Formelsammlung

1. Beschreibende Statistik

Lagerparameter

- Modalwert (x_{mod}) - am häufigsten auftrit (bei qualitativen Merkmalen)
- Mittelwert ($\bar{x}$) : $\frac{1}{n}\sum_{i=1}^{n} x_i$ - Schwerpunkt der Daten x_i
- Median ($x_{0.5}$) : $\begin{cases} x_{\frac{n+1}{2}} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) \end{cases}$ liegt in der Mitte der sortierten Daten

Streuungsparameter

- Spannweite : $\max x_i - \min x_i$
- Stichprobenvarianz ($var(x)$) : $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right)$

  die gemittelte Summe der quadratischen Abweichungen vom $\bar{x}$
- Stichprobenstandardabweichung ($sd(x)$) :

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} x_i = n \cdot \bar{x}_i$$

- Quartilabstand = $x_{p_1} - x_{p_2}$

p-Quantile

p-Quantil x_p $(0 < p < 1)$ ($quantile(x,p)$) teit die sortierten Daten im Verhältnis: $p:(1-p)$, d.h. $\hat{F}(x_p) \approx p$

$$x_p = \begin{cases} x_{floor(np)+1} & , np \notin \mathbb{N} \\ \frac{1}{2}\left(x_{np} + x_{np+1}\right) & , np \in \mathbb{N} \end{cases}$$

0.25 - Quantil : 1. Quartil
0.5 - Quantil : 2. Quartil
0.75 - Quantil : 3. Quartil

Korrelation

- Empirische Kovarianz s_{xy} ($cov(x,y)$) :

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}\right)$$

$s_{xy} > 0 \Rightarrow$ steigende Punktwolke
$s_{xy} < 0 \Rightarrow$ fallende ...

- Empirische Korrelationskoeffizient r ($cor(x,y)$) $r = \frac{s_{xy}}{s_x \cdot s_y}$
- Regressionsgerade : $y = mx + t$, $m = r\frac{s_y}{s_x}$ $\wedge$ $t = \bar{y} - m\bar{x}$

Chebuschev-Ungleichung:

Sei $\{x_1, x_2, \dots x_n\}$ eine Stichprobe mit $\bar{x}$ und $s>0$

Es sei $S_k = \{i,\ 1 \le i \le n : |x_i - \bar{x}| < ks\}$ $(k \in \mathbb{N})$ mit $N(S_k)$ - die Anz der Elemente in S_k

Dann gilt: $\frac{N(S_k)}{n} > 1 - \frac{1}{k^2}$ → relative Häufigkeit für Daten, die in S_k liegen

- $k=2 \Rightarrow \frac{N(S_k)}{n} > 1 - \frac{1}{2^2} = 0.75$: Mehr als 75% der Daten liegen in dem $2s$-Bereich um $\bar{x}$