Formelsammlung

Deskriptive Statistik

Absolute Häufigkeit

$H(a_j) :=$ "Anzahl der Fälle, in denen a_j auftritt", $\quad j = 1,\ 2, ..., k$.

Relative Häufigkeit

$$h(a_j) := \frac{1}{n} \cdot H(a_j), \qquad j = 1,\ 2, ..., k .$$

Empirische Verteilungsfunktion (Nichtgruppiert)

$$F(x) := \begin{cases} 0 & \text{für} \quad x \le a_1 \\ \sum_{i=1}^{j} h(a_i) & \text{für} \quad a_j < x \le a_{j+1} \\ 1 & \text{für} \quad x > a_k \end{cases} .$$

Wichtige Eigenschaften der Verteilungsfunktion

1.

$$F(x) = A(X < x).$$

2.

$$0 \le F(x) \le 1$$

3.

$$\forall\, x_1, x_2 : x_1 < x_2 \quad \Rightarrow \quad F(x_1) \le F(x_2)$$

4.

$$A(x_1 \le X < x_2) = F(x_2) - F(x_1)$$

5.

$$x \to -\infty \qquad F(x) \to 0$$
$$x \to +\infty \qquad F(x) \to 1 .$$

6.

Eine empirische Verteilungsfunktion ist mindestens linksseitig stetig und hat höchstens endlich viele Sprungstellen.

Absolute Klassenhäufigkeiten

$H_i :=$ "Anzahl der Beobachtungswerte in der Klasse K_i ", $\quad i = 1,\ 2, ..., p$.

Relative Klassenhäufigkeiten

$$h_i := \frac{H_i}{n},\ i = 1,\ 2, ..., p .$$

Empirische Verteilungsfunktion (Gruppiert)

$$F(x) := \begin{cases} 0 & \text{für} \quad x \le m_1 \\ \sum_{i=1}^{j} h(a_i) & \text{für} \quad m_j < x \le m_{j+1} \\ 1 & \text{für} \quad x > m_k \end{cases}$$

Lagemaße	*Ungruppiert*	*Gruppiert*
Arith. Mittel	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $= \frac{1}{n}\sum_{j=1}^{k} a_j \cdot H_n(a_j)$	$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{p} m_i \cdot H_i$ $m_i := \frac{g_i + G_i}{2}, \quad i = 1,\ 2, \ldots, p$ $\frac{1}{n}\sum_{i=1}^{p} g_i \cdot H_i \le \bar{x} \le \frac{1}{n}\sum_{i=1}^{p} G_i \cdot H_i$
Median	$Me := \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{für } n: \text{ ungerade} \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right) & \text{für } n: \text{ gerade} \end{cases}$	$\begin{cases} \sum_{j=1}^{i-1} h_j < 0.5 \\ \sum_{j=1}^{i} h_j \ge 0.5 \end{cases}$ $Me \approx g_i + \frac{0.5 - \sum_{j=1}^{i-1} h_j}{h_i} \cdot b_i$
α **– Quantil**	$\tilde{x}_\alpha := \begin{cases} x_{[k]} & \text{fall } n \cdot \alpha \text{ keine ganze Zahl ist. } (k \text{ ist dann die auf } n \cdot \alpha \text{ folgen} \\ \frac{1}{2}\left(x_{[k]} + x_{[k+1]}\right) & \text{falls } n \cdot \alpha \text{ eine ganze Zahl ist. (Es is dann } k = n \cdot \alpha.) \end{cases}$	$\begin{cases} \sum_{j=1}^{i-1} h_j < \alpha \\ \sum_{j=1}^{i} h_j \ge \alpha \end{cases}, \quad (0 < \alpha < 1)$ $\tilde{x}_\alpha \approx g_i + \frac{\alpha - \sum_{j=1}^{i-1} h_j}{h_i} \cdot b_i$

Lagemaße	*Ungruppiert*	*Gruppiert*
Modus	Als *Modus* bezeichnet man diejenige Merkmalsausprägung, die eindeutig am häufigsten vorkommt.	Im Falle eines gruppierten Datenmaterials wird er in der Klasse K_i mit der maximalen Häufigkeit folgendermaßen berechnet: $M :\approx g_i + \frac{h_i - h_{i-1}}{2h_i - h_{i-1} - h_{i+1}} \cdot b_i$
Geometrisches Mittel	$\bar{x}_g := \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$	

Streuungsmaße	*Ungruppiert*	*Gruppiert*
Spannweite	$R := x_{\max} - x_{\min}$	$R \approx G_p - g_1$
Quartilsabstand	$QA := \tilde{x}_{0.75} - \tilde{x}_{0.25}$	$QA := \tilde{x}_{0.75} - \tilde{x}_{0.25}$
Mittlerer Quartilsabstand	$\bar{Q} := \frac{1}{2}\left(\tilde{x}_{0.75} - \tilde{x}_{0.25}\right)$	$\bar{Q} := \frac{1}{2}\left(\tilde{x}_{0.75} - \tilde{x}_{0.25}\right)$
Varianz	$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ (bei einer *Grundgesamtheit)* $\left(= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right).$ $s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}.$ (Bei einer *Stichprobe)*	$\sigma^2 :\approx \frac{1}{N}\sum_{i=1}^{p}(\bar{x}_i - \mu)^2 \cdot H_i,$ (bei einer *Grundgesamtheit*) $s^2 :\approx \frac{1}{n-1}\sum_{i=1}^{p}(\bar{x}_i - \bar{x})^2 \cdot H_i,$ (bei einer *Stichprobe*) Falls $\bar{x}_i$ nicht existieren, werden sie durch m_i ersetzt.
Standardabweichung	$s > 0$ bzw. $\sigma > 0$	$s > 0$ bzw. $\sigma > 0$
Variationskoeffizient	$v := \frac{s}{\bar{x}}$ bzw. $v := \frac{\sigma}{\mu}$	$v := \frac{s}{\bar{x}}$ bzw. $v := \frac{\sigma}{\mu}$

Korrelation

Korrelationskoeffizient (*Bravais - Pearson*):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Regression:

Normalgleichungen zur Bestimmung der ***einfachen linearen Regressionsfunktion***

$$y^* = a_0 + a_1 x$$

$$\begin{cases} n \cdot a_0 \qquad + a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \cdot \sum_{i=1}^{n} x_i + a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

Korrelationskoeffizient (linear)

$$r_{yx} := \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i\right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i\right)}}$$

Bestimmtheitsmaß(linear)

$$B_{yx} := r_{yx}^2 .$$

Bestimmtheitsmaß(allgemein)

$$r_{yx}^2 := \frac{\sum_{i=1}^{n}\left(y_i^* - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \quad .$$