1. Beschreibende Statistik

Lernziele:

- Sie kennen den Unterschied zw. Balkendiagramm und Histogramm.
- Sie können Daten sinnvoll klassifizieren.
- Sie berechnen für eine Stichprobe Lage- und Streuungsmaße.
- Sie verstehen die Bedeutung eines p-Quantils.
- Sie sind in der Lage Statistiken zu „lesen“ bzw. graphische Darstellungen zu interpretieren.
- Sie treffen Aussagen mit Hilfe der Berechnung von Korrelationskoeffizienten.

Literatur:

- Teschl Band 2, Kap. 25.2
- Arens et al., Kap 36.1 - 36.4
- Zucchini, Kap. 2

1.1 Begriffe

- **Beschreibende (deskriptive) Statistik**
  Beobachtete Daten werden durch geeignete statistische Kennzahlen charakterisiert und durch geeignete Grafiken anschaulich gemacht.
- **Schließende (induktive) Statistik**
  Aus beobachteten Daten werden Schlüsse gezogen und diese im Rahmen vorgegebener Modelle der Wahrscheinlichkeitstheorie bewertet.
- **Grundgesamtheit**
  Ω: Grundgesamtheit (z.B. Studienbewerber Rosenheim WS 16/17)
  ω: Element bzw. Objekt der Grundgesamtheit
- **Merkmal**
  $X : \Omega \longrightarrow M$: Merkmal (z.B. gewählter Studiengang)
  $X(\omega) = x$: Ausprägung des Merkmals (z.B. INF, WIF, WMA)
  - qualitativ — quantitativ ($M \subset \mathbb{R}^p$)
  - diskret (<30 Ausprägungen) — "stetig" (≥ 30 Ausprägungen)
  - univariat ($p = 1$) — multivariat ($p > 1$)

1.2 Darstellung diskreter Merkmale

STUDIENGANG	absolut	relativ
BW	947	0.2119
WI	497	0.1112
MGW	443	0.0991
INF	306	0.0685
MB	298	0.0667
WIF	297	0.0665
HA	252	0.0564
INN	245	0.0548
HT	197	0.0441
EIT	185	0.0414
MEC	177	0.0396
EGT	172	0.0385
IAB	171	0.0383
WMA	171	0.0383
KT	111	0.0248

Abbildung: Absolute und relative Häufigkeiten h_i u. f_i

`x <- read.table(...)`

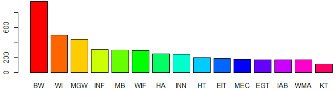


Abbildung: Höhe $\widehat{=}$ Häufigkeit

`barplot(x,col=rainbow(length(x)))`

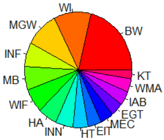


Abbildung:

`pie(x,col=rainbow(length(x)))`

Achtung: Grafiken können bewusst oder fahrlässig verfälscht sein!

s. *Arens, Hettlich, Karpfinger et al.: Mathematik (p. 1360, Kap. 36.2)*

1.3 Darstellung quantitativer, stetiger Merkmale

Beispiel: Abschlusssnoten von $n = 10$ Schülern

```
x <- c(1.97,2.33,3.51,5.11,1.2,2.59,4.18,2.81,3.27,1.50)
```

Klassenbildung:
n sei Anzahl der Ausprägungen

- Möglichst gleiche Breite Δx_i (bis auf erste und letzte)
- 5 - 20 Klassen, aber $\leq \sqrt{n}$
- Klassen links abgeschlossen, rechts offen

```
class <- c(1,1.5,2.5,3.5,4.5,6)
```

Histogramm: Flächentreue Darstellung der Häufigkeitsverteilung

- Höhe $\widehat{=}$ Dichte (Häufigkeit pro Klassenbreite)
- Fläche $\widehat{=}$ Häufigkeit h_i bzw. f_i

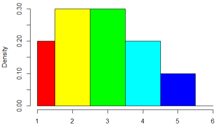


Abbildung: `hist(x,breaks=class)`

Empirische Verteilungsfunktion

Frage: Wie hoch ist der Anteil der Schüler mit einer Durchschnittsnote besser 4?

Klasse	Intervall	absolute Häufigkeit	relative Häufigkeit	relative Summenhäufigkeit
i		h_i	f_i	$F_i = \sum_{k<i} f_k$
1	$[1.0, 1.5[$	1	0.1	0
2	$[1.5, 2.5[$	3	0.3	0.1
3	$[2.5, 3.5[$	3	0.3	0.4
4	$[3.5, 4.5[$	2	0.2	0.7
5	$[4.5, 5.5[$	1	0.1	0.9
6	$[5.5, 6.0[$	0	0	1
Summe		$n = 10$	1	

Empirische Verteilungsfunktion

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

- $\widehat{F}(x) = 0$ für $x < x_1$
- $\widehat{F}(x) = 1$ für $x > x_n$
- rechtsseitig stetige Treppenfunktion

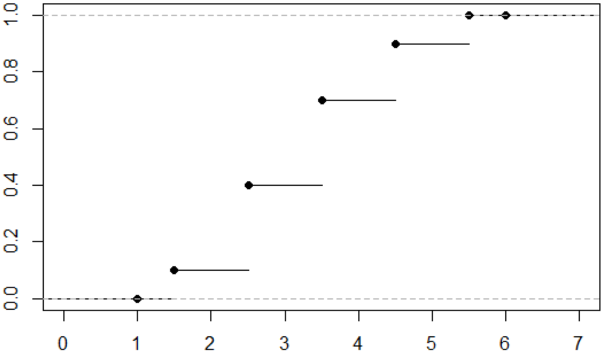


Abbildung: `plot(ecdf(x))`

1.4 Kenngrößen
1.4.1 Lagemaße

- **Modalwert(e)** x_{mod}:
  Am häufigsten auftretende Ausprägungen (insbesondere bei qualitativen Merkmalen)

- **Mittelwert** (Durchschnitt, arithmetisches Mittel) $\bar{x}$: `mean(x)`
  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
  Schwerpunkt der Daten x_i (empfindlich gegenüber Ausreißern)

- **Median** $x_{0.5}$: `median(x)`
  $$x_{0.5} = \begin{cases} x_{\frac{n+1}{2}}, & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & \text{falls } n \text{ gerade} \end{cases}$$
  Liegt in der Mitte der sortierten Daten x_i (robust gegenüber Ausreißern)

1.4.2 Streuungsmaße

- **Spannweite** $\max_i x_i - \min_i x_i$
- **Sichprobenvarianz** s^2: `var(x)`
  $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$
  Gemittelte Summe der quadratischen Abweichungen vom Mittelwert
- **Stichprobenstandardabweichung** s: `sd(x)`
  $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  Streuungsmaß mit gleicher Einheit wie beobachtete Daten x_i

Beispiel 1.4.1

Gegeben: Notenverteilung

Note i	h_i
1	2
2	4
3	3
4	4
5	3
6	0
	$n = 16$

Berechnen Sie Mittelwert, Median und Modalwert und außerdem die Stichprobenvarianz.

Beobachtungswerte $\{\underset{\substack{\shortparallel \\ x_1}}{1}, 1, 2, 2, 2, 2, 3, \underset{\substack{\shortparallel \\ x_8}}{3}, 3, 4, 4, 4, 4, 5, 5, \underset{\substack{\shortparallel \\ x_{16}}}{5}\}$

$$\bar{x} = \frac{1}{16}\left(2\cdot 1 + 4\cdot 2 + 3\cdot 3 + 4\cdot 4 + 3\cdot 5\right) = 3.125$$

$$x_{0.5} = \frac{1}{2}\left(x_8 + x_9\right) = 3$$

$$x_{mod} \in \{2, 4\}$$

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

$$= \frac{1}{15}\left(2\cdot 1^2 + 4\cdot 2^2 + 3\cdot 3^2 + 4\cdot 4^2 + 3\cdot 5^2 - 16\cdot 3.125^2\right)$$

$$= 1.85$$

1.4.3 p-Quantile

p-**Quantil** x_p $(0 < p < 1)$: `quantile(x,p)`
Teilt die sortierten Daten x_i (ungefähr) im Verhältnis $p : (1-p)$,
d. h. $\widehat{F}(x_p) \approx p$

Typ 2: R-Funktion `quantile(x,p,type=2)`

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

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

Der Quartilsabstand $x_{0.75} - x_{0.25}$ ist ein weiterer Streuungsparameter.

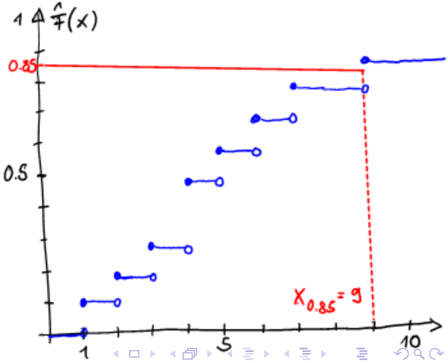
Beispiel 1.4.2

Berechnen Sie für die Beobachtungsdaten $\{1, 2, 3, 4, 4, 5, 6, 7, 9, 10\}$ das $x_{0.5}$-Quantil und das $x_{0.85}$-Quantil nach der Typ 2-Definition.

$n = 10$, $\text{floor}(10 \cdot 0.85) = 8$

$x_{0.5} = \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(4+5) = 4.5$

$x_{0.85} = x_{\text{floor}(8.5)+1} = x_9 = 9$



Boxplot

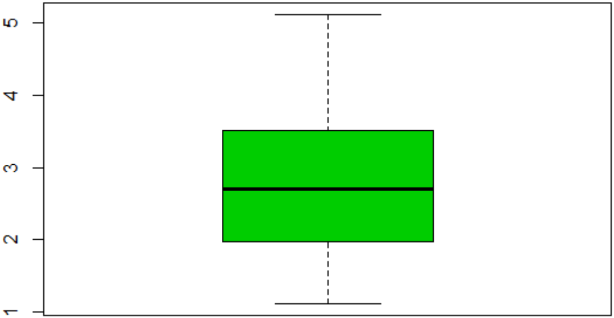


Abbildung: Abschlussnoten `boxplot(x)`

- Der grüne Bereich umfasst die mittleren 50% der Daten zwischen 1. und 3. Quartil.
- Der dicke schwarze Strich markiert den Median.
- Die "Antennen" geben die Spannweite an.

1.5 Ungleichung von Chebyshev

Satz: Ungleichung von Chebyshev
Sei $\bar{x}$ der Durchschnitt und $s > 0$ die Stichproben-Standardabweichung von Beobachtungswerten $x_1, \ldots, x_n$.
Sei $S_k = \{i, 1 \leq i \leq n : |x_i - \bar{x}| < k \cdot s\}$ und $N(S_k)$ die Anzahl der Elemente in der Menge S_k.
Dann gilt die Ungleichung:

$$\frac{N(S_k)}{n} > 1 - \frac{1}{k^2} \quad \text{für alle} \quad k \geq 1$$

In Worten:
Für eine beliebige reelle Zahl $k \geq 1$ liegen mehr als $100 \cdot (1 - \frac{1}{k^2})$ Prozent der Daten im Intervall von $\bar{x} - ks$ bis $\bar{x} + ks$.
Speziell:
Für $k = 2$ liegen mehr als 75% der Daten im $2s$-Bereich um $\bar{x}$.
Für $k = 3$ liegen mehr als 89% der Daten im $3s$-Bereich um $\bar{x}$.

Bemerkung zu Chebyshev:
Die Ungleichung liefert nur eine sehr grobe Abschätzung, ist aber unabhängig von der Verteilung der Daten.

Empirische Regeln für näherungsweise nnormalverteilte" Daten (d.h. das zugehörige Histogramm nimmt seine Maximalstelle beim Median an und verläuft links und rechts davon symmetrisch glockenförmig nach unten):

1. Ungefähr 68% der Daten liegen im Bereich um $\bar{x} \pm s$.
2. Ungefähr 95% der Daten liegen im Bereich um $\bar{x} \pm 2s$.
3. Ungefähr 99.7% der Daten liegen im Bereich um $\bar{x} \pm 3s$.

Beispiel 1.5

Ein Bäcker weiß, dass die durchschnittlich verkaufte Anzahl an Croissants an einem Sonntag bei 500 liegt bei einer Stichproben-Standardabweichung von 50.

a) Geben Sie eine Abschätzung an, wie viel Prozent der Verkaufszahlen zwischen 400 und 600 liegen.

b) Wie groß müsste die Standardabweichung sein, damit dieser Prozentsatz mindestens 90% beträgt?

Gegeben: $\bar{x} = 500$, $s = 50$

$$[400, 600] = [\bar{x} - 2s, \bar{x} + 2s] \Longrightarrow k = 2$$

a) $$\frac{N(S_2)}{n} \underset{\text{Chebyshev}}{\geq} 1 - \frac{1}{k^2} = 0.75$$

b) Bedingung: $$1 - \frac{1}{k^2} = 0.9 \iff 0.1 = \frac{1}{k^2}$$

$$\overset{k>1}{\Longrightarrow} k = \sqrt{10}$$

$$[400, 600] = [500 - \sqrt{10}\,s, 500 + \sqrt{10}\,s]$$

$$\Longrightarrow \sqrt{10}\,s = 100$$

$$\Longrightarrow s = 10\sqrt{10} \approx 31.6$$

1.6 Korrelation

Grafische Darstellung des Zusammenhangs zwischen multivariaten Daten x und y durch ein Streudiagramm:

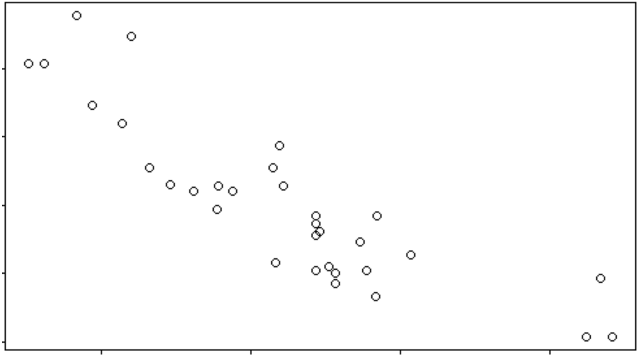


Abbildung: `plot(x,y), cov(x,y) = -5.116685, cor(x,y) = -0.8676594`

Kennzahlen zur Untersuchung des Zusammenhangs:

- **Empirische Kovarianz** s_{xy}: `cov(x,y)`
  $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}\right)$
  Für $s_{xy} > 0$ hat Punktewolke steigende, für $s_{xy} < 0$ fallende Tendenz.

- **Empirischer Korrelationskoeffizient** r: `cor(x,y)`

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

  Näherungsweise lin. Zusammenhang zw. x und y, falls $|r| \approx 1$.

- **Regressionsgerade** $y = mx + t$ mit

$$m = r \cdot \frac{s_y}{s_x} \text{ und } t = \bar{y} - m \cdot \bar{x}$$

Beispiel 1.6

Gegeben: Platzierung und Körpergewicht von Marathonteilnehmern

Läufer i	Platzierung (x_i)	Gewicht (y_i)
1	4	83
2	1	82
3	3	48
4	2	62
5	5	45

Berechnen Sie den Zusammenhang von Platzierung und Körpergewicht.
Wie lässt sich dieser Zusammenhang interpretieren?

$$\bar{x} = 3 \quad , \quad \bar{y} = 64$$

$$s_x^2 = \frac{1}{4}\left(16+1+9+4+25-5\cdot 3^2\right) = 2.5$$

$$s_y^2 = \frac{1}{4}\left(83^2+82^2+48^2+62^2+45^2-5\cdot 64^2\right) = 326.5$$

$$s_{xy} = \frac{1}{4}\left(4\cdot 83 + 82 + 3\cdot 48 + 2\cdot 62 + 5\cdot 45 - 5\cdot 3\cdot 64\right)$$
$$= -13.25$$

$$r = \frac{-13.25}{\sqrt{2.5\cdot 326.5}} \approx -0.464$$

$|r| \ll 1 \implies$ kein linearer Zusammenhang

$r < 0$, d.h.
bis auf den Außreißer
$(4, 83)$ wächst tendenziell
mit besserer (kleinerer)
Platzierung x_i das Körpergewicht y_i.