Die tabellarische Darstellung von Daten

Ein Beispiel

Wir betrachten die Zufallsvariable X (Gewicht in kg). $n = 20$ SchülerInnen einer Schulklasse werden gewogen. Die Werte x_i sind (aufgerundet auf die nächstgelegene ganze Zahl):

$$36, 41, 48, 36, 43, 32, 36, 43, 38, 32, 44, 38, 35, 29, 33, 38, 41, 44, 38, 41.$$

X ist *quantitativ*, d.h. ihre Werte sind Zahlen, und sie ist *diskret* (die Anzahl der Ausprägungen von X ist endlich).

Wir betrachten folgende Tabelle:

x_i'	f_i	r_i	f_i^+	r_i^+
29	1	0.05	1	0.05
32	2	0.1	3	0.15
33	1	0.05	4	0.2
35	1	0.05	5	0.25
36	3	0.15	8	0.4
38	4	0.2	12	0.6
41	3	0.15	15	0.75
43	2	0.1	17	0.85
44	2	0.1	19	0.95
48	1	0.05	20	1
Σ	20	1		

Die Bezeichnungen der Spalten sind:

- x_i': der Größe nach geordnete Ausprägungen der Variable X, wobei $i = 1, \ldots, k$
- f_i absolute Häufigkeiten der x_i'. Die Summe der f_i muss gleich n sein.

- r_i: relative Häufigkeit der $x_i^{'}$, d.h.

$$r_i = \frac{f_i}{n}.$$

  Die Summe der r_i muss 1 sein:

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} \frac{f_i}{n} = \frac{1}{n} \sum_{i=1}^{n} f_i = \frac{1}{n} * n = 1.$$

- f_i^+ ist die absolute kumulative Häufigkeitsfunktion; sie bezeichnet die Anzahl der Ausprägungen, die kleiner oder gleich einem bestimmten Wert sind, d.h.

$$f_i^+ = \sum_{l=1}^{i} f_l$$

$$f_k^+ = \sum_{l=1}^{k} f_l = n$$

  falls f_l die absolute Häufigkeit der Variable ist. In unserem Beispiel ist $f_3^+ = 4$, d.h. 4 Ausprägungen sind kleiner oder gleich $x_3^{'} = 33$ (nämlich $29, 32, 32, 33$).

- r_i^+ ist die relative kumulative Häufigkeitsfunktion:

$$r_i^+ = \frac{f_i^+}{n}$$

  wobei

$$r_k^+ = \sum_{i=1}^{k} \frac{f_i}{n} = \frac{1}{n} \sum_{i=1}^{k} f_i = 1$$

  r_3^+ ist 0.2, d.h. 20% der Ausprägungen sind kleiner oder gleich 33.

Mithilfe der Häufigkeitstabelle kann man statistische Maßzahlen wie z.B. den Mittelwert ermitteln:

- Mittelwert $\bar{x}$:

Den Mittelwert kann man natürlich direkt aus der Urliste berechnen:

$$\begin{aligned}
\bar{x} &= \frac{1}{20} * (36 + 41 + 48 + 36 + 43 + 32 + 36 + 43 + 38 + 32 \\
&\quad +44 + 38 + 35 + 29 + 33 + 38 + 41 + 44 + 38 + 41) \\
&= \frac{1}{20} * 766 = 38.3
\end{aligned}$$

Eine andere Möglichkeit ist die Berechnung des Mittelwertes mithilfe der Häufigkeitstabelle:

$$\boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{k} f_i x_i^{'}}$$

In unserem Beispiel ergibt dies:

$$\begin{aligned}
\bar{x} &= \frac{1}{20}(1 * 29 + 2 * 32 + 1 * 33 + 1 * 35 + 3 * 36 + 4 * 38 \\
&\quad +3 * 41 + 2 * 43 + 2 * 44 + 1 * 48) \\
&= \frac{1}{20} * 766 = 38.3
\end{aligned}$$

- Stichprobenvarianz s^2:

  Die Stichprobenvarianz berechnet sich durch den *Verschiebungssatz für Varianzen*:

$$\begin{aligned}
s^2 &= \frac{1}{n} \sum_i (x_i - \bar{x})^2 = \frac{1}{n} \sum_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n} \sum_i x_i^2 - 2\frac{1}{n} \underbrace{\sum_i x_i}_{=n\bar{x}} \bar{x} + \frac{1}{n} \sum_i \bar{x}^2 = \frac{1}{n} \sum_i x_i^2 - 2\frac{1}{n} n\bar{x}\bar{x} + \frac{1}{n} n\bar{x}^2 \\
&= \frac{1}{n} \sum_i x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n} \sum_i x_i^2 - \bar{x}^2 = \frac{1}{n}(\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n})
\end{aligned}$$

  In unserem Beispiel ergibt sich:

$$\begin{aligned}
s^2 &= \frac{1}{20}((36^2 + 41^2 + \ldots + 41^2) - \frac{(36 + 41 + \ldots + 41)^2}{20}) \\
&= \frac{1}{20}(29784 - \frac{766^2}{20}) = 22.31
\end{aligned}$$

Mithilfe der Häufigkeitstabelle berechnet sich die Stichprobenvarianz folgendermaßen:

$$\boxed{\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{k} f_i(x_i^{'} - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} f_i(x_i^{'})^2 - \bar{x}^2 \\ &= \frac{1}{n}(\sum_{i=1}^{k} f_i(x_i^{'})^2 - \frac{(\sum_i f_i x_i^{'})^2}{n}) \end{aligned}}$$

In unserem Beispiel ergibt sich:

$$\begin{aligned} s^2 &= \frac{1}{20}(1 * 29^2 + 2 * 32^2 + \ldots + 1 * 48^2 \\ &\quad + \frac{(1 * 29 + 2 * 32 + \ldots + 1 * 48)^2}{20}) \\ &= \frac{1}{20}(29784 - \frac{766^2}{20}) = 22.31 \end{aligned}$$

Anmerkung: Um eine Einteilung in Klassen zu ermöglichen, kann es notwendig sein die stetige Variable X zu 'diskretisieren'. Falls das Körpergewicht auf den Gramm genau gemessen wird, so hat die Darstellung in kg drei Nachkommastellen (z.B. 23462g=23.462kg). Um eine übertrieben feine Klasseneinteilung zu vermeiden, kann man etwa einfach die Nachkommastellen weglassen oder auf die nächstgrössere ganze Zahl runden.
Ein Beispiel: Alle Werte von 23.500 bis 24.499 werden auf 24 gerundet.