Kap. 1. Graphische Darstellungen von Messdaten

1.1. Darstellungsformen, unterschiedliche Mittelwerte.

Empirische Untersuchungen / Befragungen:

Unterschiedliche Merkmale als Gegenstand, z.B.

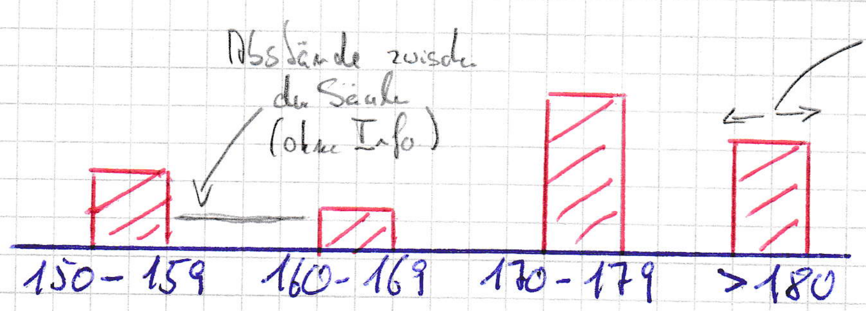
- metrisch messbare Merkmale (wie Längen...)
- nominale Merkmale (Studiengang, Haarfarbe...)
- ordinale Merkmale (nominale Merkmale, die geordnet werden können, z.B. Skalen wie ... eher schlecht ... eher gut ...)

Im Folgenden sind die Merkmalsausprägungen in der Regel metrisch messbar.

Beispiel. Die Größe von 10 Personen wurde gemessen als:

# Personen	Größe in cm
1	153
1	158
1	162
2	171
1	173
1	178
1	185
1	186
1	185

Darstellung z.B. als Säulendiagramm



Breite enthält keine Information

Insgesamt nicht alle Informationen dargestellt.

→ Vorsicht vor Manipulationsmöglichkeit

Weitere Darstellungsarten (Histogramm, Stamm-Blatt-Darstellung etc.) in den Übungen

1.3

Darstellungen, die besonders charakteristische Kennzahlen einer Datenreihe beinhalten?

Beispiele. i.) Das arithmetische Mittel (oder der Mittelwert)

N Messdaten x_i: $$\bar{x} = \frac{1}{N}\left(x_1+x_2+\dots+x_{N-1}+x_N\right).$$

(Lagemaß: $\ell(x_1+a, x_2+a, \dots, x_N+a) = \ell(x_1, x_2, \dots x_n)+a$, geom. Mittel $\sqrt[N]{x_1 x_2 \dots x_N}$ kein Lagemaß)

Notation. Nur zur Abkürzung: Σ

$$a_1+a_2+\dots+a_n = \sum_{k=1}^{n} a_k .$$

Endwert

Summand

Startwert (muss nicht 1 sein)

Laufindex

Beispiele

$$\sum_{k=1}^{n} a_k = \sum_{l=1}^{n} a_l$$

↳ egal wie man den nennt

$$\sum_{k=1}^{n} a_k = \sum_{j=2}^{n+2} a_{j-1} = \sum_{i=-3}^{n-4} a_{i+4}$$

↳ Umnummerieren

$$\sum_{m=0}^{2} b_{2m} = b_0 + b_2 + b_4$$

↳ gerade

$$\sum_{m=0}^{2} b_{2m+1} = b_1 + b_3 + b_5$$

↳ ungerade

alle Summanden gleich

$$\sum_{k=1}^{n} a = n \cdot a$$

$$\sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k = \sum_{k=1}^{n} (a_k + b_k) \quad , \quad c \sum_{k=1}^{n} a_k = \sum_{k=1}^{n} (c a_k).$$

$$\boxed{\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

Weitere Beispiele Kennzahlen

Vorstellung: $\frac{1}{2} = 0{,}5 \hat{=} 50\%$

ii.) **Empirischer Median** $x_{1/2}$ (auch Zentralwert genannt):

- Die Daten werden der Reihenfolge nach geordnet: Z.B.

  Ist $x_1 = 185$, $x_2 = 179$, $x_3 = 171$, $x_4 = 195$,

  $x_5 = 186$, $x_6 = 153$, $x_7 = 171$, $x_8 = 173$, $x_9 = 162$, $x_{10} = 158$,

  so schreibt man

  $$\begin{array}{ccccccccccccccccccc} x_6 & \le & x_{10} & \le & x_9 & \le & x_3 & \le & x_7 & \le & x_8 & \le \; x_2 = 179 \; \le & x_1 & \le & x_5 & \le & x_4 \\ \| & & \| & & \| & & \| & & \| & & \| & & \| & & \| & & \| \\ 153 & & 158 & & 162 & & 171 & & 171 & & 173 & & 185 & & 186 & & 195 \end{array}$$

- Die Daten werden umnummeriert mit der neuen Bezeichnung, hier: (vgl. Beispiel Größe von 10 Personen)

  $x_1 = 153$, $x_2 = 158$, $x_3 = 162$, $x_4 = 171$, $x_5 = 171$,

  $x_6 = 173$, $x_7 = 179$, $x_8 = 185$, $x_9 = 186$, $x_{10} = 195$.

- Der Zentralwert ist

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

z.B. bei geordneter Reihe

$$x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5 \leq x_6 \leq x_7 \leq x_8 \leq x_9 \leq x_{10}$$

N = 10 : gerade

$$x_{1/2} = \frac{1}{2}\left(x_5 + x_6\right)$$

oder bei der geordneten Reihe

$$x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5 \leq x_6 \leq x_7 \leq x_8 \leq x_9$$

N = 9 : ungerade

$$x_{1/2} = x_5.$$

Zur Lage des Medians: "Messwerte ≥ & Messwerte ≤" halten sich genau die Waage.

mind. 50% ≤
mind. 50% ≥ →

Vorstellung: z.B. $p = 0.65 \triangleq 65\%$

iii.) Es sei $0 < p < 1$. Das empirische p-Quantil wird für eine wie oben geordnete Datenreihe berechnet:

Weiter sei zunächst definiert

$$[y] := \max\{k \in \mathbb{Z} : k \leq y\}.$$

Kleinste ganze Zahl kleiner oder gleich y.

Dann ist das empirische p-Quantil:

$$x_p := \begin{cases} x_{\lceil Np+1 \rceil}, & \text{falls } Np \notin \mathbb{N} \\ \frac{1}{2}\left(x_{Np} + x_{Np+1}\right), & \text{falls } Np \in \mathbb{N}. \end{cases}$$

($Np \notin \mathbb{N} \Rightarrow Np+1 \notin \mathbb{N}$, $\lceil Np+1 \rceil$ kleinste ganze Zahl größer oder gleich Np,

(Prozent-)Übung: $p = \frac{1}{2} \Rightarrow$ beide Definitionen stimmen überein)

Beispiel geordnete Reihe.

$$x_1 \le x_2 \le x_3 \le x_4 \le x_5 \le x_6 \le x_7 \le x_8 \le x_9 \le x_{10}$$

$N = 10$

Bsp. → $p = 0{,}3$, $Np = 3 \in \mathbb{N}$

$p = 0{,}5$, $Np = 5 \in \mathbb{N}$

$p = 0{,}75$, $Np = 7{,}5 \notin \mathbb{N}$, $\lceil Np+1 \rceil = 8$

Namensgebung.

- $p = 0.25$ $\qquad x_{0.25}$: unteres Quartil
- $p = 0.75$ $\qquad x_{0.75}$: oberes Quartil
- $x_{0.75} - x_{0.25}$: Quartilsabstand.

Im Beispiel Größe von 10 Personen ist:

$$x_{1/2} = \frac{1}{2}(171 + 173) = 172,$$

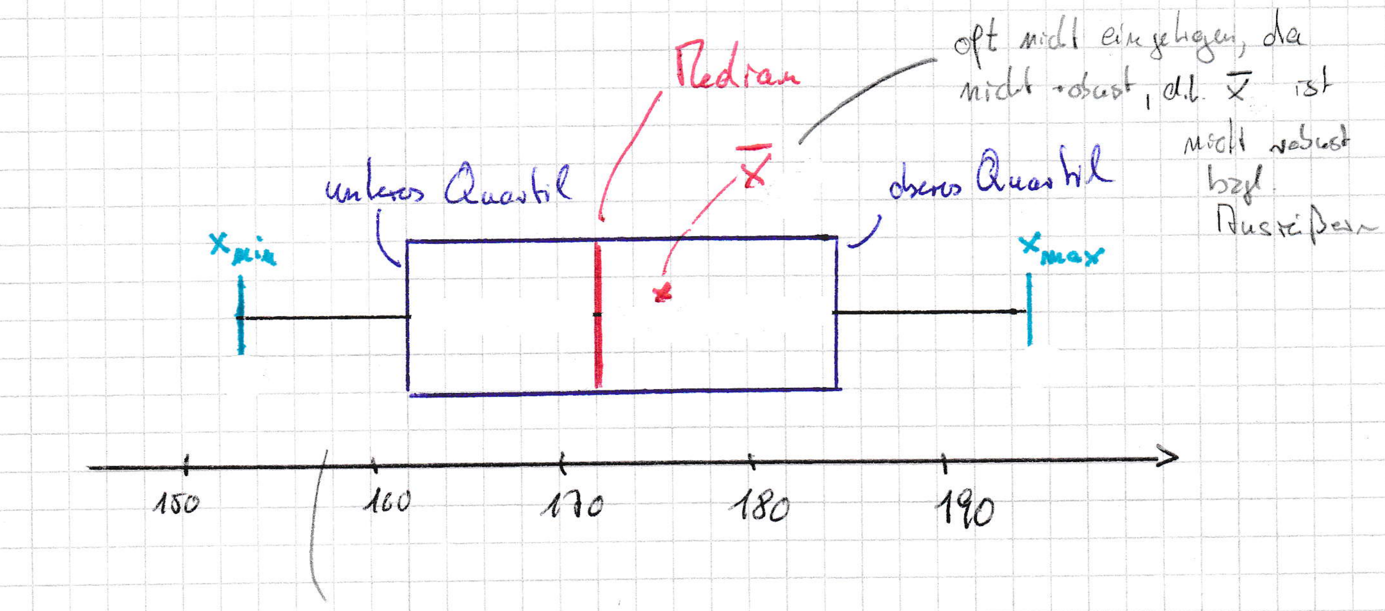
$$x_{0.25} = x_3 = 162,$$

$$x_{0.75} = x_8 = 185,$$

$$x_{0,75} - x_{0,25} = 23.$$

Weiter ist $\quad \overline{x} = 173{,}3.$

Darstellung als sogenannter Boxplot:



Bemerkung: Man kann auch geeignete "Antennen" an die Box definieren. Werte außerhalb dieses Bereiches sind Ausreißer.

1.2. Bemerkungen zur weiteren Analyse.

Wie kann man die Streuung von Messdaten beschreiben?

Ein mögliches Hilfsmittel: Empirische Varianz

(oder Stichprobenvarianz oder Varianz der Messreihe)

Definition 1. Man betr. eine Messreihe aus N Daten.

i.) Dann ist die Varianz s_x^2 mithilfe der quadratischen Abweichung vom Mittelwert definiert:

$$s_x^2 := \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_{N-1}-\bar{x})^2+(x_N-\bar{x})^2}{N-1}$$

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

ii.) Die (empirische) Standardabweichung ist

$$s_x := \sqrt{s_x^2}.$$

Bemerkungen. i.) Positive und negative Abweichungen vom Mittelwert werden hier in gleicher Weise berücksichtigt.

ii.) Manchmal wird der Vorfaktor $1/N$ statt $1/(N-1)$ verwendet. Die Wahl $\frac{1}{N-1}$ hat etwas mit der sogenannten "Erwartungstreue von Schätzverfahren" zu tun.

↳ Erwartungswert ⟷ wahrer Wert.

In dem Beispiel ist $s_x^2 \approx 176{,}23$, $s_x \approx 13{,}28$.

Man rechnet weiter leicht nach (⇝ Präsenzübungen)

Satz 1. (Verschiebungssatz für die Stichprobenvarianz)

Es gilt
$$s_x^2 = \frac{1}{N-1}\left[\left(\sum_{i=1}^{N} x_i^2\right) - N\bar{x}^2\right].$$

Bemerkung. Natürlich ist man an der relativen Streuung interessiert, d.h. für $x_n > 0$ betrachtet man den empirischen Variationskoeffizienten $v = \frac{s_x}{\bar{x}}$.

(Im Beispiel $v = 0{,}077$ ⇝ 7,7 %)

Beispiel Kolbenhubpipette: Vgl. [Ho].