Beschreibende Statistik

In der beschreibenden Statistik geht es darum, einen Datensatz möglichst übersichtlich und gleichzeitig möglichst präzise und aussagekräftig darzustellen. Dies kann einerseits durch Diagramme und andererseits durch statistische Kennzahlen erfolgen.

1 Merkmale

Die Menge aller für die Untersuchung relevanten Elemente bezeichnet man als Grundgesamtheit. Oftmals ist es nicht möglich, alle Elemente der Grundgesamtheit zu untersuchen (z. B. weil es zu teuer und aufwändig wäre, alle Wahlberechtigten zu befragen). Man untersucht stattdessen häufig einen kleinen Teil der Grundgesamtheit – eine sogenannte Stichprobe.

Die erhobenen Eigenschaften der untersuchten Elemente, bezeichnet man als Merkmale. Die möglichen Werte, die ein Merkmal aufweisen kann, nennt man Merkmalsausprägungen.

Beispiel: *Zur Untersuchung des Zusammenhangs zwischen schulischen Leistungen und privaten Einflussfaktoren wird ein Fragebogen entwickelt. Konkret werden folgende Merkmale erhoben: „Geschlecht", „Anzahl der Geschwister", „Dauer des Schulwegs", „eigener Computer", „häufigste im Haushalt gesprochene Sprache"und „Anzahl der Personen im Haushalt". Es soll jeweils geklärt werden, welche Merkmalsausprägungen die einzelnen Merkmale haben können. Außerdem soll erläutert werden, welche Probleme bei ankreuzbaren Optionen auftreten können und was bei offenen Feldern problematisch sein könnte.*

Nachfolgend wird jedes Merkmal erläutert:

- Beim Merkmal „Geschlecht" gibt es grundsätzlich die beiden Merkmalsausprägungen „männlich" und „weiblich". In vielen Ländern gibt es jedoch bereits eine dritte anerkannte Option.
- Theoretisch kann das Merkmal „Anzahl der Geschwister" jeder natürlichen Zahl inklusive 0 entsprechen.
- Die „Dauer des Schulwegs" kann theoretisch jeder nichtnegativen reellen Zahl entsprechen, wobei der Wert 0 nur bei Schulen mit Schülerheim sinnvoll ist.
- Das Merkmal „eigener Computer" kann nur die beiden Ausprägungen „ja" und „nein" aufweisen.
- Die „häufigste im Haushalt gesprochene Sprache" kann jeder existierenden Sprache entsprechen.
- Das Merkmal „Anzahl der Personen im Haushalt" kann jeder natürlichen Zahl entsprechen, wobei hier im Gegensatz zur Anzahl an Geschwistern 0 nicht möglich ist, weil man zumindest selbst im Haushalt lebt.

Bei ankreuzbaren Optionen ist es schwierig, alle möglichen Ausprägungen zu berücksichtigen. Beim Geschlecht würden sich neben „männlich" und „weiblich" die Optionen „anderes Geschlecht" und „keine Angabe" anbieten. Die Merkmale „Anzahl der Geschwister" und „Anzahl der Personen im Haushalt" sind theoretisch unbeschränkt. Man könnte jedoch als letzte Option beispielsweise „fünf oder mehr Geschwister" verwenden (wodurch jedoch Information verloren geht). Die Dauer des Schulwegs kann nur in Form von Intervallen erhoben werden, wie beispielsweise „zwischen 5 und 10 Minuten". Und bei „häufigste im Haushalt gesprochene Sprache" können kaum alle Merkmalsausprägungen abgedeckt werden.

Die Verwendung offener Felder ist jedoch auch problematisch. Einerseits kann dadurch die Datenerhebung leichter manipuliert werden, indem man beispielsweise 1000 Geschwister angibt, was die im Anschluss ermittelten statistischen Kennzahlen massiv verzerrt. Man könnte solche offensichtlich falschen Werte zwar ausfiltern, jedoch ist hier wieder zu klären, ab wann ein Wert offensichtlich falsch ist. Auch die Möglichkeiten von unklaren Antworten (beispielsweise beim Merkmal „eigener Computer") oder ungewünschten Mehrfachnennungen (beispielsweise beim Merkmal „häufigste im Haushalt gesprochene Sprache") sind bei der Erstellung des Fragebogens zu beachten. □

1.1 Skalenniveaus

Grundsätzlich lassen sich Merkmale den nachfolgenden vier Skalenniveaus zuordnen. Je nachdem, welchem Skalenniveau ein Merkmal angehört, lassen sich damit unterschiedliche mathematische Operationen auf sinnvolle Weise durchführen.

Nominalskala: Bei nominalskalierten Merkmalen kann nur die Anzahl der jeweiligen Merkmalsausprägung ermittelt werden. Beispiele dafür sind Geschlecht, Augenfarbe und Automarke.

Ordinalskala: Ordinalskalierte Merkmale können nicht nur gezählt, sondern auch geordnet werden. Die Abstände zwischen jeweils zwei Merkmalsausprägungen sind jedoch nicht vergleichbar, weshalb eine Abstandsmessung nicht sinnvoll ist. Zum Beispiel gehören Schulnoten oder Platzierungen bei Wettbewerben zu dieser Kategorie. Bei Schulnoten ist der Abstand zwischen „Nicht genügend" und „Genügend" nicht vergleichbar mit dem Abstand zwischen „Befriedigend" und „Gut". Bei Wettbewerben ist eine reine Angabe des Positionsunterschieds ebenfalls nicht aussagekräftig, da Teilnehmer mit gleichem Positionsunterschied sehr unterschiedliche Leistungen oder auch fast identische Leistungen haben könnten.

Intervallskala: Bei einer Intervallskala können auch Abstände zwischen den einzelnen Merkmalsausprägungen berechnet werden. Die Skala hat jedoch keinen natürlichen Nullpunkt, weshalb Verhältnisse wenig Sinn ergeben. Vertreter dieses Skalenniveaus sind die Temperaturskala in Grad Celsius oder die christliche Jahreszählung. Beispielsweise macht es wenig Sinn, zu sagen, dass +20 °C minus viermal so heiß ist wie –5 °C.

Verhältnisskala: Eine Verhältnisskala hat einen natürlichen Nullpunkt, der nicht unterschritten werden kann und bietet damit auch die Möglichkeit zur Berechnung von Verhältnissen. Beispiele hierfür sind Masse, Länge, Temperatur (in Kelvin) oder Punkte bei einer Schularbeit.

Bei vielen Merkmalen ist eine eindeutige Zuordnung zu diesen vier Skalenniveaus schwierig. Oftmals vertreten unterschiedliche Quellen andere Standpunkte.

1.2 Diskrete und stetige Merkmale

Neben den vier im vorherigen Abschnitt beschrieben Skalenniveaus gibt es auch die Unterscheidung zwischen diskreten oder stetigen Merkmalen.

Diskrete Merkmale: Hierbei handelt es sich um Merkmale, bei denen die Anzahl an Merkmalsausprägungen abzählbar ist. Es kann sich somit entweder um eine endliche Menge oder um eine (theoretisch) unendlich große Menge, deren Elemente jedoch durchnummeriert werden können, handeln. Beispiele dafür sind das Geburtsjahr oder die Anzahl an Geschwistern.

Stetige Merkmale: Bei dieser Art von Merkmalen ist die Menge der Merkmalsausprägungen überabzählbar. Häufig handelt es sich dabei um physikalische Größen, welche innerhalb eines bestimmten Intervalls jeder beliebigen reellen Zahl entsprechen können. Beispiele dafür wären die Körpergröße, die Geschwindigkeit eines Fahrzeugs oder die Masse eines Apfels.

Zwei Fehler werden in diesem Zusammenhang häufig gemacht:

1. Es ist auch möglich, dass Dezimalzahlen Ausprägungen eines diskreten Merkmals darstellen. Beispielsweise handelt es sich bei der Menge aller Geldbeträge, welche mittels Bargeld gebildet werden können, um ein diskretes Merkmal. Beginnend bei 0 €, 0.01 €, 0.02 € usw. können alle möglichen Ausprägungen nummeriert werden, da es dazwischen keine weiteren Werte gibt.
2. Nur weil die Skala eines Messgerätes diskret ist, kann die gemessene Größe dennoch stetig sein. Die Temperatur ist beispielsweise eine stetige Größe, auch wenn sie mit einem Thermometer gemessen wird, dessen Display nur eine Nachkommastelle anzeigt. Ist der angezeigte Wert beispielsweise 21,8 °C, so kann die Temperatur im Intervall $[21.75, 21.85)$ dennoch jeden beliebigen Wert haben. Man kann sie nur nicht genauer messen.

2 Häufigkeiten

Das Ermitteln von Häufigkeiten ist nur bei diskreten Merkmalen sinnvoll. Es wird dabei jeder möglichen Merkmalsausprägung eine Zahl zugeordnet. Man unterscheidet zwischen absoluter Häufigkeit und relativer Häufigkeit.

Absolute Häufigkeit: Die absolute Häufigkeit einer Merkmalsausprägung entspricht einer natürlichen Zahl, die angibt, wie oft diese Merkmalsausprägung in der Stichprobe bzw. in der Grundgesamtheit vorkommt.

Relative Häufigkeit: Die relative Häufigkeit gibt den Anteil dieser Merkmalsausprägung an der Größer der Stichprobe bzw. der Grundgesamtheit an. Es handelt sich dabei also um eine rationale Zahl zwischen 0 und 1, welche man auch in Prozent angeben kann. Bei einer Angabe in Prozent wird auch gelegentlich von der prozentualen bzw. prozentuellen Häufigkeit gesprochen.

Zusätzlich kann die sogenannte Häufigkeitssumme berechnet werden. Dazu werden die relativen Häufigkeiten aller bereits aufgeschriebenen Merkmalsausprägungen addiert und das Ergebnis meistens in Prozent angegeben. Dieser Wert ist nur sinnvoll, wenn die Merkmalsausprägungen zuvor sortiert wurden, was bei nominalskalierten Merkmalen nicht möglich ist.

Beispiel: *Bei einer Mathematikschularbeit, an der 18 Schüler teilnahmen, gab es folgende Noten: 4, 4, 4, 3, 5, 3, 3, 5, 5, 4, 3, 4, 3, 2, 4, 1, 3, 5. Das betrachtete Merkmal ist hier die Note und die Merkmalsausprägungen sind die Werte 1 bis 5. Es sollen die absoluten und relativen Häufigkeiten, sowie die Häufigkeitssumme (beginnend bei der Note 1) in Form einer Tabelle dargestellt werden.*

Im ersten Schritt wird gezählt, wie oft jede Merkmalsausprägung vorkommt. Dividiert man diese absoluten Häufigkeiten durch die Gesamtanzahl von 18 Noten, so erhält man die relativen Häufigkeiten.

Note	absolute Häufigkeit	relative Häufigkeit	Häufigkeitssumme
1	1	$\frac{1}{18} = 0{,}0\dot{5} = 5{,}\dot{5}\,\%$	$\frac{1}{18} = 0{,}0\dot{5} = 5{,}\dot{5}\,\%$
2	1	$\frac{1}{18} = 0{,}0\dot{5} = 5{,}\dot{5}\,\%$	$\frac{2}{18} = 0{,}\dot{1} = 11{,}\dot{1}\,\%$
3	6	$\frac{6}{18} = 0{,}\dot{3} = 33{,}\dot{3}\,\%$	$\frac{8}{18} = 0{,}\dot{4} = 44{,}\dot{4}\,\%$
4	6	$\frac{6}{18} = 0{,}\dot{3} = 33{,}\dot{3}\,\%$	$\frac{14}{18} = 0{,}\dot{7} = 77{,}\dot{7}\,\%$
5	4	$\frac{4}{18} = 0{,}\dot{2} = 22{,}\dot{2}\,\%$	$\frac{18}{18} = 1{,}0 = 100\,\%$

□

3 Lagemaße

Lagemaße sind statistische Kennzahlen, welche es ermöglichen, den Wert eines Merkmals zu beschreiben. Viele Lagemaße geben beispielsweise an, wo sich die Mitte eines Datensatzes befindet. Dabei gibt es verschiedene Methoden, um den Begriff der „Mitte“ zu definieren. Andere Lagemaße geben an, wo sich der kleinste bzw. der größte Wert befindet.

3.1 Minimum und Maximum

Das Minimum x_{min} entspricht dem kleinsten Wert, der in einem Datensatz vorkommt. Das Maximum x_{max} entspricht analog dazu dem größten Wert eines Datensatzes.

3.2 Modalwert

Der Modalwert bzw. Modus ist der am häufigsten vorkommende Wert eines Datensatzes. Kommen mehrere Werte gleich häufig vor, so gibt es mehrere Modalwerte. Sinnvoll ist diese statistische Kenn-

zahl nur bei diskreten Merkmalen. Meistens wird es bei nominalskalierten Merkmalen verwendet, da hier keine anderen Lagemaße existieren.

3.3 Median

Der Median $\tilde{x}$ teilt den geordneten Datensatz in zwei Hälften. Bei der Ermittlung des Medians müssen zwei Fälle unterschieden werden:

- Wenn der Datensatz eine ungerade Anzahl an Elementen enthält, so existiert ein mittleres Element, dessen Wert dem Median entspricht.
- Enthält der Datensatz eine gerade Anzahl an Elementen, so gibt es kein mittleres Element. Es werden daher üblicherweise die beiden mittleren Elemente addiert und durch zwei dividiert. Grundsätzlich könnte man aber jeden beliebigen anderen Wert dazwischen als Median verwenden.

Beispiel: *Acht Schüler geben bei der Dauer des Schulweges folgende Werte an (jeweils in Minuten): 5, 8, 13, 11, 4, 20, 5, 35. Es soll der Median dieses Datensatzes bestimmt werden.*

Dazu wird zuerst eine geordnete Liste erstellt: 4, 5, 5, 8, 11, 13, 20, 35. Es spielt dabei keine Rolle, ob auf- oder absteigend sortiert wird. Üblicherweise erfolgt die Sortierung jedoch aufsteigend.

Bei insgesamt acht Werten sind die beiden mittleren Werte der vierte und der fünfte Wert, also 8 und 11. Von diesen beiden Werten wird nun folgendermaßen die Mitte berechnet:

$$\frac{8+11}{2} = 9{,}5$$

Der Median des Datensatzes lautet somit 9,5 Minuten. □

3.4 Quartile

Die Quartile q_1, q_2 und q_3 teilen den Datensatz in vier Viertel. Das zweite Quartil q_2 ist identisch mit dem im vorherigen Abschnitt beschriebenen Median $\tilde{x}$. Die verbalen Definitionen von q_1 und q_3 lauten gemäß BHS-Formelsammlung (Stand: 9. September 2019) folgendermaßen:

- Mindestens 25 % der Werte sind kleiner oder gleich q_1, zugleich sind mindestens 75 % der Werte größer oder gleich q_1.
- Mindestens 75 % der Werte sind kleiner oder gleich q_3, zugleich sind mindestens 25 % der Werte größer oder gleich q_3.

Es existieren für q_1 und q_3 jedoch zahlreiche verschiedene Definitionen, welche zu unterschiedlichen Ergebnissen führen können. In diesem Skriptum wird eine relativ anschauliche und praktikable Definition verwendet, deren Ergebnisse mit den Ergebnissen des Computerprogramms GeoGebra übereinstimmen. Wie bereits beim Median ist das entscheidende Kriterium, ob der Datensatz eine gerade bzw. eine ungerade Anzahl an Elementen enthält:

- Bei einer ungeraden Anzahl werden die Bereiche links und rechts vom mittleren Wert betrachtet, wobei der mittlere Wert (der Median) selbst überhaupt nicht verwendet wird.

  Von diesen beiden Teilmengen wird nun mit der oben beschriebenen Methode der Median bestimmt, welcher dem ersten Quartil q_1 bzw. dem dritten Quartil q_3 entspricht.

– Bei einer geraden Anzahl an Werten fällt es leicht, sie in zwei Hälften zu teilen:

Von diesen beiden Teilmengen wird nun der Median bestimmt, welcher dem ersten Quartil q_1 bzw. dem dritten Quartil q_3 entspricht.

In MICROSOFT EXCEL stehen die beiden Funktionen QUARTILE.EXKL und QUARTILE.INKL zur Verfügung. Diese verwenden jedoch beide eine andere Definition als GEOGEBRA.

Achtung: Die in diesem Skriptum verwendete Definition stimmt nicht immer mit der verbalen Definition der BHS-Formelsammlung überein (siehe folgendes Beispiel).

Beispiel: *Es soll der Median und die Quartile q_1 und q_3 der Zahlen 1, 2, 3, 4, 5, 6, 7, 8, 9 bestimmt werden.*

Der Median einer ungeraden Anzahl entspricht dem mittleren Wert. Somit erhält man $\tilde{x} = 5$.

Das Quartil q_1 entspricht gemäß der oben beschriebenen Definition dem Median der Liste 1, 2, 3, 4. Daher ergibt sich $q_1 = 2{,}5$. Analog erhält man $q_3 = 7{,}5$.

Zu beachten ist, dass dies nicht mit der Definition der BHS-Formelsammlung verträglich ist, denn diese verlangt, dass mindestens 25 % der Werte kleiner oder gleich q_1 sind. Hier sind aber nur $\frac{2}{9} = 22{,}\dot{2}\,\%$ der Werte kleiner oder gleich. Um die Bedingungen der BHS-Formelsammlung zu erfüllen, müsste man stattdessen $q_1 = 3$ und $q_3 = 7$ verwenden. □

Beispiel: *Das Alter der elf Spieler einer Fußballmannschaft (in Jahren) ist durch folgenden Datensatz gegeben: 23, 25, 19, 31, 35, 24, 21, 28, 36, 19, 25. Es sollen die drei Quartile bestimmt werden.*

Dazu wird zuerst eine geordnete Liste erstellt:

$$19,\ 19,\ 21,\ 23,\ 24,\ 25,\ 25,\ 28,\ 31,\ 35,\ 36$$

Da es sich bei elf Werten um eine ungerade Anzahl handelt, ist der Median der mittlere aller Werte (was hier dem sechsten Wert entspricht). Man erhält $\tilde{x} = q_2 = 25$ Jahre.

Bei einer ungeraden Anzahl wird für die Ermittlung von q_1 und q_3 der Median weggelassen und die Bereiche links und rechts davon untersucht. Man erhält die Teildatensätze 19, 19, 21, 23, 24 und 25, 28, 31, 35, 36. Von diesen jeweils fünf Werten wird abschließend ebenfalls der Median bestimmt. Somit ergibt sich $q_1 = 21$ Jahre und $q_3 = 31$ Jahre. □

3.5 Arithmetischer Mittelwert

Der arithmetische Mittelwert $\bar{x}_{\text{ari}}$ der Daten $x_1, x_2, ..., x_n$ ist folgendermaßen definiert:

$$\bar{x}_{\text{ari}} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Es werden also alle Werte addiert und das Ergebnis anschließend durch die Anzahl geteilt. Man kann dabei gleiche Werte zusammenfassen, wie das folgende Beispiel zeigen wird.

Beispiel: *In einer Schulklasse mit 21 Schülern haben sechs Schüler keine Geschwister, zehn Schüler haben einen Bruder bzw. eine Schwester, drei Schüler haben zwei Geschwister und jeweils ein Schüler hat drei bzw. vier Geschwister. Es soll der arithmetische Mittelwert der Anzahl an Geschwistern berechnet werden.*

Bei der Ermittlung der Summe werden die einzelnen Merkmalsausprägungen mit deren absoluter Häufigkeit multipliziert:

$$\bar{x}_{\text{ari}} = \frac{6 \cdot 0 + 10 \cdot 1 + 3 \cdot 2 + 1 \cdot 3 + 1 \cdot 4}{21} = \frac{23}{21} \approx 1{,}095$$

Durchschnittlich hat jeder Schüler 1,095 Geschwister. □

Ebenso kann der arithmetische Mittelwert berechnet werden, wenn die relativen Häufigkeiten bekannt sind. Hier muss nicht durch die Anzahl dividiert werden, da die Summe aller relativen Häufigkeiten 1 ergibt.

Beispiel: *Folgende Tabelle zeigt, wie viele Fernseher pro Haushalt zur Verfügung stehen:*

Anzahl der Fernseher	*0*	*1*	*2*	*3*	*4*	*5*
Anteil der Haushalte	*2,9 %*	*51,3 %*	*31,5 %*	*9,8 %*	*3,4 %*	*1,1 %*

Es soll der arithmetische Mittelwert der Fernseher pro Haushalt berechnet werden.

Dazu wird jeweils die Merkmalsausprägung mit der zugehörigen relativen Häufigkeit multipliziert. Eine Division durch die Anzahl ist hier nicht nötig, da die Summe aller relativen Häufigkeiten 100 % und somit 1 ergibt.

$$\bar{x}_{\text{ari}} = 0{,}029 \cdot 0 + 0{,}513 \cdot 1 + 0{,}315 \cdot 2 + 0{,}098 \cdot 3 + 0{,}034 \cdot 4 + 0{,}011 \cdot 5 = 1{,}628$$

Durchschnittlich gibt es in jedem Haushalt 1,628 Fernseher. □

Oft kommt es vor, dass Werte nachträglich zu einem bereits berechneten Mittelwert hinzugefügt werden sollen. Dazu muss der bereits vorhandene Mittelwert entsprechend gewichtet werden. Dies ist nur möglich, wenn bekannt ist, aus wie vielen Werten er gebildet wurde. Deshalb sollte man in Protokollen zusätzlich zu statistischen Kennzahlen auch die Stichprobengröße angeben.

Beispiel: *Der arithmetische Mittelwert der Körpergröße von 17 Schülern beträgt 1,64 m. Zwei Schüler mit den Körpergrößen 1,82 m und 1,75 m waren bei der Messung krank und sollen nachträglich in diese Statistik eingearbeitet werden.*

Der neue arithmetische Mittelwert wird folgendermaßen berechnet:

$$\bar{x}_{\text{ari}} = \frac{17 \cdot 1{,}64 + 1{,}82 + 1{,}75}{19} \approx 1{,}66$$

Der neue arithmetische Mittelwert beträgt daher ca. 1,66 m. □

3.6 Geometrischer Mittelwert

Der geometrische Mittelwert $\bar{x}_{\text{geo}}$ der Daten $x_1, x_2, ..., x_n$ ist folgendermaßen definiert:

$$\bar{x}_{\text{geo}} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Da Wurzeln im Allgemeinen für negative Werte kein reelles Ergebnis liefern, ist die Anwendung nur bei Merkmalen mit Verhältnisskala sinnvoll. Verwendet wird dieser Mittelwert beispielsweise bei der Berechnung der Durchschnittszinsen. Wie schon bei der Berechnung des arithmetischen Mittelwerts können auch beim geometrischen Mittelwert mehrfach vorkommende Werte zusammengefasst werden.

Beispiel: *Ein Konto wird drei Jahre lang mit 2,4 % verzinst, anschließend fünf Jahre lang mit 1,7 % und zuletzt ein Jahr lang mit 3,5 %. Es soll der durchschnittliche Zinssatz berechnet werden.*

Der durchschnittliche Zinssatz entspricht dem geometrischen Mittelwert der Zinssätze der einzelnen Jahre:

$$\bar{x}_{\text{geo}} = \sqrt[9]{1{,}024^3 \cdot 1{,}017^5 \cdot 1{,}035} \approx 1{,}0213$$

Das Ergebnis lautet somit ca. 2,13 %. □

Auch hier kann des vorkommen, dass nachträglich Werte zu einem geometrischen Mittelwert hinzugefügt werden sollen. Die Vorgehensweise wird durch folgendes Beispiel veranschaulicht.

Beispiel: *Der geometrische Mittelwert von zwölf Zahlen beträgt ca. 15,3. Es sollen nachträglich die Zahlen 13, 17 und 22 hinzugefügt werden.*

Die Berechnung des neuen geometrischen Mittelwerts erfolgt folgendermaßen:

$$\bar{x}_{\text{geo}} = \sqrt[15]{15{,}3^{12} \cdot 13 \cdot 17 \cdot 22} \approx 15{,}6$$

Weil die alte Stichprobe 12 Werte umfasste, musste der alte Mittelwert mit 12 potenziert werden. Da die neue Stichprobe nun 15 Werte enthält, wurde die entsprechende Wurzel gezogen. □

3.7 Harmonischer Mittelwert

Der Abschnitt „Harmonischer Mittelwert" ist für die sRDP nicht relevant.

Der harmonische Mittelwert $\bar{x}_{\text{har}}$ der Daten $x_1, x_2, x_3, ..., x_n$ wird durch folgende Formel berechnet:

$$\bar{x}_{\text{har}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + ... + \frac{1}{x_n}}$$

Grundsätzlich können nur Werte ungleich Null in die Formel eingesetzt werden. Sobald ein Wert des Datensatzes gegen Null geht, ist auch der Grenzwert des harmonischen Mittelwertes gleich Null. Somit ist es sinnvoll, $\bar{x}_{\text{har}} = 0$ festzulegen, falls ein Wert des Datensatzes gleich 0 ist.

3.8 Mittelwertungleichung

Für einen Datensatz, dessen Werte allesamt nicht negativ sind, ist stets folgende Ungleichung erfüllt:

$$x_{\text{min}} \leq \bar{x}_{\text{har}} \leq \bar{x}_{\text{geo}} \leq \bar{x}_{\text{ari}} \leq x_{\text{max}}$$

4 Streumaße

Durch die verschiedenen Lagemaße alleine werden Datensätze oftmals nicht ausreichend beschrieben. Man kennt dann zwar dessen ungefähre Größe, weiß jedoch nicht, wie aussagekräftig dieser Wert ist. Die folgenden beiden Datensätze haben beide den arithmetischen Mittelwert 10:

1. Datensatz: 15, 18, 7, 4, 6
2. Datensatz: 9.9, 10.2, 9.5, 9.7, 10.6, 10.1

Beim ersten Datensatz ist die Streuung der Werte viel größer als beim zweiten Datensatz. Es wäre daher sinnvoll, auch dieses Merkmal durch geeignete Kennzahlen zu beschreiben.

Da in diesem Kapitel nur der arithmetische Mittelwert vorkommt und daher keine Verwechslungsgefahr mit dem geometrischen Mittelwert und dem harmonischen Mittelwert besteht, wird nur das Symbol $\bar{x}$ anstelle von $\bar{x}_{\text{ari}}$ verwendet.

4.1 Spannweite

Die Spannweite (engl. *range*) entspricht der Differenz zwischen größtem und kleinstem Wert eines Datensatzes. Man verwendet dafür gemäß der englischen Bezeichnung das Symbol R.

$$R = x_{\text{max}} - x_{\text{min}}$$

Der Nachteil der Spannweite ist, dass sie gegenüber Ausreißern sehr empfindlich ist, da bereits ein sehr großer Wert genügt, um diese Kennzahl zu verändern.

4.2 Interquartilsabstand

Etwas unempfindlicher ist der Interquartilsabstand (engl. *interquartile range*):

$$\mathrm{IQR} = q_3 - q_1$$

4.3 Mittlere absolute Abweichung

Die mittlere absolute Abweichung e der Daten $x_1, x_2, ..., x_n$ ist folgendermaßen definiert:

$$\mathrm{e} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \bar{x}|$$

4.4 Varianz (mittlere quadratische Abweichung)

Weitaus wichtiger als die mittlere absolute Abweichung ist die Varianz s^2. Es gibt hierfür zwei verschiedene Definitionen:

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{und} \qquad s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Die linke Version wird meistens verwendet, wenn es sich beim vorliegenden Datensatz um die Grundgesamtheit handelt. Man nennt sie auch die Varianz der Grundgesamtheit. Die rechte Version (die aufgrund des kleineren Nenners ein größeres Ergebnis liefert) kommt meistens bei Stichproben zum Einsatz und wird daher auch Stichprobenvarianz genannt.

Das Berechnen der in beiden Varianten vorkommenden Summe ist sehr mühsam. Eine wesentliche Vereinfachung stellt der sogenannte Verschiebungssatz dar. Dieser besagt, dass für beliebige Datensätze folgende Gleichung gilt:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2 \right) - n\bar{x}^2$$

Anstelle der ursprünglichen Summe kann also auch die Summe der Quadrate berechnet werden (was wesentlich kürzer ist). Von dieser Summe muss anschließend nur noch der Term $n\bar{x}^2$ subtrahiert werden. Die Klammer auf der rechten Seite ist nicht notwendig, verdeutlicht jedoch, dass die Subtraktion nicht innerhalb der Summe steht.

Herleitung: Ausgangsbasis bildet folgende Summe:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Im ersten Schritt wird die binomische Formel angewendet. Anschließend werden die drei Summanden zu einzelnen Summen zerlegt:

$$\sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2 \right) = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2$$

Als nächstes werden bei der zweiten Summe alle konstanten und somit von der Indexvariable i unabhängigen Faktoren vor die Summe geschrieben. Bei der dritten Summe wird die Konstante $\bar{x}$ insgesamt n-mal summiert.

Somit kann man stattdessen $n \cdot \bar{x}$ schreiben. Man erhält Folgendes:

$$\sum_{i=1}^{n} x_i^2 - 2\bar{x} \cdot \sum_{i=1}^{n} x_i + n \cdot \bar{x}^2$$

Als Zwischenschritt wird nun die Definition des arithmetischen Mittelwertes umgeformt:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \quad \Rightarrow \quad n \cdot \bar{x} = \sum_{i=1}^{n} x_i$$

Dieses Resultat wird nun statt der zweiten Summe eingesetzt, wodurch man folgendes Ergebnis erhält:

$$\sum_{i=1}^{n} x_i^2 - 2\bar{x} \cdot n \cdot \bar{x} + n \cdot \bar{x}^2$$

Fasst man nun die Terme hinter der Summe soweit wie möglich zusammen, so erhält man das zu beweisende Resultat:

$$\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2$$

■

4.5 Standardabweichung

Ein Problem der Varianz ist ihre Einheit. Da die Daten quadriert werden, hat die Varianz nicht dieselbe Einheit wie die Messwerte. Sind die Messwerte beispielsweise in der Einheit kg gegeben, so hat die Varianz die Einheit kg^2. Es wird daher häufig die Wurzel der Varianz berechnet. Man bezeichnet diese als Standardabweichung. Auch hier gibt es zwei Varianten:

$$s = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \text{und} \qquad s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Die linke Variante wird meist als Standardabweichung der Grundgesamtheit und die rechte Variante als Stichprobenstandardabweichung bezeichnet.

Beispiel: *Es soll für den Datensatz 45, 33, 29, 71, 56, 34, 47 die Standardabweichung der Grundgesamtheit berechnet werden.*

Dazu wird der oben genannte Verschiebungssatz verwendet. Im ersten Schritt wird der arithmetische Mittelwert ermittelt. Dieser lautet $\bar{x} = 45$. Als nächstes wird die Summe der Quadrate aller Werte gebildet:

$$\sum_{i=1}^{7} x_i^2 = 45^2 + 33^2 + 29^2 + 71^2 + 56^2 + 34^2 + 47^2 = 15\,497$$

Abschließend wird in die entsprechende Formel eingesetzt:

$$s = \sqrt{\frac{15\,497 - 7 \cdot 45^2}{7}} \approx 13{,}74$$

Die Standardabweichung der Grundgesamtheit beträgt somit ca. 13,74. □

5 Grafische Darstellung von Daten

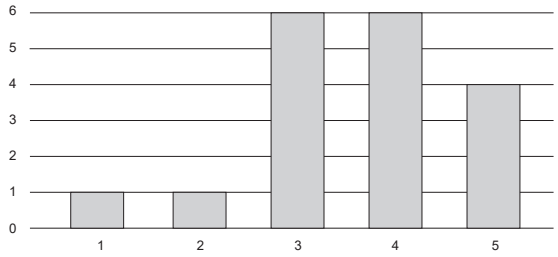
Zur graphischen Darstellung von Daten steht ein großes Spektrum an Diagrammen zur Verfügung. Die am weitesten verbreiteten Diagrammtypen werden in den folgenden Abschnitten näher erläutert.

5.1 Säulendiagramm

Das Säulendiagramm ist besonders gut geeignet, um die absoluten oder relativen Häufigkeiten der Merkmalsausprägungen eines diskreten Merkmals zu veranschaulichen. Dabei sollte die Anzahl der Merkmalsausprägungen zur besseren Übersicht einen gewissen Wert nicht überschreiten (höchstens ca. 10 bis 15).

Auf der horizontalen Achse werden die einzelnen Merkmalsausprägungen aufgelistet (üblicherweise aufsteigend sortiert) und auf der vertikalen Achse wird die absolute oder relative Häufigkeit mit einer geeigneten Skalierung aufgetragen. Die Breite der Säulen spielt keine Rolle, jedoch sollten sie sich nicht berühren, da das Säulendiagramm sonst mit einem Histogramm verwechselt werden könnte (siehe Seite 11).

Beispiel: *Das Ergebnis der Mathematikschularbeit aus dem Beispiel von Seite 3 soll durch ein Säulendiagramm dargestellt werden.*



Falls die Häufigkeiten der einzelnen Säulen nicht so klar ablesbar wären, könnte man diese auch oberhalb der Säulen dazuschreiben. □

Es gibt verschiedene Spezialformen bzw. Abwandlungen des Säulendiagramms:

- Balkendiagramm: Hier wird das Säulendiagramm gekippt, d. h. die absoluten bzw. relativen Häufigkeiten werden nun horizontal aufgetragen und die Merkmalsausprägungen vertikal. Dies ist empfehlenswert, wenn die Beschriftung der Merkmalsausprägungen viel Platz in Anspruch nimmt.
- Stabdiagramm: Säulendiagramme mit sehr schmalen Säulen werden häufig auch als Stabdiagramme bezeichnet.
- Pareto-Diagramm: Bei einem Pareto-Diagramm werden die einzelnen Werte absteigend nach ihrer Häufigkeit geordnet.

5.2 Kreisdiagramm

Relative Häufigkeiten lassen sich sehr gut durch ein Kreisdiagramm darstellen. Der volle Kreis entspricht einem Anteil von 100 %. Die Winkel α_i der Kreissektoren für die einzelnen Ausprägungen können durch die Formel

$$\alpha_i = r_i \cdot 360\,^\circ$$

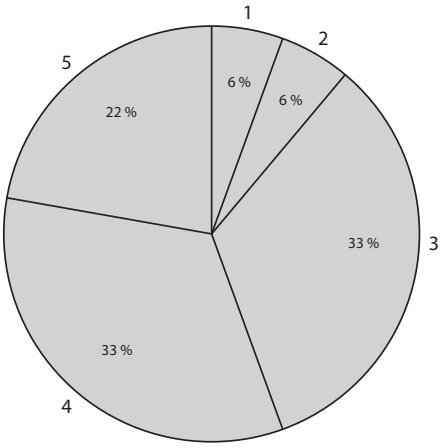
berechnet werden, wobei $r_i \in (0, 1)$ die relativen Häufigkeiten sind. Sinnvoll ist es, wenn die Sektoren nach ihrer Größe sortiert werden. Häufig wird dazu bei der oberen senkrechten Linie begonnen und im Uhrzeigersinn fortgesetzt.

Beispiel: *Das Ergebnis der Mathematikschularbeit aus dem Beispiel von Seite 3 soll durch ein Säulendiagramm dargestellt werden.*

Zunächst muss berechnet werden, welchen Winkel die einzelnen Sektoren haben. Dazu wird die entsprechende relative Häufigkeit mit 360° multipliziert. Man erhält folgende Werte:

Note	1	2	3	4	5
relative Häufigkeit	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{6}{18}$	$\frac{6}{18}$	$\frac{4}{18}$
Winkel	20°	20°	120°	120°	80°

Anschließend werden die Kreissektoren beginnend beim höchsten Punkt des Kreises im Uhrzeigersinn gezeichnet. Das Resultat sieht folgendermaßen aus:



Häufig werden außerhalb der einzelnen Sektoren die zugehörigen Merkmalsausprägungen und innerhalb die relativen Häufigkeiten angegeben. □

Falls Sektoren zu klein sind und daher nicht erkennbar wären, können sie unter „Sonstige" zusammengefasst werden. Gelegentlich wird für alle sonstigen Werte neben dem ursprünglichen Kreisdiagramm ein weiteres Kreisdiagramm erstellt.

Eine sehr nützliche Eigenschaft des Kreisdiagramms ist jene, dass Median, 1. Quartil und 3. Quartil ohne Berechnung und sogar ohne Kenntnis des Datensatzes ermittelt werden können. Die Voraussetzung dafür ist, dass es sich zumindest um ein ordinalskaliertes Merkmal (höhere Niveaus sind auch möglich) handelt und die Merkmalsausprägungen vom obersten Punkt des Kreises beginnend aufsteigend sortiert aufgetragen wurden (so wie im vorherigen Beispiel).

Ist dies der Fall, so ist nach 90° das 1. Quartil, nach 180° der Median und nach 270° das 3. Quartil ersichtlich. Es muss nur überprüft werden, in welchen Sektor diese drei Winkel fallen. Im obigen Beispiel wären $q_1 = 3,\ \tilde{x} = 4$ und $q_3 = 4$.

5.3 Histogramm

Um die Verteilung stetiger Werte graphisch darstellen zu können, werden oft Histogramme verwendet. Diese sehen ähnlich aus wie Säulendiagramme. Die wesentlichen Unterschiede sind, dass sich die Säulen hier berühren und dass deren Breite eine Rolle spielt.

Die Merkmalsausprägungen werden beim Histogramm in verschiedene Klassen zusammengefasst. Meistens verwendet man fünf bis zehn Klassen, damit einerseits eine gewisse Aussagekraft erreicht wird, aber andererseits auch noch eine ausreichende Übersichtlichkeit gegeben ist.

Die Breite der Klassen kann prinzipiell frei gewählt werden und muss nicht bei allen Klassen gleich sein. Häufig werden an Randbereichen eher breite Klassen und in der Mitte eher schmale Klassen verwendet. Die Häufigkeit wird im Gegensatz zum Säulendiagramm nicht durch die Höhe sondern durch den Flächeninhalt der Rechtecke repräsentiert, da es sonst bei unterschiedlicher Klassenbreite zu verzerrten Eindrücken kommen würde.

Bei der Erstellung eines Histogramms geht man folgendermaßen vor:

1. Die Anzahl der Klassen und deren Grenzen wird festgelegt. Es muss auch entschieden werden, ob Werte, die genau an einer Grenze liegen, zur linken oder zur rechten Klasse gehören.
2. Die Fläche der Rechtecke ist durch die absolute bzw. relative Häufigkeit vorgegeben. Die Höhe wird daher berechnet, indem man die Fläche durch die Breite dividiert.
3. Gegebenenfalls können am Ende alle Höhen mit demselben Faktor multipliziert werden.

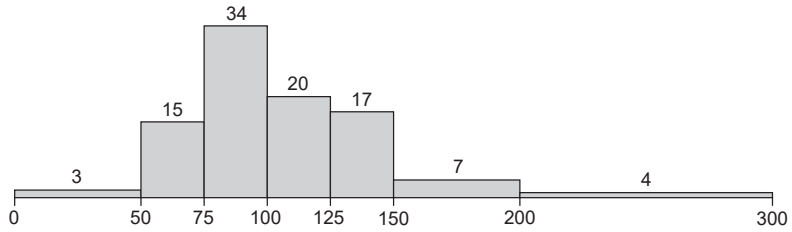
Beispiel: *Von 100 zufällig ausgewählten PKWs wurde die Leistung (in PS) ermittelt. Anschließend wurde der Datensatz in sieben Klassen unterteilt, wobei die Untergrenze jeweils noch dazugehört und die Obergrenze bereits zur nächsten Klasse gehört. Die Ergebnisse, welche in nachfolgender Tabelle dargestellt sind, sollen durch ein Histogramm veranschaulicht werden.*

Untergrenze	*0*	*50*	*75*	*100*	*125*	*150*	*200*
Obergrenze	*50*	*75*	*100*	*125*	*150*	*200*	*300*
Anzahl	*3*	*15*	*34*	*20*	*17*	*7*	*4*

Der Flächeninhalt der einzelnen Klassen entspricht der absoluten Häufigkeit. Im ersten Schritt wird für jede Klasse die Breite ermittelt. Um die Höhe der Rechtecke zu erhalten, wird anschließend der Flächeninhalt jeder Klasse durch die Klassenbreite dividiert.

Flächeninhalt	3	15	34	20	17	7	4
Breite	50	25	25	25	25	50	100
Höhe	0,06	0,60	1,36	0,80	0,68	0,14	0,04

Abschließend wird das Histogramm graphisch dargestellt. Die Höhe kann dabei beliebig skaliert werden.



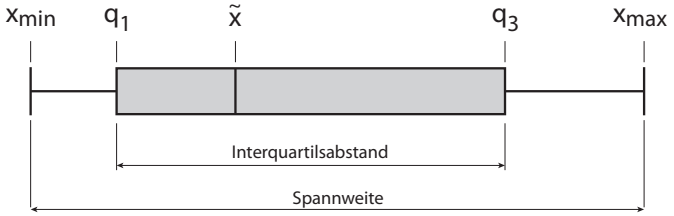
Als Beschriftung sollen auf jeden Fall die Klassengrenzen angegeben werden. Zusätzlich wurden hier die absoluten Häufigkeiten der jeweiligen Klassen angeführt. □

5.4 Boxplot

Durch einen Boxplot werden sowohl Lage- als auch Streumaße eines Datensatzes kompakt dargestellt, sodass man möglichst schnell einen Überblick erhält. Konkret kann man aus einem Boxplot die folgenden statistischen Kennzahlen ablesen: Minimum, Maximum, Median, 1. Quartil, 3. Quartil, Spannweite und Interquartilsabstand. Unterhalb des Boxplots ist meistens der entsprechende Abschnitt der Zahlengerade abgebildet.

5.4.1 Klassischer Boxplot

Unten wird zur besseren Vorstellung der nachfolgenden Beschreibungen ein Boxplot dargestellt (jedoch ohne Skala, da es sich um ein allgemeines Beispiel handelt).



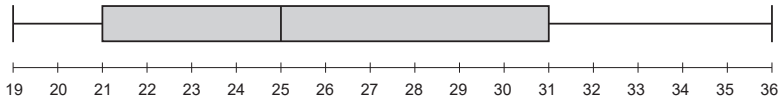
Das Rechteck im Zentrum nennt man Box (daher der Name). Es wird links und rechts begrenzt durch das 1. und das 3. Quartil. Die Länge der Box entspricht somit dem Interquartilsabstand. Die senkrechte Linie im Inneren der Box repräsentiert den Median. Die beiden Linien links und rechts der Box bezeichnet man als Antenne bzw. Whisker (dt. *Schnurrhaar*). Sie geben an, wo sich das Minimum und das Maximum des Datensatzes befinden. Die Gesamtbreite des Boxplots entspricht daher der Spannweite.

Beispiel: *Es soll ein Boxplot erstellt werden, der das Alter der Spieler der Fußballmannschaft des Beispiels von Seite 5 veranschaulicht.*

Die für diese Aufgabe benötigten statistischen Kennzahlen wurden bereits im ursprünglichen Beispiel ermittelt:

$$x_{\min} = 19, q_1 = 21, \tilde{x} = 25, q_3 = 31, x_{\max} = 36$$

Der zugehörige Boxplot sieht folgendermaßen aus:



Man kann aus dem Boxplot sofort ablesen, dass ca. ein Viertel der Mannschaft zwischen 19 und 21 Jahre alt ist. In diesem Bereich ist die Dichte der Werte ziemlich hoch. Im Vergleich dazu verteilt sich die ältere Hälfte der Mannschaft auf den Bereich von 25 bis 36 Jahre. Hier ist die Dichte der Werte eher gering. □

5.4.2 Darstellung von Ausreißern

Der Abschnitt „Darstellung von Ausreißern“ ist für die sRDP nicht relevant.

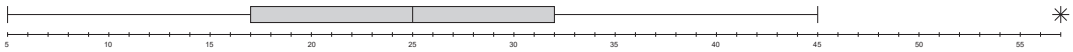
Manchmal wird auch eine andere Variante des Boxplots verwendet, welche jene Werte speziell hervorhebt, die besonders weit von der Box entfernt sind (sogenannte Ausreißer). Die Maximallänge der Antennen entspricht hierbei dem 1,5-fachen Interquartilsabstand. Das heißt, sie reichen bis zum letzten Wert, der noch innerhalb dieses Maximalabstandes liegt. Alle Werte außerhalb werden durch Sterne oder Kreise dargestellt und als Ausreißer eingestuft.

Beispiel: *Es soll der Boxplot des nachfolgenden Datensatzes erstellt werden, wobei Ausreißer speziell hervorgehoben werden sollen:*

$$5, 13, 16, 18, 23, 23, 25, 26, 29, 31, 33, 45, 57$$

Der Median beträgt 25. Für das 1. Quartil erhält man 17 und für das 3. Quartil ergibt sich der Wert 32. Der 1,5-fache Interquartilsabstand beträgt somit $15 \cdot 1{,}5 = 22{,}5$.

Auf der linken Seite sind alle Werte innerhalb dieses Bereichs, denn der Abstand zwischen Minimum und q_1 ist kleiner als 22,5. Auf der rechten Seite liegt der Wert 57 außerhalb des Maximalabstandes, denn der Abstand zwischen 57 und q_3 beträgt 25. Somit erhält man folgenden Boxplot:



Die rechte Antenne reicht bis zu 45, da dies der größte Wert ist, der innerhalb des Maximalabstandes liegt. Der Ausreißer wird durch einen Stern gekennzeichnet. □