene arithmetische Mittel kommt in der Praxis häufig vor, beispielsweise zur Berechnung der *Tagesdurchschnittstemperatur*: Zur Berechnung der Tagesdurchschnittstemperatur benutzt man vier Messwerte. Sie werden in 2 m Höhe über dem Erdboden gemessen, und zwar um 7 Uhr, 14 Uhr und 21 Uhr. Die Temperatur um 21 Uhr geht mit dem Gewichtungsfaktor 2 ein. Die Tagesmittel werden also berechnet nach der Formel

$$\frac{7^h + 14^h + 2 \cdot 21^h}{4}.$$

(Quelle: Statistisches Jahrbuch 2009 für die Bundesrepublik Deutschland. Wiesbaden 2009, S. 26.)

Hat man *Daten in gruppierter Form* vorliegen (klassierte Daten), so ist das arithmetische Mittel aller Beobachtungen leicht zu berechnen, wenn die arithmetischen Mittel in jeder Klasse bekannt sind oder berechnet werden können. Sind n Beobachtungswerte x_1, x_2, \dots, x_n gegeben und liegen s Klassen k_1, k_2, \dots, k_s vor, und bezeichnet $H_n(i)$ die Anzahl der Merkmale in der i -ten Klasse, so ist:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^s \bar{x}_i \cdot H_n(i) \quad \text{mit} \quad \bar{x}_i = \begin{cases} \frac{1}{H_n(i)} \cdot \sum_{x_i \in k_i} x_i, & \text{falls } H_n(i) \neq 0, \\ \text{sonst } \bar{x}_i = 0. \end{cases}$$

\bar{x}_i ist also das arithmetische Mittel der i -ten Klasse. Dieses ist aber häufig nicht bekannt. Als Näherung für das arithmetische Mittel kann dann der Wert

$$\bar{x} = \frac{1}{n} \sum_{i=1}^s x_i^* \cdot H_n(i)$$

genommen werden. Hierbei sind x_i^* die Klassenmitte und $H_n(i)$ die Klassenhäufigkeit der i -ten Klasse.

Didaktischer Hinweis

Die Wahl der Klassen kann die Größe des arithmetischen Mittels ganz entscheidend beeinflussen wie das folgende Beispiel zeigt.

Beispiel 1.7
(Gehaltsstatistik eines Betriebes (Monatlicher Bruttolohn))

| Gehaltsklassen (in Euro) | Anzahl der Mitarbeiter $H_n(i)$ |
|--------------------------|---------------------------------|
| von 1000 bis unter 1400 | 8 |
| von 1400 bis unter 1600 | 10 |
| von 1600 bis unter 1800 | 10 |
| von 1800 bis unter 2000 | 10 |
| von 2000 bis unter 3000 | 2 |

$$\bar{x} = 1200 \cdot \frac{8}{40} + 1500 \cdot \frac{10}{40} + 1700 \cdot \frac{10}{40} + 1900 \cdot \frac{10}{40} + 2500 \cdot \frac{2}{40} = 1640 \text{ [Euro]}.$$

Fasst man die letzten zwei Klassen zusammen, so erhält man unter Beibehaltung der anderen Klassen für dieselbe Gehaltsstatistik die folgende Tabelle:

| Gehaltsklassen (in Euro) | Anzahl der Mitarbeiter $H_n(i)$ |
|--------------------------|---------------------------------|
| von 1000 bis unter 1400 | 8 |
| von 1400 bis unter 1600 | 10 |
| von 1600 bis unter 1800 | 10 |
| von 1800 bis unter 3000 | 12 |

Bei dieser Klasseneinteilung beträgt das arithmetische Mittel
 $\bar{x} = 1760$ [Euro].

Man erkennt, dass man durch „geschickte“ Wahl der Klassen günstigere Ergebnisse erzielen kann. Das ist eine häufig genutzte Manipulationsmöglichkeit. ■

Geometrisches Mittel

Wir zeigen zunächst, dass das arithmetische Mittel für den im folgenden Beispiel angesprochenen Sachzusammenhang kein angemessener Mittelwert zur Charakterisierung der Daten ist. Aufgrund dieser Erkenntnis stellt sich dann die Frage nach einem anderen Mittelwert als Kennzahl, der die Situation besser beschreibt. Diese Überlegungen führen zum geometrischen Mittel.

Beispiel 1.8

(Bevölkerungsentwicklung) Die folgende Tabelle gibt einen fiktiv angenommenen Wachstumsprozess der Bevölkerung einer Stadt in vier aufeinanderfolgenden Jahren wieder.

| Jahr | Anzahl der Bewohner | Zuwachsraten in % |
|------|---------------------|-------------------|
| 2000 | 100 000 | – |
| 2001 | 150 000 | 50 |
| 2002 | 195 000 | 30 |
| 2003 | 214 500 | 10 |
| 2004 | 257 400 | 20 |

Wie die Tabelle erkennen lässt, beziehen sich die angegebenen prozentualen Zuwachsraten stets auf das vorangegangene Jahr als Basisjahr. Wir fragen, um wieviel Prozent die Bevölkerung im „Durchschnitt“ in jedem der vier Jahre zugenommen hat.

Bildet man als Lösung das arithmetische Mittel der Zuwachsraten, so erhält man $(50 + 30 + 10 + 20) : 4 = 27,5$ [%]. Berechnet man bei Zugrundelegung eines jährlichen Zuwachses von 27,5 % die Anzahl der Bewohner für das Jahr 2004, erhält man (ausgehend von 100 000) sukzessive für die Anzahl der Bewohner 2001: 127500, 2002: 162562, 2003: 207266, 2004: 264264. Durch diese Berechnung erhält man für das Jahr 2004 also 6864 Bewohner mehr als tatsächlich gezählt wurden. Das arithmetische Mittel 27,5 % ist zu groß. Ergebnis: Das arithmetische Mittel ist in diesem Sachzusammenhang (Wachstumsraten) offenbar nicht der angemessene Mittelwert. Denn man möchte ja bei Anwendung des *Mittelwertes*, also bei Anwendung ein und derselben Zahl, auf *alle* Bezugseinheiten dasselbe Gesamtergebnis erhalten (im Beispiel 257400) wie bei der

Anwendung der jeweils konkreten Zuwachsraten auf die einzelnen Einheiten.

Im obigen Lösungsweg wurde nicht beachtet, dass die angegebenen Wachstumsraten verschiedene Bezugspunkte haben:

Unter Berücksichtigung der verschiedenen Bezugspunkte führt das zu der Gleichung

$$1,5 \cdot 1,3 \cdot 1,1 \cdot 1,2 \cdot 100000 = 257400.$$

Der gesamte Wachstumsprozess wird also durch das Produkt der 4 Zahlen

$$1,5 \cdot 1,3 \cdot 1,1 \cdot 1,2 = 2,574$$

adäquat beschrieben. Wir suchen jetzt eine mit Hilfe der Zahlen 1,5; 1,3; 1,1 und 1,2 gebildete Zahl g , die als Ersatz für die vier verschiedenen Zahlen dasselbe Ergebnis 2,574 liefert. Das führt zum Ansatz

$$\begin{aligned} g \cdot g \cdot g \cdot g &= 1,5 \cdot 1,3 \cdot 1,1 \cdot 1,2 \\ g^4 &= 2,574 \\ g &= \sqrt[4]{2,574} = 1,26664. \end{aligned}$$

Aus den vier gegebenen *Wachstumsfaktoren* 1,5; 1,3; 1,1 und 1,2 haben wir einen neuen Wachstumsfaktor 1,26664 für *alle* vier Jahre gefunden.

Der in obiger Rechnung als Ersatz gefundene Wachstumsfaktor $g = 1,26664$ bedeutet also eine durchschnittliche Wachstumsrate (einen durchschnittlichen Zuwachs) von 0,26664 bzw. 26,664 %. Eine Probe hat für Lernende eine große Überzeugungskraft:

$$(((100000 \cdot 1,26664) \cdot 1,26664) \cdot 1,26664) \cdot 1,26664 = 257402,5.$$

Der mit Hilfe der mittleren Zuwachsrate 0,26664 errechnete Endzustand der Anzahl der Bewohner im Jahr 2004 stimmt also mit der in der Tabelle angegebenen Zahl fast überein. Die Zahl 26,664 % als jährliche mittlere prozentuale Zuwachsrate beschreibt also den Sachzusammenhang wesentlich besser als das arithmetische Mittel 27,5 %.

Die Zahl $g = \sqrt[4]{1,5 \cdot 1,3 \cdot 1,1 \cdot 1,2}$ heißt *das geometrische Mittel* der Zahlen 1,5; 1,3; 1,1; 1,2.

Aus den *Wachstumsfaktoren* x_i lassen sich natürlich sofort auch die *Wachstumsraten* r_i berechnen:

$$r_i = x_i - 1.$$

Konkret für das Beispiel erhalten wir: $r_1 = 1,5 - 1 = 0,5 = 50\%$; $r_2 = 1,3 - 1 = 0,3 = 30\%$ usw. ■

Wir fassen die dem Beispiel inliegende Struktur allgemein zusammen: Gegeben sind zeitliche Beobachtungswerte (Wachstumsraten): Gegeben ist eine Größe A , die in den Zeitpunkten $t_0, t_1, t_2, \dots, t_n$ mit $t_0 < t_1 < t_2 < \dots < t_n$ die Werte $A_0, A_1, A_2, \dots, A_n$ annimmt. Ferner gilt $A_i = x_i \cdot A_{i-1}$ mit einem Wachstumsfaktor x_i für $i = 1, 2, \dots, n$. Für A_n erhält man dann

$$A_n = (x_1 \cdot x_2 \cdot \dots \cdot x_n) A_0.$$

Der Gesamtwachstumsfaktor für den letzten Wert A_n bezogen auf A_0 ist also $x_1 \cdot x_2 \cdot \dots \cdot x_n$. Ein aus x_1, x_2, \dots, x_n gebildetes Mittel dient als Ersatz für die x_i . Man setzt:

$$A_n = g^n \cdot A_0 \quad \text{mit} \quad g^n = x_1 \cdot x_2 \cdot \dots \cdot x_n.$$

Die Zahl $g = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$ heißt das geometrische Mittel der Zahlen x_1, x_2, \dots, x_n .

Definition 1.4 (Geometrisches Mittel)

Es seien x_1, x_2, \dots, x_n n Daten eines quantitativen Merkmals mit $x_i > 0$ für $i = 1, 2, \dots, n$. Dann heißt die Zahl

$$\bar{x}_g := \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

das **geometrische Mittel** dieser Daten.

Analog zum gewogenen arithmetischen Mittel lässt sich auch hier das **gewogene geometrische Mittel** definieren:

$$\bar{x}_g := \sqrt[G]{x_1^{g_1} \cdot x_2^{g_2} \cdot \dots \cdot x_n^{g_n}} \quad \text{mit} \quad G = \sum_{i=1}^n g_i.$$



Harmonisches Mittel

Das harmonische Mittel ist wie das arithmetische und geometrische Mittel ein *errechneter* Wert und für quantitative Merkmale definiert. Es ist ein selten gebrauchter Lageparameter und ergibt sich – wie wir in den folgenden zwei Beispielen zeigen – auch aus dem Lösungsweg zur Bestimmung eines angemessenen Mittelwerts bei bestimmten Sachproblemen, ohne dass Kenntnisse über das harmonische Mittel vorausgesetzt werden müssen. Das lässt sich bei Kenntnis der Definition auch errahnen.

Definition 1.5 (Harmonisches Mittel)

Es seien n Daten x_1, x_2, \dots, x_n eines quantitativen Merkmals mit $x_i > 0$ für $i = 1, 2, \dots, n$ gegeben. Dann heißt die Zahl

$$\bar{x}_h = \frac{1}{\frac{1}{n}(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n})} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad (1.3)$$

das **harmonische Mittel** der Daten x_1, x_2, \dots, x_n . ◆

Hinweis: Wie beim arithmetischen und geometrischen Mittel lässt sich auch analog das gewogene harmonische Mittel definieren.

Didaktischer Hinweis

Die Berechnung des harmonischen Mittels erfolgt, indem man den Stichprobenumfang n durch die Summe aller Kehrwerte $\frac{1}{x_i}$ dividiert. Man kann also vermuten, dass man den Durchschnittswert der Daten eines konkreten Sachproblems, für das das harmonische Mittel ein adäquater Durchschnittswert ist, auch ohne Kenntnis der Definition bestimmen kann.

Wir betrachten dazu das folgende Beispiel „**Durchschnittsgeschwindigkeit**“:

Ein Zug fährt die ersten 100 km mit einer konstanten Geschwindigkeit von 70 km/h, die zweiten 100 km mit einer konstanten Geschwindigkeit von 110 km/h. Wie groß ist seine Durchschnittsgeschwindigkeit?

Zur *Lösung* berechnen wir zunächst die Gesamtfahrzeit des Zuges für 200 km. Die ersten 100 km legt der Zug in $\frac{100}{70}h = \frac{10}{7}h$ zurück, die zweiten 100 km in $\frac{100}{110}h$. Die Gesamtfahrzeit beträgt also: $\frac{10}{7}h + \frac{10}{11}h = \frac{180}{77}h = 2,34h$. Für die gesuchte Durchschnittsgeschwindigkeit erhält man dann:

$$200 : \frac{180}{77} \text{ km/h} \approx 85,56 \text{ km/h}.$$

Die alleinige Anwendung der Definition 1.5 liefert aber kaum einen Beitrag zur Einsicht, dass der „richtige“ Mittelwert für das Sachproblem bestimmt wurde.

Das Beispiel Durchschnittsgeschwindigkeit spricht ein typisches Problem an, bei dem zur Lösung das harmonische Mittel der angemessene Lageparameter ist. Es handelt sich um eine Mittelung von Geschwindigkeiten auf *gleichlangen* Wegstrecken.

Eine andere Situation liegt bei folgender Aufgabenstellung vor: Ein Zug fährt eine Stunde mit konstanter Geschwindigkeit von 70 km/h und eine zweite Stunde mit konstanter Geschwindigkeit von 110 km/h. Wie groß ist seine Durchschnittsgeschwindigkeit? Jetzt ist das arithmetische Mittel der angemessene Mittelwert: $\frac{70+110}{2} = 90$ [km/h].

Ein weiteres typisches Problem für die Verwendung des harmonischen Mittels ist die Berechnung des Durchschnittspreises bei vorgegebenem *gleichen* Kapitalaufwand.

Größenvergleich dieser drei Mittelwerte

Zwischen arithmetischem, geometrischem und harmonischem Mittel besteht eine interessante Größenrelation. Es gilt:

Satz 1.1

Seien x_1, x_2, \dots, x_n metrische Daten mit $x_i > 0$ für alle $i = 1, 2, \dots, n$, dann gilt stets

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x}.$$

Das Gleichheitszeichen gilt nur dann, wenn $x_1 = x_2 = \dots = x_n$ ist.

Didaktische Hinweise

1. In Aufgabe 7 des Abschnitts 1.2.8 ist Satz 1.1 für den Fall $n = 2$ zu beweisen.
2. Für den allgemeinen Fall gibt es eine Reihe von unterschiedlichen Beweisen, die aber alle nicht ganz trivial sind. Ausgehend von der Aussage in Aufgabe 7 kann man einen Beweis durch vollständige Induktion führen. Zweckmäßig beweist man zunächst die rechte Ungleichung $\bar{x}_g \leq \bar{x}$. Im Werk von Mangoldt/Knopp [120], S. 128ff ist ein besonders kurzer Beweis für Satz 1.1 angegeben. Wir weisen ferner hin auf Ostrowski [126], S. 35ff und auf Dallmann/Elster [36], S. 33.
3. Hat man $\bar{x}_g \leq \bar{x}$ bewiesen, folgt leicht $\bar{x}_h \leq \bar{x}_g$. Zunächst gilt:

$$\sqrt[n]{\frac{1}{x_1} \cdot \frac{1}{x_2} \cdot \dots \cdot \frac{1}{x_n}} = \frac{1}{\sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}},$$

d. h. das geometrische Mittel der Zahlen $\frac{1}{x_1}, \frac{1}{x_2}, \dots, \frac{1}{x_n}$ ist gleich dem reziproken Wert des geometrischen Mittels der Werte x_1, x_2, \dots, x_n . Wir betrachten jetzt die Zahlen $y_i = \frac{1}{x_i}$ ($1 \leq i \leq n$) und wenden jetzt das bewiesene Ergebnis $\bar{y}_g \leq \bar{y}$ an:

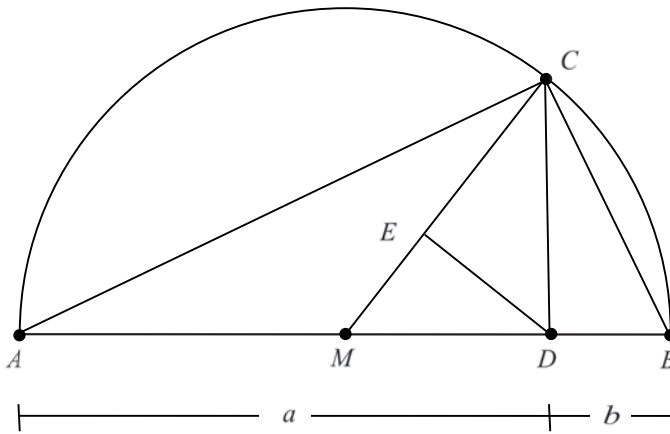
$$\begin{aligned} \bar{y}_g &\leq \bar{y} \\ \sqrt[n]{y_1 \cdot y_2 \cdot \dots \cdot y_n} &\leq \frac{1}{n}(y_1 + y_2 + \dots + y_n) \\ \Leftrightarrow \sqrt[n]{\frac{1}{x_1} \cdot \frac{1}{x_2} \cdot \dots \cdot \frac{1}{x_n}} &\leq \frac{1}{n} \left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right) \\ \Leftrightarrow \frac{1}{\sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}} &\leq \frac{1}{n} \left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right) \\ \Leftrightarrow \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} &\geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} \\ \text{d. h.} \quad \bar{x}_g &\geq \bar{x}_h. \end{aligned}$$

4. Für die Sekundarstufe I ist ein geometrischer Beweis für eine Teilaussage des obigen Satzes 1.1 interessant: Seien a und b zwei quantitative Daten mit $a > 0$ und $b > 0$ und $a \neq b$. Dann gilt

$$\frac{2 \cdot a \cdot b}{a + b} < \sqrt{a \cdot b} < \frac{a + b}{2},$$

d. h. das geometrische Mittel *zweier* positiver, ungleicher Zahlen ist kleiner als das arithmetische, aber größer als das harmonische Mittel dieser Zahlen.

Der *Beweis* kann auf geometrischem Wege geführt werden. Man betrachte folgende Zeichnung:



Auf dem Halbkreis über der Strecke \overline{AB} wird ein Punkt C gewählt; man verbindet den Mittelpunkt M von \overline{AB} mit C . Von C aus wird das Lot auf die Strecke \overline{AB} gefällt, der Lotfußpunkt sei D . Von D aus wird das Lot auf die Strecke \overline{MC} gefällt, der Lotfußpunkt sei E . Sei $a = \overline{AD}$, $b = \overline{DB}$. Dann gilt (man wendet u. a. Sätze über rechtwinklige Dreiecke an):

$$\overline{MC} = \frac{1}{2}(a + b),$$

$$\overline{CD} = \sqrt{a \cdot b},$$

$$\overline{CE} = \frac{2 \cdot a \cdot b}{a + b} = \frac{2}{\frac{1}{a} + \frac{1}{b}}.$$

Man erkennt nun an der Zeichnung

$$\frac{2 \cdot a \cdot b}{a + b} < \sqrt{a \cdot b} < \frac{a + b}{2}.$$

Die folgenden zwei Mittelwerte könnte man im Vergleich zu dem arithmetischen, geometrischen und harmonischen Mittel, die wir als *errechnete Mittelwerte* bezeichneten, als *Mittelwerte der Lage* bezeichnen: **Median** (allgemeiner Quantile) und **Modalwert**.

Median

Der **Median** (Zentralwert, englisch: *median*) ist dadurch bestimmt, dass er in der „Mitte“ der Reihe einer der Größe nach geordneten Datenmenge liegt. Mindestens 50 % der Daten sind kleiner oder gleich und mindestens 50 % der Daten sind größer oder gleich der Daten (50%-Punkt der Daten). Zur Bestimmung des Medians werden keine quantitativen Merkmale benötigt, es genügen Rangmerkmale.

Es ist üblich, Daten x_1, x_2, \dots, x_n , die der Größe nach geordnet sind, durch runde Klammern in den Indizes zu kennzeichnen. Das wird in der folgenden Definition benutzt.

Definition 1.6 (Median)

Seien $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}$ der Größe nach geordnete n Daten. Als **Median**, den man mit $x_{0,5}$ bezeichnet, wählt man

1. bei *Daten von Rangmerkmalen* die Zahl

$$x_{0,5} := \begin{cases} x_{(\frac{n+1}{2})} & \text{bei ungeradem } n \\ x_{(\frac{n}{2})} \text{ oder } x_{(\frac{n}{2}+1)} & \text{bei geradem } n \end{cases}$$

2. bei *quantitativen nicht gruppierten Daten* die Zahl

$$x_{0,5} := \begin{cases} x_{(\frac{n+1}{2})} & \text{bei ungeradem } n \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{bei geradem } n \end{cases}$$



Anmerkung zu 1. Für eine gerade Anzahl n von Daten hat sich bei Rangmerkmalen keine einheitliche Festlegung des Medians durchgesetzt. Gelegentlich wählt man auch wie bei quantitativen Merkmalen das arithmetische Mittel aus

$$x_{(\frac{n}{2})} \quad \text{und} \quad x_{(\frac{n}{2}+1)}.$$

Bei *gruppierten* Daten kann man nur die Klasse angeben, in der der Median liegt, denn man kennt ja in der Regel nicht die einzelnen Daten in der Klasse. Bei geometrischer Interpretation kann man mit Bezug auf die empirische Verteilungsfunktion sagen, dass der Median in der Klasse liegt, in der die empirische Verteilungsfunktion den Wert 0,5 erreicht.

Bei quantitativ gruppierten Daten bestimmt man häufig den Median approximativ. Ist $[x_{r-1}, x_r[$ die Medianklasse, so berechnet man den Median durch

$$x_{0,5} = x_{r-1} + \frac{0,5 - \sum_{i=1}^{r-1} h_n(k_i)}{h_n(k_r)} \cdot \Delta k_r.$$

Hierbei bedeuten

- $\sum_{i=1}^{r-1} h_n(k_i)$ die aufaddierten (kumulierten) relativen Häufigkeiten aller Klassen, die kleiner als die Klasse sind, in der der Median liegt,
- $h_n(k_r)$ die relative Häufigkeit der Klasse k_r , in der der Median liegt,
- Δk_r die Breite der Klasse k_r .

Die Bestimmung des Medians ist also recht einfach, wenn man von der Approximation bei gruppierten Daten absieht. Der Median kann durch Abzählen oder durch einfache Rechnung (arithmetisches Mittel zweier Werte) bestimmt werden. Wird der Median als arithmetisches Mittel zweier benachbarter Daten, die voneinander verschieden sind, berechnet, so entspricht dem Median natürlich kein konkreter Datenwert.

Im Beispiel 1.7 (Gehaltsstatistik eines Betriebes) liegt der Median $x_{0,5}$ in der Klasse „von 1600 bis unter 1800 Euro“. Rechnerische Bestimmung:

$$\begin{aligned} x_{0,5} &= 1600 + \frac{0,5 - (\frac{8}{40} + \frac{10}{40})}{\frac{10}{40}} \cdot 200 \\ x_{0,5} &= 1600 + 40 = 1640 \text{ [Euro]}. \end{aligned}$$

Der Median stimmt in diesem Beispiel mit dem arithmetischen Mittel überein.

Für quantitative Merkmale besitzt der Median eine wichtige Eigenschaft, die sogenannte **Minimumseigenschaft des Medians**.

Satz 1.2 (Minimumseigenschaft des Medians)

Seien x_1, x_2, \dots, x_n quantitative Daten. Die Summe der absoluten Abweichungen aller Daten x_i von ihrem Median ist kleiner oder gleich der Summe aller absoluten Abweichungen der Daten x_i von irgendeinem anderen Wert c , ist also ein Minimum. Es gilt:

$$\sum_{i=1}^n |x_i - x_{0,5}| \leq \sum_{i=1}^n |x_i - c| \quad \text{für beliebiges } c \in \mathbb{R}.$$

Der arithmetische Nachweis dieser Eigenschaft erfordert einigen Rechenaufwand. Man macht zweckmäßigerweise eine Fallunterscheidung und betrachtet die Fälle, dass die Anzahl der Daten gerade bzw. ungerade ist. (Siehe Lösung von Aufgabe 9 des Abschnitts 1.2.8.) Einen schönen graphischen Nachweis findet man in Bentz [15] und in Bentz/Borovcnik [16]. Dieser Beweis ist auch in der Sekundarstufe I möglich.

Diese Eigenschaft des Medians ist der Hintergrund für eine „klassische“ Anwendung des Medians, die auch in Schulbüchern zu finden ist. Es handelt sich darum, ein „Standortproblem“ zu lösen.

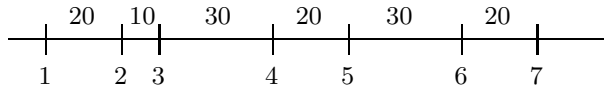
Beispiel 1.9

(Standortproblem) Ein Unternehmen muss entlang einer Straße sieben Geschäfte wöchentlich einmal beliefern. Wo ist an dieser Straße der Standort des Unternehmens mit Lager einzurichten damit die Gesamtstrecke zu allen Geschäften möglichst kurz ist?

Lösung: Bezeichnet man mit x_i ($i = 1, 2, \dots, 7$) die Lage der sieben Geschäfte, so ist eine Zahl a gesucht, so dass $\sum_{i=1}^7 |x_i - a|$ minimal ist. Nach obigem Satz besitzt der Median diese lineare Minimumseigenschaft.

Für konkrete Situationen und für eine spezielle Fragestellung lässt sich das Standortproblem im Unterricht der Sekundarstufe I elementar behandeln.

Die Lage der sieben Geschäfte 1, 2, 3, 4, 5, 6, 7 an der Straße sei so wie in nachfolgender Skizze angegeben. Zwischen den Positionen der Geschäfte sind die Entfernungen benachbarter Geschäfte in km angegeben. Wir fragen jetzt speziell, bei *welchem* Geschäft das Lager einzurichten ist, damit die Gesamtstrecke zur Belieferung aller Geschäfte minimal ist.



Nach der Minimumseigenschaft des Medians ist das Lager bei Geschäft Nr. 4 einzurichten. Schüler können das Ergebnis bei unserer speziellen Fragestellung konkret überprüfen, indem sie eine Entfernungstabelle für die Geschäfte aufstellen:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Summe |
|-------|-----|-----|-----|-----|-----|-----|-----|-------|
| 1 | – | 20 | 30 | 60 | 80 | 110 | 130 | 430 |
| 2 | 20 | – | 10 | 40 | 60 | 90 | 110 | 330 |
| 3 | 30 | 10 | – | 30 | 50 | 80 | 100 | 300 |
| 4 | 60 | 40 | 30 | – | 20 | 50 | 70 | 270 |
| 5 | 80 | 60 | 50 | 20 | – | 30 | 50 | 290 |
| 6 | 110 | 90 | 80 | 50 | 30 | – | 20 | 380 |
| 7 | 130 | 110 | 100 | 70 | 50 | 20 | – | 480 |
| Summe | 430 | 330 | 300 | 270 | 290 | 380 | 480 | |

Aus der Tabelle liest man ab, dass für den Standort des Lagers bei Geschäft Nr. 4 die Summe der Entfernungskilometer kleiner ist als bei den anderen Geschäften. ■

Modalwert

Geht es bei Untersuchungen um Krankheiten bzw. Warenfehler, so kann ein Interesse daran bestehen, die häufigste Krankheit bzw. den häufigsten Fehler einer Ware zu kennen. Der hierfür geeignete Lageparameter ist der Modalwert (im Französischen: *valeur normale*, im Englischen: *mode*):

Der Modalwert x_{Mod} ist die Merkmalsausprägung, die am häufigsten vorkommt.

Der Modalwert heißt auch Modus oder dichtester Wert.

Der Modalwert ist sehr einfach zu bestimmen und sehr wirklichkeitsnah. Der Modalwert braucht jedoch nicht eindeutig zu sein. Bei mehrgipfligen Verteilungen können zwei oder mehrere lokale Häufigkeitsstellen als lokale Modalwerte vorhanden sein.

Bei gruppierten Daten nimmt man als Modalwert den Repräsentanten (die Klassenmitte) der Klasse mit der größten Häufigkeit.

p-Quantil

Definition 1.7 (p-Quantil)

Sei $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}$ eine geordnete Messreihe. Dann heißt die Zahl x_p , für die gilt: mindestens $p \cdot 100\%$ der Daten liegen vor x_p und mindestens $(1 - p) \cdot 100\%$ der Daten liegen nach der Zahl x_p das **p-Quantil**.

Das p-Quantil wird berechnet durch:

$$\begin{aligned} x_p &:= x_{([np]+1)}, & \text{falls } np \text{ nicht ganzzahlig ist,} \\ x_p &:= \frac{1}{2}(x_{(np)} + x_{(np+1)}), & \text{falls } np \text{ ganzzahlig ist.} \end{aligned}$$

Hinweise:

1. Unter dem Symbol $[np]$ in $x_{([np]+1)}$ versteht man die größte ganze Zahl, die kleiner oder gleich np ist.
2. Für $p = 0,5$ erhält man den Median. In der Praxis treten p-Quantile häufig auf. Es sind die folgenden Bezeichnungen üblich (Auswahl):

$x_{0,25}$ heißt erstes Quartil (auch unteres Quartil),

$x_{0,5}$ heißt zweites Quartil (Median),

$x_{0,75}$ heißt drittes Quartil (auch oberes Quartil),

$x_{0,1}$ heißt erstes Dezil,

$x_{0,9}$ heißt neuntes Dezil.

Das untere Quartil $x_{0,25}$, der Median $x_{0,5}$ und das obere Quartil $x_{0,75}$ spielen im box-plot-Diagramm (graphische Darstellung einer Datenmenge) eine große Rolle.



Abschließende Bemerkungen

1. Die verschiedenen Mittelwerte besitzen unterschiedliche sachlogische Bedeutungen. Zunächst ist die Wahl unter **Berücksichtigung** der vorliegenden **Merkmalsart** zu treffen:

- Der *Modalwert* ist der einzige Mittelwert, der bei allen Typen von Merkmalen anwendbar ist. Bei qualitativen Merkmalen ist er auch der einzige.
- Der *Median* und die Quantile sind Kennziffern für Rangmerkmale und quantitative Merkmale.
- *Arithmetisches Mittel*, *geometrisches Mittel* und *harmonisches Mittel* sind bei quantitativen Merkmalen anwendbare Mittelwerte. Sie sollten nicht für Rangmerkmale benutzt werden.

Gibt es bei einer Merkmalsart mehrere Möglichkeiten, so ist für die „richtige“ Entscheidung dann das konkrete Sachproblem heranzuziehen. Wir nennen einige **typische Anwendungen für die verschiedenen Mittelwerte**:

- *Modalwert*: Größtes Verkehrsaufkommen an einem Verkehrsknotenpunkt, größte Besucherzahl einer Einrichtung, häufigster Fehler einer Ware, häufigste Todesursache in einem bestimmten Alter, häufigste Krankheit in einem Land.
- *Median*: Der Median kann als mittlerer Wert von Bedeutung sein bei Einkommensvergleichen, z. B. oberhalb und unterhalb liegen gleich viele Einkommensempfänger. Besonders seit 1995 („Jahr der Armut“) steht die „wachsende Armut“ im Brennpunkt öffentlichen Interesses. Gemeint ist stets die relative (nicht die absolute) Armut. Galt früher als arm, wer weniger als die Hälfte des durchschnittlichen Einkommens der Vollzeitbeschäftigten erhielt (arithmetisches Mittel), so wird heute der Median als Maßstab benutzt. In der EU gilt z. Zt. als arm, wer weniger als 60 % des Medians des Einkommens aller Vollzeitbeschäftigten eines Landes zur Verfügung hat.
- *Geometrisches Mittel*: Das geometrische Mittel wird angewandt, um die durchschnittliche relative Veränderung zu bestimmen, z. B.: durchschnittliche Wachstumsrate des Bruttosozialproduktes oder einer Bevölkerungsentwicklung oder prozentualer Lohnerhöhungen. Bei solchen relativen Änderungen ist es nicht sinnvoll, das arithmetische Mittel zu berechnen. Man beachte, dass die Daten bei Anwendung des geometrischen Mittels nicht Null oder negativ sein dürfen.
- *Harmonisches Mittel*: Das harmonische Mittel dient z. B. zur Bestimmung der durchschnittlichen Geschwindigkeit bei Angaben der Geschwindigkeit für gleichlange Teilstrecken und zur Ermittlung des Durchschnittspreises einer Ware mit verschiedenen Preisen aber mit gleichem Kostenaufwand.

- *Arithmetisches Mittel*: Das arithmetische Mittel wird in der Praxis wohl am häufigsten benutzt. Warum besitzt das arithmetische Mittel eine solche Vorrangstelle?
 - * Es ist leicht zu berechnen, und die Reihenfolge der Daten spielt keine Rolle. Die Daten müssen also nicht der Größe nach geordnet werden.
 - * Wenn man an die Berechnung des arithmetischen Mittels denkt, so erkennt man, dass man aus dem Mittelwert und der Anzahl der Daten die Summe der Daten berechnen kann ($n \cdot \bar{x} = \sum_{i=1}^n x_i$) oder aus der Summe der Daten und dem Mittelwert die Anzahl der Daten. Hier liegen Vorteile gegenüber dem Median und Modalwert.
 - * Das arithmetische Mittel ist *der* Mittelwert, der später zur weiteren Charakterisierung der Datenmenge durch Streuungsmaße eine wichtige Rolle spielt.
2. Ein weiterer Gesichtspunkt soll noch angesprochen werden: Das **Problem der Ausreißer**. Es handelt sich bei Ausreißern um Daten, die (extrem) weit weg isoliert von der Mehrzahl der Daten liegen. Beispiel: Wenn die monatlichen Einkommen (in Euro) von 9 Personen 1600, 1700, 1500, 2000, 2100, 1800, 1900, 1650, 7000 betragen, so können 7000 Euro als Ausreißer angesehen werden. Soll man solche Ausreißer überhaupt berücksichtigen? Wenn man begründet annehmen kann, dass ein Erhebungsfehler oder Schreibfehler vorliegt, wird man Ausreißer gegebenenfalls unberücksichtigt lassen. Dieses muss aber bei der Auswertung der Daten in jedem Fall angegeben werden. Wie wirken sich Ausreißer auf die Mittelwerte aus? *Modalwert* und *Median* reagieren auf Ausreißer überhaupt nicht. Man sagt, sie sind *unempfindlich* gegenüber Ausreißern. Das kann natürlich als Nachteil angesehen werden.
- Arithmetisches Mittel*, *geometrisches Mittel* und *harmonisches Mittel* werden aufgrund des Rechenvorgangs von jedem Einzelwert beeinflusst, also auch von Ausreißern. Das arithmetische Mittel reagiert stärker auf Ausreißer als das geometrische Mittel. Diese *Empfindlichkeit* hat jedoch auch einen Vorteil: Ein „ungewöhnlicher“ Mittelwert gibt Veranlassung, kritisch auf die Daten selbst zu schauen.
3. Wir beschließen diesen Abschnitt mit einem Hinweis auf ein interessantes Beispiel (Kundeneinzugsbereich) bei Bahrenberg/Giese [4], S. 14ff. Siehe auch Kütting [102], S. 101.

1.2.4 Streuungsparameter

Es gibt keine allgemeinen Richtlinien für die Verwendung von Mittelwerten. Oberstes Gebot sollte immer sein: Der gewählte Mittelwert sollte repräsentativ für die Datenmenge sein. Das kann er allein nicht leisten. Man benötigt noch eine Beschreibung der Streuung der Daten um den angegebenen Mittelwert. Ein

„klassisches“ Beispiel kann das Problem bewusst machen: Der Vergleich von Jahresdurchschnittstemperaturen von Quito und Peking. In Quito (in Ecuador am Äquator gelegen) herrscht „ewiger Frühling“ mit einer Temperatur stets um etwa $13\text{ }^{\circ}\text{C}$ durch das ganze Jahr, wohingegen in Peking die Temperaturen in der Jahreszeit schwanken zwischen fast $30\text{ }^{\circ}\text{C}$ und $-6\text{ }^{\circ}\text{C}$. Aber auch hier beträgt die Jahresdurchschnittstemperatur etwa $13\text{ }^{\circ}\text{C}$.

Anmerkung: Das Äquatordenkmal in der Umgebung von Quito verfehlt um etwa 8 km den Äquator.

Ganz allgemein bedeutet Streuung in einer Datenmenge die Abweichung der Messwerte voneinander, oder auch spezieller die Abweichung der Messwerte einer Datenmenge von einem Mittelwert der Datenmenge als Bezugspunkt. Beide Gesichtspunkte führen zu spezifischen Streuungsmaßen. Der erste Gesichtspunkt (keine Berücksichtigung von Mittelwerten als Bezugswerte) führt zu Begriffen wie **Spannweite** und **Quartilabstand**. Der zweite Gesichtspunkt findet in der **mittleren absoluten Abweichung**, der **empirischen Varianz** und der **empirischen Standardabweichung** seine Berücksichtigung. Es sind mindestens Rangmerkmale vorausgesetzt, in der Regel quantitative Merkmale. Nominalskalierte Merkmale entziehen sich hier der Aufbereitung.

Spannweite

Die Spannweite SW (englisch: *range*) ist das einfachste und wohl auch anschaulichste Streuungsmaß für Daten. Es berücksichtigt noch nicht Mittelwerte als Bezugspunkte für die Berechnung der Streuung.

Definition 1.8 (Spannweite)

Die Differenz $SW := x_{(\max)} - x_{(\min)}$ zwischen dem größten $x_{(\max)}$ und dem kleinsten $x_{(\min)}$ Merkmalswert einer geordneten Datenmenge heißt **Spannweite** SW . Die Spannweite wird auch **Variationsbreite** genannt. ♦

Im Beispiel „Kundeneinzugsbereich“ betrug die Spannweite $67,6\text{ km} - 0,1\text{ km} = 67,5\text{ km}$.

Der Begriff der Spannweite ist leicht verständlich, und die Spannweite ist ohne großen Rechenaufwand bestimmbar. Diesen Vorteilen stehen aber auch Nachteile gegenüber:

- Die Aussagekraft der Spannweite ist gering, denn die Spannweite wird nur durch den größten und kleinsten Wert bestimmt, wird also stark durch Extremwerte (Ausreißer) beeinflusst
- Die Spannweite gibt keine Auskunft darüber, wie sich die Daten innerhalb des Intervalls $[x_{(\min)}, x_{(\max)}]$ verteilen.
- Die Spannweite ändert sich in der Messreihe nur, wenn ein Wert auftritt, der kleiner als der bisher kleinste oder größer als der bisher größte Wert ist.

Quartilabstand

Während durch die Spannweite ein Bereich festgelegt ist, in dem 100 % der Merkmalswerte liegen, wird durch den Quartilabstand ein Bereich definiert, in dem 50 % aller Messwerte liegen, und in dem auch der Median $x_{0,5}$ liegt.

Definition 1.9 (Quartilabstand)

Es seien $x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}$ geordnete Daten. Dann heißt die Differenz $QA := x_{0,75} - x_{0,25}$ zwischen dem oberen (dritten) Quartil $x_{0,75}$ und dem unteren (ersten) Quartil $x_{0,25}$ der Daten der **Quartilabstand** QA . ♦

Der Quartilabstand ist also ähnlich einfach zu bestimmen wie die Spannweite. Der Median $x_{0,5}$ liegt zwar immer gemäß Definition in dem Bereich, der durch den Quartilabstand festgelegt ist, bei asymmetrischen Verteilungen liegt der Median aber nicht in der Mitte des Quartilsintervalls $[x_{0,25}, x_{0,75}]$.

Durch den Quartilabstand werden die Daten praktisch in drei Bereiche eingeteilt:

1. 25 % der Werte, die kleiner als das untere Quartil sind;
2. 50 % der Werte, die im Quartilintervall $[x_{0,25}, x_{0,75}]$ liegen;
3. 25 % der Werte, die größer als das obere Quartil sind.

Ergänzungen zu graphischen Darstellungen

1. Fünf-Zahlen-Zusammenfassung

Eine gegebene Datenmenge wird gelegentlich durch die fünf Kennzahlen $x_{0,5}, x_{0,25}, x_{0,75}, x_{(\min)}$ und $x_{(\max)}$ beschrieben. Man spricht von einer **Fünf-Zahlen-Zusammenfassung**. Man ordnet die fünf Zahlen im Schema folgendermaßen an:

| | |
|--------------|--------------|
| $x_{0,5}$ | |
| $x_{0,25}$ | $x_{0,75}$ |
| $x_{(\min)}$ | $x_{(\max)}$ |

2. Box-plot-Diagramm

Das **box-plot-Diagramm** (Kastenschaubild), das die Fünf-Zahlen-Zusammenfassung aufgreift und den Quartilabstand benutzt, gewinnt in wissenschaftlichen Publikationen immer mehr an Bedeutung.

Wir beschreiben diese Darstellung an einem Beispiel.

Beispiel 1.10

(**Körpergröße**) Aus den im folgenden Stengel-Blatt-Diagramm fiktiven Daten über die Körpergröße (in cm) von 62 Personen berechnen wir zunächst einige wichtige Werte, die wir für die Konstruktion des box-plot-Diagramms benötigen.

Sie reichen bis zum kleinsten bzw. größten beobachteten Wert *innerhalb* eines Quartilsabstandes QA , jeweils gemessen von den Enden des Kastens aus (im Beispiel: links bis 83, rechts bis 143). Die Fühlerenden sind die Grenzen eines sogenannten inneren Zaunes, nach links bis maximal $x_{0,25} - 1QA$, nach rechts bis maximal $x_{0,75} + 1QA$.

Außerhalb der Fühlerenden jeweils bis $1,5QA$ (gemessen jeweils von den Kastenenden) liegende Werte werden als Kreise \circ eingezeichnet (im Beispiel bedeuten die zwei untereinander gesetzten Kreise \circ_\circ , dass der Wert 144 zweimal auftritt). Weiter als $1,5QA$ vom Kasten entfernt liegende Werte werden als fette \bullet Punkte eingetragen (im Beispiel 163). Diese Werte liegen unter $x_{0,25} - 1,5QA$ bzw. über $x_{0,75} + 1,5QA$. Man könnte sie als Ausreißer bezeichnen.

■

Didaktische Hinweise:

1. Innerhalb des Kastens, also zwischen dem unteren Quartil $x_{0,25}$ und dem oberen Quartil $x_{0,75}$ liegen 50 % der Daten.
2. Die Festlegung der maximalen Länge der Fühler ist nicht einheitlich. Statt $1QA$ wählt man häufig auch $1,5QA$, dann werden Werte, die mehr als anderthalb Kastenlängen außerhalb liegen, mit \circ bezeichnet. Werte, die um mehr als drei Kastenlängen außerhalb liegen (Extremwerte), werden durch einen fetten Punkt gekennzeichnet.

In keinem Fall sollten aber die Fühlerenden jeweils bis $x_{(\min)}$ bzw. $x_{(\max)}$ reichen. Denn dadurch geht viel an Informationen über die Daten verloren (Ausreißer, Streuungen). Denn die Visualisierung der Daten durch das box-plot-Diagramm lässt gut Ausreißer, Symmetrien und auch Streuungen erkennen.

3. Box-plot-Diagramme geben einen sehr guten Überblick über die Verteilung der Daten und ermöglichen in der empirischen Forschung einen zuverlässigen Vergleich zwischen verschiedenen Datenmengen.

Mittlere absolute Abweichung

Bei diesem Streuungsmaß handelt es sich um ein Maß für die Abweichungen der Daten von einem Mittelwert als Bezugswert. Bezugswert ist meistens das arithmetische Mittel.

Didaktische Vorbemerkung

Wenn in Schulversuchen ein Streuungsmaß für die Abweichungen der Daten vom arithmetischen Mittel gefunden werden sollte, schlugen die Schüler wiederholt vor, die (algebraische) Summe aller Abweichungen vom Mittelwert zu bilden und dann diese Summe durch die Anzahl der Daten zu dividieren. Denn man will ja einen Mittelwert für die Abweichungen bestimmen. Zur Überraschung der Schüler ergab sich bei verschiedenen Datenmengen stets Null als Ergebnis.

Die inhaltliche Bedeutung des arithmetischen Mittels musste den Schülern erst wieder präsent werden. Es gilt ja stets $\sum_{i=1}^n (x_i - \bar{x}) = 0$. Man schlug vor, die Abweichungen der Daten vom arithmetischen Mittel absolut zu wählen. Das führte dann zur Definition der **mittleren (linearen) absoluten Abweichung** vom arithmetischen Mittel.

Definition 1.10 (Mittlere (lineare) absolute Abweichung)

Seien $x_1, x_2, x_3, \dots, x_n$ Merkmalsausprägungen eines quantitativen Merkmals. Sei \bar{x} das arithmetische Mittel dieser Daten. Dann heißt

$$d_{\bar{x}} := \frac{1}{n} \cdot \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{n} (|x_1 - \bar{x}| + \dots + |x_n - \bar{x}|)$$

die **mittlere (lineare) absolute Abweichung** vom arithmetischen Mittel \bar{x} . ♦

Hinweis:

Analog kann man auch die mittlere absolute Abweichung vom Median einführen:

$$d_{x_{0,5}} := \frac{1}{n} \sum_{i=1}^n |x_i - x_{0,5}| = \frac{1}{n} (|x_1 - x_{0,5}| + \dots + |x_n - x_{0,5}|).$$

Empirische Varianz, empirische Standardabweichung

Auch bei der Berechnung dieses Streuungsmaßes ist das arithmetische Mittel der Daten die Bezugsgröße. Während bei der Bildung der mittleren absoluten Abweichung die positiven und negativen Abweichungen durch die Betragsbildung zu absoluten Abweichungen wurden (sie konnten sich so nicht mehr insgesamt wechselseitig aufheben), erreicht man dieses bei der Bildung der empirischen Varianz durch Quadratbildung der jeweiligen Differenz. Die Summe dieser Quadrate teilt man zur Mittelwertbildung aber nicht durch die Anzahl n der Daten (Summanden), sondern durch $n - 1$ (vgl. hierzu die späteren Anmerkungen).

Definition 1.11 (Empirische Varianz)

Bezeichnen $x_1, x_2, x_3, \dots, x_n$ die Merkmalsausprägungen eines quantitativen Merkmals, und bezeichnet \bar{x} das arithmetische Mittel dieser Daten, so bezeichnet man als **empirische Varianz** s^2 die Zahl

$$s^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad n \geq 2.$$

♦

Es handelt sich bei der empirischen Varianz um ein *quadratisches* Abstandsmaß. Man kann in Annäherung sagen, dass die Varianz das arithmetische Mittel der Abweichungsquadrate ist. Wie bei der mittleren absoluten Abweichung vom arithmetischen Mittel werden auch hier bei der empirischen Varianz die Abweichungen aller Daten vom arithmetischen Mittel berücksichtigt. Durch das Quadrieren werden größere Abweichungen vom arithmetischen Mittel in starkem Maße berücksichtigt.

Die empirische Varianz hat als Streuungsparameter wegen der Quadrate eine andere Einheit als die Merkmalsausprägungen. Sind z. B. die Merkmalsausprägungen in kg gemessen, so wird s^2 in $(\text{kg})^2$ gemessen. Man definiert deshalb als weiteres Maß die **empirische Standardabweichung** s (englisch: *standard deviation*), indem man die Quadratwurzel aus s^2 zieht.

Definition 1.12 (Standardabweichung)

Die Zahl s mit

$$s := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad n \geq 2$$

heißt **empirische Standardabweichung**. ◆

Dadurch hat das Streuungsmaß wieder die ursprüngliche Einheit. Empirische Standardabweichung s und empirische Varianz s^2 werden in den Anwendungen am häufigsten gebraucht.

Didaktische Anmerkungen:

1. Die Frage, warum man bei der empirischen Varianz bei der Mittelwertbildung der quadratischen Abweichungen durch $n-1$ und nicht durch n dividiert, kann in der Sekundarstufe I nicht überzeugend beantwortet werden.

In der Sekundarstufe II kann im Rahmen der Schätztheorie die Begründung für die Division durch $n-1$ statt durch n gegeben werden. Die empirische Varianz $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ist ein sogenannter erwartungstreuer Schätzer für die Varianz σ^2 , während $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ kein erwartungstreuer Schätzer wäre (siehe Pöppelmann [133]).

2. In der (didaktischen) Literatur findet man bei obigen Definitionen auch die Division durch n statt durch $n-1$. Auf Taschenrechnern sind häufig beide Implementationen gebräuchlich. Deshalb sollte vorher überprüft werden, ob durch n oder durch $n-1$ dividiert wird.
3. Bei großem Stichprobenumfang n ist der Unterschied zwischen der „Division durch n “ und der „Division durch $n-1$ “ jedoch unerheblich.
4. Bei Anwendungen (insbesondere in den Naturwissenschaften) gibt man arithmetisches Mittel \bar{x} und Standardabweichung s häufig nicht getrennt an, sondern in der Form $\bar{x} \pm s$.

Hat man annähernd normalverteilte Daten, dann gilt:

- a) Ca. 68 % der Daten liegen im Bereich $\bar{x} \pm s$, also im Intervall zwischen $\bar{x} - s$ und $\bar{x} + s$.
- b) Ca. 96 % der Daten liegen im Bereich $\bar{x} \pm 2s$.
- c) Ca. 99 % der Daten liegen im Bereich $\bar{x} \pm 3s$.

Das bedeutet, dass im Durchschnitt etwa 68 % bzw. 96 % bzw. 99 % um höchstens eine Standardabweichung bzw. zwei Standardabweichungen bzw. drei Standardabweichungen vom Mittelwert abweichen. Diese anschauliche Interpretation der empirischen Standardabweichung steht in Korrespondenz zu den drei Sigma-Regeln bei der Normalverteilung. Die Begriffe Normalverteilung und Sigma-Regeln werden im Kapitel 8, Abschnitte 8.5 und 8.5.4, erklärt.

Beispiel 1.11

(Körpergewicht von Kindern) Bei einer medizinischen Untersuchung wurden bei 30 Kindern folgende Körpergewichte (in kg) notiert (Urliste):

| | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 35 | 27 | 36 | 42 | 50 | 32 | 35 | 29 | 44 | 40 |
| 36 | 38 | 45 | 40 | 42 | 34 | 38 | 43 | 45 | 42 |
| 37 | 45 | 51 | 48 | 31 | 34 | 46 | 30 | 38 | 35 |

1. $n = 30$, $\bar{x} = 38,93$ [kg]. Bei der Berechnung von s^2 und s mittels Division durch $n - 1 = 29$ erhält man:

$$s^2 = \frac{1}{29} \sum_{i=1}^{30} (x_i - 38,93)^2 = 39,09 \text{ [kg]}^2;$$

$$s \approx 6,25 \text{ [kg]};$$

$$\bar{x} \pm s = 38,93 \pm 6,25 \text{ [kg]}.$$

2. $n = 30$, $\bar{x} = 38,93$ [kg]. Bei der Berechnung von s^2 und s dividieren wir jetzt durch $n = 30$:

$$s^2 = \frac{1}{30} \sum_{i=1}^{30} (x_i - 38,93)^2 \approx 37,79 \text{ [kg]}^2;$$

$$s \approx 6,15 \text{ [kg]};$$

$$\bar{x} \pm s = 38,93 \pm 6,15 \text{ [kg]}.$$

■

Abschließende didaktische Bemerkungen

1. Da Mittelwerte allein nicht aussagekräftig sind, bedürfen sie zur sachgemäßen Interpretation als Ergänzung der Streuungsmaße. Denn wenn ein See eine durchschnittliche Tiefe von 0,80 m hat, so ist es dennoch nicht ratsam zu versuchen, den See aufrecht gehend zu durchqueren. Der See könnte ja an einer zu durchquerenden Stelle 3 m tief sein.
2. Die Konstruktion der Streuungsparameter erfolgte nach zwei unterschiedlichen Prinzipien:
 - Die Maßzahl wird durch den Abstand zweier Rangmerkmale bestimmt (vgl. Spannweite, Quartilabstand).
 - Die Maßzahl wird durch die Abstände der Daten von einem Lageparameter bestimmt (vgl. mittlere absolute Abweichung, empirische Varianz).
3. Die Aussagekraft des Quartilabstandes ist größer als die der Spannweite, da sich der Quartilabstand nicht nur auf den größten und kleinsten Wert stützt. Durch den Quartilabstand werden die Daten in drei Bereiche aufgeteilt.
4. Im Zusammenhang mit der Behandlung der Quadratwurzel sollte auch die Behandlung der empirischen Standardabweichung in jedem Falle angestrebt werden.
5. Stärker als bisher sollte auch die Fünf-Punkte-Darstellung für Daten in der Schule genutzt werden.

1.2.5 Lineare Regression

Bisher haben wir uns ausschließlich mit der Datenaufbereitung *eines* Merkmals befasst. Von Interesse und von Bedeutung für die Praxis sind aber auch Erkenntnisse über *statistische* Zusammenhänge zwischen *zwei* oder *mehr* Merkmalen innerhalb derselben statistischen Masse. Es geht also in diesem Abschnitt um das Entdecken von Zusammenhängen.

Wir beschränken uns auf bivariate (zweidimensionale) Verteilungen. Wir beobachten und vergleichen also Daten von zwei Merkmalen, die gleichzeitig an einer statistischen Einheit erhoben worden sind, z. B. Körpergröße und Körpergewicht bei Personen, Bruttoeinkommen und Kapitalvermögen bei Familien, Geschwindigkeit und Bremsweg bei Autos, Nettoeinkommen und Mietkosten für das Wohnen, Alter von Männern und Alter von Frauen bei Ehepaaren usw.

Wir beschreiben den Zusammenhang der zwei Variablen X und Y zunächst durch eine Funktion und beschränken uns auf den einfachen Fall des linearen Zusammenhangs und bestimmen die *Regressionsgeraden*. Dabei ist zu bedenken, dass der errechnete funktionale Zusammenhang zwischen den zwei Größen natürlich nur eine mathematische Modellbeschreibung für ein gegebenes Sach-

problem ist. Eine eventuell tatsächlich vorhandene kausale Abhängigkeit der zwei Größen voneinander kann nicht aus dem mathematischen Modell gefolgert werden. Hier ist der Fachmann für das jeweilige Sachproblem gefordert. Das gilt auch für den anschließend behandelten *Korrelationskoeffizienten*. Dieser ist ein Maß für die Stärke des linearen Zusammenhangs.

Das Wort Regression (lateinisch *regressus*: Rückkehr, Rückzug) ist von seinem Wortsinn her eine zunächst durchaus merkwürdig erscheinende Bezeichnung für den durch die Bezeichnung heute in der beschreibenden Statistik gemeinten Sachverhalt. Grob gesagt geht es in der beschreibenden Statistik bei der Regression um eine Beschreibung einer Variablen als Funktion einer anderen Variablen. Es sollen also stochastische Zusammenhänge (Abhängigkeiten) zweier Variablen beschrieben werden.

Die Bezeichnung Regression ist historisch bedingt und geht auf *Sir Francis Galton* (1822 – 1911) zurück. In seinen Studien zur Vererbungslehre stellte Galton fest, dass einerseits große Väter häufig große Nachkommen haben, dass aber andererseits die durchschnittliche Größe der Nachkommen kleiner ist als die der Väter. Analog verhielt es sich mit der Kleinheit. Kleinere Väter hatten häufig kleine Nachkommen, aber die Durchschnittsgröße der Nachkommen war größer als die der Väter. Es ist insgesamt eine Tendenz zur Durchschnittsgröße der Nachkommen gegeben, d. h. es liegt ein Zurückgehen (eine Regression) bezüglich der Größe der Nachkommen auf den Durchschnitt vor. Allgemeiner formuliert: Eine Eigenschaft des Menschen wird von den Nachkommen zwar übernommen, aber nur in einem geringeren Maße. Bezüglich der Eigenschaft tritt also eine (langsame) Rückbildung ein. Galton sprach von einer Regression. Die Merkmalsausprägung aller Individuen einer Art schwankt um einen Mittelwert.

Doch lassen wir Galton selbst zu Wort kommen. In der Einleitung zu der zweiten Ausgabe von 1892 seines Werkes *Hereditary Genius* schreibt Galton: „In der *Natürlichen Vererbung* habe ich gezeigt, daß die Verteilung von Eigenschaften in einer Bevölkerung nicht konstant bleiben kann, wenn *durchschnittlich* die Kinder ihren Eltern ähnlich sehen. Ist dies der Fall so würden die Riesen (in bezug auf irgend eine geistige oder physische Eigentümlichkeit) in jeder folgenden Generation noch riesiger und die Zwerge noch zwerghafter werden. Die gegenwirkende Tendenz ist die, welche ich ‘Regression’ nenne.“ (Siehe: Galton, F.: *Genie und Vererbung*. Autorisierte Übersetzung von O. Neurath und A. Schapire-Neurath. Leipzig 1910. S. XVIII. Die 1. Auflage von *Hereditary Genius* erschien 1869. – Die *Natürliche Vererbung* hat im Original den Titel *Natural Inheritance*. Die 1. Auflage dieses Werkes erschien 1889.) Bei Galton wird also eine Denkweise deutlich, die an Quetelet (siehe Abschnitt 1.1.3) erinnert, nämlich das Bemühen, Durchschnittstypen zu erkennen und aufzustellen.

Wie schon in der Einführung bemerkt, beschränken wir uns im Folgenden auf die Behandlung zweier Variablen.

Sind zweidimensionale Verteilungen (X, Y) gegeben, z. B. die gemeinsame Verteilung der Merkmale Körpergröße X und Körpergewicht Y bei n Personen, so können die Beobachtungswerte dargestellt werden durch Paare von reellen Zahlen $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$. Dieses ist die Urliste. Stellt man diese Datenpaare in einem Koordinatensystem dar, so erhält man eine **Punkt- wolke (Scatter-Diagramm, Streudiagramm)**. Die Punktwolke kann ganz unterschiedlich aussehen.



Abb. a

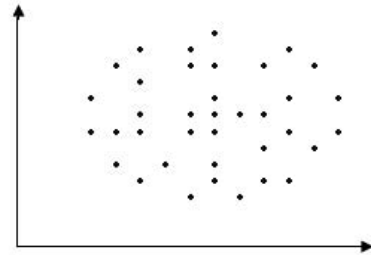


Abb. b

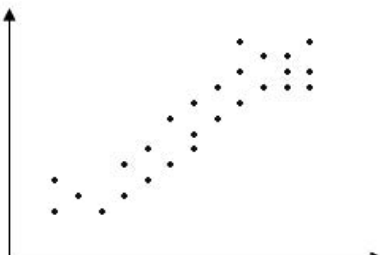


Abb. c

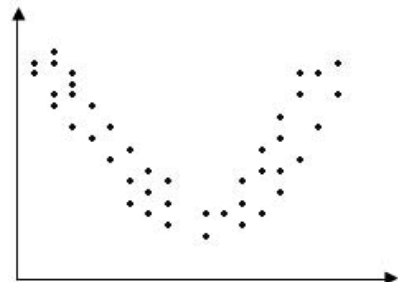


Abb. d

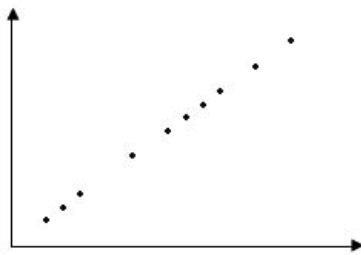


Abb. e

Man versucht, die Punktwolken durch mathematische Funktionen näherungsweise zu beschreiben. Es interessiert die Art (Form) des Zusammenhangs zwischen den beiden Variablen X und Y , falls überhaupt eine Zusammenhangsbeziehung durch die Punktwolke nahegelegt wird. So könnte man bei den Punktwolken in den Abbildungen a und c je einen linearen Zusammenhang, bei der

Punktwolke in Abbildung d einen quadratischen Zusammenhang vermuten. Dass alle Messwerte exakt auf einer Geraden liegen wie in Abbildung e wird man nicht erwarten können. Dagegen sind auch andere Zusammenhangsbeziehungen, wie z. B. ein exponentieller Zusammenhang, denkbar. Die Punktwolke in Abbildung b lässt keinen Zusammenhang erkennen.

In den folgenden Ausführungen beschränken wir uns auf den linearen Fall. Das führt zur Aufstellung der sogenannten Regressionsgeraden. Wir gehen von einem Beispiel aus.

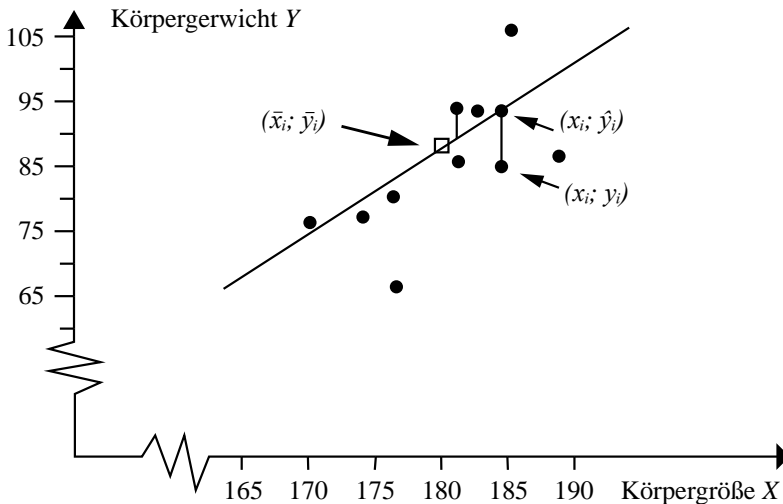
Beispiel 1.12

(Körpergröße/Körpergewicht) Gegeben sei die gemeinsame Verteilung der Merkmale Körpergröße X (in cm) und Körpergewicht Y (in kg) von 10 Personen. Es handelt sich um fiktive Daten.

Die Urliste besteht aus 10 Datenpaaren $(x_i; y_i)$:

(188; 88,5), (177,5; 86,5), (183; 102), (182; 93), (170; 81,5), (185,5; 83,5),
(175,5; 82,5), (175,5; 69), (183; 87,5), (173; 79,5).

Bei Darstellung dieser Paare in einem Koordinatensystem erhält man die folgende Punktwolke (das folgende Scatter-Diagramm):



Wir versuchen, zu dieser Punktwolke eine Gerade, die sogenannte Regressionsgerade, zu bestimmen, die sich der Punktwolke, also den Paaren $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_{10}, y_{10})$, „besonders gut anpasst“. Mathematisch bedeutet dieses das Aufstellen einer Geradengleichung. Je nachdem, ob wir x bzw. y als unabhängige Variable ansehen, müssen wir die Geradengleichung $y = a + bx$ bzw. $x = c + dy$ bestimmen. Statt von unabhängiger Variable und abhängiger Va-

riable zu sprechen, sollte man besser von Einflussgröße und Zielgröße sprechen. Diese Bezeichnungen treffen die Sache und vermeiden Missverständnisse. ■

Bei der mathematischen Berechnung der Regressionsgeraden muss man zunächst klären, was es bedeuten soll, wenn man sagt, die Regressionsgerade hat sich der Punktwolke „besonders gut“ anzupassen. Bezeichnen wir den zu x_i tatsächlich gemessenen Wert mit y_i und den gemäß $y = a + bx$ für x_i theoretisch errechneten Wert mit \hat{y}_i , so sollte die Abweichung der theoretisch berechneten Werte \hat{y}_i von den gemessenen Werten y_i möglichst klein sein. Diese angestrebte Minimierung kann auf verschiedenen Wegen erfolgen. Wir nennen zwei Möglichkeiten:

- a) Die Summe der absoluten Abweichungen, also die Summe $\sum_{i=1}^n |\hat{y}_i - y_i|$, soll minimiert werden.
- b) Die Summe der Quadrate der Abweichungen der \hat{y}_i von den y_i , also die Summe $\sum_{i=1}^n (\hat{y}_i - y_i)^2$, soll minimiert werden.

Zur Bestimmung der Regressionsgeraden ist die unter Punkt b) genannte Möglichkeit die günstigste. Sie wird als „Methode der kleinsten Quadrate“ bezeichnet und geht auf *Carl Friedrich Gauß* (1777 – 1855) zurück. Die Methode bestimmt eindeutig die Variablen a und b , legt also rechnerisch eindeutig die Regressionsgerade fest.

Didaktische Hinweise:

1. Häufig spricht man bei der „Methode der kleinsten Quadrate“ statt von Abweichungen auch von Abständen. Dann ist zu beachten, dass die „Abstände“ der Messpunkte *parallel* zur y -Achse genommen werden und nicht – wie man beim Wort Abstand meinen könnte – die Länge des Lotes vom gemessenen Punkt auf die Regressionsgerade.
2. Wählt man y als Einflussgröße (unabhängige Variable), so hat man analog zur Bestimmung der Geraden $x = c + dy$ die Abstände der Messpunkte parallel zur x -Achse zu nehmen.

Sei also jetzt (x_i, y_i) das gemessene Paar, und sei das Paar (x_i, \hat{y}_i) das Paar, das den Punkt auf der Regressionsgeraden kennzeichnet. Wir suchen die Gerade $y = a + bx$ zu bestimmen unter der Bedingung, dass $\sum_{i=1}^n (\hat{y}_i - y_i)^2$ minimal ist (siehe Abbildung). Da $\hat{y}_i = a + bx_i$ ist, folgt

$$\hat{y}_i - y_i = a + bx_i - y_i.$$

Für die Summe S der Quadrate erhalten wir

$$S(a, b) = \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (a + bx_i - y_i)^2. \quad (1.4)$$

Die zu minimierende Funktion S ist also eine Funktion der zwei Variablen a und b . Mit Hilfe der Differentialrechnung zweier Variablen sind die Variablen a und b so zu bestimmen, dass die Funktion S minimal wird.

Im folgenden hinreichenden Kriterium für den Nachweis der Existenz eines relativen Minimums treten partielle Ableitungen auf. Wir geben deshalb vorab einige formale Hinweise.

Im Gegensatz zum geraden d bei der Differentiation von Funktionen einer Veränderlichen benutzt man bei der Differentiation von Funktionen zweier (oder mehrerer) Veränderlichen ein rundes geschwungenes ∂ . Die erste partielle Ableitung nach x (man betrachtet dabei y als fest) einer Funktion $f(x, y)$ mit den zwei Variablen x und y schreibt man als $\frac{\partial f}{\partial x}$, und für die erste partielle Ableitung der Funktion f nach y (man betrachtet x als konstant) schreibt man $\frac{\partial f}{\partial y}$. Bei den partiellen Ableitungen zweiter Ordnung schreibt man analog $\frac{\partial^2 f}{\partial x^2}$ (zweimaliges differenzieren nach x bei festgehaltenem y). Das Symbol $\frac{\partial^2 f}{\partial x \partial y}$ steht für die gemischte zweite partielle Ableitung, in der zunächst nach x (bei festgehaltenem y) und anschließend nach y (bei festgehaltenem x) differenziert wird. Analog besagt das Symbol $\frac{\partial^2 f}{\partial y \partial x}$, dass zunächst f partiell nach y und anschließend partiell nach x differenziert wird. Statt $\frac{\partial f}{\partial x}$ schreibt man auch f_x und entsprechend auch für die anderen Fälle, z. B. $\frac{\partial^2 f}{\partial x \partial y} = f_{xy}$.

Nun das hinreichende Kriterium für den Nachweis der Existenz eines relativen Minimums, das wir ohne Beweis der Analysis entnehmen:

Kriterium für relatives Minimum

Falls die Funktion $f(x, y)$ mit zwei Variablen zweimal stetig partiell differenzierbar in einer Umgebung von (x_0, y_0) ist, dann besitzt $f(x, y)$ in (x_0, y_0) ein relatives Minimum, wenn gilt:

- a) $\frac{\partial f}{\partial x}(x_0, y_0) = 0$;
- b) $\frac{\partial f}{\partial y}(x_0, y_0) = 0$;
- c) $\frac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0$;
- d) $\frac{\partial^2 f}{\partial x^2} \cdot \frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 > 0$ in (x_0, y_0) .

Die Gleichungen unter a) und b) formulieren notwendige Bedingungen.

Bestimmung des relativen Minimums für die Funktion S mittels des obigen Kriteriums:

Notwendig für die Existenz eines Minimums der Funktion S ist also, dass die beiden ersten partiellen Ableitungen der Funktion S eine gemeinsame Nullstelle (a_0, b_0) haben:

Partielle Differentiation von S nach a (b wird als Konstante angesehen) ergibt

$$\frac{\partial S}{\partial a} = -2 \cdot \sum_{i=1}^n (y_i - a - bx_i).$$

Partielle Differentiation nach b (a wird als Konstante angesehen) ergibt

$$\frac{\partial S}{\partial b} = -2 \cdot \sum_{i=1}^n x_i (y_i - a - bx_i).$$

Wir ermitteln die Werte a_0 und b_0 , für die die beiden partiellen Ableitungen Null werden:

$$\sum_{i=1}^n (y_i - a_0 - b_0 x_i) = 0 \quad (1.5)$$

$$\sum_{i=1}^n (y_i x_i - a_0 x_i - b_0 x_i^2) = 0. \quad (1.6)$$

Wir erhalten aus (1.5)

$$\begin{aligned} \sum_{i=1}^n y_i - na_0 - b_0 \cdot \sum_{i=1}^n x_i &= 0 \\ \Leftrightarrow n\bar{y} - na_0 - nb_0\bar{x} &= 0. \end{aligned}$$

Also:

$$\bar{y} = a_0 + b_0\bar{x}, \quad (1.7)$$

d. h. (\bar{x}, \bar{y}) liegt auf der Regressionsgeraden. Das ist ein interessantes Zwischenergebnis. Der Punkt (\bar{x}, \bar{y}) heißt **Schwerpunkt**.

Aus (1.6) erhalten wir

$$\begin{aligned} \sum_{i=1}^n x_i y_i - \sum_{i=1}^n a_0 x_i - \sum_{i=1}^n b_0 x_i^2 &= 0 \\ \Leftrightarrow \sum_{i=1}^n x_i y_i - na_0\bar{x} - b_0 \cdot \sum_{i=1}^n x_i^2 &= 0. \end{aligned}$$

Einsetzen von $a_0 = \bar{y} - b_0\bar{x}$ (Gleichung (1.7)) liefert

$$\begin{aligned} \sum_{i=1}^n x_i y_i - n\bar{x}(\bar{y} - b_0\bar{x}) &= b_0 \cdot \sum_{i=1}^n x_i^2 \\ \Leftrightarrow \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} &= b_0 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right), \end{aligned}$$

also

$$b_0 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}.$$

Man erhält also insgesamt:

$$\begin{aligned} a_0 &= \bar{y} - b_0 \bar{x} \\ b_0 &= \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}. \end{aligned} \quad (1.8)$$

Wir können diese beiden Gleichungen kürzer schreiben. Dazu beachten wir, dass für die empirische Varianz der x_i -Werte gilt (siehe Aufgabe 6 in Abschnitt 1.2.8)

$$s_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right), \quad \text{für } n \geq 2.$$

Erweitert man in (1.8) den Bruch für b_0 mit $\frac{1}{n-1}$, erhält man

$$b_0 = \frac{\frac{1}{n-1} (\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y})}{\frac{1}{n-1} (\sum_{i=1}^n x_i^2 - n \bar{x}^2)}.$$

Definieren wir außerdem

$$s_{xy} := \text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

als die **empirische Kovarianz** der x - und y -Werte und zeigen, dass

$$s_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right)$$

gilt, so erhält man

$$\boxed{b_0 = \frac{s_{xy}}{s_x^2}} \quad \text{und} \quad \boxed{a_0 = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x}}$$

für $s_x^2 \neq 0$.

Um zu beweisen, dass S in (a_0, b_0) tatsächlich ein Minimum hat, ist noch zu zeigen, dass

$$\frac{\partial^2 S}{\partial a^2} \quad \text{an der Stelle } (a_0, b_0) \quad \text{positiv ist} \quad (1.9)$$

und

$$\frac{\partial^2 S}{\partial a^2} \cdot \frac{\partial^2 S}{\partial b^2} - \left(\frac{\partial^2 S}{\partial a \partial b} \right)^2 \quad \text{an der Stelle } (a_0, b_0) \quad \text{positiv ist} \quad (1.10)$$

Beweis zu (1.9):

$$\frac{\partial^2 S}{\partial a^2} = -2 \cdot \sum_{i=1}^n (-1) = 2n,$$

d. h. $\frac{\partial^2 S}{\partial a^2} > 0$ für alle a, b , da unabhängig von a und b .

Beweis zu (1.10):

$$\begin{aligned}\frac{\partial^2 S}{\partial b^2} &= -2 \cdot \sum_{i=1}^n (-x_i^2) = 2 \cdot \sum_{i=1}^n x_i^2, \\ \frac{\partial^2 S}{\partial a \partial b} &= -2 \cdot \sum_{i=1}^n (-x_i) = 2 \cdot \sum_{i=1}^n x_i = 2n\bar{x}.\end{aligned}$$

Also:

$$\begin{aligned}\frac{\partial^2 S}{\partial a^2} \cdot \frac{\partial^2 S}{\partial b^2} - \left(\frac{\partial^2 S}{\partial a \partial b} \right)^2 &= 4n \cdot \sum_{i=1}^n x_i^2 - 4n^2 \bar{x}^2 \\ &= 4n \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = 4n(n-1)s_x^2,\end{aligned}$$

d. h.

$$\frac{\partial^2 S}{\partial a^2} \cdot \frac{\partial^2 S}{\partial b^2} - \left(\frac{\partial^2 S}{\partial a \partial b} \right)^2 > 0 \quad \text{für alle } a \text{ und } b.$$

Damit ist der Nachweis erbracht, dass die Funktion $S(a, b)$ in (a_0, b_0) ein Minimum hat.

Analog kann man auch die Regressionsgerade bei einer vermuteten Abhängigkeit der x -Werte von den y -Werten herleiten und zeigen, dass auch diese Regressionsgerade $x = c + dy$ durch den Schwerpunkt (\bar{x}, \bar{y}) geht.

Für das Beispiel 1.12 (Körpergröße/Körpergewicht) erhält man

- für den Schwerpunkt (\bar{x}, \bar{y}) die Werte $\bar{x} = 179,3$ und $\bar{y} = 85,3$;
- für die Regressionsgerade $y = a + bx$ die Gleichung $y = 40,53 + 0,25x$.

Didaktische Hinweise und Ergänzungen

1. Eine einfachere – vor allem für die Schule geeignete – Bestimmung der Regressionskoeffizienten, die ohne die partielle Differentiation auskommt, erhält man, wenn man von der *Voraussetzung* ausgeht, dass der Schwerpunkt (\bar{x}, \bar{y}) , das Paar der arithmetischen Mittel, auf der Regressionsgeraden liegen soll. (Man beachte, dass wir diese Forderung oben als Ergebnis erhielten.) Unter dieser Annahme lässt sich die gesuchte Regressionsgerade in der Punkt-Steigungs-Form durch die Gleichung

$$y - \bar{y} = b(x - \bar{x})$$

bei zu optimierendem *Steigungskoeffizienten* b beschreiben, falls nicht alle x_i -Werte in der Urliste gleich sind.

Man bestimmt den Steigungskoeffizienten b wiederum derart, dass die Summe der Quadrate der Abweichungen der theoretischen (berechneten) Werte $\hat{y}_i = f(x_i)$ von den empirischen Werten y_i

$$\sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (b(x_i - \bar{x}) + \bar{y} - y_i)^2$$

minimal ist. D. h. wenn wir die Summe der Fehlerquadrate als Funktion von b betrachten, suchen wir das Minimum der Funktion S einer Variablen mit

$$S(b) = \sum_{i=1}^n (b(x_i - \bar{x}) + \bar{y} - y_i)^2.$$

2. Der folgende kurz angedeutete Weg kann ebenfalls im Unterricht vertreten werden. Man setzt voraus, dass die Summe aller Differenzen der theoretischen von den erhobenen Daten Null ist:

$$\sum_{i=1}^n (\hat{y}_i - y_i) = 0.$$

Man sagt: Die Summe aller **Residuen** $e_i = \hat{y}_i - y_i$ ist Null. (Der Beweis dieser Annahme ist in Aufgabe 13 a) im Abschnitt 1.2.8 zu führen.) Mit dieser Aussage zeigt man, dass (\bar{x}, \bar{y}) auf der gesuchten Geraden liegt und errechnet die Regressionsgerade wie unter Punkt 1.

3. Die Regressionsgeraden werden auch Ausgleichsgeraden genannt.
4. Befasst man sich mit Zeitreihen (man beobachtet die Entwicklung einer Größe über längere Zeitspannen: Geburtenentwicklung, Produktionsentwicklung, Entwicklung der Zahl der Arbeitslosen etc.), kann die Regressionsgerade (Ausgleichsgerade) zur Beschreibung eines *Trends* herangezogen werden.
5. Bei der *Interpretation* der Regressionsgeraden ist Vorsicht geboten. Es ist zunächst zu beachten, dass sich die Regressionsgeraden immer bestimmen lassen, also auch dann, wenn die Punktwolke die Annahme eines *linearen* Zusammenhangs eigentlich verbietet. Bei der mathematischen Modellbildung darf man also nie die empirischen Daten aus dem Blick verlieren. Ferner ist auch beim mathematischen Umgang mit einer konkreten Regressionsgeraden Vorsicht geboten. Aus der Regressionsgeraden $y = 40,53 + 0,25x$ für das Beispiel „Körpergröße/Körpergewicht“ kann nicht geschlossen werden, dass sich das Körpergewicht y für eine vorgegebene Körpergröße x exakt nach der Gleichung $y = 40,53 + 0,25x$ berechnen lässt. Wenn das so wäre, hätten ja alle 1 m großen Personen dasselbe Körpergewicht von 65,53 kg, und eine Person der Größe 0 cm hätte ein Gewicht von 40,53 kg. Man erkennt, dass diese Interpretation der Regressionsgeraden sinnlos und unzulässig ist. Ausgangspunkt für das Aufstellen

der Regressionsgeraden waren gegebene Punktepaaire aus einem bestimmten Bereich, z. B. Körpergrößen von 170 cm bis 188 cm. Nur für diesen Bereich kann die Regressionsgerade als zusammenfassende Beschreibung des Zusammenhangs zwischen den Größen X und Y angesehen werden. Die Regressionsgerade könnte eine andere Lage haben, wenn weitere Daten zur Verfügung stehen würden. Mit *Vorhersagen* muss man also sehr vorsichtig sein.

6. Die beiden Regressionsgeraden $y = a_0 + b_0x$ und $x = a_1 + b_1y$ fallen zusammen genau dann, wenn $b_0 = \frac{1}{b_1}$ gilt (siehe Aufgabe 14 im Abschnitt 1.2.8).
7. In der didaktischen Literatur werden zahlreiche Vorschläge gemacht, die lineare Regression und Korrelation auf Schulniveau zu behandeln. Zur ausführlichen Diskussion dieses Themenkreises verweisen wir an dieser Stelle auf weitere Literatur. Hingewiesen sei insbesondere auf die *Kommentierte Bibliographie zum Thema „Regression und Korrelation“* von Borovcnik und König ([24]). Weiter weisen wir hin auf: Borovcnik [23], [25], Engel/Sedlmeier [50], Heilmann [65], Hui [70], Ineichen/Stocker [75], Koßwig [82], v. Pape/Wirths [128], Reichel [136], Vohmann [176], Wirths [185], [186], Wolf [188].

1.2.6 Korrelation

Mit dem Aufstellen der Regressionsgeraden ist die einfache Beschreibung des linearen Zusammenhangs der Variablen X und Y erreicht. Das Beispiel „Körpergröße/Körpergewicht“ und seine weitere Bearbeitung zeigen jedoch, dass die Beschreibung eine Vereinfachung mit Informationsverlust bedeutet. Wir hatten ein lineares Modell zugrundegelegt, und unter dieser Modellannahme ist die gefundene Geradengleichung die beste. Insgesamt kann aber das Sachproblem durch die Geradengleichung immer noch sehr schlecht beschrieben sein – denn die Geradengleichung kann ja immer bestimmt werden.

Wir suchen deshalb nach einem Maß der Korrelation, also nach einem Maß für die *Stärke (Güte)* des *linearen* Zusammenhangs der beiden Merkmale. Diese wird durch eine Zahl, den *Korrelationskoeffizienten*, beschrieben. Wir besprechen nur den **Korrelationskoeffizient nach Bravais-Pearson** (August Bravais (1811 – 1863), Karl Pearson (1857 – 1936)).

Um ein Maß für die Stärke des linearen Zusammenhangs zu finden, berücksichtigt man die Streuung der Punkte um die Regressionsgeraden. Genauer: Man vergleicht die Varianz der \hat{y}_i -Werte (auf der Regressionsgeraden) mit der Varianz der tatsächlichen y_i -Werte aus der Erhebung. Bei einem starken linearen Zusammenhang müssten beide Varianzen in etwa übereinstimmen.

Gegeben seien also n Datenpaare $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Wir wissen, dass die Regressionsgerade $y = a + bx$ durch den Schwerpunkt (\bar{x}, \bar{y}) geht. Hierbei ist $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ und $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Bezeichnen wir wieder die y -Werte auf der Regressionsgeraden mit \hat{y}_i , so gilt

$$\begin{aligned}\hat{y}_i - \bar{y} &= b(x_i - \bar{x}), \\ \hat{y}_i &= b(x_i - \bar{x}) + \bar{y} = bx_i - b\bar{x} + \bar{y}.\end{aligned}\quad (1.11)$$

Wir betrachten die Varianz $\frac{1}{n-1} \cdot \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2$ der \hat{y}_i -Werte bezogen auf ihr arithmetisches Mittel $\bar{\hat{y}}$.

Bevor wir diese Varianz umformen, versuchen wir $\bar{\hat{y}}$ durch \bar{y} auszudrücken. Es gilt gemäß (1.11):

$$\begin{aligned}\bar{\hat{y}} &= \frac{1}{n} \sum_{i=1}^n (bx_i - b\bar{x} + \bar{y}), \quad \text{und es folgt} \\ \bar{\hat{y}} &= \frac{1}{n} \cdot \left(\sum_{i=1}^n bx_i - n \cdot b\bar{x} + n\bar{y} \right), \\ \bar{\hat{y}} &= \frac{1}{n} \cdot \left(\sum_{i=1}^n bx_i - nb \cdot \frac{1}{n} \sum_{i=1}^n x_i - n\bar{y} \right), \\ \bar{\hat{y}} &= \frac{1}{n} \cdot n\bar{y}, \\ \bar{\hat{y}} &= \bar{y},\end{aligned}$$

d. h. das arithmetische Mittel \bar{y} der beobachteten Werte y_i ist gleich dem arithmetischen Mittel $\bar{\hat{y}}$ der mittels der Regressionsgeraden errechneten Werte \hat{y}_i .

Mit diesem interessanten Ergebnis erhalten wir:

$$\begin{aligned}\frac{1}{n-1} \cdot \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 &= \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (bx_i - b\bar{x} + \bar{y} - \bar{y})^2 \\ &= \frac{b^2}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= b^2 \cdot s_x^2.\end{aligned}$$

Hierbei bedeutet s_x^2 die empirische Varianz der x -Werte. Berücksichtigen wir, dass gemäß Abschnitt 1.2.5 für die empirische Kovarianz s_{xy} der x - und y -Werte die Beziehung $s_{xy} = b \cdot s_x^2$ gilt, so folgt

$$\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = b^2 \cdot s_x^2 = \frac{s_{xy}^2 \cdot s_x^2}{s_x^4} = \frac{s_{xy}^2}{s_x^2}.$$

Diese Varianz vergleichen wir mit der Varianz der y -Werte. Wir bezeichnen letztere analog mit s_y^2 . Wir berechnen den Quotienten und erhalten:

$$\frac{s_{xy}^2}{s_x^2} : s_y^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2}.$$

Diese Zahl ist ein Maß für die Stärke der linearen Abhängigkeit der beiden Verteilungen.

Definition 1.13 (Korrelationskoeffizient nach Bravais-Pearson)

Die Zahl $r := \frac{s_{xy}}{s_x \cdot s_y}$ mit $s_x \neq 0$ und $s_y \neq 0$ heißt der **Korrelationskoeffizient nach Bravais-Pearson**. ◆

Durch Einsetzen der Werte für s_{xy} , s_x und s_y erhalten wir

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Der Zähler s_{xy} bestimmt das Vorzeichen von r . Ist beispielsweise x_i größer (bzw. kleiner) als das arithmetische Mittel \bar{x} , und ist y_i größer (bzw. kleiner) als das arithmetische Mittel \bar{y} , dann sind die Abweichungen $(x_i - \bar{x})$ und $(y_i - \bar{y})$ beide positiv (bzw. negativ), und folglich ist ihr Produkt $(x_i - \bar{x})(y_i - \bar{y})$ positiv. In den anderen Fällen, wenn also die Abweichungen $(x_i - \bar{x})$ und $(y_i - \bar{y})$ entgegengesetzte Vorzeichen haben, ist das Produkt $(x_i - \bar{x})(y_i - \bar{y})$ negativ. Die Zahl r ist also dann positiv, wenn die positiven Werte der Produkte $(x_i - \bar{x})(y_i - \bar{y})$ in den n Messwerten überwiegen.

Man kann zeigen, dass stets gilt:

$$-1 \leq r \leq +1.$$

Dieser Nachweis ist in Aufgabe 15 im Abschnitt 1.2.8 zu erbringen.

Die Messdaten $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ liegen genau dann auf einer Geraden, wenn der zugehörige Korrelationskoeffizient r gleich 1 oder gleich -1 ist. Wie wir schon früher bemerkten, wird das bei realen Daten wohl nie eintreten.

Zur Interpretation des Korrelationskoeffizienten r ist folgende *Sprechweise* geeignet:

| r | Korrelation |
|------------|--|
| 0 | keine (lineare) Korrelation; es kann andere Zusammenhänge geben |
| 1 | perfekte Korrelation; steigende Werte der unabhängigen Variablen entsprechen steigenden Werten der abhängigen Variablen |
| -1 | perfekte Korrelation, allerdings negative lineare Abhängigkeit; steigende Werte der unabhängigen Variablen entsprechen fallenden Werten der abhängigen Variablen |
| 0 bis 0,5 | schwache (positive) Korrelation |
| 0,8 bis 1 | starke Korrelation |
| 0 bis -0,5 | schwache (negative) Korrelation |

Eine quantitative Interpretation ist grundsätzlich schwierig. Denn bei einer Korrelation von 0,95 wissen wir ohne Kenntnis des Scatterdiagramms oder der Regressionsgeraden nicht, ob die „Zunahme“ steil oder flach verläuft (Steigung!).

Im Beispiel „Körpergröße/Körpergewicht“ beträgt der Korrelationskoeffizient r

$$r = \frac{27,25}{5,84 \cdot 8,69} \approx 0,54.$$

Anmerkungen und Ergänzungen:

1. Der Korrelationskoeffizient r ist nur auf lineare Zusammenhänge bezogen. Das macht auch seine Herleitung deutlich.
2. Beim Korrelationskoeffizienten r nach Bravais-Pearson wird nicht zwischen Einflussgröße (unabhängiger Variable) und Zielgröße (abhängiger Variable) unterschieden. Man schaue sich unter diesem Aspekt noch einmal die Definition von r an!
3. Der Korrelationskoeffizient nach Bravais-Pearson ist das geometrische Mittel der Steigungen der beiden Regressionsgeraden.
4. Der Korrelationskoeffizient r nach Bravais-Pearson ist nicht definiert, wenn die empirische Standardabweichung s_x oder s_y gleich Null ist.
5. Das Vorzeichen des Korrelationskoeffizienten r nach Bravais-Pearson drückt die Richtung des linearen Zusammenhangs aus, der absolute Betrag von r drückt die Stärke des linearen Zusammenhangs aus.

6. Bei der *Interpretation* ist äußerste Vorsicht geboten. Der lineare funktionale Zusammenhang zwischen den zwei Größen ist eine mathematische Modellbeschreibung eines Sachproblems, nicht mehr. Scatterdiagramm und Regressionsgerade sagen nichts aus über die Stärke des Zusammenhangs. Das macht der Korrelationskoeffizient. Aber auch bei einer starken Korrelation darf daraus nicht auf eine kausale Abhängigkeit der zwei Größen geschlossen werden. Der Nachweis einer kausalen Beziehung kann nicht aus dem mathematischen Modell gefolgert werden, sondern nur aus der Sache selbst. Es ist ein Sachproblem. Ein Beispiel kann dieses verdeutlichen: In schwedischen Landkreisen beobachtete man eine Abnahme der Störche und gleichzeitig eine Abnahme der Geburten. Ein kausaler Zusammenhang ist aber trotz hoher Korrelation auszuschließen. Das ist das „klassische“ Beispiel einer „nonsense“ Korrelation. Perfekte Korrelation sagt nur, dass sich Daten zweier Größen (linear) gleichzeitig verändern, aber nicht, dass sie ursächlich miteinander gekoppelt sind. W. Krämer ([83], S. 145) nennt ein anderes interessantes Beispiel: In den 1960er- und 1970er-Jahren hat man „eine erstaunliche negative Korrelation zwischen Rocklänge in der Damenwelt und dem Dow-Jones-Aktienindex festgestellt, wofür wohl nur der Zufall als Erklärung bleibt.“ (Siehe auch W. Krämer [85], Kapitel 14.) Neben den sinnlosen Korrelationen gibt es auch noch die „scheinbaren“ Korrelationen zwischen zwei Datenmengen, bei der die Korrelation nur mittelbar (also indirekt) über eine dritte Variable gegeben ist. So glauben z. B. auch einige Forscher nicht ausschließen zu können, dass die „unsinnige“ Korrelation „Störche/Geburten“ in Wirklichkeit vielleicht doch eine „scheinbare“ Korrelation ist, indem nämlich ein drittes Merkmal „zunehmende Industrialisierung“ sowohl die Abnahme der Störche als auch die Abnahme der Geburten bedingt.

Im Rahmen dieses Buches können wir weitere interessante Themen aus der Beschreibenden Statistik wie z. B. *Konzentrationsphänomene im wirtschaftlichen Bereich (Monopolbildung)* nicht behandeln. Wir verweisen auf entsprechende Literatur. Einen guten Einstieg liefern Lehn u. a. [115].

1.2.7 Fehler und Manipulationsmöglichkeiten

Da es in der Statistik um die Erhebung, Aufbereitung und Interpretation von Daten geht, können auf *jeder* dieser Stufen Fehler gemacht werden.

Zu dieser Thematik gibt es zahlreiche Literatur. Auch wir haben uns dazu geäußert, so dass wir uns an dieser Stelle auf Literaturhinweise beschränken: Kütting [102] (Kapitel VII), [104], [99], W. Krämer [85]. In den genannten Publikationen befinden sich zahlreiche weitere Literaturhinweise zu dieser Thematik.

1.2.8 Aufgaben und Ergänzungen

1. Formulieren Sie das in Abschnitt 1.2.3 an einem Beispiel beschriebene „Verfahren zur Berechnung des arithmetischen Mittels durch Korrektur eines geschätzten arithmetischen Mittels“ allgemein und begründen Sie dieses.
2. Die folgenden Daten geben die (fiktiven) Körpergrößen von Neugeborenen in einer Klinik an:

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| 40 | 41 | 48 | 52 | 52 | 49 | 49 | 50 |
| 46 | 51 | 49 | 51 | 51 | 48 | 52 | 49 |
| 51 | 50 | 53 | 52 | 53 | 50 | 51 | 50 |
| 54 | | | | | | | |

- a) Stellen Sie die Daten in einem Stengel-Blatt-Diagramm dar.
 - b) Berechnen Sie das arithmetische Mittel und den Median der Daten.
 - c) Berechnen Sie die empirische Standardabweichung und den Quartilabstand.
 - d) Stellen Sie die Verteilung der absoluten Häufigkeiten der Daten in einem Histogramm mit jeweils der Klassenbreite 3 (cm) dar.
3. Bei einem Sportfest wurden die folgenden Weitsprungleistungen (in cm) gemessen:

| | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|
| 340 | 417 | 525 | 495 | 530 | 340 | 430 | |
| 553 | 373 | 450 | 485 | 510 | 492 | 387 | |
| 420 | 505 | 495 | 407 | 482 | 533 | 447 | 530 |

- a) Stellen Sie die Daten in einem Stengel-Blatt-Diagramm dar.
 - b) Stellen Sie die Verteilung der Daten in einem Kasten-Schaubild (box-plot-Diagramm) dar.
4. Die Preissteigerungen für ein elektronisches Gerät betrugen in fünf aufeinanderfolgenden Jahren 5 %, 7 %, 12 %, 6 % und 4 %.
 - a) Geben Sie mit kurzer Begründung an, welcher Mittelwert die durchschnittliche Preissteigerung für den angegebenen Zeitraum am besten beschreibt.
 - b) Wie groß ist die durchschnittliche Preissteigerung?
 5. Bei einer medizinischen Untersuchung einer Schulklasse wurden u. a. folgende Körpergewichte (in kg) festgestellt:

| | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 35 | 27 | 36 | 42 | 50 | 32 | 35 | 29 | 44 | 40 |
| 36 | 38 | 45 | 40 | 42 | 34 | 38 | 43 | 45 | 42 |
| 37 | 45 | 52 | 48 | 31 | 34 | 46 | 30 | 38 | 35 |

- a) Bilden Sie Klassen der Breite 3 beginnend mit der Klasse $K_1 = [27, 30[$ für die Merkmalsausprägungen und bestimmen Sie die kumulierten relativen Häufigkeiten.

b) Stellen Sie die empirische Verteilungsfunktion der klassierten Daten graphisch dar.

6. Zeigen Sie: Für die empirische Varianz s^2 gilt:

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right), \quad n \geq 2.$$

7. Gegeben seien zwei reelle Zahlen $a > 0$, $b > 0$. Seien \bar{x}_h das harmonische, \bar{x}_g das geometrische und \bar{x} das arithmetische Mittel dieser Daten.

a) Zeigen Sie:

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x}.$$

b) Wann gilt das Gleichheitszeichen?

8. Gegeben seien die quantitativen Daten x_1, x_2, \dots, x_n . Beweisen Sie: Die Summe der Quadrate der Abweichungen aller n Daten von ihrem arithmetischen Mittel \bar{x} ist kleiner als die Summe der Quadrate der Abweichungen aller Messwerte von einem beliebigen anderen reellen Wert c :

$$\sum_{i=1}^n (x_i - \bar{x})^2 < \sum_{i=1}^n (x_i - c)^2, \quad c \in \mathbb{R}, \quad c \neq \bar{x}.$$

(Minimumeigenschaft des arithmetischen Mittels)

9. Beweisen Sie die Minimumeigenschaft des Medians (Satz 1.2): Für beliebiges $c \in \mathbb{R}$ gilt:

$$\sum_{i=1}^n |x_i - x_{0,5}| \leq \sum_{i=1}^n |x_i - c|.$$

10. Gegeben seien n Daten x_1, x_2, \dots, x_n eines quantitativen Merkmals mit dem arithmetischen Mittel \bar{x} . Zeigen Sie: Unterwirft man alle diese Daten x_i ($i = 1, 2, \dots, n$) einer linearen Transformation $x_i \rightarrow a + b \cdot x_i$ ($a, b \in \mathbb{R}$, $b \neq 0$), so erhält man das arithmetische Mittel \bar{x}_t der transformierten Daten durch dieselbe lineare Transformation aus dem arithmetischen Mittel \bar{x} der ursprünglichen Daten.

11. Gegeben seien n quantitative Daten x_1, x_2, \dots, x_n mit dem arithmetischen Mittel \bar{x} .

a) Zeigen Sie (ohne Verwendung der Differentialrechnung): Die Funktion $f: \mathbb{R} \rightarrow \mathbb{R}$ mit $f(x) = \sum_{i=1}^n (x_i - x)^2$ hat an der Stelle \bar{x} ein Minimum.

b) Beweisen Sie a) mit Methoden der Differentialrechnung.

12. Sei $d_{\bar{x}}$ die mittlere absolute Abweichung vom arithmetischen Mittel \bar{x} , sei $d_{x_{0,5}}$ die mittlere absolute Abweichung vom Median (siehe Definition 1.10) und sei s die Standardabweichung der Daten (siehe Definition 1.12). Dann gilt

$$d_{x_{0,5}} \leq d_{\bar{x}} \leq s.$$

13. Gegeben seien n Datenpaare $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Es sei \bar{x} das arithmetische Mittel der x_i -Werte ($i = 1, 2, \dots, n$), und es sei \bar{y} das arithmetische Mittel der y_i -Werte ($i = 1, 2, \dots, n$). Die y -Werte auf der Regressionsgeraden werden mit \hat{y}_i ($i = 1, 2, \dots, n$) bezeichnet. Mit $\bar{\hat{y}}$ wird das arithmetische Mittel der \hat{y}_i -Werte bezeichnet. Als **Residuen** e_i werden die Differenzen $e_i := \hat{y}_i - y_i$ ($i = 1, 2, \dots, n$) bezeichnet.
- a) Beweisen Sie: Die Summe aller Residuen ist Null.
 - b) Beweisen Sie (auf einem anderen Weg als in Abschnitt 1.2.6) die Gültigkeit der Gleichung $\bar{\hat{y}} = \bar{y}$.
14. Beweisen Sie: Die beiden Regressionsgeraden $y = a_x + b_x x$ und $x = a_y + b_y y$ fallen genau dann zusammen, wenn gilt: $b_x = \frac{1}{b_y}$.
15. Beweisen Sie: Der Korrelationskoeffizient r nach Bravais-Pearson (s. Definition 1.13) nimmt nur Werte aus dem Intervall $[-1, +1]$ an.

Elementare Stochastik

Mathematische Grundlagen und didaktische Konzepte

Kütting, H.; Sauer, M.J. - Padberg, F. (Hrsg.)

2011, XII, 414 S. 103 Abb., Softcover

ISBN: 978-3-8274-2759-5