Grundlagen der Statistik

Von 10 Schülerinnen und Schüler wurde das Geschlecht notiert. Es ergab sich folgende Stichprobe: m, w, w, m, w, m, w, w, m, w (w=weiblich, m=männlich)

Diese Stichprobe möchten wir statistisch auswerten. Was können wir tun?

1) Die **absolute Häufigkeit** berechnen:
Wir zählen ab und stellen fest, dass es 4 Jungen und 6 Mädchen waren:

Geschl.	absolute Häufigkeit
m	4
w	6

2) Wir können die **relative Häufigkeit** berechnen. Hierzu muss man jeweils die absolute Häufigkeiten durch die Anzahl der Stichprobenweite (hier 10) teilen:

relative Häufigkeit für m: $\frac{4}{10} = \frac{2}{5}$

relative Häufigkeit für w: $\frac{6}{10} = \frac{3}{5}$

Diese kann man als Bruch, als Kommazahl oder in % angeben:

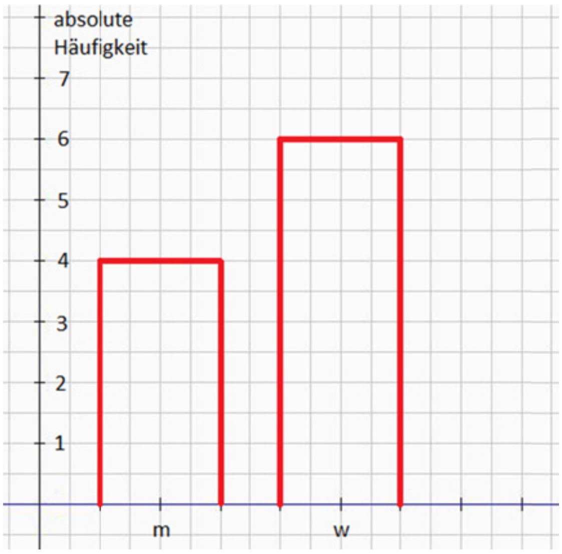
$$\frac{4}{10} = 0{,}4 = 40\%$$

Für die Kommazahl muss man nun Zähler (oben) durch Nenner (unten) dividieren (teilen), also 4:10=0,4. Für den %-Wert muss man nur noch mit 100 multiplizieren, oder das Komma zweimal nach rechts verschieben.

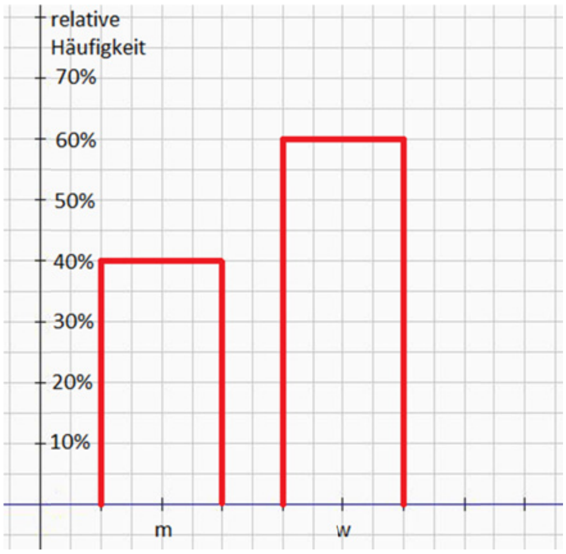
$$0{,}4 = 0{,}4 \cdot 100\% = 40\%$$

Geschl.	abs. H.	rel. H.
m	4	0,4
w	6	0,6

3) Wir können die absolute oder relative Häufigkeit in einem Diagramm darstellen z.B. in einem **Balkendiagramm**:



oder

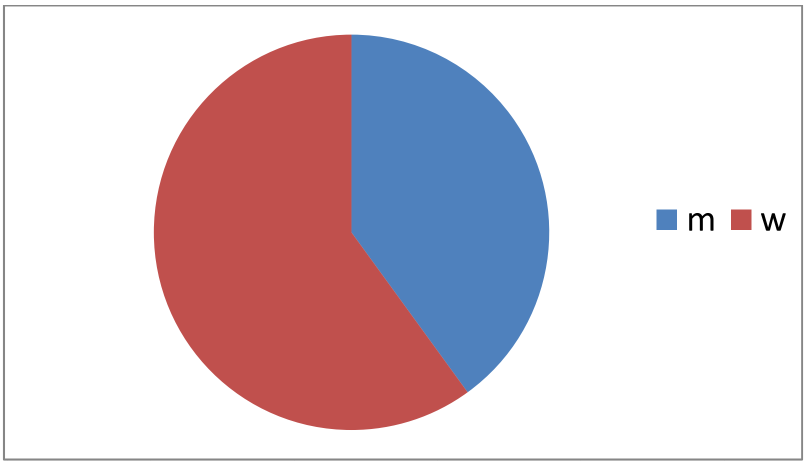


4) Man kann auch ein **Kreisdiagramm** erstellen.

Datei entspricht die Gradzahl des Tortenstücks der %-Zahl:

Für Jungen: $0{,}4 \cdot 360° = 144°$ $\quad (= \text{ relative Häufigkeit} \cdot 360°)$

Für Mädchen: $0{,}6 \cdot 360° = 216°$



Man kann hier keinen **Mittelwert** (Symbol: $\overline{\mathrm{x}}$) berechnen, denn die Daten sind nicht „nummerisch“ bzw. man sagt hier nicht-metrisch.

Nun haben wir noch eine Stichprobe bestimmt, dabei haben wir die Körpergröße in cm von 5 Jungen und Mädchen notiert:

151, 152, 148, 160, 157

Der **Mittelwert** ist die Summe der Werte, dividiert durch die Anzahl (die allgemein mit n bezeichnet wird):

$$\overline{\mathrm{x}} = (151 + 152 + 148 + 160 + 157) : 5 = 768 : 5 = 153{,}6$$

Also sind die 5 Personen im Mittel 153,6 cm groß.

Formel für den Mittelwert: $\overline{x} = (x_1 + x_2 + x_3 + \cdots + x_n)/n$

Eine weitere Kenngröße ist der **Median**.

Hier muss man den mittleren Wert der sortierten Stichprobe bestimmen:

Sortierte Stichprobe:

148, 151, 152, 157, 160

In der Mitte liegt die 152 ⇒ Median = 152 cm

Wie bestimmt man die Mitte bzw. die Position des mittleren Wertes?

Man dividiert die Anzahl n = 5 durch 2 und rundet auf 5:2 = 2,5 ⇒ aufrunden ergibt 3, also der dritte Wert der sortierten Stichprobe.

Nun weiß man, dass mindestens 50% der Personen bis zu 152cm groß sind.

Was ist, wenn es nur 4 Werte in der Stichprobe gewesen wären z. B. bei der Stichprobe:

160, 168, 170, 176.

Hier würde sich n = 4 und n:2 = 4:2 = 2 ergeben.

Wenn n gerade ist, gibt es aber keinen mittleren Wert, denn es liegen praktisch zwei Werte in der Mitte, hier die 168 und die 170. Hier wird dann der Mittelwert aus den beiden Werten in der Mitte gebildet:

$$\text{Median} = \frac{168\text{cm} + 170\text{cm}}{2} = 169\text{cm}.$$

Es gibt auch noch Kenngrößen für die Streuung, wie die **Varianz** s^2 oder die **Standardabweichung** s.

Wir betrachten die Stichprobe von drei Schülerinnen und Schülern, die nach ihrem Körpergewicht in kg gefragt wurden:

43, 50, 57

Der erste Stichprobenwert wird wie üblich mit x_1, der zweite mit x_2, …., bezeichnet:

$x_1 = 43$
$x_2 = 50$
$x_3 = 57$

Für die Berechnung der Varianz muss erst der Mittelwert $\overline{x}$ berechnet werden:

$$\overline{x} = (43 + 50 + 57)/3 = 50$$

Nun muss von jedem Stichprobenwert der Mittelwert subtrahiert und das Ergebnis quadriert werden. Die Summe dieser Werte dividiert durch n ist dann die Varianz der Stichprobe s^2:

$$\begin{aligned} s^2 &= ((43-50)^2 + (50-50)^2 + (57-50)^2)/3 \\ &= (49+0+49)/3 \approx 32{,}67 \end{aligned}$$
(mit Einheit wäre $s^2 \approx 32{,}67$ cm²)

Die Standardabweichung s ist die Wurzel aus der Varianz:

$$s = \sqrt{s^2} \approx \sqrt{32{,}67} \approx 5{,}72$$
(mit Einheit wäre $s \approx 5{,}72$ cm)

Je stärker die Stichprobenwerte vom Mittelwert abweichen, umso größer ist die Standardabweichung (oder die Varianz).

Formel
$$s^2 = ((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x}^2))/n$$
$$s = \sqrt{s^2}$$

(In der schließenden Statistik wird oft durch (n - 1) statt durch n geteilt.)

Bemerkung: Für was benötigt man die Varianz oder die Standardabweichung?
Bei gleichem Mittelwert von zwei Stichproben kann es große Unterschiede bei der Streuung der Stichprobenwerte geben.

Es wurden drei Schülerinnen und Schüler nach ihrem Taschengeld gefragt:

Sie gaben in € an: 19, 20, 21

Der Mittelwert beträgt $\overline{x} = (19 + 20 + 21)/3 = 20$

Also beträgt der Mittelwert 20 €. Wenn man die Stichprobe nicht kennt, kann man aber alleine am Mittelwert noch nicht viel über die Personen sagen, obwohl dieser oft als tatsächlicher Wert interpretiert wird. Bei der oberen Stichprobe ist die Streuung noch relativ klein, denn die Werte liegen nahe beim Mittelwert ($s^2 = 2/3$, $s \approx 0{,}82$). Man würde aber denselben Mittelwert erhalten, wenn 2 Schülerinnen und Schüler 0 € Taschengeld angegeben hätten und einer Schüler 60 € erhalten würde. Die Stichprobe 0, 0, 60 hat auch den Mittelwert 20 und man würde sagen, im Mittel hat jeder 20 € Taschengeld. Hier wäre dann aber die Streuung deutlich größer ($s^2 = 2400$, $s \approx 49$).