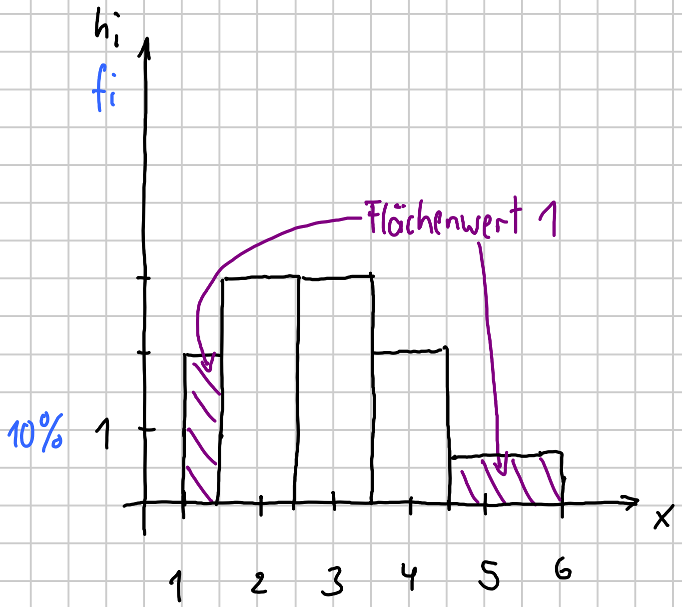
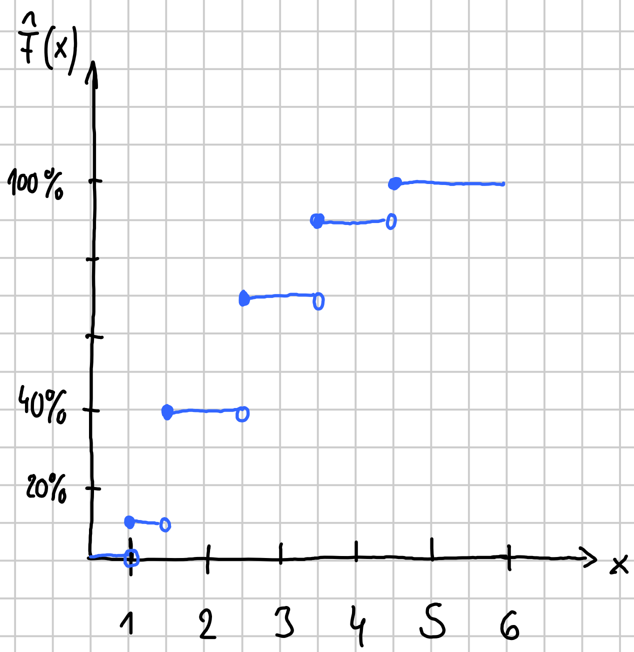




Eigenschaften der empirischen Verteilungsfunktion $\hat{F}(x) = \sum_{i:x_i<x} f_i$

(kumulierte rel. Häufigkeiten)

(1) $\hat{F}(x) = 0$ für $x < x_1$ (x_1: 1 Stichprobenwert)

(2) $\hat{F}(x) = 1$ für $x \geq x_n$ (x_n: letzter Stichprobenwert)

(3) rechtsseitig stetige Treppenfunktion

1.4 Kenngrößen

- Berechnung von Größen, die Lage und Streuung der Daten beschreiben
- Nur für quantitative Merkmale möglich

1.4.1 Lagemaße: Mittelwert, Median, Modalwert

Definition 12:

Sei $\{x_1, x_2, \ldots, x_n\}$ eine Stichprobe vom Umfang n.

(1) Der Mittelwert $\bar{x}$ berechnet sich aus

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

(2) Der Median $x_{\frac{1}{2}}$ ist gegeben durch

$$x_{\frac{1}{2}} = \begin{cases} x_{\frac{n+1}{2}} & , \text{ falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & , \text{ falls } n \text{ gerade} \end{cases}$$

d.h. er liegt in der Mitte der (aufsteigend geordneten) Stichprobenwerte x_i.

(3) Die Modalwerte x_{mod} sind die Merkmalsausprägungen a_i $(i = 1, 2, \ldots, k)$, die in einer Stichprobe am häufigsten auftreten.
Sie sind das einzig mögliche Lagemaß für qualitative Merkmale.

Bemerkungen:

(1) R-Funktionen:
- Mittelwert: mean(x)
- Median : median(x)

(2) Minimumseigenschaften

- Mittelwert $\sum_{i=1}^{n} (x_i - \bar{x})^2 \leq \sum_{i=1}^{n} (x_i - c)^2$, $c \in \mathbb{R}$ beliebig

  Minimum der Summe der Abstandsquadrate der Stichprobenwerte von einem $c \in \mathbb{R}$

- Median $\sum_{i=1}^{n} |x_i - x_{\frac{1}{2}}| \leq \sum_{i=1}^{n} |x_i - c|$, $c \in \mathbb{R}$ bel.

Bsp.: Notenverteilung

a_i	h_i
1	2
2	4
3	3
4	4
5	3
6	0

Berechnen Sie die drei Lagemaße $\bar{x}$, $x_{1/2}$ und x_{mod}.

Stichprobenumfang: $n = 16$

Modalwerte $x_{mod} = \{2, 4\}$

Mittelwert: $\bar{x} = \frac{1}{16}\left(2\cdot 1 + 4\cdot 2 + 3\cdot 3 + 4\cdot 4 + 3\cdot 5\right)$

$= 3.125$

Median: $x_{\frac{1}{2}} = \frac{1}{2}\left(x_8 + x_9\right) = \frac{1}{2}\left(3+3\right) = 3$

Der Mittelwert wird durch Ausreißer in den Daten beeinflusst, während das beim Median nicht der Fall ist.

1.4.2 Streuungsmaße: Stichprobenvarianz und -standardabweichung

Streuungsmaße beschreiben die Variabilität der Stichprobenwerte um ein Lagezentrum, z.B. den Mittelwert $\bar{x}$.

Definition 1.3:

Sei $\{x_1, x_2, \dots, x_n\}$ eine Stichprobe vom Umfang n.

(1) Die Stichprobenvarianz s^2 ist die gemittelte Summe der quadratischen Abweichungen vom Mittelwert $\bar{x}$:

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

(2) Die Stichprobenstandardabweichung s, die im Gegensatz zu s^2 die gleiche Dimension bzw. Einheit hat wie die Werte x_i, berechnet sich aus $s = \sqrt{s^2}$:

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

Bemerkungen:

(1) R-Funktionen

Stichprobenvektor

- Stichprobenvarianz: var(x)
- Stichprobenstandardabweichung: sd(x)

(2) Der Faktor $\frac{1}{n-1}$ in s^2 liefert eine sog. "erwartungstreue" Schätzung der Varianz $\longrightarrow$ Erklärung später bei schließender Statistik

Satz 1.1: Verschiebungssatz

Die Stichprobenvarianz (empirische Varianz) s^2 lässt sich mit folgender Formel oft einfacher berechnen:

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

Beweis:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) =$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\cdot\underbrace{\sum_{i=1}^{n} x_i}_{n\cdot\bar{x}} + \underbrace{\sum_{i=1}^{n}\bar{x}^2}_{n\cdot\bar{x}^2} = \qquad \left(\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i\right)$$

$$= \sum_{i=1}^{n} x_i^2 - 2n\,\bar{x}^2 + n\,\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2$$

Beispiel: Berechnen Sie mit den beiden Formeln s^2 für das Beispiel der Notenverteilung

$x_1 = x_2 = 1$, $x_3 = x_4 = x_5 = x_6 = 2$, $x_7 = x_8 = x_9 = 3$

$x_{10} = \ldots = x_{13} = 4$, $x_{14} = x_{15} = x_{16} = 5$

$\bar{x} = 3.125$

(1) $$s_2 = \frac{1}{15}\Big(2\cdot(1-3.125)^2 + 4\cdot(2-3.125)^2 + 3\cdot(3-3.125)^2 + 4\,(4-3.125)^2 + 3\cdot(5-3.125)^2\Big) = 1.85$$

(2) $$s_2 = \frac{1}{15}\left(2\cdot 1 + 4\cdot 4 + 3\cdot 9 + 4\cdot 16 + 3\cdot 25 - 16\cdot 3.125^2\right) = 1.85$$

1.4.3 p-Quantile

Unter einem p-Quantil (mit $0 \le p \le 1$) versteht man einen Wert x_p, der die Stichprobenwerte x_i $(i = 1, 2, \dots, n)$ (ungefähr) im Verhältnis $p : (1-p)$ teilt, d.h.

empirische Verteilungsfkt.

$$\frac{\text{Anzahl}\left(\{x_i \le x_p\}\right)}{n} \approx p \iff \hat{F}(x_p) \approx p$$

Bemerkungen:

(1) R-Funktionen

empirische Verteilungsfunktion: ecdf(x)

empirical cumulative distribution function

p-Quantil : quantile(x, p, [type = 1 ⋮ 9])

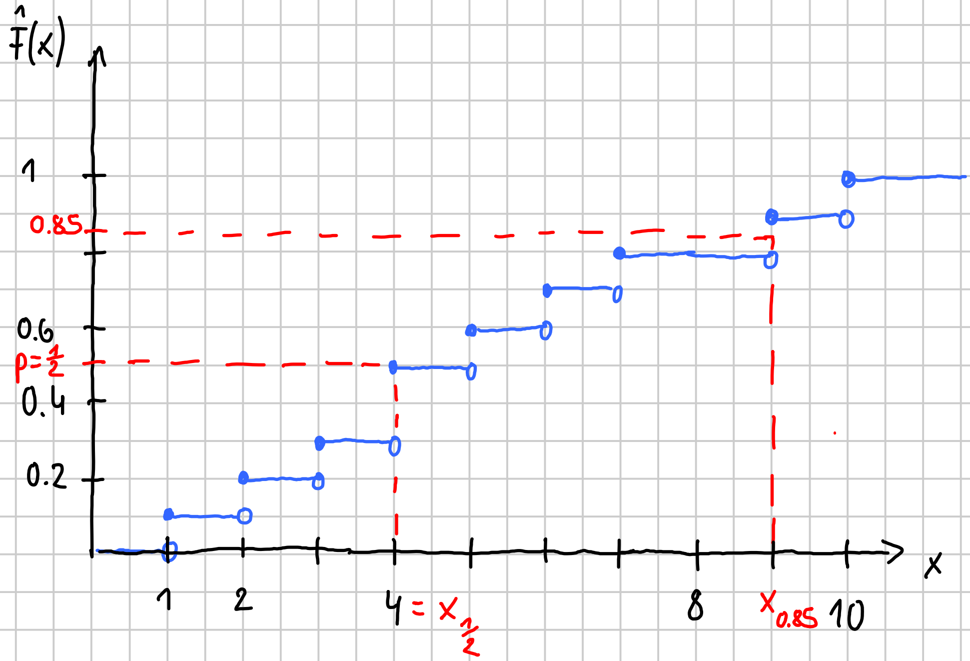
optional: Angabe der verwendeten Definition zur Berechnung des Quantils

Beispiel: Stichprobe $\{1, 2, 3, 4, 4, 5, 6, 7, 9, 10\}$, $n = 10$

Typ 1: Das p-Quantil x_p ist der minimale Stichprobenwert x_i, für den gilt: $\hat{F}(x_i) \ge p$

$$x_p = \min\{x_i : \hat{F}(x_i) \ge p\}$$

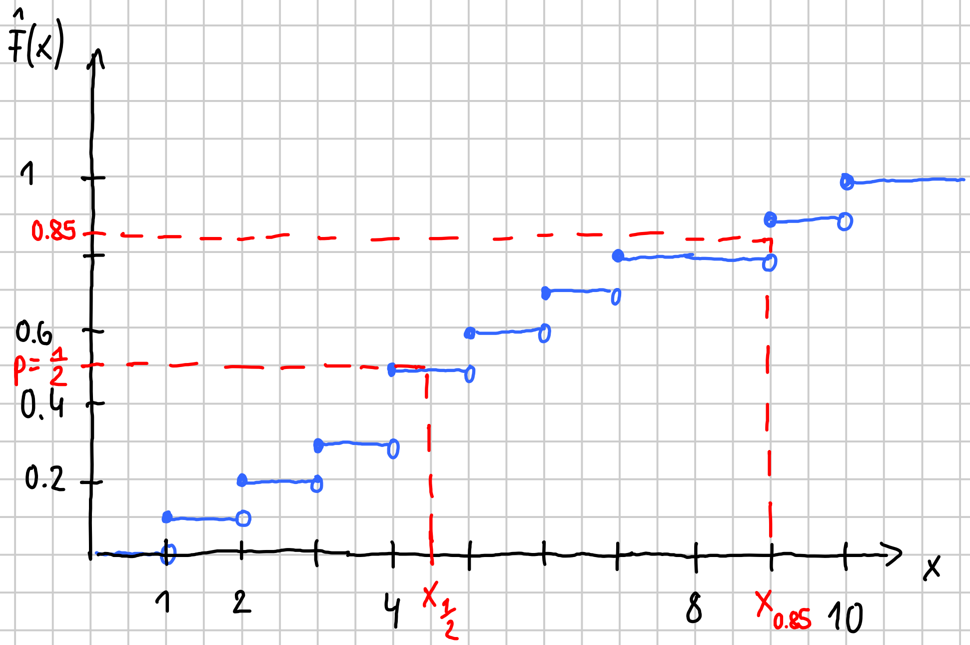
13



Typ 2:

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

$$x_{\frac{1}{2}} = \frac{1}{2}\left(x_5 + x_6\right) \quad , \quad x_{0.85} = x_9$$



- Wichtige Quantilswerte:

  (1) 0.25 - Quantil : 1. Quartil
  0.5 - Quantil : Median
  0.75 - Quantil : 3. Quartil

  umfassen die mittleren 50% aller Daten

- Grafische Darstellung, die die wichtigsten statistischen Größen zusammenfasst : Boxplot

  R- Funktion : boxplot (x)