Statistik

Definition

In der Statistik werden Daten bzw. Datensätze untersucht und ihre Eigenschaften beschrieben. Wichtige Begriffe sind hierbei das arithmetische Mittel $\bar{x}$, der Median $\tilde{x}$, der Modus, die Spannweite, die Varianz V sowie die Standardabweichung σ einer Stichprobe, die auf bestimmte Merkmale untersucht wird. Ebenso ist es wichtig zu verstehen, dass Daten auf verschiedene Weisen dargestellt werden können, etwa in einem Kreis-, Säulen-, Pikto-, Stab-, Boxplot- oder Stänglblattdiagramm.

Grundbegriffe

Grundgesamtheit … die Menge aller für eine statistische Untersuchung relevanten Objekte. Beispiel: alle Schüler und Schülerinnen einer Volksschule in Fresach.

Stichprobe … die Teilmenge der Grundgesamtheit, für die statistische Merkmale erhoben werden. Sie sollte möglichst repräsentativ für die Grundgesamtheit stehen. Beispiel: zehn zufällig gewählte Schüler der oben genannten Volksschule.

Merkmal … die statistisch relevante Variable die bei der Stichprobe untersucht wird. Beispiel: die Körpergröße der oben genannten zehn zufällig gewählten Schüler.

Es gibt verschiedene Kategorien von Merkmalen, für welche unterschiedliche Eigenschaften interessant sind:

- **Nominalskala**

  Merkmale werden entsprechend einer Kategorie geordnet: Nationalität, Geschlecht, Sprache usw.

- **Ordinalskala**

  Merkmale werden nach Größe oder Rang geordnet: Bildungsniveau, Dienstgrad, Platzierung bei einem Wettbewerb.

- **Metrische Skala**

  - **Diskrete Merkmale** … aufzählbare Variablen $\in \mathbb{N}$: Alter, Einwohnerzahl
  - **Stetige Merkmale** … durch Messung bestimmte Variablen $\in \mathbb{R}$: Länge, Gewicht

Absolute Häufigkeit … die absolute Anzahl einer statistischen Variable ohne Relation zur Stichprobe. Beispiel: 4 Schüler sind 8 Jahre alt.

Relative Häufigkeit … die Anzahl einer statistischen Variable in Relation zur Stichprobe. Beispiel: 4 von 10 Schülern sind 8 Jahre alt ($4 \div 10 = 40\%$)

Darstellungsformen

Urliste

Am Anfang jeder statistischen Untersuchung steht eine Urliste, welche die gefundenen Merkmale der Stichprobe auflistet. Beispiel: die Körpergröße (in cm) der zehn zufällig gewählten Schülerinnen und Schüler einer Volksschule in Fresach:

$$\{153, 164, 112, 160, 160, 210, 155, 153, 112, 153\}$$

Stänglblattdiagramm

Im ersten Schritt würde man diese Urliste ordnen, wozu man ein Stänglblattdiagramm als Hilfe anfertigen kann. In einem Stänglblattdiagramm werden in einer Tabelle in der ersten Spalte konstantere Stellen der statistischen Variablen angegeben und in der zweiten Spalte variierendere Stellen.

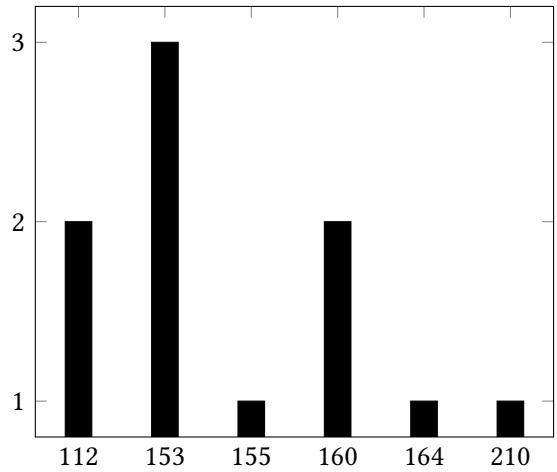
11	2, 2
15	3, 3, 3, 5
16	0, 0, 4
21	0

Aus diesem Stänglblattdiagramm kann man folglich die geordneten Merkmale auslesen und in eine neue, geordnete Urliste geben:

$$\{112, 112, 153, 153, 153, 155, 160, 160, 164, 210\}$$

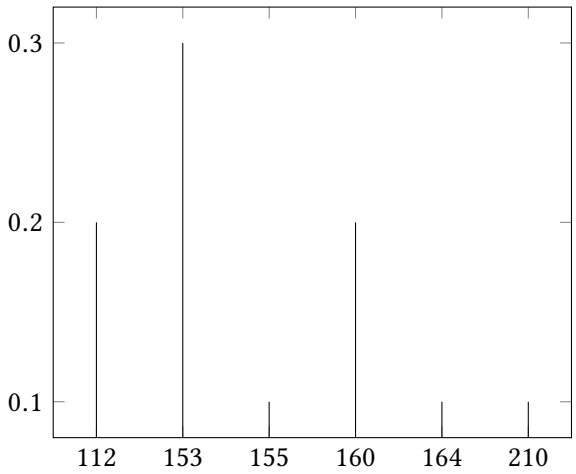
Säulendiagramm

Ein Säulendiagramm stellt die absoluten Häufigkeiten der einzelnen statistischen Variablen als Säulen auf einer Skala dar.



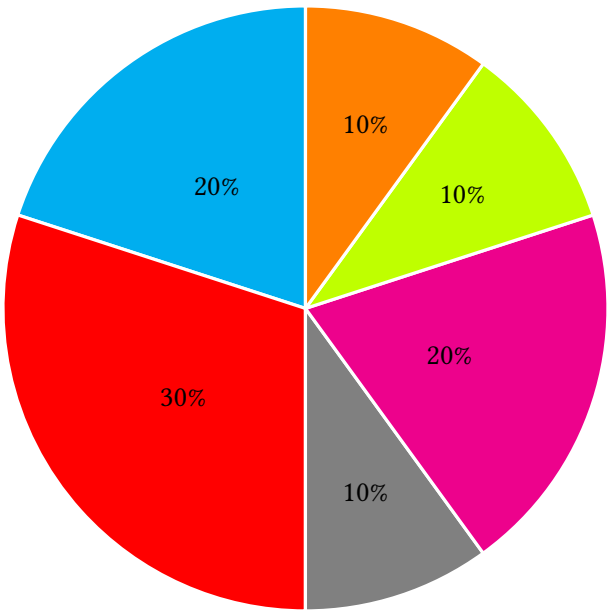
Stabdiagramm

Gegensätzlich zum Säulendiagramm bildet ein Stabdiagramm nicht die absoluten, sondern die relativen Häufigkeiten der Daten ab.



Kreisdiagramm

Ein Kreisdiagramm stellt ebenso wie ein Stabdiagramm die relativen Häufigkeiten einzelner Daten eines Datensatzen dar. Dabei steht der ganze Kreis für 100% der Häufigkeit. Die relative Fläche einzelner Kreissektoren ist dabei im Verhältnis zur Gesamtfläche des Kreises equivalent zu den relativen Häufigkeiten der Daten im Verhältnis zum Gesamtdatensatz.



Eigenschaften

Arithmetisches Mittel

Das arithmetische Mittel $\bar{x}$ eines Datensatzes, auch Mittel- oder Durchschnittswert genannt, wird berechnet, in dem man die Summe aller Daten durch ihre Anzahl dividiert:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Ebenso kann man das arithmetische Mittel als die Summe aller erhobenen Werte, multipliziert mit ihrer absoluten Häufigkeit, dividiert durch die Gesamtanzahl sehen:

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

Man kann sich die Division durch N auch sparen, in dem man mit den relativen Häufigkeiten arbeitet, da in diesen die Relativität schon inbegriffen ist:

$$\bar{x} = \sum_{i=1}^{n} h_R(i) \cdot x_i$$

Für den oben beschriebenen Datensatz der Körpergrößen der zehn Schülerinnen und Schüler einer Volksschule in Fresach wäre das arithmetische Mittel somit:

$$\bar{x} = \frac{112 + 112 + 153 + 153 + 153 + 155 + 160 + 160 + 164 + 210}{10} = 153.2$$

oder

$$\bar{x} = \frac{2 \cdot 112 + 3 \cdot 153 + 1 \cdot 155 + 2 \cdot 160 + 1 \cdot 164 + 1 \cdot 210}{10} = 153.2$$

oder

$$\bar{x} = 0.2 \cdot 112 + 0.3 \cdot 153 + 0.1 \cdot 155 + 0.2 \cdot 160 + 0.1 \cdot 164 + 0.1 \cdot 210 = 153.2$$

Median

Der Median $\tilde{x}$ eines Datensatzes ist jener Wert, der in der Mitte der geordneten Urliste steht. Bei ungerader Gesamtanzahl n ist gibt es einen definitiven Wert in der Mitte bei Position $\left\lceil \frac{n}{2} \right\rceil$. Bei gerader Anzahl muss das arithmetische Mittel zwischen den beiden mittleren Werten genommen werden ($d(x)$ sei die Datensatzfunktion):

$$\tilde{x} = \frac{d(\frac{n}{2}) + d(\frac{n}{2} + 1)}{2}$$

Der Volksschülerdatensatz hat eine gerade Gesamtanzahl von 10, der Median liegt also beim arithmetischen Mittel aus den Werten bei $10 \div 2 = 5$ und $10 \div 2 + 1 = 6$.

$$\tilde{x} = \frac{153 + 155}{2} = 154$$

Generell ist der Median dann ein sichererer, aussagekräftigerer Wert als das arithmetische Mittel, wenn es im Datensatz Ausreißer gibt, die das arithmetische Mittel stark beeinflussen könnten, den Median aber nicht.

Modus

Der Modus, oft Modalwert genannt, bzw. die Modi oder Modalwerte eines Datensatzes sind jene Werte, die am häufigsten vorkommen. Der Modus des vorhin genannten Datensatzes ist 153, weil dieser Wert mit einer absoluten Häufigkeit von 3 am öftesten gefunden wurde.

Spannweite

Die Spannweite einer Stichprobe ist die Differenz zwischen dem Maximum und Minimum der Werte. Sie beschreibt, wie sehr die Werte maximal von einander abweichen. Bei den Volkschülern beträgt die Spannweite $210 - 112 = 98$.

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

Varianz und Standardabweichung

Die Standardabweichung σ beschreibt die durschnittliche Abweichung der Daten einer Stichprobe vom arithmetischen Mittel $\bar{x}$. Sie legt fest, in welchem Intervall $[\bar{x} - \sigma; \bar{x} + \sigma]$ sich der Großteil der Daten befindet, bzw. um welchen Wert die Daten um das arithmetische Mittel streuen. Die Differenzen zwischen den einzelnen Werten und $\bar{x}$ werden quadriert, um ihr Vorzeichen aufzuheben.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}}$$

Ebenso kann man zuerst den Mittelwert aus den quadrierten Werten berechnen und dann die Differenz zwischen diesem Mittelwert und dem quadrierten arithmetischen Mittel berechnen:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2} = \sqrt{\frac{x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2}{n} - \bar{x}^2}$$

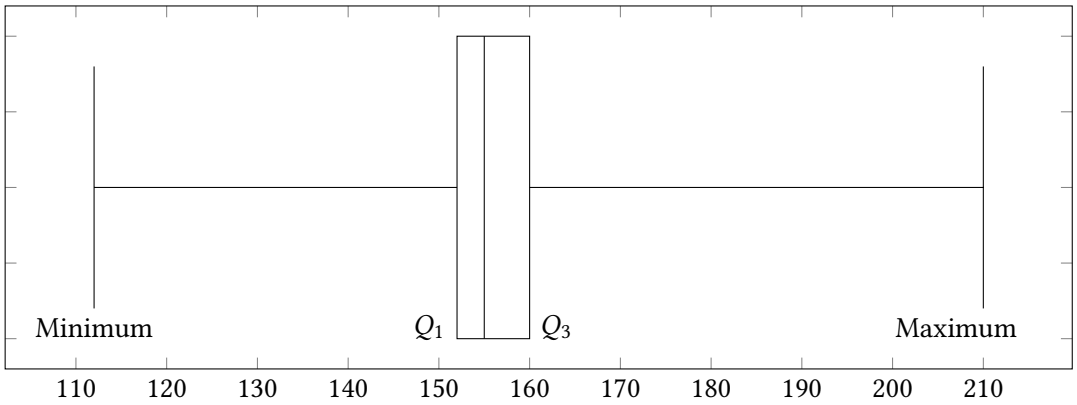
Die Varianz V bzw. σ^2 ist das Quadrat der Standardabweichung: $V = \sigma^2$

Boxplotdiagramme

Letztlich sollten noch Boxplotdiagramme untersucht werden. Sie stellen alle bis hierhin beschriebenen Begriffe in einem Diagramm bildlich dar. Man nehme wieder den Datensatz der Volksschüler aus Fresach:

$$\{112, 112, 153, 153, 153, 155, 160, 160, 164, 210\}$$

Das dazugehörige Boxplotdiagramm wäre:



Auf dieser Skala werden Maximum und Minimum als Grenzen des Diagramms links und rechts eingezeichnet. Der mittlere Strich im Kasten ist der Median, die beiden weiteren Striche links und rechts vom Median sind die Quartile Q_1 und Q_2. Also:

Minimum … untere Grenze des Boxplotdiagramms

Maximum … obere Grenze des Boxplotdiagramms

Erstes Quartil Q_1 … der Median der Datenreihe zwischen Minimum und Median $\tilde{x}$. Zwischen dem Minimum und dem ersten Quartil liegen 25% der Werte, darüber die restlichen 75%. Zwischen Q_1 und dem Median liegen auch 25%.

Zweites Quartil Q_2 … der Median $\tilde{x}$ der Datenreihe wird zwischen Q_1 und Q_3 eingezeichnet. Über und unter dem Median liegen 50 Prozent der Werte. Der Median $\tilde{x}$ teilt den Datensatz also in zwei gleich große Hälften.

Drittes Quartil Q_3 … der Median der Datenreihe zwischen Maximum und Median $\tilde{x}$. Zwischen Maximum und diesem Quartil liegen 25% der Werte, zwischen Minimum und Q_3 die unteren 75%.

Quartilsabstand … die Differenz zwischen Q_1 und Q_3. In diesem Intervall liegen die Hälfte (50%) aller Werte.