

# 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**

$\Omega$ : Grundgesamtheit (z.B. Studienbewerber Rosenheim WS 16/17)

$\omega$ : 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" ( $\geq 30$  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: Höhe  $\hat{=}$

Häufigkeit

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

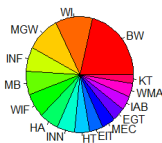


Abbildung: Absolute und relative Häufigkeiten  $h_i$  u.  $f_i$

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

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:** Abschlussnoten 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:**

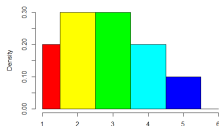
$n$  sei Anzahl der Ausprägungen

- Möglichst gleiche Breite  $\Delta x_i$  (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  $\hat{=}$  Dichte (Häufigkeit pro Klassenbreite)
- Fläche  $\hat{=}$  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  $\hat{F}(x) = \sum_{i: x_i \leq x} f_i$

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

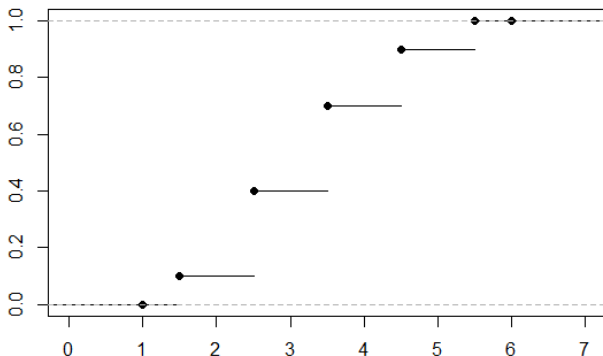


Abbildung: `plot(ecdf(x))`

## 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**  $\max_i x_i - \min_i x_i$

- **Stichprobenvarianz**  $s^2$ :  $\text{var}(\mathbf{x})$

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

Gemittelte Summe der quadratischen Abweichungen vom Mittelwert

- **Stichprobenstandardabweichung**  $s$ :  $\text{sd}(\mathbf{x})$

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

Streuungsmaß mit gleicher Einheit wie beobachtete Daten  $x_i$



# Beispiel



## 1.4.3 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.  $\hat{F}(x_p) \approx p$

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

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

- 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



# Boxplot

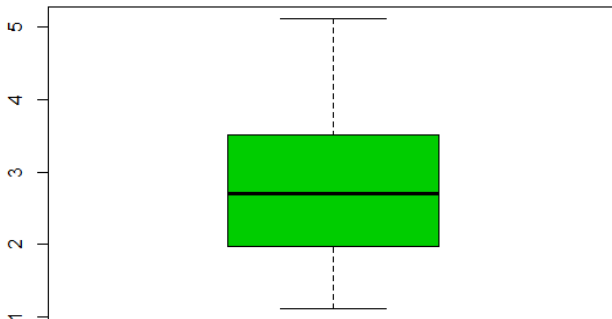


Abbildung: Abschlussnoten `boxplot(x)`

- 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





## 1.6 Korrelation

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

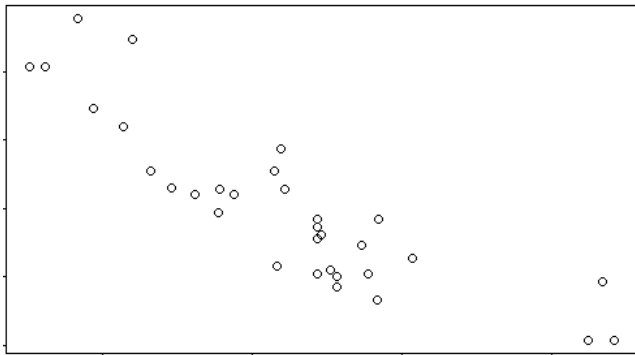


Abbildung: `plot(x,y)`,  $\text{cov}(x,y) = -5.116685$ ,  $\text{cor}(x,y) = -0.8676594$

## Kennzahlen zur Untersuchung des Zusammenhangs:

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

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

Für  $s_{xy} > 0$  hat Punktwolke steigende, für  $s_{xy} < 0$  fallende Tendenz.

- **Empirischer Korrelationskoeffizient**  $r$ :  $\text{cor}(x, y)$

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

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

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

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