Kapitel 9: Beschreibende Statistik

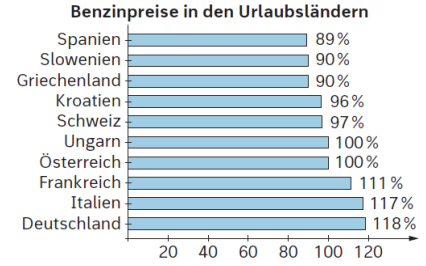
Absolute Häufigkeit
Die **absolute Häufigkeit H** gibt an, wie oft ein bestimmtes Merkmal in der Stichprobe vorkommt

Relative Häufigkeit
Die **relative Häufigkeit h** gibt als Verhältnis an, wie oft ein bestimmtes Merkmal in Bezug auf die Gesamtanzahl vorkommt.
Die relative Häufigkeit ist daher ein Bruch mit einem Wert zwischen 0 und 1.
Alternativ ist auch eine Angabe in Prozent möglich.

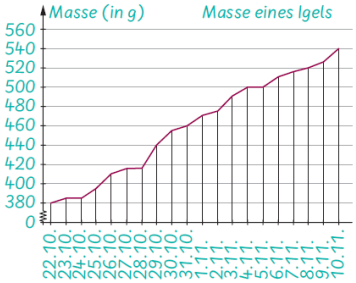
Qualitative Merkmale		Qualitative Merkmale	
Nominalskala Keine Ordnung in der Merkmalsausprägungen	**Ordinalskala** Merkmalsausprägung besitzen eine natürliche Ordnung, allerdings sind die Abstände nicht vergleichbar.	**Metrisch-diskrete Skala** Die Werte werden durch Zählen bestimmt.	**Metrisch-stetige Skala** Die Werte werden durch Messen bestimmt
Beispiel: Geschlecht, Familienstand, Augenfarbe, Eissorten,...	Beispiel: Beurteilungen (Film, Schularbeiten, ...), Energieverbrauchsklassen von Elektrogeräten,...	Beispiel: Alter, Einkommen,...	Beispiel: Körpergröße, Masse, Sprungweite, Füllmenge,...

Diagrammarten

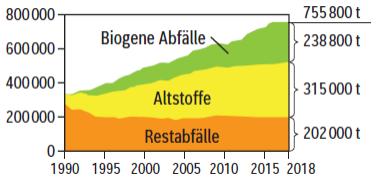
Balkendiagramm



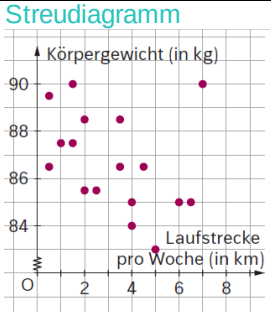
Liniendiagramm



Flächendiagramm



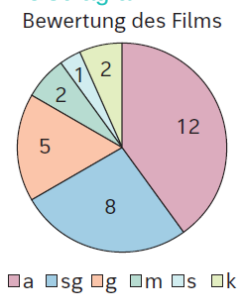
Streudiagramm



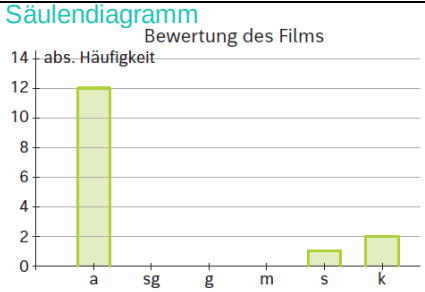
Stängel-Blatt-Diagramm

0	0 2 4 7 8
1	0 0 0 2 2 3 4 5
2	2 3 3 4 6 9
3	3
4	8

Kreisdiagramm



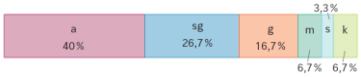
Säulendiagramm



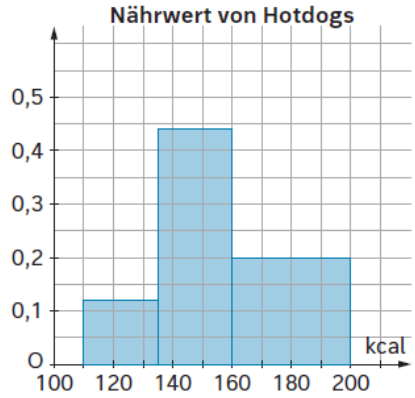
Strichliste

Merkmalsausprägung	
Ausgezeichnet (a)	

Prozentstreifen



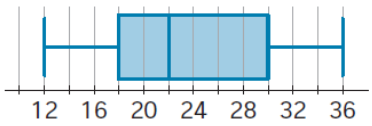
Histogramm



Ein **Histogramm** ist analog zu einem Säulendiagramm zu sehen, jedoch mit dem wesentlichen Unterschied, dass nicht die Höhe wie im Säulendiagramm die **Flächen** der Säulen die (absoluten oder relativen) **Häufigkeiten wiedergeben**.

Für die Ordinatenwerte in einem Histogramm gilt:
Ordinatenwert = $\frac{\text{Häufigkeit}}{\text{Klassenbreite}}$

Kastenschaubild=Boxplot



Zentral- und Streuungsmaße

Spannweite (Range)
Die Spannweite ist die Differenz zwischen dem größten Wert (Maximum) und dem kleinsten Wert (Minimum) der Stichprobe

$$R = x_{MAX} - x_{MIN}$$

Beispiel:
Datenliste: 1, 2, 3, 4, 4, 5, 5, 5, 5, 6, 7, 12, 15 ⇒ $R = 15 - 1 = 14$

Modus (Modalwert)
Der Modus ist der häufigste Wert, der in einer Stichprobe vorkommt.
Kommen mehrere Werte gleich häufig vor, dann ist der Modus nicht definiert.

Beispiel:
Datenliste: 1, 2, 3, 4, 4, 5, 5, 5, 5, 6, 7, 12, 15 ⇒ Modus = 5

Median (Zentralwert)
Sortiert man die Werte (Merkmalsausprägungen) der Größe nach, so ist der Median jener Wert, der in der Mitte der Liste steht.
Im Falle einer ungeraden Anzahl von Werten gibt es immer einen Median, im Fall einer geraden Anzahl von Werten wird das arithmetische Mittel der beiden mittleren Zahlen ermittelt.

Beispiel:
- Geordnete Datenliste mit ungerader Anzahl von Werten: 1, 2, 4, **6**, 7, 10, 12 ⇒ Median = 6
- Geordnete Datenliste mit gerader Anzahl von Werten: 1, 2, 4, **6**, **7**, 10, 12, 14 ⇒ Median = $\frac{6+7}{2} = 6{,}5$

Arithmetisches Mittel, Mittelwert, $\bar{x}$ oder μ
Eine Liste von Daten $x_1, x_2, x_3, \ldots, x_n$ ist gegeben.
Das arithmetische Mittel wird dann berechnet durch: $\bar{x} = \frac{x_1+x_2+x_3+\ldots+x_n}{n}$
Wenn Werte mit der absoluten Häufigkeit H_i bzw. relativen Häufigkeit h, auftreten, dann gilt:
$\bar{x} = \frac{x_1 \cdot H_1 + x_2 \cdot H_2 + \ldots + x_k \cdot H_k}{n} = x_1 \cdot h_1 + x_2 \cdot h_2 + \ldots + x_k \cdot h_k$, für $k \leq n$

Beispiel:
Datenliste: 4, 5, 6, 3, 1, 2, 7 ⇒ arithmetisches Mittel $\bar{x} = \frac{4+5+6+3+1+2+7}{7} = 4$

Empirische Varianz σ^2
Eine Liste von Daten $x_1, x_2, x_3, \ldots, x_n$ ist gegeben.
Die Varianz σ^2 wird berechnet durch: $\boldsymbol{\sigma^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n}}$

Wenn Werte mit der absoluten Häufigkeit H, bzw. relativen Häufigkeit h, auftreten, dann gilt:
$\sigma^2 = \frac{(x_1-\bar{x})^2\cdot H+(x_2-\bar{x})^2\cdot H+\ldots+(x_n-\bar{x})^2\cdot H_k}{n} = (x_1-\bar{x})^2\cdot h_1+(x_2-\bar{x})^2\cdot h_2+\ldots(x_n-\bar{x})^2\cdot h_k$, für $k \leq n$

Beispiel:
Datenliste: 4, 5, 6, 3, 1, 2, 7 $\Rightarrow \bar{x} = \frac{4+5+6+3+1+2+7}{7} = 4$

$$\sigma^2 = \frac{(4-4)^2+(5-4)^2+(6-4)^2+(3-4)^2+(1-4)^2+(2-4)^2+(7-4)^2}{7} = 4$$

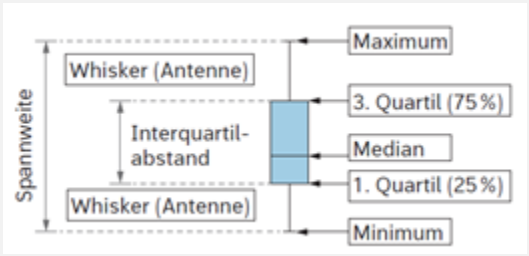
Empirische Standardabweichung σ
Eine Liste von Daten $x_1, x_2, x_3, \ldots, x_n$ ist gegeben.
Die empirische Standardabweichung wird dann berechnet durch $\sigma = \sqrt{\sigma^2}$.

Beispiel:
Datenliste: 4, 5, 6, 3, 1, 2, 7 $\Rightarrow \quad \sigma^2 = 4 \quad \Rightarrow \sigma = \sqrt{4} = 2$

Boxplot (Kastendiagramm, Kastenschaubild)

Ein **Boxplot** ist eine graphische Zusammenfassung der folgenden fünf Punkte:

- Minimum (0%-Quartil)
- 25 %-Quartil Q_1 bzw. q_1 bzw. Q_{25}
- Median (50%-Quartil) bzw. q_2
- 75%-Quartil Q_3 bzw. q_3 bzw. Q_{75}
- Maximum (100%-Quartil)

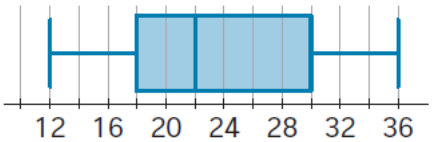


Quartil stammt aus dem Lateinischen und heißt wörtlich „Viertelwert". **Quartile** zerlegen daher eine sortierte Datenreihe in vier (annähernd) gleich große Abschnitte oder Klassen. Das erste Quartil (Q_1) teilt die geordnete Datenreihe in das untere Viertel und das obere Dreiviertel. Das dritte Quartil (Q_3) teilt die geordnete Datenreihe in das untere Dreiviertel und das obere Viertel.

Ganz allgemein gilt, dass eine geordnete Datenreihe durch Quartile als Lagemaß in beliebige Abschnitte geteilt werden kann. Ein p-Quartil gibt daher an, bei welchem Wert einer Verteilung p% der Werte unterhalb dieses Wertes liegen, z. B. gibt der Wert eines 10%-Quartils an, welche Werte der geordneten Datenreihe zu den unteren 10% bzw. oberen 90% gehören.

Für den Aufbau gilt, dass die Box (Kasten) zwischen dem 25%-Quartil und dem 75%-Quartil aufgespannt ist. In ihr wird der **Median** durch einen Querstrich markiert. Die Striche (**Whiskers**) außerhalb der Box gehen bis zum kleinsten bzw. größten Wert. Boxplots werden dann aussagekräftig, wenn die Anzahl der Werte „nicht zu klein" ist, im Allgemeinen sollte daher die Stichprobe mindestens den Umfang 20 besitzen.

Beispiel:
$x_{Min} = 12;\ x_{Max} = 36;\ R = 24;\ q_1 = 18;\ q_2 = 22;\ q_3 = 30$



Interpretationsmöglichkeiten:

- Mindestens 25% aller Werte haben einen Wert von 18 oder kleiner.
- Mindestens 75% aller Werte haben einen Wert von 18 oder größer.
- Mindestens 25% aller Werte haben einen Wert von 30 oder größer.
- Mindestens 75% aller Werte haben einen Wert von 30 oder kleiner.
- Mindestens 50% aller Werte haben einen Wert zwischen 18 und 30.
- Mindestens 50% aller Werte haben einen Wert von 22 oder kleiner.
- Mindestens 50% aller Werte haben einen Wert von 22 oder größer.