Schweizerische Studienstiftung: Intellectual Tool

Essentielle Konzepte der Statistik

Prof. Dr. Bernhard Schipp

Balsthal, März 2016

Übersicht

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

1. Basiswissen Statistik

Grundlagen

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Struktur einer **Beobachtungsstudie**
  - Grundgesamtheit: Menge aller Individuen/Untersuchungsobjekte.
    $\hookrightarrow$ Stichprobenziehung: Mehrere Verfahren möglich.
  - erzeugt Stichprobe, im Idealfall repräsentativ.
    $\hookrightarrow$ Informationsaggregation von Merkmalen/Variablen
  - Überblick anhand deskriptiver Statistiken bzw. graphischer/numerischer explorativer Datenanalyse.
  - Basis zur inferentiellen Statistik, welche Rückschlüsse auf die Grundgesamtheit zulässt.

Grundlagen

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Struktur eines **geplanten statistischen Experiments**:
  - Festlegung der Grundgesamtheit (Population), worauf bezieht sich die Studie?
    - ↪ Stichprobenziehung, Sicherstellung der Repräsentativität durch Auswahlverfahren.
    - ↪ Randomisierte Gruppenzuweisung (verschiedene Möglichkeiten)
    - ↪ Verblindung (reduziert systematische Fehler)
  - generiert i.d.R. homogene Behandlungsgruppen (Kontrollgruppe, Gruppe A, Gruppe B,...). von Interesse: Effekte der Gruppen (Faktoren) auf die interessierenden Zielgrößen.
    - ↪ Datenmaterial
  - erlaubt Überblick anhand deskriptiver Statistiken und graphischer/numerischer explorativer Datenanalyse.
  - dient als Basis der inferentiellen Statistik
  - kann helfen, Kausalitäten zwischen Faktoren und Response-Variablen zu erarbeiten.

Grundlagen

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Variablen und Datentypen
  - Variablen $:= [A, X]$.
    Realisationen $:= [a_1, ...a_n, x_1, ...x_n]$.

$\hookrightarrow$ Diskret: Zähldaten wie Anzahl der Arztbesuche, Nachkommen
$\hookrightarrow$ Stetig: Körpergröße, Zeitdauer, Konzentrationsfähigkeit

$\hookrightarrow$ Qualitativ (Kategorial), nicht-numerische Ausprägungen, immer diskret.
$\hookrightarrow$ Quantitativ, numerische Ausprägungen, diskret oder stetig

$\hookrightarrow$ Nominal: z.B. Geschlecht, Blutgruppe können nicht angeordnet werden.
$\hookrightarrow$ Ordinal: Rangordnungen, können angeordnet werden, Differenzenbildung macht aber i.d.R. keinen Sinn, z.B. Schulnoten
$\hookrightarrow$ Intervall: Differenzenbildung macht Sinn, Verhältnisbildung nicht, z.B. Zeit, Temperatur
$\hookrightarrow$ Verhältnis: Intervalldaten mit absolutem Nullpunkt, z.B. Preise

Grundlagen

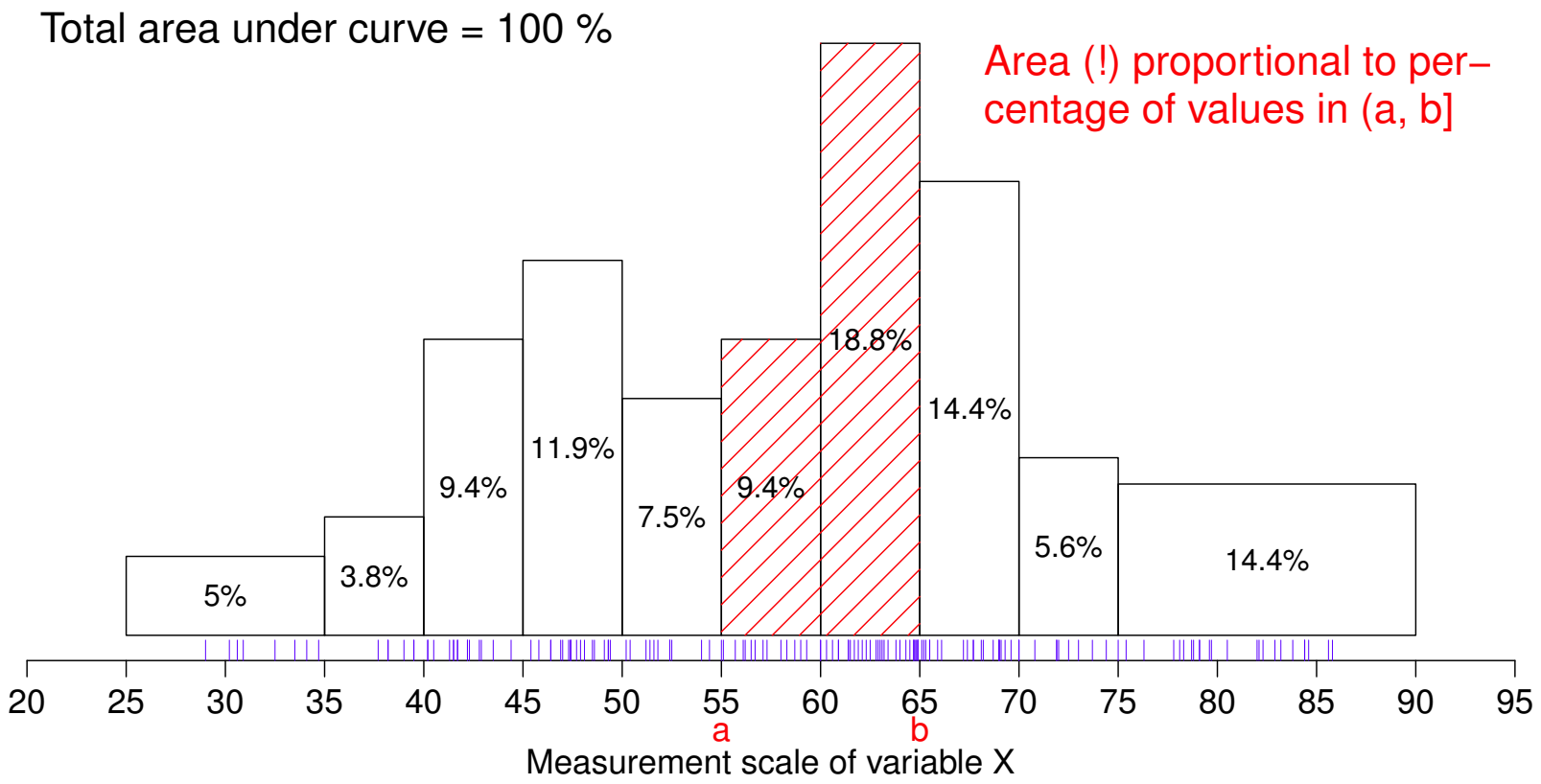
1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Von Interesse: Verteilung der Daten: mögl. Ausprägungen und deren proportionales Auftreten:
  - $\hookrightarrow$ Stichprobenverteilung (empirische Verteilung) $\rightarrow$ bekannt.
  - $\hookrightarrow$ Populationsverteilung $\rightarrow$ unbekannt, muss oft durch eine Annahme festgelegt werden oder wird aus den Daten geschätzt.
- Darstellungsmethoden: Tabellen, Grafiken

Univariate Analyse

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

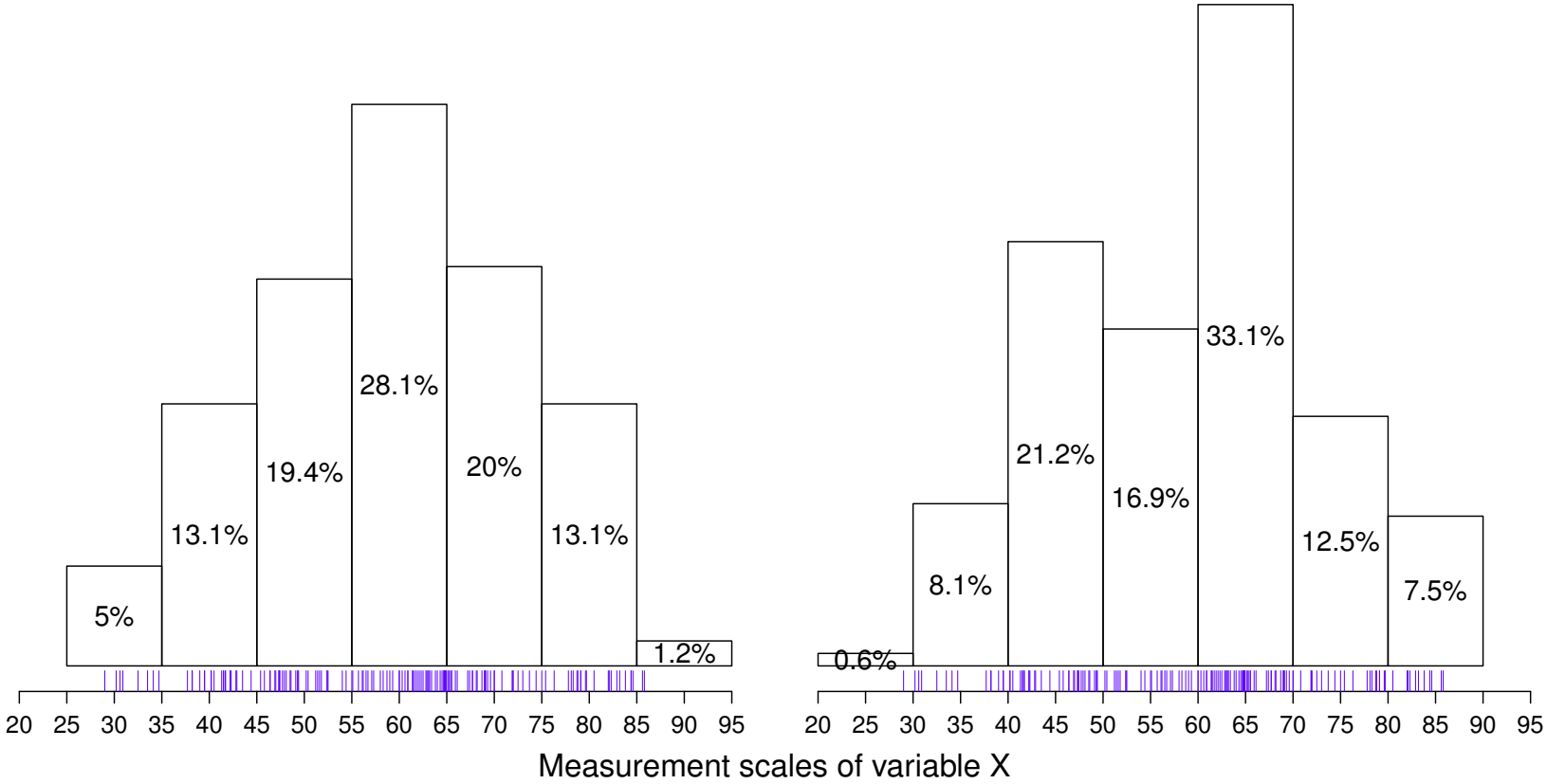
- Relative Häufigkeitsverteilung (Histogramm) der Daten (blau) nach Klassifizierung in Gruppen.



Histogramm

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Histogramme reagieren sensibel auf die (subjektive gewählte) Anzahl, Position und Breite der Klassen.



Deskriptive Kenngrößen

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Lageparameter der Stichprobe $x_1, ..., x_n$
  - Arithmetischer Mittelwert

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

  - Median

$$\tilde{x} := \begin{cases} \tilde{x} := x_{(0.5(n+1))} & \text{falls } n \text{ ungerade} \\ \tilde{x} := 0.5(x_{(0.5n)} + x_{(0.5n+1)}) & \end{cases}$$

  - Modus: Häufigster Wert

Deskriptive Kenngrößen

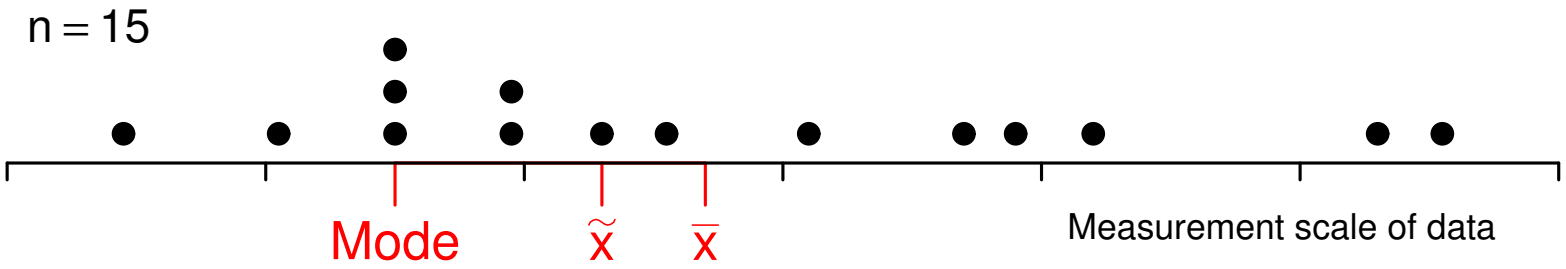
1. Basiswissen Statistik
- Grundlagen
- Univariate Analyse
- Bivariate Analyes
- Korrelation
- Rangkorrelation
- Bayes-Theorem
- Kontingenztafeln

2. Epidemiologie

3. Statistische Inferenz

Anhang: Wichtige Verteilungen

- Graphische Illustration:



- Schranken geordneter Daten:

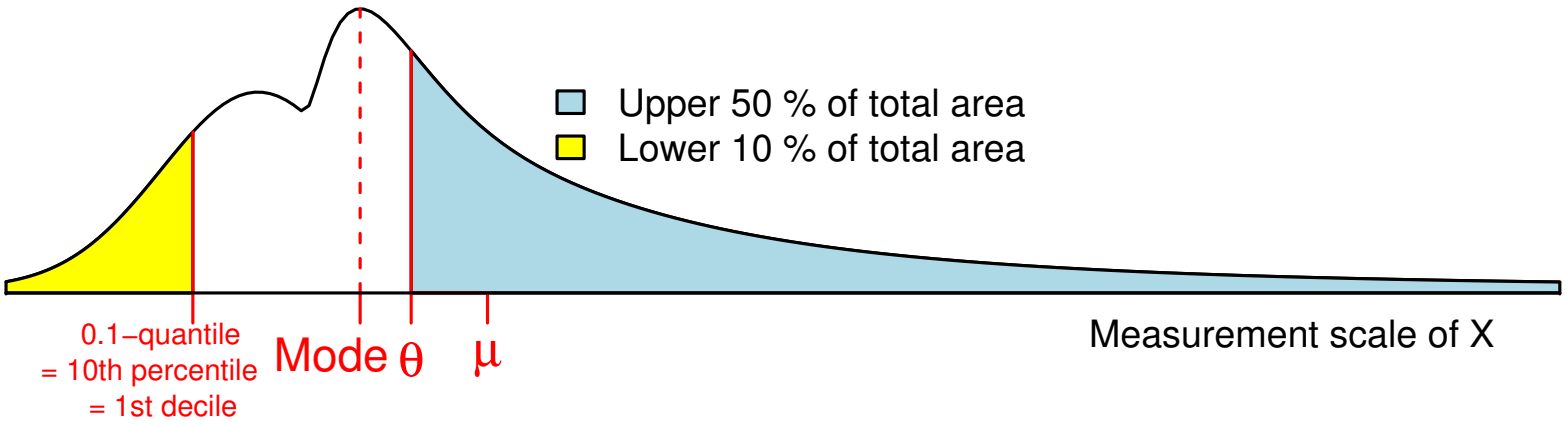
α-Quantile	teilt die Daten in	beliebige Anteile
Perzentile	”	Hunderstel
Quartile	”	Viertel
insbesondere:		
	$Q_{0.25}$	Unteres Quartil
	$Q_{0.75}$	Oberes Quartil

- Median = 0.5-Quantil usw.

Kenngrößen der Grundgesamtheit

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Lageparameter der Grundgesamtheit:
  - Erwartungswert $E(X) = \mu$
  - Median θ mit $F_X(x \leq \theta) = 0.5$
  - Modus (Maximalstelle der Dichtefunktion der Grundgesamtheit)
- Graphische Illustration



Deskriptive Kenngrößen

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Ein gern gemachter Fehler:
  - Gegeben: $\bar{x}$ der Stichprobe $x_1, ..., x_n$ und $\bar{y}$ der Stichprobe $y_1, ..., y_m$.

  $\hookrightarrow$ Wie lautet der Mittelwert der gemeinsamen Stichprobe $x_1, ..., x_n, y_1, ..., y_n$?

  $\hookrightarrow$ Ist es der Mittelwert $\bar{z}*$ der Mittelwerte $\bar{x}$ und $\bar{y}$? I.A. nicht, da

$$\bar{z} := \frac{1}{n+m}\left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i\right) = \frac{n\bar{x} + m\bar{y}}{n+m}$$

  $\hookrightarrow$ Es gilt nur $\bar{z} = 0.5(\bar{x} + \bar{y}) =: \bar{z}*$ falls $m = n$.

  - Beispiel: $n = 2, m = 10$

x_i	2.5	3									$\bar{x} = 2.75$
y_i	1.1	1.2	0.9	1.0	1.1	0.8	1.2	0.9	1.0	0.9	$\bar{y} = 1.01$
											$\bar{z}* = 1.88$
											$\bar{z} = 1.3$

Deskriptive Kenngrößen

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Ein weiterer gern gemachter Fehler:
  - Beobachtung des konsekutiven Wachstums

Populationsgröße Start: 110	Wachstumsrate per Periode (%)	Wachstumsfaktor per Periode
nach Periode 1: 132	$r_1 = +20 \quad \%$	$f_1 = 1 + \frac{r_1}{100} = \frac{6}{5}$
nach Periode 2: 165	$r_2 = +25 \quad \%$	$f_2 = 1 + \frac{r_2}{100} = \frac{5}{4}$
nach Periode 3: 110	$r_3 = -33\frac{1}{3} \ \%$	$f_3 = 1 + \frac{r_3}{100} = \frac{2}{3}$

$\hookrightarrow$ Wie lautet die durchschnittliche Wachstumsrate per Periode?

$\hookrightarrow$ $3^{-1}(r_1 + r_2 + r_3) = 3.89\%$ pro Periode ist offensichtlich grober Unfug.

Deskriptive Kenngrößen

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Lösung: Nutze den geometrischen Mittelwert der Wachstumsfaktoren, um die durchschnittliche Wachstumsrate zu erhalten.

  i) Berechne das geometrische Mittel der Wachstumsfaktoren:

$$\bar{f}^{(geom)} = \sqrt[n]{f_1 \cdot \ldots \cdot f_n}$$

  ii) Der „richtige" Durchschnitt der Wachstumsrate ist sodann:

$$r^* = \left(\bar{f}^{(geom)} - 1\right) \cdot 100\ \%$$

- Beispiel (Fortsetzung): $\bar{f}^{(geom)} = \sqrt[3]{\frac{6}{5} \cdot \frac{5}{4} \cdot \frac{2}{3}} = 1 \Rightarrow r^* = 0\ \%$

Deskriptive Kenngrößen: Streuung

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Empirische Variationsmaße:

  i) Standardabweichung

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

  ii) Spannweite (geordnete Stichprobe $x_{(1)} < \ldots < x_{(n)}$)

$$R := x_{(n)} - x_{(1)}$$

  iii) Interquartilsabstand (geordnete Stichprobe $x_{(1)} < \ldots < x_{(n)}$)

$$IQR := Q_{0.75} - Q_{0.25}$$

  wobei Q_α das α-Quantil mit $\alpha \in [0, 1]$ ist.

Deskriptive Kenngrößen: Streuung

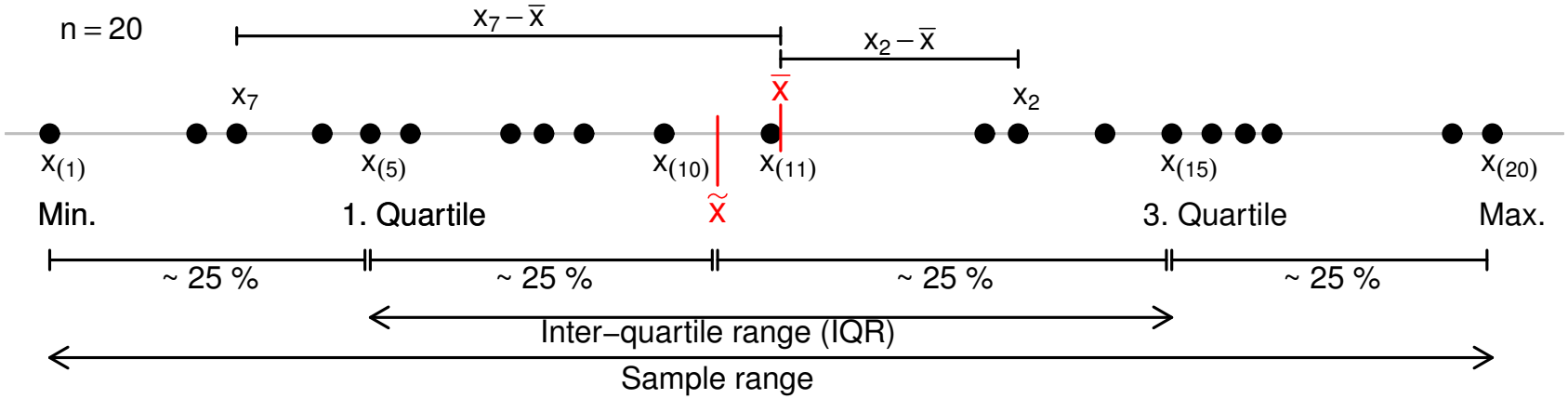
1. Basiswissen Statistik
- Grundlagen
- Univariate Analyse
- Bivariate Analyes
- Korrelation
- Rangkorrelation
- Bayes-Theorem
- Kontingenztafeln

2. Epidemiologie

3. Statistische Inferenz

Anhang: Wichtige Verteilungen

- Graphische illustration:



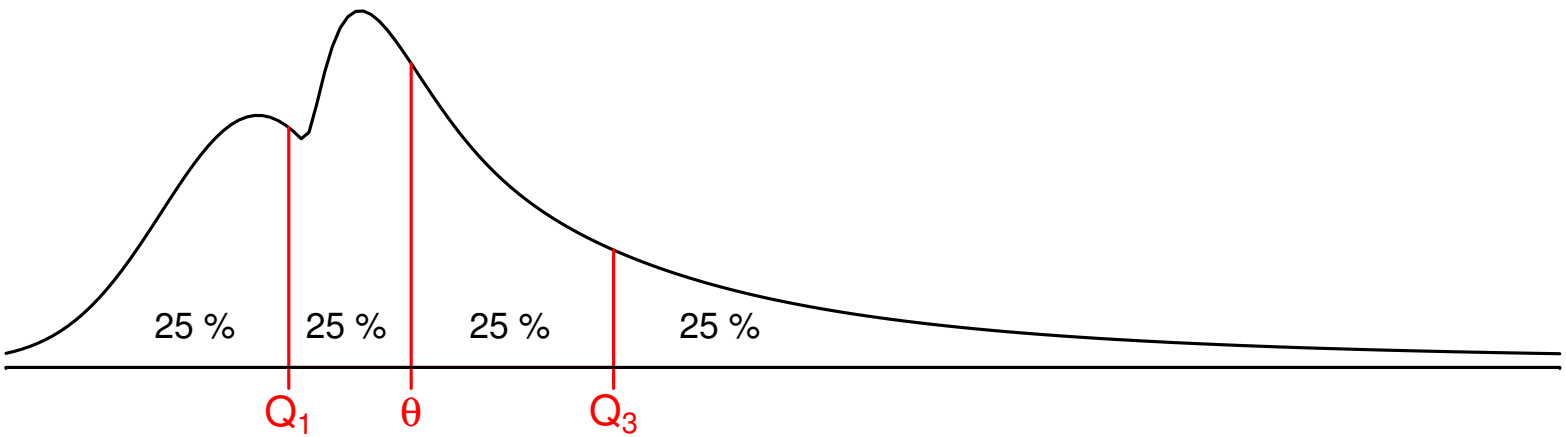
Streuungsmaße in der Grundgesamtheit

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Variationsmaße der Population:
  - i) Standardabweichung/Varianz: $\sigma = ?$ (unbekannt), $\sigma^2 = Var(X)$
  - ii) Spannweite: Existiert nur, falls die Populationswerte eine obere und untere Grenze aufweisen.
  - iii) Interquartilsabstand:

$$IQR_P := \theta_{0.75} - \theta_{0.25}$$

wobei $F_X(x \leq \theta_\alpha) = \alpha$ und $\alpha \in [0, 1]$.



Variationskoeffizient

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Variationskoeffizient (Streuung relativ zu arithm. Mittelwert)
- ist definiert durch

$$vc = \frac{s}{\bar{x}} \cdot 100(\%)$$

Interpretation (Faustregel):

- $(0 < vc \leq 20)$: arithm. Mittel präzise
- $(20 < vc \leq 40)$: arithm. Mittel moderat präzise
- $(40 < vc \leq 60)$: arithm. Mittel eher ungenau
- $(60 < vc)$: arithm. Mittel extrem ungenau

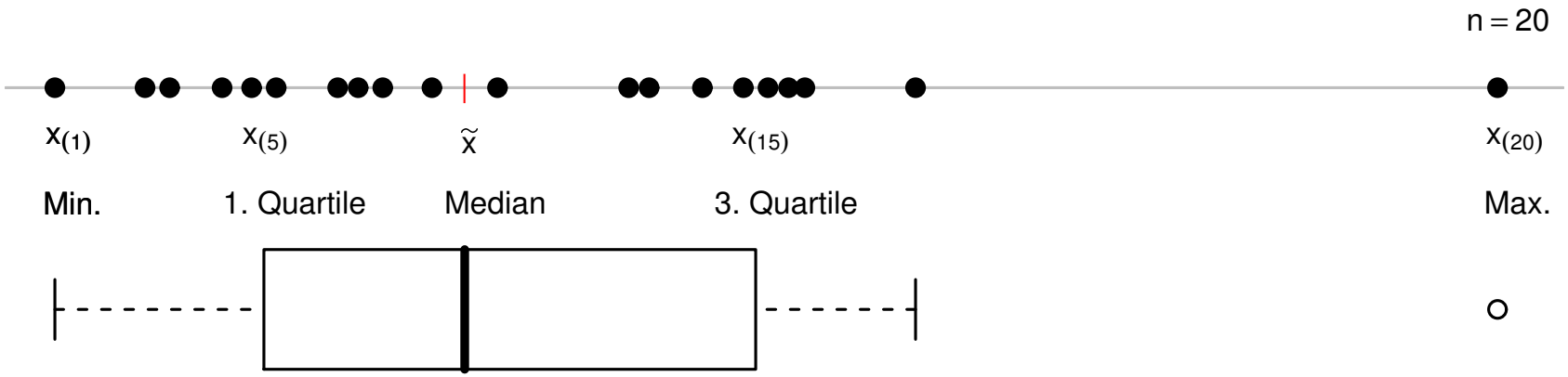
Box-Plot

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Boxplots für Stichprobendaten:
  - i) Nutze die „5 Kenngrößen”

    Minimum, $Q_{0,25}$, Median, $Q_{0,75}$, Maximum
  - ii) Ausreißeridentifikation:

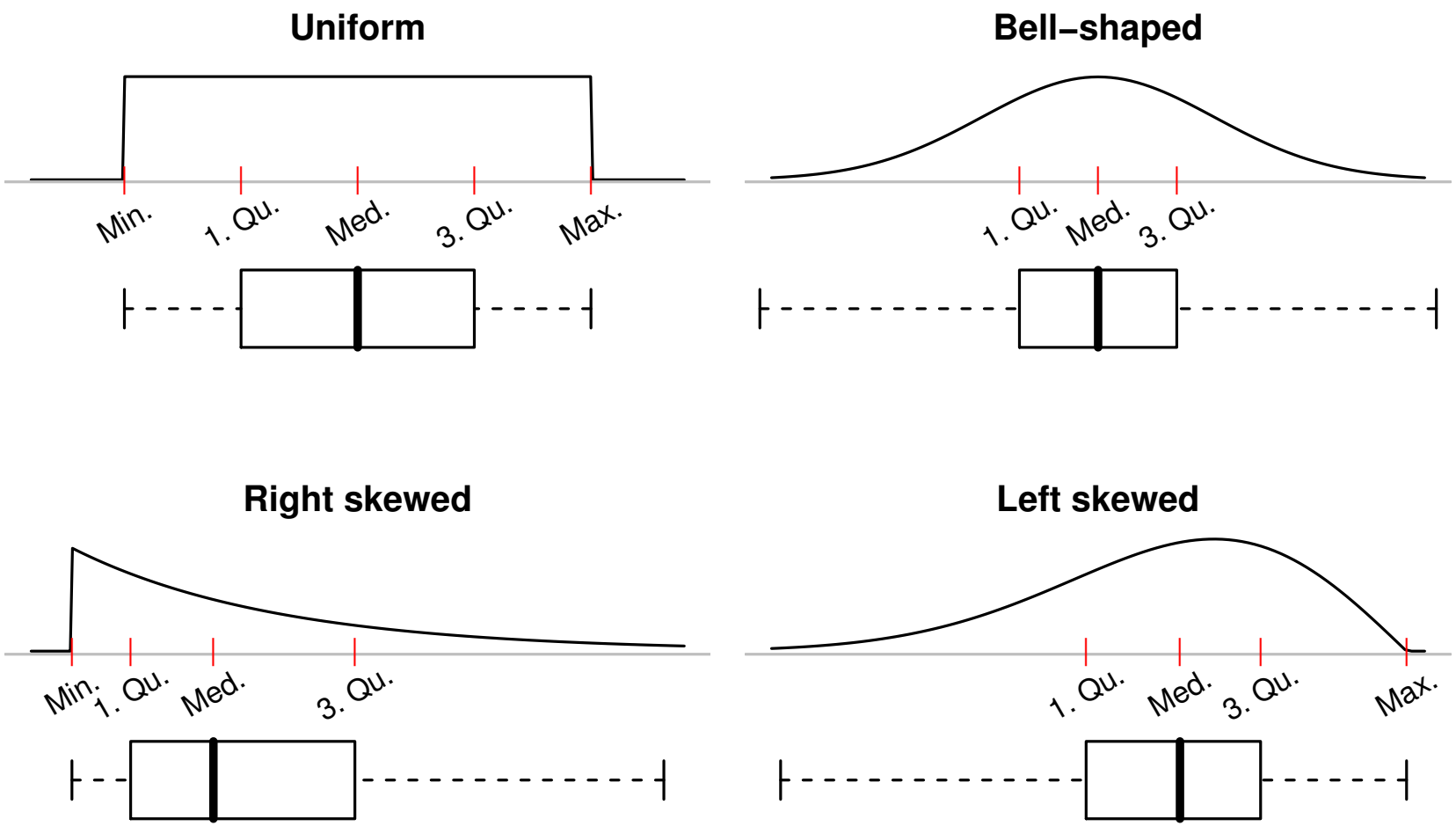
    $x_i > Q_{0,75} + 1.5 \cdot \text{IQR}$ oder $x_i < Q_{0,25} + 1.5 \cdot \text{IQR}$
  - iii) Boxplotkonstruktion:



Box-Plot

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Boxplots und Verteilungen:

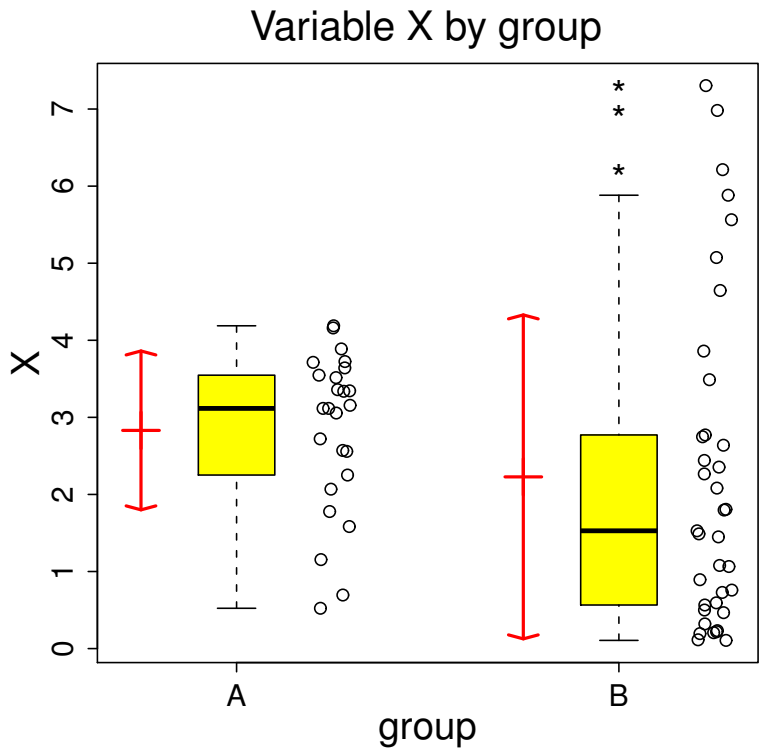


Box-Plot

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Für jede der Gruppen A, B

- Boxplots, gelb und mittig.
- $\bar{x} + s$, rot und links.
- Rohdaten im Dot-Chart-Format.



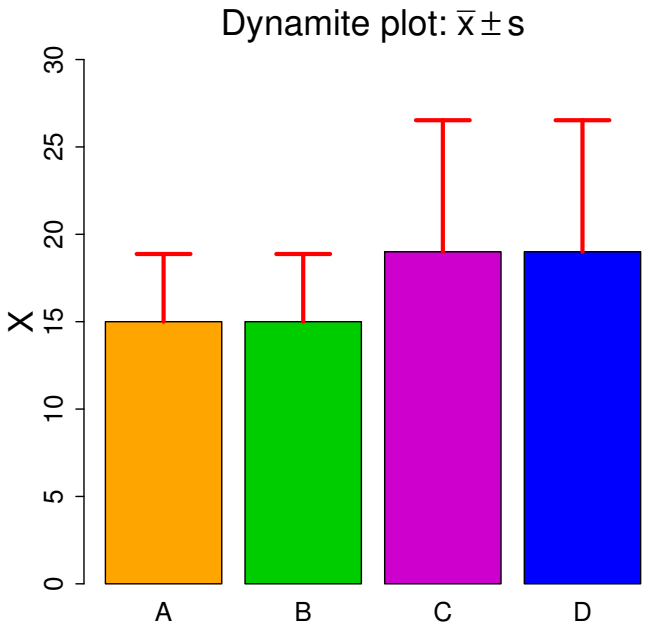
Empfehlung: Überlagere Dot-Chart und Boxplot

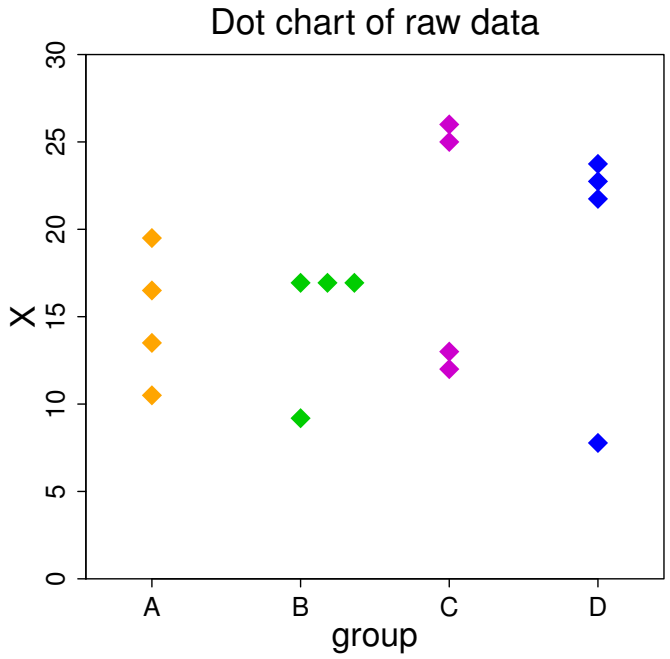
- Warnung: $\bar{x} \pm s$ verbirgt wichtige Informationen insb. bei kleinen Datenmengen oder schiefer Verteilung.

Bad Graphs

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Es gibt viele ungeeignete Darstellungen: Illustration





- Offensichtlich gilt $\bar{x}_A = \bar{x}_B$ und $s_A = s_B$, ebenso für die Gruppen C und D.

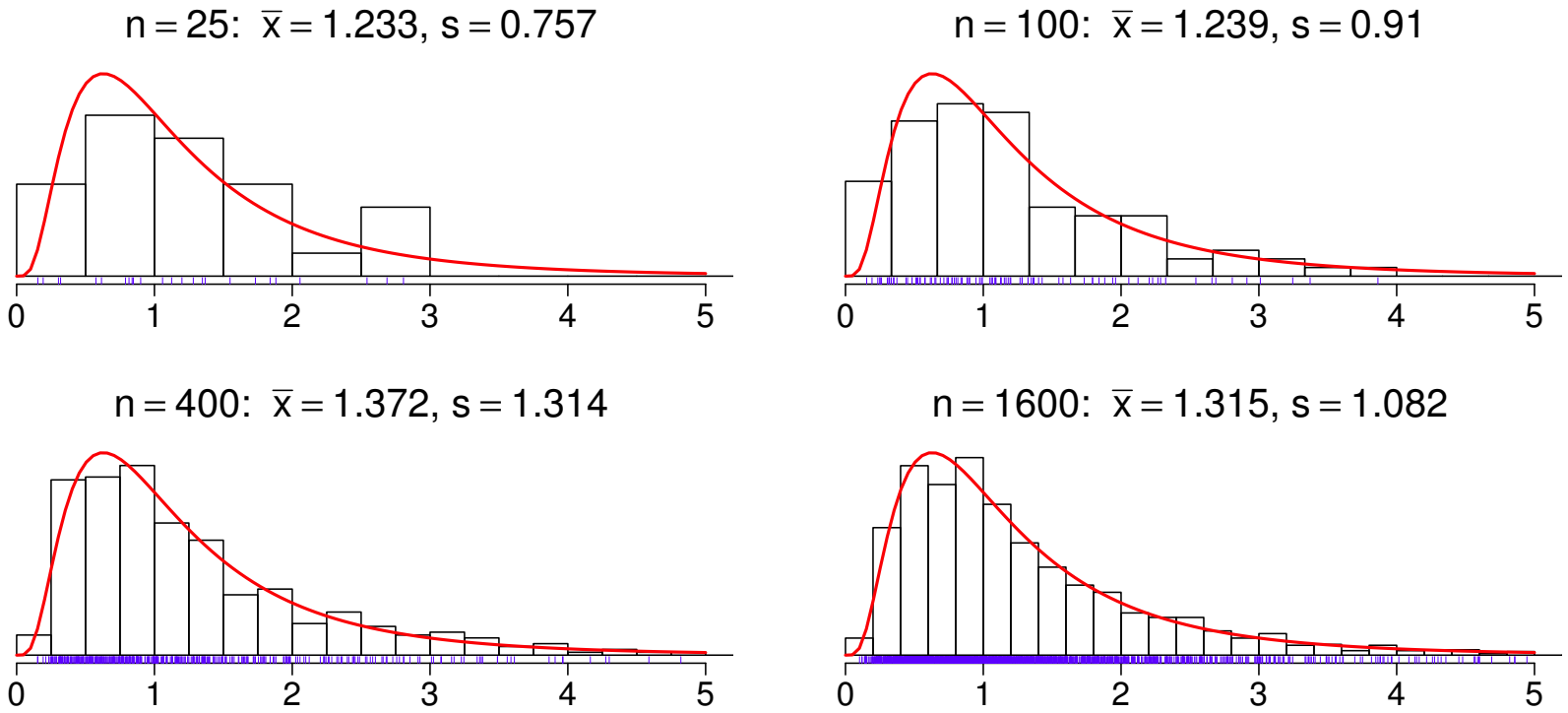
$\hookrightarrow$ Der Dot-Chart rechts zeigt die tatsächliche Verteilungssituation. Siehe auch: web links zu „Bad graph contest“.

Univariate Analyse

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Die Stichprobenverteilung approximiert für großes n die unbekannte Verteilung der Grundgesamtheit.
- Je größer die Stichprobe desto besser ist die Approximation (Ausnahmen exisitieren auch hier).

Simulated sampling with increasing n: convergence of histograms & descriptive statistics with $\mu = 1.3$ and $\sigma = 1$



Bivariate Analyse

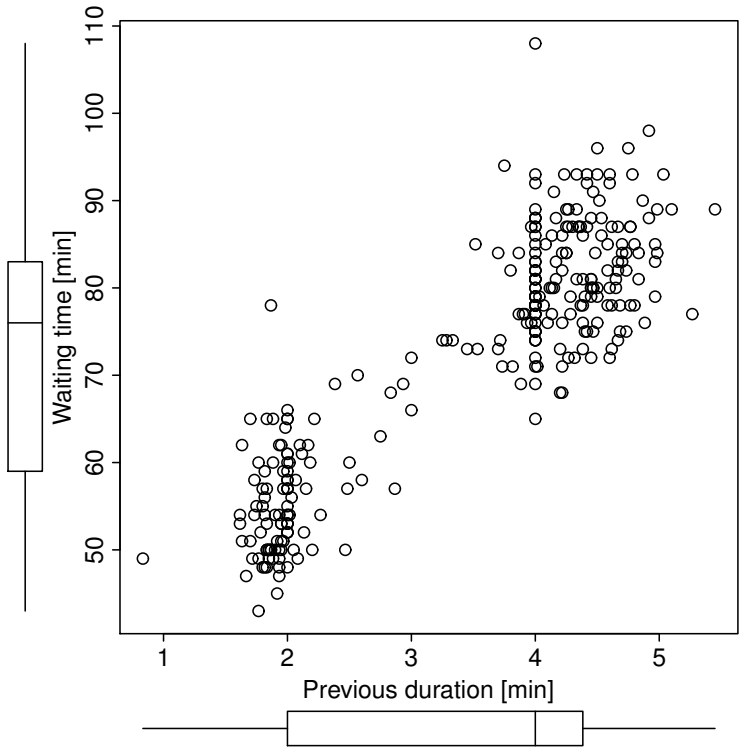
1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Verteilung der Paare $(x_1, y_1), ..., (x_1, y_1)$: Der Scatterplot

= gemeinsame Verteilung der Paare (x_i, y_i).

= Beschreibung der Streuung der Punkte (x_i, y_i) im 2D-Raum.

Hier noch mit den jeweiligen 1D-Boxplots erweitert.



Empfehlung: Überlagere Dot-Chart und Boxplot

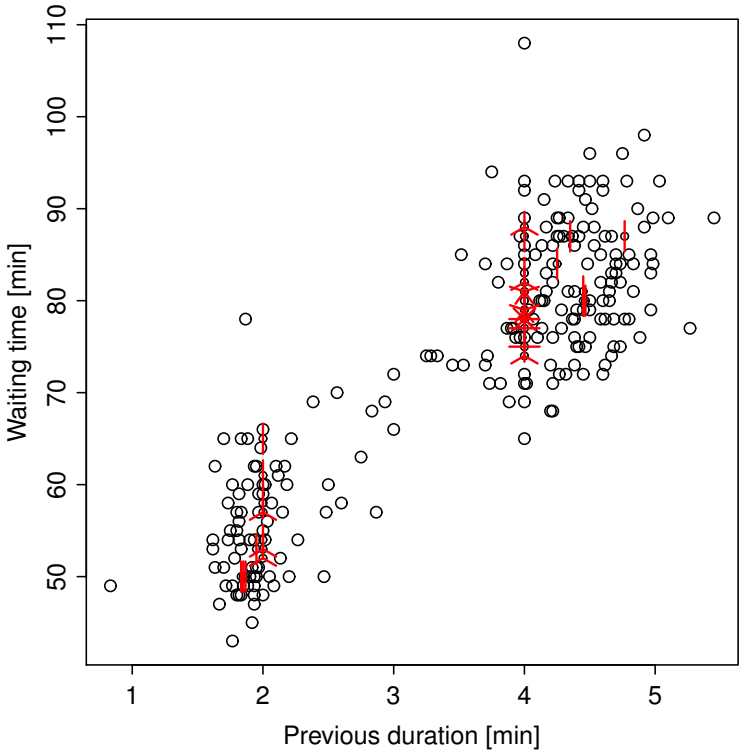
Bivariate Grafiken

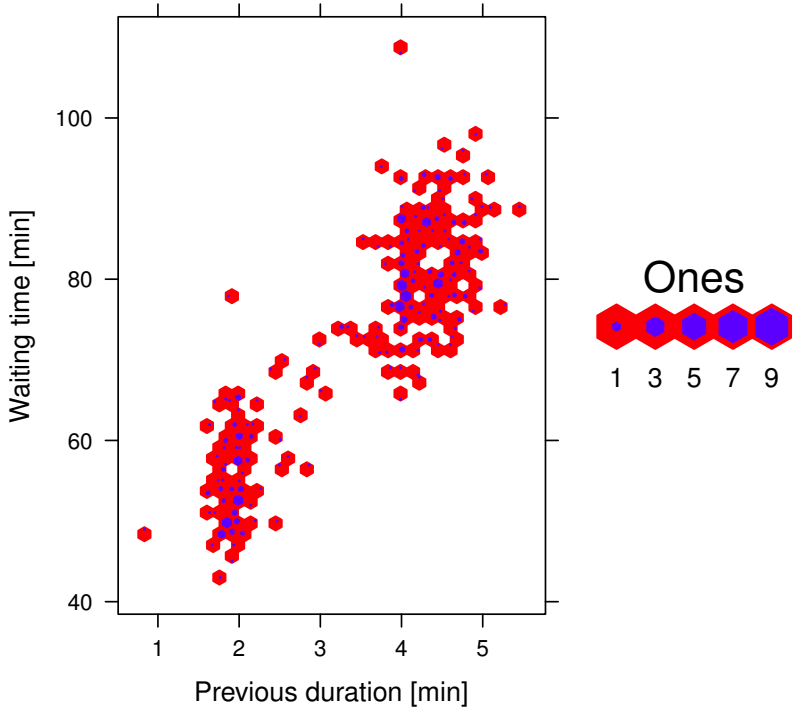
1. Basiswissen Statistik
- Grundlagen
- Univariate Analyse
- Bivariate Analyes
- Korrelation
- Rangkorrelation
- Bayes-Theorem
- Kontingenztafeln

2. Epidemiologie

3. Statistische Inferenz

Anhang: Wichtige Verteilungen





- Links: „Sunflowerplot" um doppelt belegete Punkte zu visulisieren. Rechts: Mehrfachbelegung durch hexagonale Zellenstruktur visualisiert.

Bivariate Grafiken

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Eigenschaften von Interesse:
  - i) Die Gestalt bivariater Verteilung kann analog zu den univariaten Verteilungen klassifiziert werden, zeigt aber i.A. komplexeres Verhalten bezüglich der Modalitäten und Asymmetrien.
  - ii) Der Interessenfocus liegt generisch auf einfachen Eigenschaften der Punktewolke:
    - Lage
    - Gerichtete Streuung
    - Statistische Abhängigkeit (Kovarianz, Korrelation)
    - Funktionsartige Strukturen (Im Sinne von Trends z.B. Regression von y auf x).

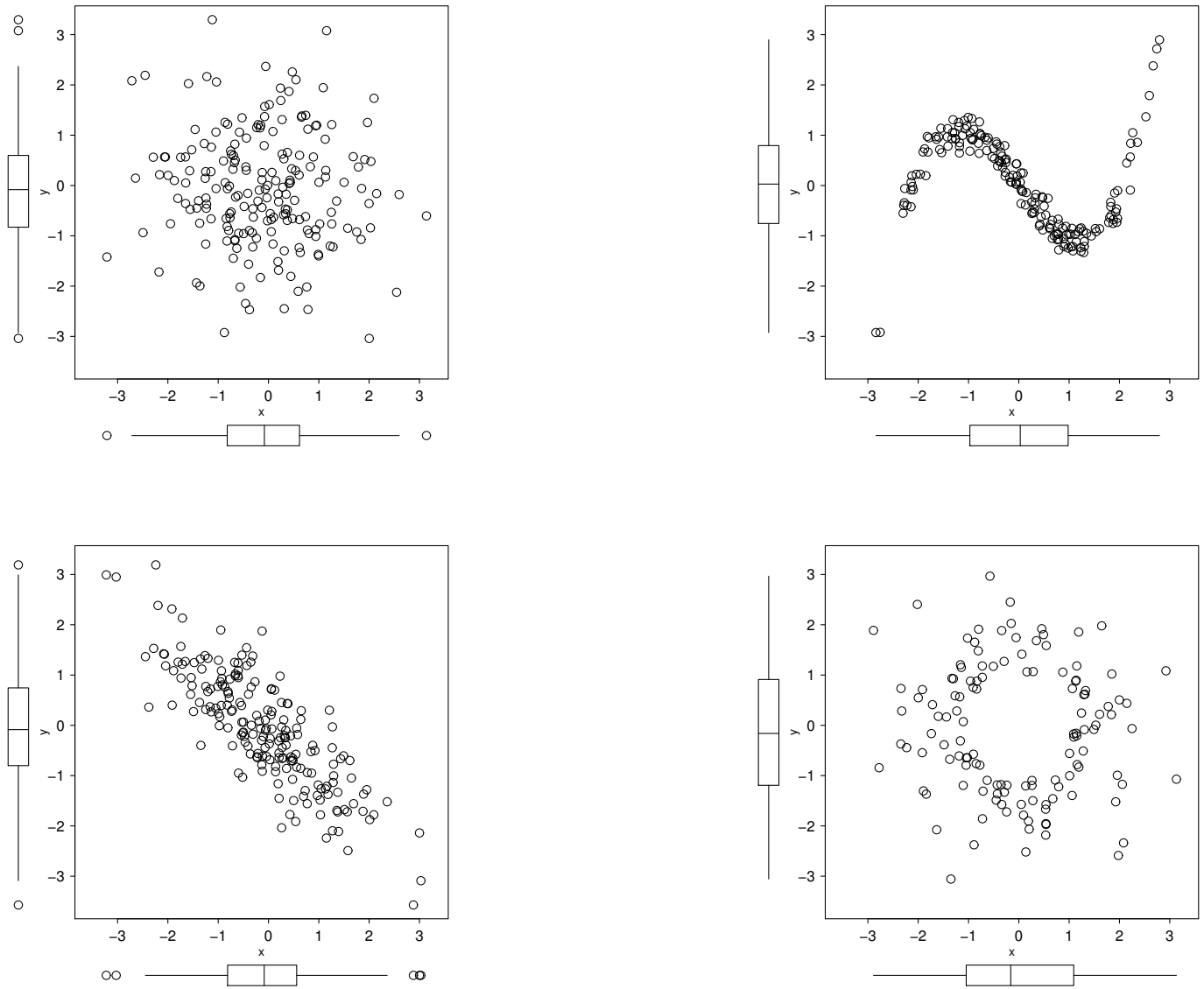
Bivariate Grafiken

1. Basiswissen Statistik
- ❖ Grundlagen
- ❖ Univariate Analyse
- ❖ Bivariate Analyes
- ❖ Korrelation
- ❖ Rangkorrelation
- ❖ Bayes-Theorem
- ❖ Kontingenztafeln

2. Epidemiologie

3. Statistische Inferenz

Anhang: Wichtige Verteilungen



Bivariate Analyse: Korrelation

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Korrelation: Beschreibt eine Beziehung zwischen zwei oder mehreren Merkmalen, Ereignissen oder Zuständen, die nicht notwendigerweise ein Kausalzusammenhang darstellen muss.
- Für metrische Daten exisitieren unterschiedliche Typisierungen.

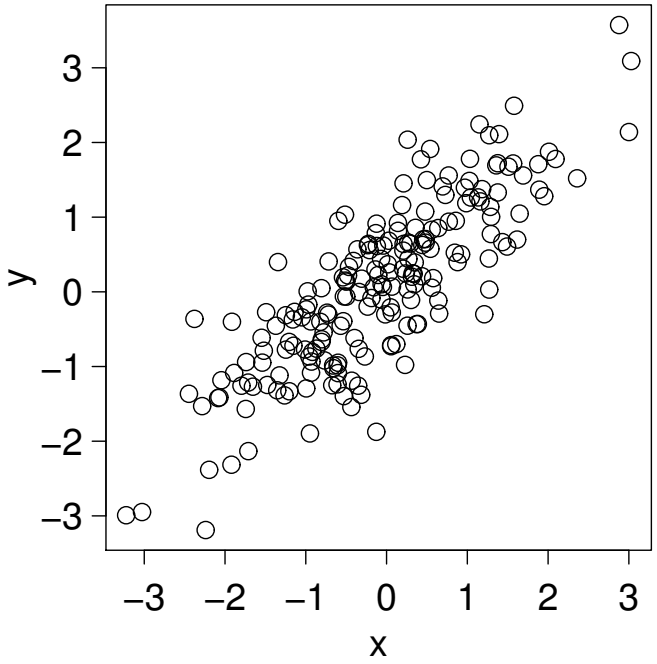


Abbildung 1: Linear

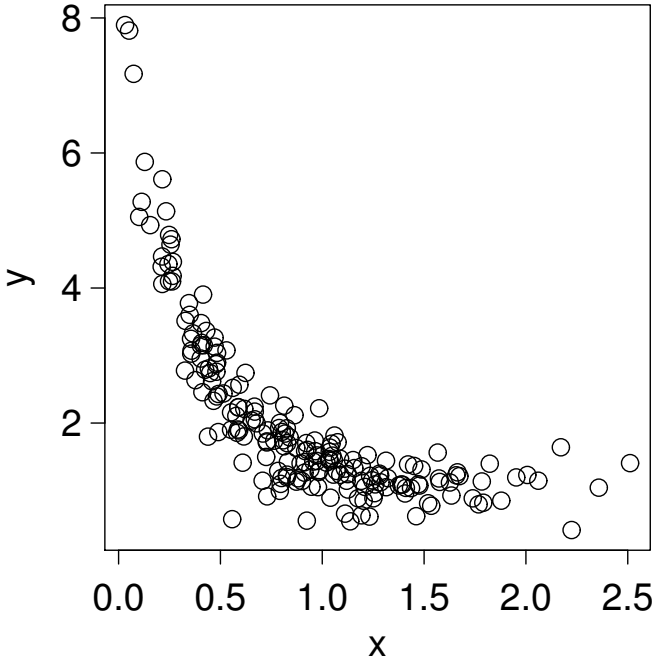


Abbildung 2: Monoton

Bivariate Analyse: Korrelation

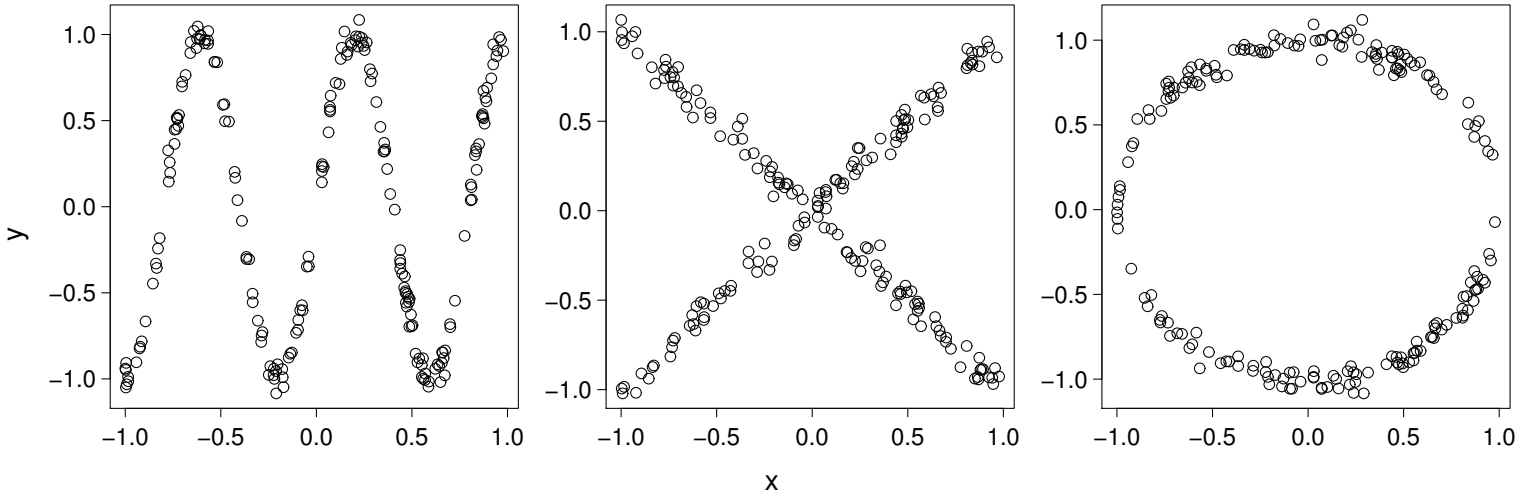
1. Basiswissen Statistik
- ❖ Grundlagen
- ❖ Univariate Analyse
- ❖ Bivariate Analyes
- ❖ Korrelation
- ❖ Rangkorrelation
- ❖ Bayes-Theorem
- ❖ Kontingenztafeln

2. Epidemiologie

3. Statistische Inferenz

Anhang: Wichtige Verteilungen

- Oder eher pathologische Beispiele



Bivariate Analyse: Korrelation

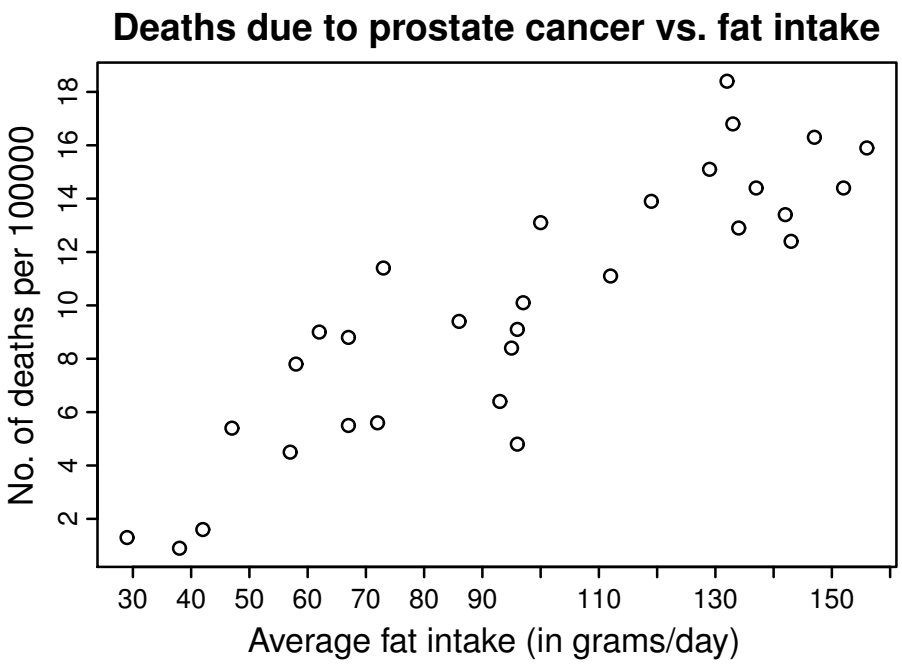
1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Beispiel:

  $n = 30$ Länder
  x Täglicher Fettverzehr
  y Tod durch Prostatakrebs

- Ziel:

  Ein Maß der (linearen) Abhängigkeit zwischen den Größen x und y.



- Lösung: Korrelationskoeffizient nach Karl Pearson:

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Bivariate Analyse: Korrelation

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- $-1 \leq r_P \leq +1$ (Skaleninvariant, Einheitenfrei)
- Je mehr sich $|r_P|$ der 1 nähert desto eher liegen die Punkte (x_i, y_i) auf einer „imaginären" Geraden.
  - $r_P = 1$ alle Punkte liegen auf einer steigenden Geraden.
  - $r_P = -1$ alle Punkte liegen auf einer fallenden Geraden.

  $\hookrightarrow$ Funktional deterministische Beziehung zwischen y und x. Praktisch unmöglich für reale Datensituationen.
- r_P sagt nichts über die Steigung der „imaginären" Geraden aus.

  $\hookrightarrow$ Dies ist die grundsätzliche Fragestellung innerhalb der Regressionsanalyse.
- $r_P = 0$ zeigt nur die Abwesenheit einer linearen statistischen Beziehung zwischen den Größen y und x auf.

Bivariate Analyse: Korrelation

1. Basiswissen Statistik
- ❖ Grundlagen
- ❖ Univariate Analyse
- ❖ Bivariate Analyes
- ❖ Korrelation
- ❖ Rangkorrelation
- ❖ Bayes-Theorem
- ❖ Kontingenztafeln

2. Epidemiologie

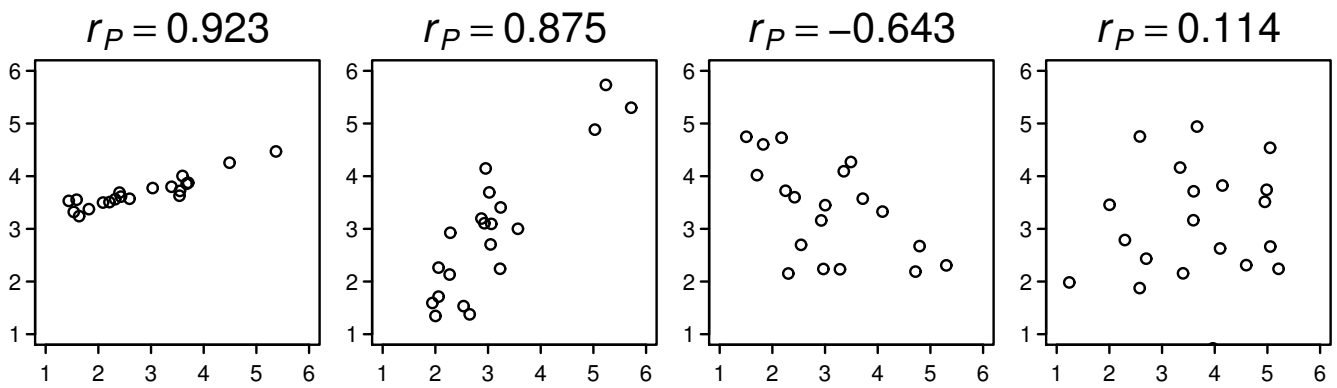
3. Statistische Inferenz

Anhang: Wichtige Verteilungen

- Einstufungen (Faustregel!)

$$\left.\begin{array}{rcccl} & & |r_P| & = & 0 & : & \text{keine} \\ 0 & < & |r_P| & \leq & 0.5 & : & \text{schwache} \\ 0.5 & < & |r_P| & \leq & 0.8 & : & \text{mittlere} \\ 0.8 & < & |r_P| & < & 1 & : & \text{starke} \\ & & |r_P| & = & 1 & : & \text{perfekte} \end{array}\right\} \text{(Pearson) Korrelation}$$

- Beispiele: Korrelationen mit unterschiedlicher Größenausprägungen



- Beachten Sie: Die Steigung der ersten Geraden ist kleiner als die der Zweiten, obwohl die Korrelation im ersten Plot deutlich größer ist.

Bivariate Analyse: Korrelation

1. Basiswissen Statistik
- ❖ Grundlagen
- ❖ Univariate Analyse
- ❖ Bivariate Analyes
- ❖ Korrelation
- ❖ Rangkorrelation
- ❖ Bayes-Theorem
- ❖ Kontingenztafeln

2. Epidemiologie

3. Statistische Inferenz

Anhang: Wichtige Verteilungen

- $r_P = 0$ schließt keine nicht-linearen Beziehungen aus:

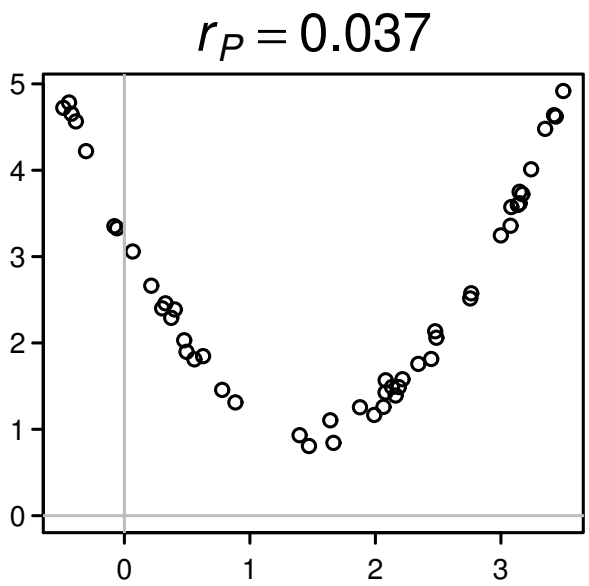


Abbildung 3: Standard

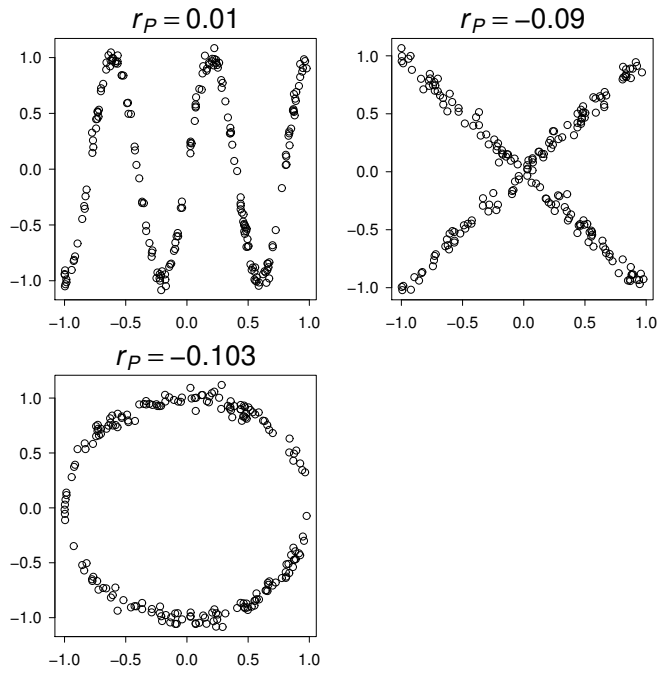


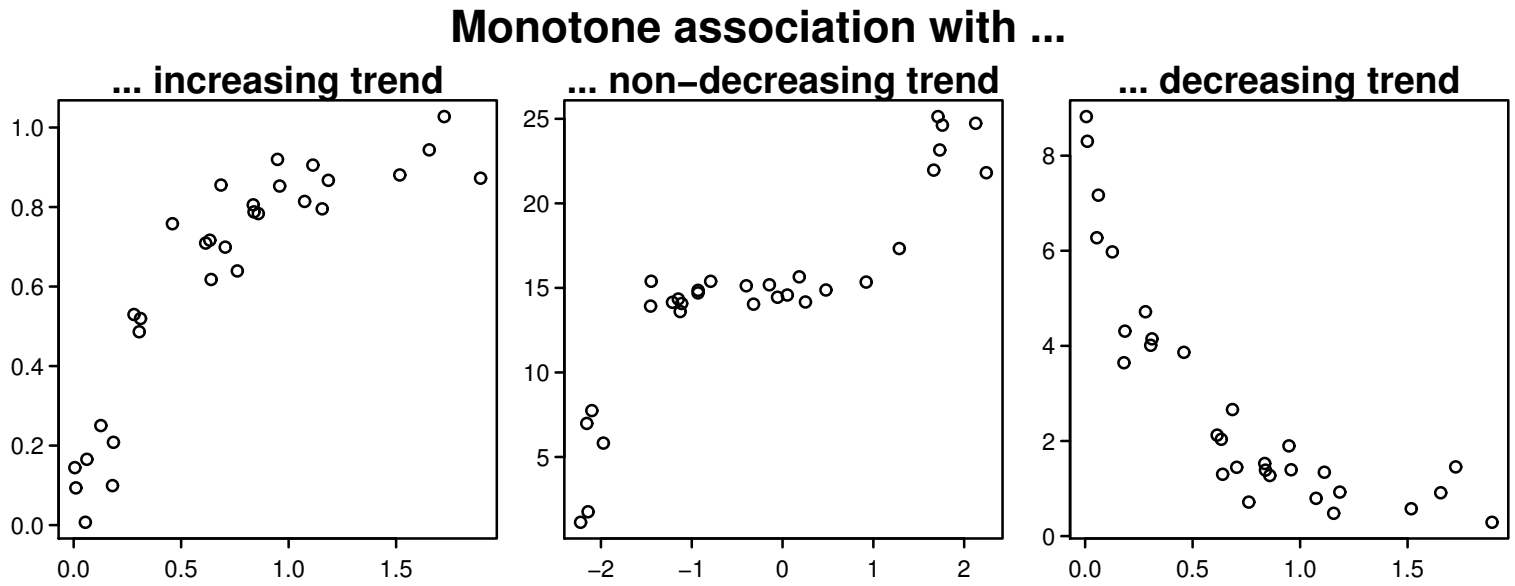
Abbildung 4: Pathologisch

↪ Merke: Erstellen Sie, falls möglich, einen Scatterplot, um nicht-lineare Beziehungen zwischen den Größen y und x zu entdecken.

Bivariate Analyse: Rangkorrelation

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Problem: Messung der Stärke einer monotonen Beziehung zwischen metrisch bivariaten Daten (Monotonie ist eine schwächere Eigenschaft als Linearität).
- Grafische Illustration:



Bivariate Analyse: Rangkorrelation

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Lösung:
  - ↪ Nicht die eigentlichen Datenpaare (x_i, y_i) werden genutzt, sondern die damit assoziierten Rangpaare $(R(x_i), R(y_i))$.
  - ↪ die Ränge werden separat für x_i und y_i ermittelt: Das kleinste x_i bekommt $R(x_i) = 1$ zugewiesen, das zweitkleinste x_i bekommt $R(x_i) = 2$ zugewiesen etc. Das gleiche Verfahren erfolgt für y_i.
  - ↪ Im Fall gleicher Werte innerhalb der x_i Datenreihe (bzw. in y_i), also bei sog. „ties", wird dann ein mittlerer Rang vergeben.
  - ↪ Rangkorrelation nach Spearman:

$$r_S = \frac{\sum_{i=1}^{n} \left(R(x_i) - \frac{n+1}{2}\right)\left(R(y_i) - \frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n} \left(R(x_i) - \frac{n+1}{2}\right)^2} \cdot \sqrt{\sum_{i=1}^{n} \left(R(y_i) - \frac{n+1}{2}\right)^2}}$$

Bivariate Analyse: Rangkorrelation

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Eigenschaften von r_S
  - $\hookrightarrow$ $-1 \leq r_S \leq +1$
  - $\hookrightarrow$ Je näher $|r_S|$ an 1 liegt, desto eher folgen (x_i, y_i) einem monotonen Trend.
  - $\hookrightarrow$ $r_S = 1 \Leftrightarrow$ Perfekte positive Rangkorrelation.
  - $\hookrightarrow$ $r_S = -1 \Leftrightarrow$ Perfekte negative Rangkorrelation.
  - $\hookrightarrow$ $r_S = 0 \Leftrightarrow$ Keine monotone Beziehung
  - $\hookrightarrow$ Der Rangkorrelationkoeffizient sagt nichts (bis auf das Vorzeichen) über die Steigung des monotonen Trends aus.

Bivariate Analyse: Rangkorrelation

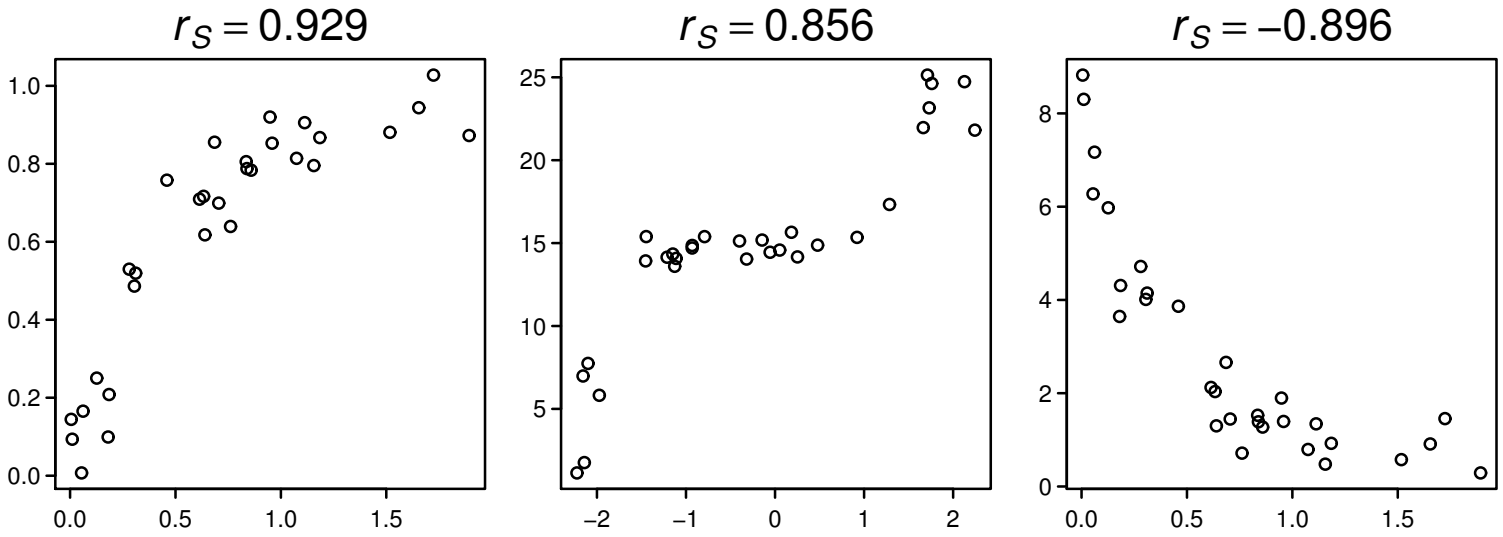
1. Basiswissen Statistik
- Grundlagen
- Univariate Analyse
- Bivariate Analyes
- Korrelation
- Rangkorrelation
- Bayes-Theorem
- Kontingenztafeln

2. Epidemiologie

3. Statistische Inferenz

Anhang: Wichtige Verteilungen

- Einstufungen (Faustregel!): Siehe r_P
- Die vorherigen Beispiele nun mit r_S



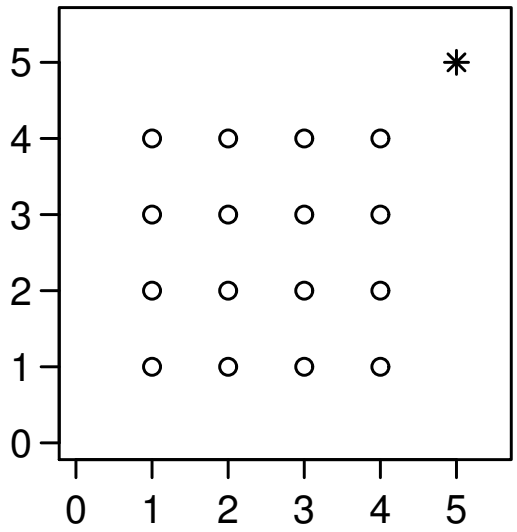
- $r_S = 0$ schließt eine nicht-monotone Beziehung nicht aus; machen Sie sich das anhand der pathologischen Beispiel klar.

Bivariate Analyse: Rangkorrelation

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- r_S reagiert robust gegenüber sog. Ausreißern.
- Beispiel:

16 Datenpunkte, $r_P = r_S = 0$



17. Datenpunkt hinzugefügt

point	$(5,5)$	$(10,10)$	$(20,20)$	$(50,50)$	$(100,100)$
r_P	0.2273	0.7258	0.9351	0.9908	0.9978
r_S	0.1753	0.1753	0.1753	0.1753	0.1753

Bedingte Wahrscheinlichkeit und Bayes-Theorem

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Gesucht: Wahrscheinlichkeit eines Ereignisses unter der Bedingung, dass ein anderes bereits eingetreten ist.
Formal:

$$P(A|B) = \frac{P(A \text{ und } B)}{P(B)}, P(B) > 0$$

Eigenschaften:

$P(A \text{ und } B) = P(A)P(B)$ falls A und B (stochastisch) unabhängig sind
$P(A \text{ und } B) = P(A) + P(B) - P(A \text{ oder } B)$
$P(A \text{ und } B) = P(A|B)P(B)$

Bedingte Wahrscheinlichkeit und Bayes-Theorem

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Seien A und B Ereignisse, mit den Gegenereignissen $\bar{A}$ und $\bar{B}$, dann gilt

$$P(B) = P(B \text{ und } A) + P(B \text{ und } \bar{A})$$

was gleichbedeutend ist mit

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

Damit erhält man den **Satz von Bayes**:

$$P(A|B) = \frac{P(A \text{ und } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

Kontingenztafel; Beispiel

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Beispiel: Ebola (Ebola-Disease); Sensitivität, Spezifität des Ebola-Tests und die Prävalenz von Ebola in einer Population seien bereits bekannt.

- Sensitivität: $P(T|K) = \dfrac{P(T \text{ und } K)}{P(K)} = 0.94$
- Spezifität: $P(\bar{T}|\bar{K}) = \dfrac{P(\bar{T} \text{ und } \bar{K})}{P(\bar{K})} = 0.93$
- Prävalenz (gegeben): $P(Ebola) = 0.000002$
- Bayes: $P(K|T) = \dfrac{P(T|K)P(K)}{P(T|K)P(K) + P(T|\bar{K})P(\bar{K})}$
- d.h. $P(K|T) = 0.000027$!!!
  und damit ist der Test ungeeignet.

Kontingenztafel; Beispiel

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Die Prävalenz ändert sich i.d.R. bei Verdachtsfällen (Stichprobe):

- Vierfeldertafel (two-way table): Verdachtsfälle von Ebola

	T	$\bar{T}$	
K	126	8	134
$\bar{K}$	7	97	104
	133	105	238

- P(Ebola bei Verdachtsfällen) $= \frac{134}{238} = 0.56$

- d.h. $P(K|T) = 0.95$!!!
  und damit ist der Test in Verdachtsfällen zur Absicherung durchaus geeignet.

Validität diagnostischer Verfahren

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Um zu beurteilen, wie gut ein Labortest für die Diagnose der Krankheit ist, wird bei jedem Patienten dessen tatsächlicher Gesundheitszustand mit dem Ergebnis des Tests verglichen. Dabei können vier mögliche Fälle auftreten:
  - ***Richtig positiv***: Der Patient ist krank und der Test hat dies richtig angezeigt.
  - ***Falsch negativ***: Der Patient ist krank, aber der Test hat ihn fälschlicherweise als gesund eingestuft.
  - ***Falsch positiv***: Der Patient ist gesund, aber der Test hat ihn fälschlicherweise als krank eingestuft.
  - ***Richtig negativ***: Der Patient ist gesund und der Test hat dies richtig angezeigt.

Im ersten und letzten Fall war die Diagnose also richtig, in den anderen beiden Fällen liegt ein Fehler vor.

Kontingenztafeln

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

I.d.R. müssen die Wahrscheinlichkeiten durch eine Datenerhebung geschätzt werden. Beispiel: ***absolute Häufigkeiten***

Personen	krank	gesund	Summe
Test positiv	98	40	138
Test negativ	1	1960	1961
Summe	99	2000	2099

Schätzung durch die ***relative Häufigkeiten***:

Personen	krank	gesund	Summe
Test positiv	98/2099=0.047	40/2099=0.019	138/2099=0.066
Test negativ	1/2099=0.0005	1960/2099=0.934	1961/2099=0.934
Summe	99/2099=0.047	2000/2099=0.953	1

Daraus kann z.B. ermittelt werden:
P(krank|Test positiv)=
P(krank und Test positiv)/P(Test positiv)=0.047/0.066=0.71
alternativ: P(krank|Test positiv)=
Anzahl(krank und Test positiv)/Anzahl(Test positiv)=98/138=0.71

Kontingenztafelanalyse

1. Basiswissen Statistik
❖ Grundlagen
❖ Univariate Analyse
❖ Bivariate Analyes
❖ Korrelation
❖ Rangkorrelation
❖ Bayes-Theorem
❖ Kontingenztafeln
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Bezeichnungen: T = Test positiv, K = Patient krank
  $\bar{T}$ = Test negativ, $\bar{K}$ = Patient nicht krank

Kontingenztafel:

Personen	krank $(r_p + f_n)$	gesund $(f_p + r_n)$
Test positiv $(r_p + f_p)$	richtig positiv (r_p)	falsch positiv (f_p)
Test negativ $(f_n + r_n)$	falsch negativ (f_n)	richtig negativ (r_n)

$\hookrightarrow$ ***Sensitivität***: Wie gut erkennt der Test Personen, die tatsächlich krank sind? $P(T|K)$, geschätzt durch $r_p/(r_p + f_n)$

$\hookrightarrow$ ***Spezifität***: Wie gut erkennt der Test Personen, die tatsächlich gesund sind? $P(\bar{T}|\bar{K})$, geschätzt durch $r_n/(r_n + f_p)$

$\hookrightarrow$ ***Positiver prädiktiver Wert***: $P(K|T)$, geschätzt durch $r_p/(r_p + f_p)$

$\hookrightarrow$ ***Negativer prädiktiver wert***: $P(\bar{K}|\bar{T})$, geschätzt durch $r_n/(r_n + f_n)$

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

2. Epidemiologie

Relatives Risiko (RR)

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Eine andere Form der Kontingenztafel ermöglicht eine Untersuchung des Zusammenhangs zwischen einem Ereignis (z.B. krank/Nicht krank) mit einer Exposition (z.B. Raucher/Nichtraucher)

- Das ***relative Risiko*** beschreibt dann die Erkrankungswahrscheinlichkeit einer exponierten Gruppe relativ zur Erkrankungswahrscheinlichkeit einer nicht exponierten Gruppe.
- RR = (Inzidenz der Exponierten) / (Inzidenz der Nichtexponierten)
- Die Inzidenz einer Erkrankung ist die Anzahl neuer Krankheitsfälle, die in einem bestimmten Zeitraum auftreten, bezogen auf die Bevölkerung mit gleichem Erkrankungsrisiko.

		Ereignis		
		Ja	Nein	
Exposition	Ja	a	b	a+b
	Nein	c	d	c+d

- Relatives Risiko: $\dfrac{a/(a+b)}{c/(c+d)}$

Relatives Risiko (RR)

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Beispiel:

		Lungenkrebs		
		Ja	Nein	
Raucher	Ja	90	9910	10000
	Nein	40	19960	20000

- Risiko der Raucher: 90 von 10000 oder 0.009%
- Risiko der Nichtraucher: 40 von 20000 oder 0.002%
- Verhältnis des Risikos der Raucher zu dem der Nichtraucher ist das relative Risiko
- Relatives Risiko: $[90/10.000]/[40/20000] = 4.5$
- ***Interpretation***: Raucher haben das 4.5-fache Risiko eines Nichtrauchers, Lungenkrebs zu bekommen.

RR: Einsatz und Interpretation

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Einsatz

- Das relative Risiko ist nur in prospektiven Studien (Kohortenstudien) eine zulässige Größe.
- In retrospektiven Case-Control-Studien ist es nicht zulässig, das relative Risiko zu berechnen.

Interpretation

- RR = 1 Das Risiko der Exponierten ist gleich groß wie das Risiko der Nichtexponierten
- RR $>$ 1 Das Risiko der Exponierten ist größer als das Risiko der Nichtexponierten (Exposition = Risikofaktor)
- RR $<$ 1 Das Risiko der Exponierten ist kleiner als das Risiko der Nichtexponierten (Exposition = Protektiver Faktor)

Chancenverhältnis: Odds Ratio (OR)

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Das OR beschreibt die Quote des Expositionsfaktors unter den Erkrankten relativ zur Quote unter den Gesunden
- OR = (Expositionsquote der Erkrankten) / (Expositionsquote der Gesunden)
- Die Expositionsquote ist das Verhältnis der Anzahl Exponierter zur Anzahl Nichtexponierter (unter den Erkrankten bzw. unter den Gesunden).
- Odds Ratio: $\frac{a/c}{b/d}$

Odds Ratio: Beispiel

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Beispiel:

		Lungenkrebs	
		Ja	Nein
		Fall(Case)	Kontrolle)
Raucher	Ja	60	50
	Nein	40	150

- Die Raucherquote unter den Lungenkrebspatienten beträgt $60 : 40$ oder $3 : 2$.
- Die Raucherquote unter den Gesunden beträgt $50 : 150$ oder $1 : 3$.
- Das Odds Ratio ist das Verhältnis der Raucherquote der Erkrankten zur Raucherquote der Gesunden.
- Numerisch: Odds Ratio = $(60/40)/(50/150) = 4.5$
- *Interpretation*: Die Chance, unter Lungenkrebspatienten einen Raucher anzutreffen, ist 4.5mal so hoch wie unter den Gesunden.

Zusammenhang: Relatives Risiko (RR) und Odds Ratio (OR)

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

- Das Odds Ratio kann sowohl in prospektiven Kohortenstudien wie auch in retrospektiven Case-Control-Studien berechnet werden.
- Das Odds Ratio dient häufig als Schätzung für das relative Risiko.
- Das Odds Ratio ist nicht dasselbe wie das relative Risiko.

↪ ***Achtung***: In Fall-Kontroll-Studien kann das relative Risiko gar nicht berechnet werden, da die Erkrankungshäufigkeit (Anzahl der Fälle) vom Untersucher vorgegeben ist.

Wann ist OR ein guter Schätzer für RR?

- Wenn die Fälle und Kontrollen im Hinblick auf die Expositionsanamnese repräsentativ sind für alle Menschen mit bzw. ohne diese Erkrankung in der Population.
- Wenn die untersuchte Krankheit selten auftritt.

RR vs. OR: Beispiel

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

		Erkrankung		
		Ja	Nein	
Exposition	Ja	200	9800	10000
	Nein	100	9900	10000

- Relatives Risiko: $\frac{200/10000}{100/10000} = 2$
- Odds Ratio: $\frac{200/100}{9800/9900} = 2.02$

Warum?

Das RR: $\frac{\frac{a}{a+b}}{\frac{c}{c+d}}$ ist annähernd gleich dem OR: $\frac{\frac{a}{b}}{\frac{c}{d}}$

... falls a sehr viel kleiner als b und c sehr viel kleiner als d ist.

RR vs. OR

1. Basiswissen Statistik
2. Epidemiologie
❖ Grundbegriffe
3. Statistische Inferenz
Anhang: Wichtige Verteilungen

Bei häufigen Krankheiten ist das Odds Ratio kein guter Schätzer für das relative Risiko.

Beispiel:

		Erkrankung		
		Ja	Nein	
Exponiert	Ja	50	50	100
	Nein	25	75	100

- Relatives Risiko: $[50/100]/[25/100] = 2$
- Odds Ratio: $[50/25]/[50/75] = 3$

PS: Die Erläuterungen zu RR und OR sind angelehnt an ein Schnellrepetitorium von E. Kvas, Hermesoft, Biostatistik und Data Management, office@hermesoft.at

1. Basiswissen Statistik

2. Epidemiologie

3. Statistische Inferenz
- Zufallsvariablen
- Verteilungen
- QQ-Plot
- Parameterschätzung
- Zentraler Grenzwertsatz
- Konfidenzintervalle
- Hypothesentests
- Stichprobenumfang
- Teststärke
- Wahl des Testverfahrens
- Anpassungstest
- Unabhängigkeitstest
- Parametrische Tests
- Nichtparametrische Tests
- Varianzanalyse
- Multiple Vergleiche
- Regression
- KQ-Residuen
- Modellgüte
- Standardfehler
- t-Tests
- Modellvarianten
- Box-Cox-Modelle
- Logistische Regression

3. Statistische Inferenz

Zufallsvariablen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- **Definition**:

  Eine **Zufallsvariable (ZV)** ist eine Größe deren Ausprägung durch den Zufall bestimmt wird.

  Interpretation: Verallgemeinert den generischen Variablenbegriff zu jedweder Art von Variable

  Notation: Üblicherweise in Großbuchstaben X, Y...

- **Beispiele**: ZV's als Output mathematischer Funktionen deren Input zufälliger Natur ist.

  i) $\log(C)$ wobei C Glukosekonzentration im Blut.

  ii) $BMI := M/H^2$ für Größe und Gewicht einer zufällig ausgewählten Person.

  iii) Mittelwert und Standardabweichung von n zufällig ausgewählten Personen.

Verteilungsfunktion

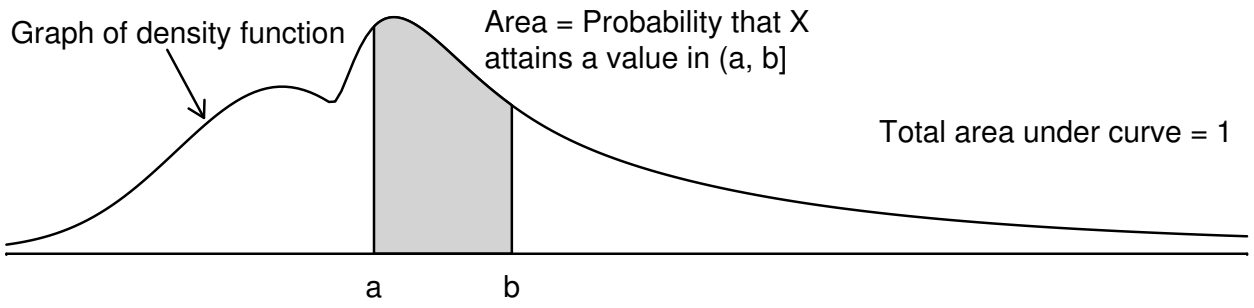
1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- **Definititon**:

  Die **Wahrscheinlichkeitsverteilung** $F_X(x)$ einer ZV X weist den Ausprägungen respektive den Ausprägungsintervallen besagter ZV eine Zahl zwischen 0 und 1 zu.

$\hookrightarrow$ **Diskrete ZV**: Nimmt nur abzählbar viele Zustände/Ausprägungen an. Das heißt, die Output-Menge ist durch $\mathbb{N}, \mathbb{Z}$ oder jede endliche Menge gegeben. Einzelne Werte können mit einer Wahrscheinlichkeit belegt werden

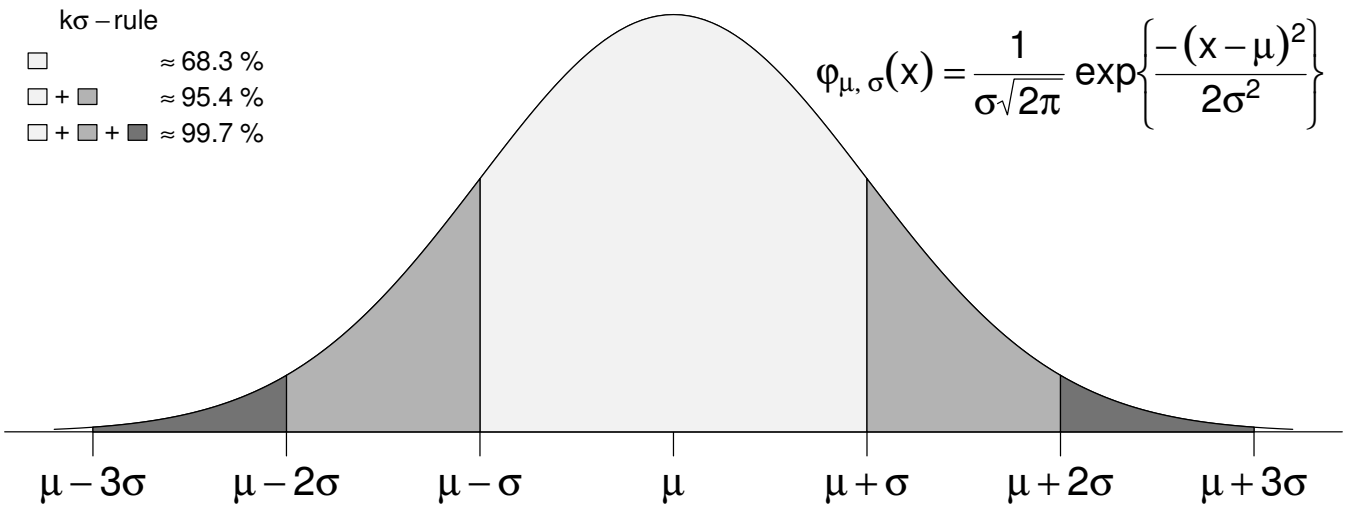
$\hookrightarrow$ **Stetige ZV**: Nimmt überabzählbar viele Zustände/Ausprägungen an. Das heißt, die Output-Menge ist durch $\mathbb{R}, \mathbb{C}$ gegeben. Nur Intervalle können mit einer Wahrscheinlichkeit belegt werden.



Normalverteilung

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Viele -aber nicht alle- realen Zufallsphänomene weisen eine **Dichtefunktion** mit der Gestalt einer Gauss'schen Dichte auf.



- Notation $X \sim N(\mu, \sigma^2)$
- $N(\mu, \sigma^2)$ ist eindeutig durch die Parameter bestimmt.
- $N(0, 1)$ heißt **Standardnormalverteilung**.

Normalverteilung

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- **Anmerkung**: Die Normalverteilung spielt innerhalb der statistischen Theorie eine große Rolle, da diese Verteilung oft per Annahme vorausgesetzt wird und sich zudem die Qualität vieler statistischer Verfahren auf diese Annahme stützt.
- Explorative Werkzeuge um Normalität aufzuzeigen:
  i) **Histogramm**: sollte annähernd Glockengestalt aufweisen.
  ii) **Boxplot**: sollte nahezu symmetrisch sein und nur wenige Ausreißer aufweisen.
  iii) **Normal QQ-Plot**: sollte nahezu linearen Verlauf haben.

Normalverteilung

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

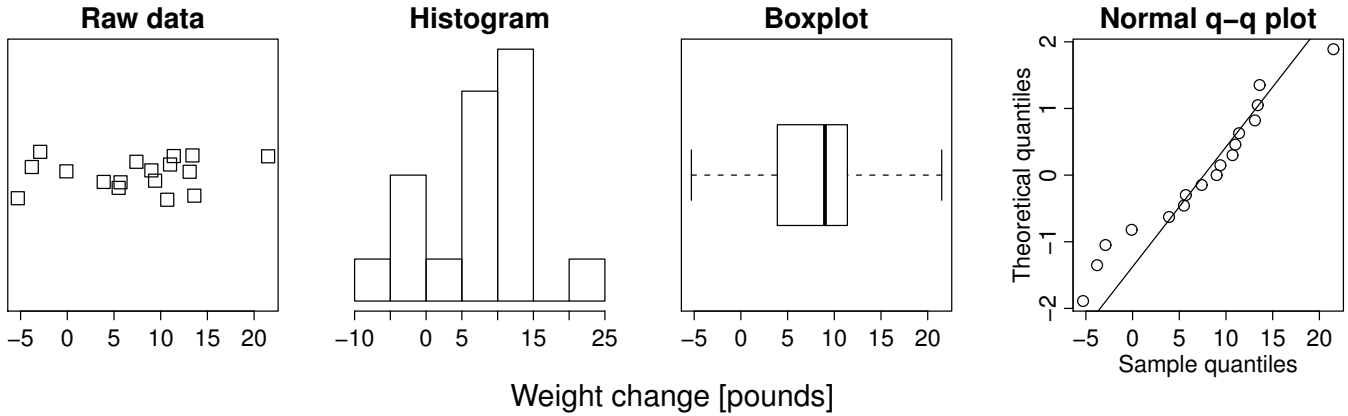
- **Realdaten-Beispiel: Anorexia nervosa**

  Therapieeffekt ist durch Gewichtsveränderung (in Pfund) bei weiblichen Patienten gegeben, wobei

$$\text{Gewichtsänderung} = \text{Gewicht post-therapeutisch} - \text{Gewicht prä-therapeutisch}$$

Daten

n	Min.	1. Q.	Med.	Mean	3. Q.	Max.	St. dev.
17	-5.3	3.9	9.0	7.3	11.4	21.5	7.2



QQ-plot

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

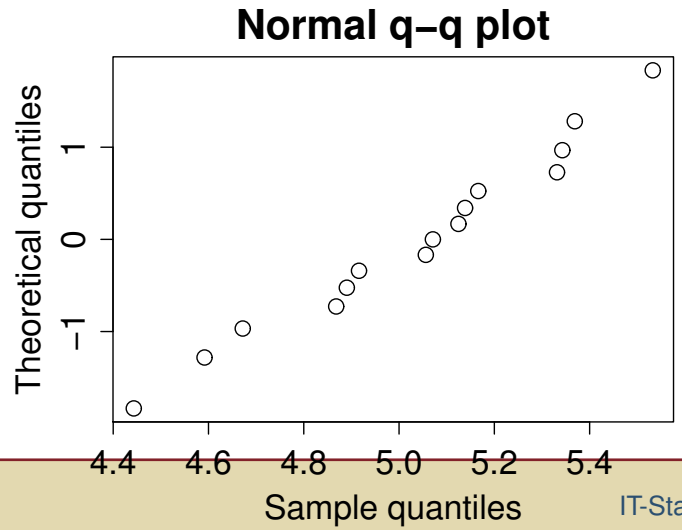
- Falls $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ geordnete Beobachtungen von unabhängig normal-verteilten ZV's darstellen, dann gibt es eine approximativ lineare Beziehung zwischen dem empirischen i/n Quantil $x_{(i)}$ und dem theoretischen $(i-0.5)/n$ Quantil einer standard-normal-verteilten ZV.

$\hookrightarrow$ Ein Plot, der empirische gegen theoretische Quantile abträgt, sollte eine Punktewolke zeigen, die nahezu einen linearen Verlauf hat, falls die Beobachtungen aus einer $N(\mu, \sigma^2)$-verteilten Grundgesamtheit stammen.

Details: Scatter plot bezüglich

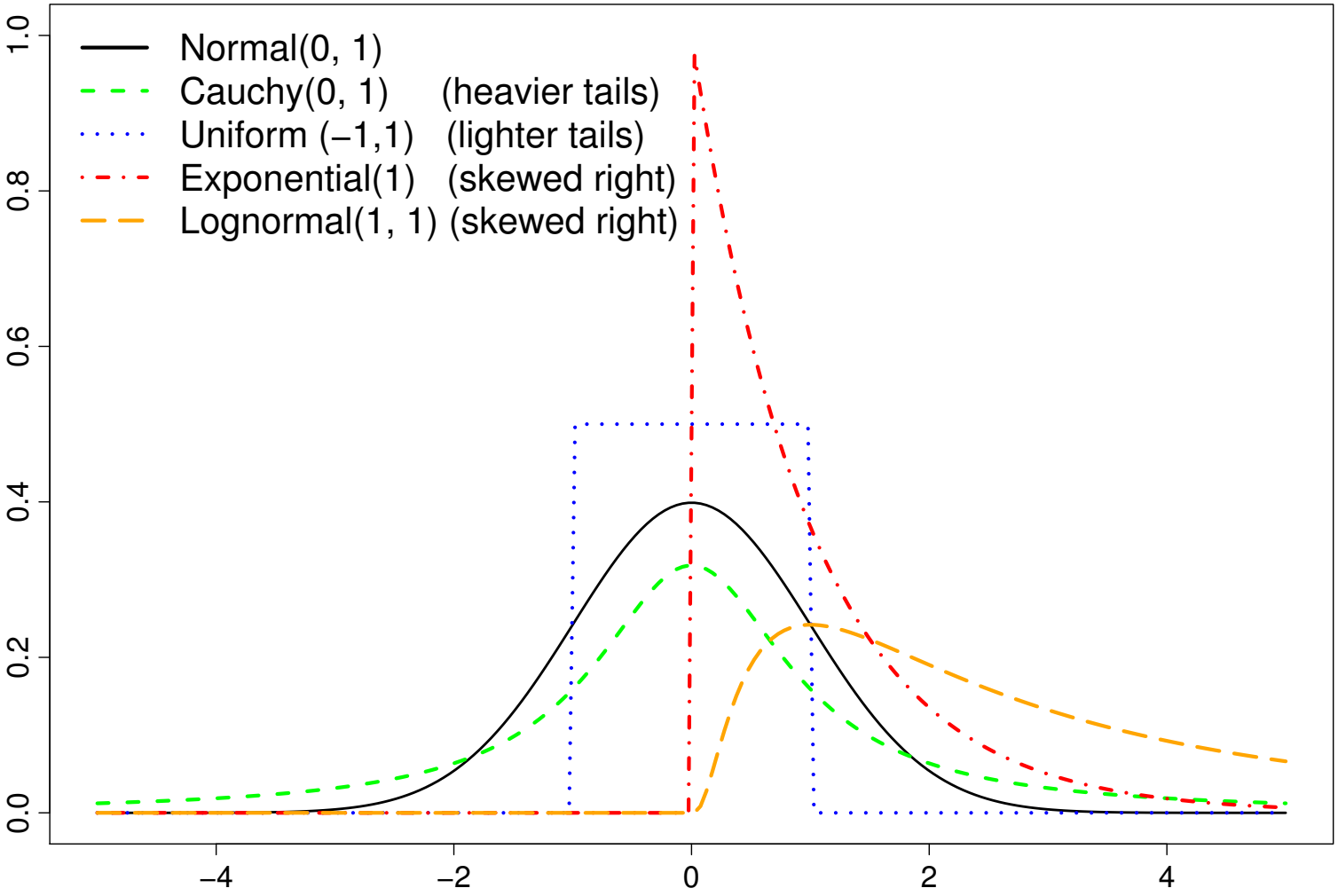
$$\left(x_{(i)}, \Phi^{-1}\left(\frac{i-0.5}{n}\right)\right)$$

wobei Φ^{-1} die Inverse der Verteilungsfunktion der Standardnormalverteilung darstellt.



Andere Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
- Zufallsvariablen
- Verteilungen
- QQ-Plot
- Parameterschätzung
- Zentraler Grenzwertsatz
- Konfidenzintervalle
- Hypothesentests
- Stichprobenumfang
- Teststärke
- Wahl des Testverfahrens
- Anpassungstest
- Unabhängigkeitstest
- Parametrische Tests
- Nichtparametrische Tests
- Varianzanalyse
- Multiple Vergleiche
- Regression
- KQ-Residuen
- Modellgüte
- Standardfehler
- t-Tests
- Modellvarianten
- Box-Cox-Modelle
- Logistische Regression



Parameterschätzung

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- **Szenario**: Eine zufällig gezogene Stichprobe aus einer $N(\mu, \sigma^2)$-Verteilung mit unbekanntem μ und σ^2 liegt vor.
- **Ziel**: Schätze μ und behandle σ^2 wie einen Störparameter.
- **Bekannt**: $\bar{X}$ ist ein guter **Punktschätzer**, da

$$\bar{X} \xrightarrow{n\to\infty} \mu, \quad Var(\bar{X}) \xrightarrow{n\to\infty} 0$$

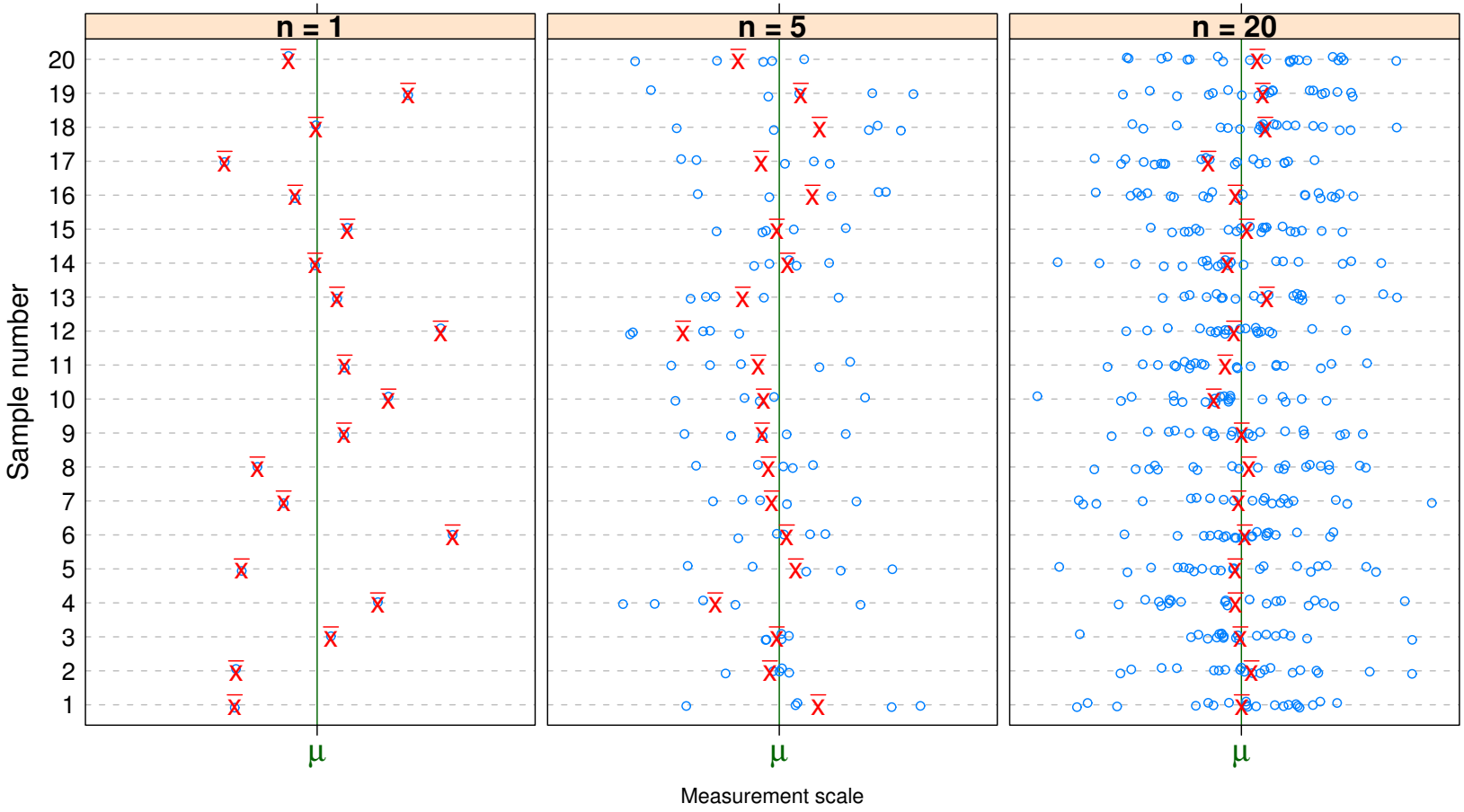
- Wie gut $\bar{X}$ das macht, zeigt das folgende Beispiel einer simulierten Stichprobenverteilung für $\bar{X}$ für größer werdendes n.

Parameterschätzung

1. Basiswissen Statistik

2. Epidemiologie

3. Statistische Inferenz
- Zufallsvariablen
- Verteilungen
- QQ-Plot
- Parameterschätzung
- Zentraler Grenzwertsatz
- Konfidenzintervalle
- Hypothesentests
- Stichprobenumfang
- Teststärke
- Wahl des Testverfahrens
- Anpassungstest
- Unabhängigkeitstest
- Parametrische Tests
- Nichtparametrische Tests
- Varianzanalyse
- Multiple Vergleiche
- Regression
- KQ-Residuen
- Modellgüte
- Standardfehler
- t-Tests
- Modellvarianten
- Box-Cox-Modelle
- Logistische Regression



Zentraler Grenzwertsatz

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- $X_1, X_2, \ldots, X_n$ ... n unabhängige identisch verteilte ZV mit Erwartungswert μ und Varianz σ^2
- wenn n gegen unendlich wächst ($n \to \infty$), konvergiert die Verteilung des Stichprobenmittels $\bar{X}$ gegen eine Normalverteilung mit Erwartungswert μ und Varianz σ^2/n.

$$\bar{X}_{n\to\infty} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Folgerung:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

- Wird σ durch s geschätzt ändert sich die Grenzverteilung etwas

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{s} \sim t_{n-1}$$

Konfidenzintervalle

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Man sieht, dass in den meisten Fällen $\bar{X} \neq \mu$ gilt. Selbst wenn $\bar{X} = \mu$ wäre, würde man es nicht wissen, da μ unbekannt ist.
- Frage: Wie groß ist $|\bar{X} - \mu|$
- Ziel: Entwickle einen **Intervallschätzer**, der den wahren Parameter der Population mit einer vorspezifizierten Wahrscheinlichkeit $1 - \alpha$ beinhaltet.
- Lösung: Konfidenzintervall (KI) mit Irrtumswahrscheinlichkeit α, wobei α üblicherweise mit $0.1, 0.05, 0.01$ angesetzt wird.
- Anmerkung: Die Breite des KI's zeigt den Schätzfehler auf.

Konfidenzintervalle

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Szenario: Eine zufällig gezogene Stichprobe aus einer $N(\mu, \sigma^2)$-Verteilung mit unbekanntem μ und σ^2 liegt vor.
- Exaktes KI für den Parameter μ

$$KI_{\mu}^{(1-\alpha)} := \left[\overline{X} - t_{n-1,1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \; ; \; \overline{X} + t_{n-1,1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

wobei

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

und $t_{k,\gamma}$ das γ-Quantil der t-Verteilung mit $k := n - 1$ Freiheitsgraden ist.

Konfidenzintervalle

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Eigenschaften von $KI_{\mu}^{(1-\alpha)}$
  - i) Zentriert und symmetrisch um $\bar{X}$
  - ii) Breite $= 2 \cdot t_{n-1,1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$ hängt ab von $1-\alpha$, $\hat{\sigma}$ and n:

$$\text{Je größer} \left\{ \begin{array}{ll} 1-\alpha, & \text{desto größer} \\ \hat{\sigma}, & \text{desto größer} \\ n, & \text{desto kleiner} \end{array} \right\} \text{ ist die Breite.}$$

  - iii) Das theoretische KI ist zufällig, da $\bar{X}, \hat{\sigma}$ ZV's sind. Das numerische KI ist deterministisch, da $\bar{x}, s$ Realisationen der ZV's $\bar{X}, \hat{\sigma}$ sind.

Konfidenzintervalle

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Beispiel (Anorexia nervosa):

  Punktschätzungen:

n	$\overline{x}$	s
17	7.3	7.2

  Ziel: 90 % KI bezüglich μ (erwartete Gewichtsänderung), wobei $t_{n-1,1-\alpha/2} = t_{16,0.95} \doteq 1.746$

  Es ergibt sich:

$$\left[\overline{x} - t_{n-1,1-\alpha/2} \cdot \frac{s}{\sqrt{n}}\ ;\ \overline{x} + \ldots\right]$$

$$= \left[7.3 - 1.746 \cdot \frac{7.2}{\sqrt{17}}\ ;\ 7.3 + \ldots\right] \doteq [4.2; 10.4]$$

- Rundungsregel für KI's: Untere/Obere KI-Grenze wird immer ab/aufgerundet.

Konfidenzintervalle

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Warnung: Eine Interpretation der Art

  „Das wahre μ ist mit einer Wahrscheinlichkeit von 90% zwischen 4.2 and 10.4"

  ist nicht zulässig, da das numerische KI eine Realisation des Konfidenzintervallschätzers ist.

- Entweder enthält das konkrete KI den wahren Parameter der Population oder nicht, d.h. die Wahrscheinlichkeit ist entweder 1 oder 0, aber nicht 90%.

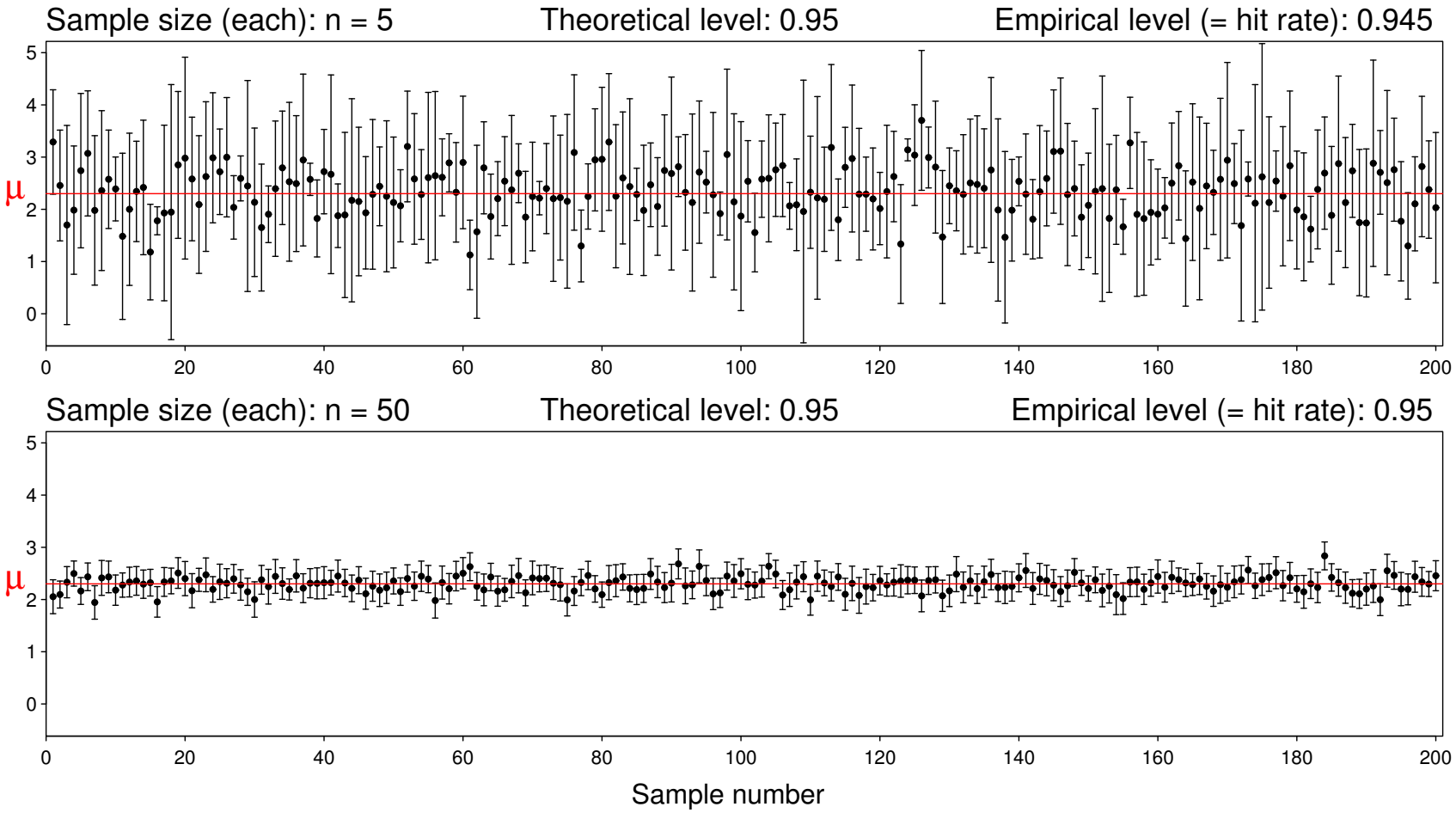
- Korrekt wäre:

  $(1-\alpha)\cdot 100$ % aller KI's bezüglich **aller** Stichproben überdecken den wahren Parameter der Grundgesamtheit.

Konfidenzintervalle

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
- Zufallsvariablen
- Verteilungen
- QQ-Plot
- Parameterschätzung
- Zentraler Grenzwertsatz
- Konfidenzintervalle
- Hypothesentests
- Stichprobenumfang
- Teststärke
- Wahl des Testverfahrens
- Anpassungstest
- Unabhängigkeitstest
- Parametrische Tests
- Nichtparametrische Tests
- Varianzanalyse
- Multiple Vergleiche
- Regression
- KQ-Residuen
- Modellgüte
- Standardfehler
- t-Tests
- Modellvarianten
- Box-Cox-Modelle
- Logistische Regression

- Trefferrate des KI's für 2×200 simulierten $N(\mu, \sigma^2)$-Verteilungen



Konfidenzintervalle

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Szenario: Eine zufällig gezogene Stichprobe aus einer $N(\mu, \sigma^2)$-Verteilung mit unbekanntem μ und σ^2 liegt vor.
- Exaktes KI für den Parameter σ^2

$$KI_{\sigma^2}^{(1-\alpha)} := \left[\frac{(n-1)\hat{\sigma}^2}{\chi^2_{n-1;1-\alpha/2}} \; ; \; \frac{(n-1)\hat{\sigma}^2}{\chi^2_{n-1;\alpha/2}} \right]$$

  wobei $\chi^2_{k;\gamma}$ das γ-Quantil der χ^2-Verteilung mit k Freiheitsgraden darstellt.

- Interessante Anwendung:

  Die obere Grenze von $KI_{\sigma^2}^{(1-\alpha)}$ (Pilotstudie) kann als σ^* für eine Stichprobenumfangsabschätzung genutzt werden.

Hypothesentests, Allgemeines

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Eher informell:
  - 1a Eine Vermutung wird abgelehnt, um auf das Gegenteil zu schließen.
  - 1b Sammle sinnvolle aber zufallsbehaftete Informationen.
  - 1c Sinnvolles Auswerten der Informationen, um ggfs. Evidenz für die Gegenvermutung zu erhalten.
  - 1d Entscheide, ob genug Evidenz vorliegt, um die Vermutung abzulehnen.
- Eher formell:
  - 2a Teste sich gegenseitig ausschließende Hypothesen H_0 vs. H_1.
  - 2b Adäquate Entwicklung eines Experimentes oder sorgfältige durchgeführte Beobachtungsstudie mit gut definierten ZV's.
  - 2c Verwende eine Teststatistik $T(X_1, X_2, ...)$, deren zufälliges Verhalten unter H_0 bekannt ist und sich von dem unter H_1 unterscheidet.
  - 2d Entscheide, ob die Ausprägung von $T(X_1, X_2, ...)$ groß genug ist und damit Abweichungen des unter H_0 hypothesierten Wertes nicht mehr als zufällig zu erachten sind.

Hypothesentests, Allgemeines

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Versuch einer Analogie: Strafrechtlicher Prozess

  Es gilt die Unschuldsvermutung.

  Der Staatsanwalt versucht, die Schuld des Angeklagten zu beweisen.

  Indizien (Evidenz) werden gesammelt.

  Das Gericht entscheidet mit Hilfe der gesammelten Indizien, ob der Angeklagte schuldig ist oder nicht.

- Fundamentales Prinzip: Nicht genug Indizien führen zum Freispruch. „In dubio pro reo” entspricht hier also der Nullhypothese.

Hypothesentests, Fehlertypen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

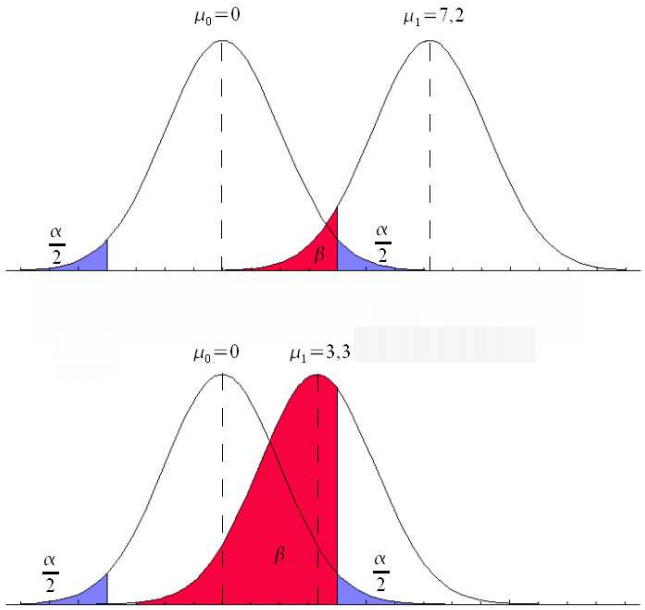
		Unknown truth	
		H_0 is true	H_0 is false
Test decision	Reject H_0	Type I error $^{*)}$	$\surd$
	Do not reject H_0	$\surd$	Type II error $^{**)}$

- Zu $*$:
  - Lehne H_0 ab, obwohl die Hypothese wahr ist (Falscher Alarm).
  - Wahrscheinlichkeit eines Typ I Fehlers entspricht genau α.
  - Die Irrtumswahrscheinlichkeit ist unter Kontrolle des Anwenders.
- Zu $**$:
  - Lehne H_0 nicht ab, obwohl die Nullhypothese nicht wahr ist.
  - Wahrscheinlichkeit eins Typ II Fehlers wird mit β notiert. Bei $1-\beta$ spricht man auch von der Stärke eines statistischen Signifikanztests.
  - β ist meistens nicht unter Kontrolle des Anwenders und gewöhnlich unbekannt.

Hypothesentests, Fehlertypen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Illustration des Fehlers 2. Art (rot) und des Fehlers 1. Art (blau)



Der Fehler 2. Art hängt von μ ab. Je dichter das wahre μ bei dem unter der Nullhypothese angenommen Wert liegt, desto größer ist dieser Fehler. Da μ bei Annahme der Alternativhypothese nicht bekannt ist, kann die Wahrscheinlichkeit für den Fehler zweiter Art nicht berechnet werden.

Hypothesentests: Interpretation

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Nehmen Sie an, dass α hinreichend klein ist:

  $\hookrightarrow$ Schließen sie auf die Richtigkeit von H_1, wenn H_0 abgelehnt werden kann. Die Entscheidung ist verlässlich, da die Irrtumswahrscheinlichkeit klein ist. Das Testergebnis ist statistisch signifikant.

  $\hookrightarrow$ Falls H_0 nicht abgelehnt werden kann, schließen Sie daraus, dass nicht genug Evidenz vorliegt, welche H_1 unterstützt. Diese Entscheidung ist nicht zuverlässig, da der assoziierte Fehler unbekannt und womöglich groß ist.

- Zusammenfassung:

  Ein statistischer Signifikanztest ist i.d.R. ein Werkzeug, um eine Nullhypothese zu verwerfen, aber nicht um diese zu bestätigen. Die Aussage, die bestätigt werden soll, sollte in der Alternativhypothese enthalten sein. Merke auch: ein statistischer Signifikanztest ist kein formaler Beweis.

Hypothesentests, Durchführung

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Wählen Sie zwei sich ausschließende Arbeitshypothesen und spezifizieren Sie α.
- Erheben Sie Daten und prüfen Sie, ob annähernd normalverteilte Daten vorliegen (EDA;Chi-Quadrat-Test).
- Wählen Sie dann auf Basis der Verteilung einen parametrischen oder nichtparametrischen Test. Dieser ist in Form einer Teststatistik gegeben, welche eine bekannte Verteilung unter H_0 hat und dazu tendiert, (absolut) große Werte anzunehmen, falls H_1 wahr ist.
- Berechnen Sie den konkreten Wert der Teststatistik und vergleichen sie ihn mit dem kritischen Wert der Verteilung der Teststatistik.
- Alternativ: geben Sie den zugehörigen p-Wert an (p-Wert: minimale Irrtumswahrscheinlichkeit, bei der H_0 gerade noch abgelehnt werden kann).
- Lehnen Sie H_0 ab, falls der p-Wert kleiner als α ist.

Stichprobenumfang

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Zur Erinnerung:

$$\left[\overline{X} - t_{n-1,1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \; ; \; \overline{X} + \ldots\right]$$

- Frage: Wie viele Beobachtungen sind notwendig, um eine vorspezifizierte Breite Δ einzuhalten.

$$2t_{n-1,1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \Delta, \quad \text{umgestellt} \quad n \geq \left(2t_{n-1,1-\alpha/2} \cdot \frac{\hat{\sigma}}{\Delta}\right)^2$$

- Probleme:
  - Nur implizite Lösung bezüglich n.
  - $\hat{\sigma}$ ist zufällig und daher im Vorfeld unbekannt.
  - Selbst wenn für $\hat{\sigma}$ ein „gutes” Surrogat existiert und n determiniert werden kann, würde nur der Erwartungswert des immer noch zufälligen KI's kleiner Δ sein.

Stichprobenumfang, Beispiel

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

↪ Lege α und $1-\beta$ (Power) fest. Standard: $\alpha = 5\%$ und $(1-\beta) = 80\%$

$$n \geq \left(2\left(t_{n-1,1-\alpha/2} + t_{n-1,1-\beta/2}\right) \cdot \frac{\hat{\sigma}}{\Delta}\right)^2$$

↪ Praxis: ersetze $\hat{\sigma}$ durch die Schätzung s, d.h.

$$n \geq \left(2\left(t_{n-1,1-\alpha/2} + t_{n-1,1-\beta/2}\right) \cdot \frac{s}{\Delta}\right)^2$$

- Beispiel:

↪ Wie viele Wassermelonen sind notwendig, um ein KI bezüglich des erwarteten Gewichts zu erhalten, das $1-\alpha = 0.9$ und $1-\beta = 0.8$ für $\Delta \leq 0.3kg$ aufweist?

$n > 295$

Stichprobenumfang, Zusammenfassung

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Folgende Größen werden zur Fallzahlschätzung benötigt:
  - primäre Variable
  - Null- und Alternativhypothese
  - Testverfahren
  - minimal zu erreichender relevanter Unterschied
  - Abschätzung der Standardabweichung für die primäre Variable
  - Fehler 1. Art
  - Power
- Der nachzuweisende Unterschied basiert auf einer fachlichen Einschätzung des minimalen klinisch relevanten Effektes oder des erwarteten Effektes einer neuen Behandlung.
- Die Fallzahl ist umso höher, je kleiner der Fehler 1. Art, je höher die Power, je kleiner der nachzuweisende Unterschied und je größer die Standardabweichung.

Teststärke

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Zur Erinnerung: Für jeden statistischen Test H_0 vs. H_1 gilt:
  - α Wahrscheinlichkeit eines Typ I Fehlers.
  - β Wahrscheinlichkeit eines Typ II Fehlers.
- Ziel: Die Wahrscheinlichkeit bezüglich jeder Fehlentscheidung so gering wie möglich halten.
- Definition: Die Teststärke (Power) ist durch $1 - \beta$ gegeben, also durch die Wahrscheinlichkeit H_0 abzulehnen, wenn H_1 tatsächlich wahr ist.
- Fakten:
  - Der Typ I Fehler ist unter Kontrolle.
  - Der Typ II Fehler ist nicht unter Kontrolle.
  - Hohe Teststärke $\Leftrightarrow$ Kleine Wahrscheinlichkeit eines Typ II Fehlers.

Teststärke

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Für ein festgesetztes α und einem typischerweise unbekanntem σ beantwortet die Powerfunktion im Prinzip drei Fragen:
  - i) Sei n und das wahre $\Delta \neq 0$ vorgelegt: Wie groß ist die Power $1-\beta$, um dieses wahre Δ aufzudecken?
  - ii) Sei eine Teststärke von $1-\beta$ und das wahre $\Delta \neq 0$ vorgelegt: wie groß muss n sein, um das wahre Δ mit einer Wahrscheinlichkeit von $1-\beta$ aufzudecken?
  - iii) Sei n und das wahre $1-\beta$ vorgelegt: Welches ist das kleinste Δ, das mit einer Wahrscheinlichkeit von $1-\beta$ aufgedeckt werden kann.
- Anmerkung: Auch hier wird eine Grobschätzung von σ benötigt.

Teststärke, Beispiel

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Beispiel: Gewicht von Wassermelonen

  Hypothese: $\mu_0 = 5.3$kg

  Die Nullhypothese kann bei einem arithmetischen Mittel von 5.1 zum Niveau 0.05 nicht abgelehnt werden ($p = 0.2598$). Wie hoch ist die Power für diesen Ein-Stichproben-t-Test?

  - Geforderter Mittelwert $|\Delta| := |\mu_0| \geq 0.2$.
  - Stichprobe: $n = 20$, $\alpha = 0.05$ und $\sigma^* = 0.77$

$\hookrightarrow$ Power für vorgelegtes $n = 20$ und $|\Delta| = 0.2$ (mit $\sigma = 0.77$ and $\alpha = 5\ \%$) ist 0.197.

$\hookrightarrow$ Stichprobengröße für vorgelegtes $1 - \beta = 80\ \%$ und $|\Delta| = 0.2$ (with $\sigma = 0.77$ and $\alpha = 5\ \%$) ist 119.

$\hookrightarrow$ Kleinstes Δ für vorgelegtes $1 - \beta = 80\ \%$ and $n = 20$ (with $\sigma = 0.77$ and $\alpha = 5\ \%$) ist 0.51.

- Berechnungen erfolgen etwa mit G∗Power oder via R mit Hilfe des Befehls `power.t.test(···)`.

Tests der Lageparameter

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- 1 Stichprobe (mind. ordinales Messniveau):

NV	nicht NV
Ein-stichproben t-test	Wilcoxons Vorzeichen rang-test

- mehrere Stichproben (mind. ordinales Messniveau):

	Anzahl der Stichproben			
	2		$p > 2$	
	unabh.	abh.	unabh.	abh.
NV	Zwei-stichproben-t-Test	Ein-stichproben-t-Test	ANOVA	ANOVA mit Mess-wiederholungen
nicht NV	Wilcoxons Rangsummen-test	Wilcoxons Vorzeichen-rang-test	Kruskal-Wallis test	Friedman-test

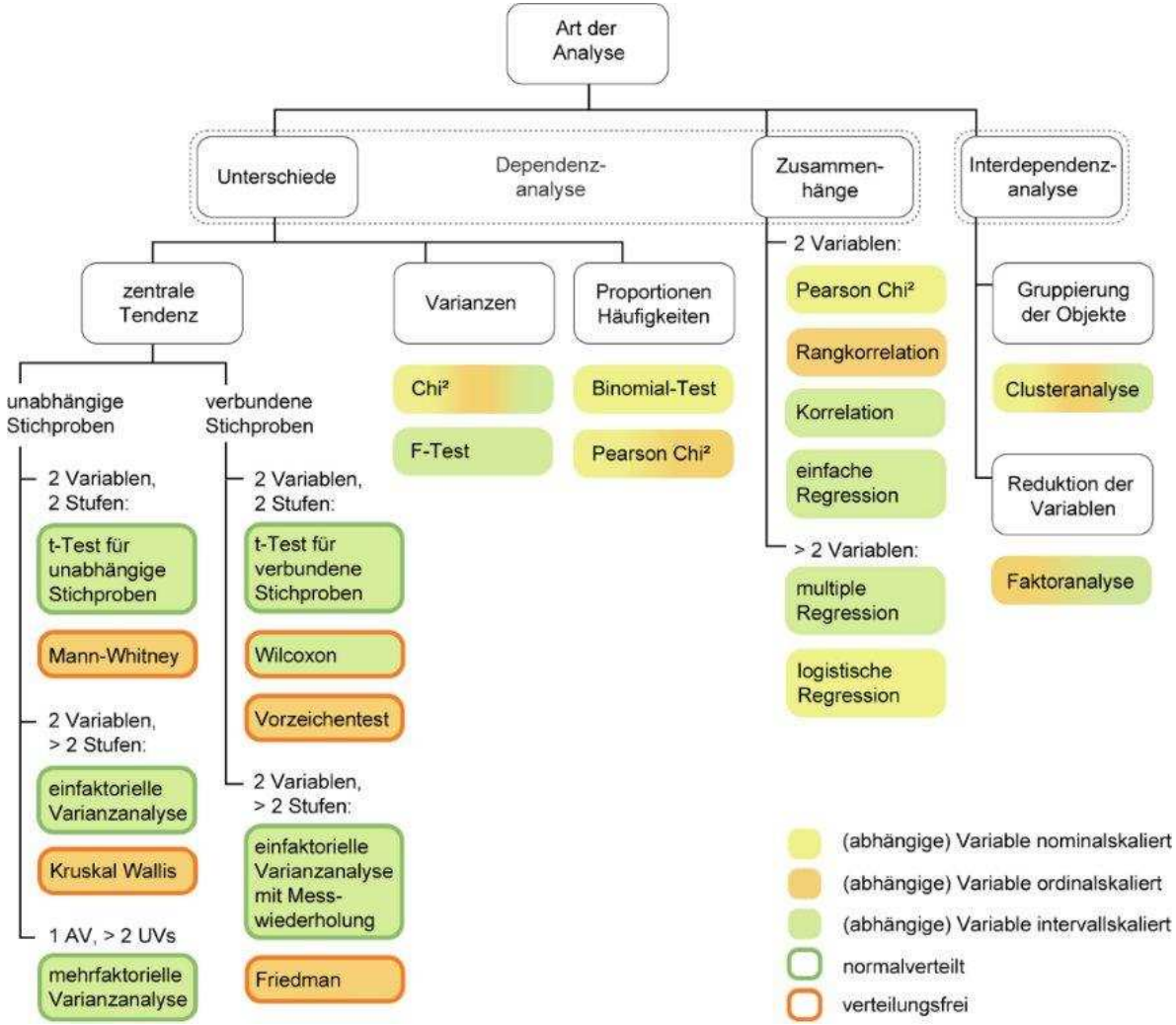
Allgemein

1. Basiswissen Statistik

2. Epidemiologie

3. Statistische Inferenz

- Zufallsvariablen
- Verteilungen
- QQ-Plot
- Parameterschätzung
- Zentraler Grenzwertsatz
- Konfidenzintervalle
- Hypothesentests
- Stichprobenumfang
- Teststärke
- Wahl des Testverfahrens
- Anpassungstest
- Unabhängigkeitstest
- Parametrische Tests
- Nichtparametrische Tests
- Varianzanalyse
- Multiple Vergleiche
- Regression
- KQ-Residuen
- Modellgüte
- Standardfehler
- t-Tests
- Modellvarianten
- Box-Cox-Modelle
- Logistische Regression



Chi-Quadrat-Anpassungstest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Fragestellung: Ist die Annahme eines bestimmten Verteilungsmodells gerechtfertigt oder nicht?

Vorgehensweise:

- Vergleich von beobachteten und erwarteten Häufigkeiten:
  O_i ... beobachtete Anzahl bei Ausprägung i
  E_i ... erwartete Anzahl aus der Verteilungsannahme
- Modifikation der Differenz $O_i - E_i$ in eine χ^2-verteilte Variable:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

- der χ^2-Wert ist stets positiv
- große Werte von χ^2 offenbaren nicht vernachlässigbare Abweichungen von der unterstellten Verteilung

Unabhängigkeitstest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Fragestellung: Sind 2 kategorielle Variablen mit je 2 Ausprägungen als unabhängig anzusehen?
Beispiel (Vierfeldertafel): Unterscheidet sich in einer Firma die Arbeitszufriedenheit von Frauen und Männern?

Beobachtete Häufigkeiten:

	Zufrieden	Nicht zufrieden
Frauen	102 (a)	234 (b)
Männer	89 (c)	170 (d)

Randhäufigkeiten:

	Zufrieden	Nicht zufrieden	Gesamt
Frauen	102	234	336
Männer	89	170	259
Gesamt	191	404	595

Unabhängigkeitstest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression
Anhang: Weitere Verteilungen

Erwartete Häufigkeiten:

	Zufrieden	Nicht zufrieden
Frauen	(191*336)/595=108	(404*336)/595=228
Männer	(191*259)/595=83	(404*259)/595=176

Berechnung der Teststatistik:

$$\chi^2 = [(102-108)^2/108] + [(89-83)^2/83] + [(234-228)^2/228] + [(170-176)^2/176] = 1.12$$

Vereinfachte Formel:

$\chi^2 = \frac{N(|ab-cd|)}{(a+b)(c+d)(a+c)(c+d)}$, bzw. mit Stetigkeitskorrektur nach Yates (bei kleinen Stichprobenumfängen)

$$\chi^2 = \frac{N(|ab-cd|-\frac{N}{2})}{(a+b)(c+d)(a+c)(c+d)}.$$

Einstichproben-t-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Beispiel:

Behauptung: (Nullhypothese)
Wassermelonen wiegen im Schnitt 5.5 kg
d.h. $H_0 : \mu_0 = 5.5$

Mögliche Alternativen:

$$\begin{aligned} A: &\quad H_1: \quad \mu \neq 5.5 \\ B: &\quad H_1: \quad \mu < 5.5 \\ C: &\quad H_1: \quad \mu > 5.5 \end{aligned}$$

Einstichproben-t-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Beispiel:

$$H_0 : \mu = 5.5 = \mu_0$$
$$H_1 : \mu \neq 5.5$$

Irrtumswahrscheinlichkeit: $\alpha = 0.05$

- Vorgehensweise: ZV X: “Gewicht“ , $n = 20$
- Annahme:

$$X \sim N(\mu_0; \sigma^2)$$

Rechtfertigt die Information in der Stichprobe diese Aussage der Nullhypothese über μ?

Einstichproben-t-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Lösung:
  Nach dem Zentralen Grenzwertsatz (ZGWS) ist:

$$\bar{X} \sim N(5.5; \frac{\sigma^2}{n})$$

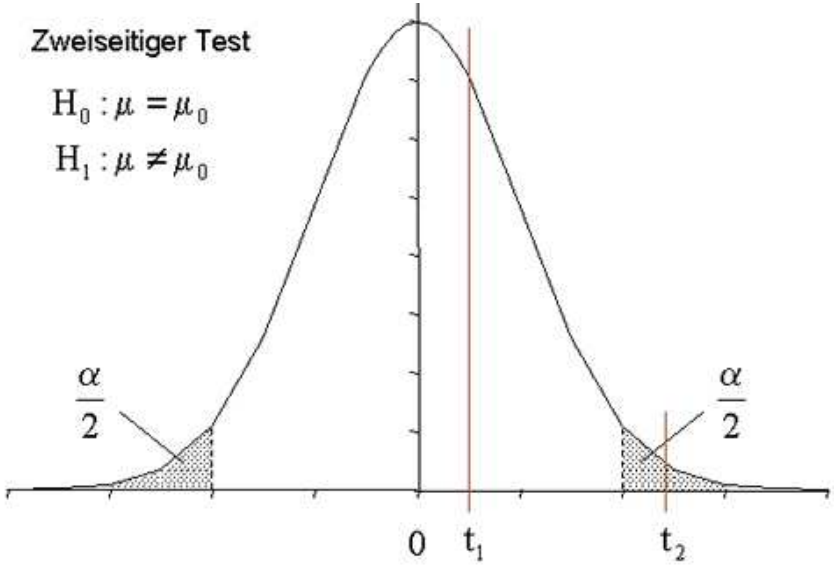
  Schätze die unbekannte Varianz σ^2 durch s^2. Im Beispiel: $s^2 = 0.6$
  Es gilt:

$$T = \sqrt{n}\left(\frac{\bar{X} - \mu_0}{s}\right) \text{ ist } t_{(n-1)}\text{-verteilt.}$$

- Intuitv:
  - lehne H_0 ab, falls $\bar{X}$ (Mittelwert der Stichprobe) „zu weit“ von $\mu_0 = 5.5$ entfernt ist.
    $\rightarrow$ oder was identisch ist:
  - lehne H_0 ab, falls $\left|\sqrt{n}\left(\frac{\bar{X}-\mu_0}{s}\right)\right|$ „zu groß“

Einstichproben-t-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
- Zufallsvariablen
- Verteilungen
- QQ-Plot
- Parameterschätzung
- Zentraler Grenzwertsatz
- Konfidenzintervalle
- Hypothesentests
- Stichprobenumfang
- Teststärke
- Wahl des Testverfahrens
- Anpassungstest
- Unabhängigkeitstest
- Parametrische Tests
- Nichtparametrische Tests
- Varianzanalyse
- Multiple Vergleiche
- Regression
- KQ-Residuen
- Modellgüte
- Standardfehler
- t-Tests
- Modellvarianten
- Box-Cox-Modelle
- Logistische Regression



- Hier: Teststatistik $T = \sqrt{20}\left(\frac{5.1 - 5.5}{0.77}\right) = |-2.32|$
- kritischer Wert: $t_{(n-1,\alpha/2)} = t_{(19,0.025)} = 2.093$
  d.h. $|T|$ „zu groß“.
- p-Wert: $p = 0.0314$
- KI bezüglich des Erwartungswertes: $[4.73; 5.46]$.

$\Rightarrow$ lehne H_0 ab: D.h. es kann nicht davon ausgegangen werden, dass das Durchschnittsgewicht der Wassermelonen gleich 5.5 kg ist.

Einstichproben-t-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- **Achtung:** Je weniger Daten, desto unsicherer die Entscheidung.
  - Beispiel:

$$n = 10 \ , \ \bar{X} = 5.1 \ , \ s^2 = 0.6$$

$$T = |-1.64| < t_{(9,0.025)} = 2.26$$

d.h. lehne H_0 nicht ab.

Zweistichproben-t-Test

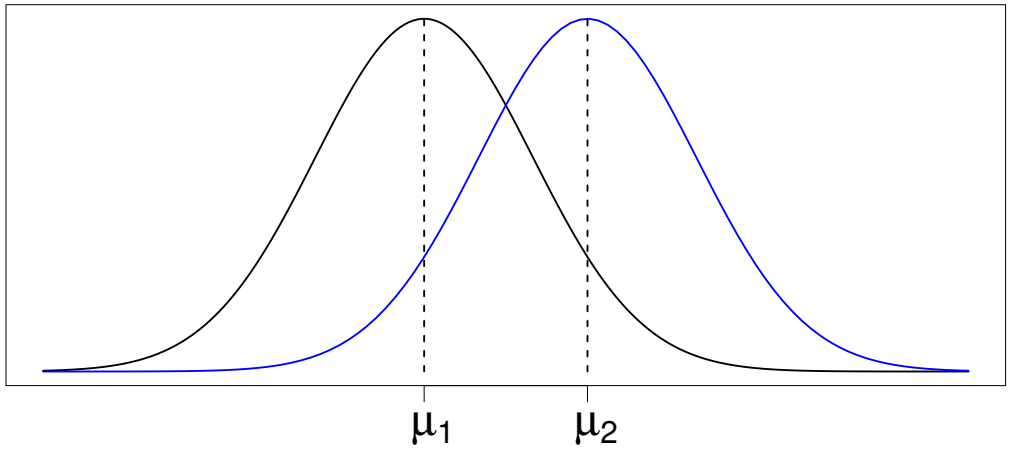
1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Effekt einer toxischen Substanz auf den Erwartungswert des Nierengewichts von Ratten.
- Zwei Zufallsstichproben von $n_1 = 13$ (Placebo-Gruppe) und $n_2 = 18$ (Verum-Gruppe).
- Annahme: Das individuelle Nierengewicht ist in beiden Gruppen $i = 1, 2$ unabhängig $N(\mu_i, \sigma^2)$-verteilt.
- Hypothesen informal: H_0: Verum hat keinen Effekt vs. H_1 : Verum hat einen Effekt.
- Hypothesen formal: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.
- Irrtumswahrscheinlichkeit $\alpha = 0.05$.

Zweistichproben-t-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Mögliches Szenario für Normaldichten unter H_1.



- Für zwei unabhängige, normalverteilte Stichproben mit gleicher Varianz (Homoskedastie) eignet sich der zweiseitige Zwei-Stichproben-t-Test

$$T := \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}} \sim t(n_1 + n_2 - 2)$$

Zweistichproben-t-Test

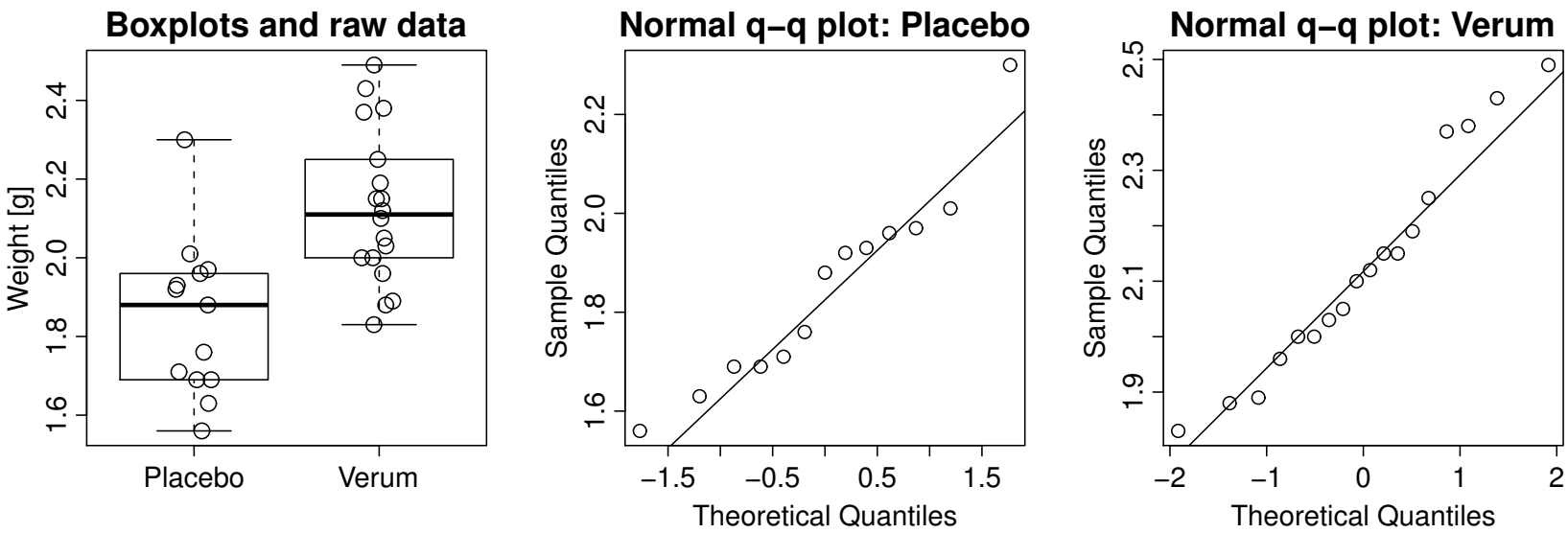
1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Datentabelle

ID	Group	Weight	ID	Group	Weight
R01	Verum	2.15	R17	Verum	2.10
R02	Placebo	1.56	R18	Placebo	1.88
R03	Verum	1.96	R19	Verum	1.88
R04	Verum	2.49	R20	Placebo	1.93
R05	Placebo	2.30	R21	Placebo	1.96
R06	Verum	1.83	R22	Verum	2.19
R07	Placebo	2.01	R23	Verum	2.37
R08	Verum	2.12	R24	Verum	2.38
R09	Placebo	1.71	R25	Placebo	1.97
R10	Verum	1.89	R26	Verum	2.00
R11	Verum	2.03	R27	Verum	2.25
R12	Verum	2.43	R28	Verum	2.00
R13	Placebo	1.69	R29	Verum	2.05
R14	Placebo	1.92	R30	Placebo	1.63
R15	Placebo	1.69	R31	Verum	2.15
R16	Placebo	1.76			

Zweistichproben-t-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
- Zufallsvariablen
- Verteilungen
- QQ-Plot
- Parameterschätzung
- Zentraler Grenzwertsatz
- Konfidenzintervalle
- Hypothesentests
- Stichprobenumfang
- Teststärke
- Wahl des Testverfahrens
- Anpassungstest
- Unabhängigkeitstest
- Parametrische Tests
- Nichtparametrische Tests
- Varianzanalyse
- Multiple Vergleiche
- Regression
- KQ-Residuen
- Modellgüte
- Standardfehler
- t-Tests
- Modellvarianten
- Box-Cox-Modelle
- Logistische Regression



- EDA:

Placebo	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	1.560	1.690	1.880	1.847	1.960	2.300

Verum	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	1.830	2.000	2.110	2.126	2.235	2.490

- $\hat{\sigma}_1 = 0.1995990$ (Placebo) und $\hat{\sigma}_2 = 0.1953018$ (Verum)

Zweistichproben-t-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- EDA zeigt, dass die Annahmen bezüglich des zweiseitigen Zwei-Stichproben-t-Test erfüllt sind.
- Teststatistik:

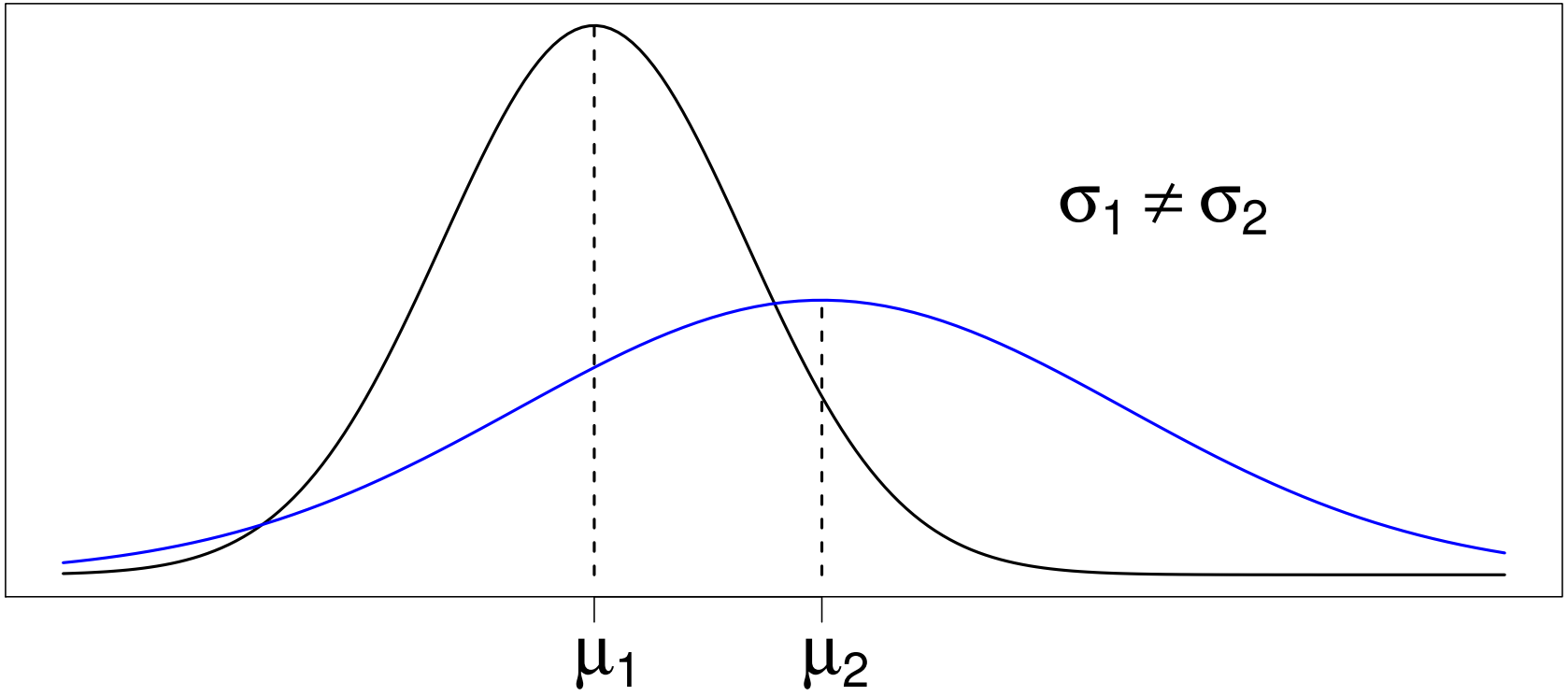
$$T := \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}} = -3.8919$$

- Freiheitsgrade: $n_1 + n_2 - 2 = 29$.
- p-value: $p = 0.0005362$.
- KI bezüglich der Differenz der Erwartungswerte: $[-0.43; -0.13]$.
- Testentscheidung: Lehne H_0 ab, da $p < \alpha$, respektive weil 0 nicht im KI bezüglich der Differenz der Erwartungswerte enthalten ist.

Welch-Satterthwaite-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Annahme: Population ist normalverteilt aber heteroskedastisch (verschiedene Streuung)



- Falls $\mu_1 - \mu_2$ von Interesse, können Sie den modifizierten Welch-Test nutzen.

Welch-Satterthwaite-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Irrtumswahrscheinlichkeit $\alpha = 0.05$.
- Teststatistik

$$t := \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -3.8777$$

- Freiheitsgrade (Welch-Satterthwaite-Gleichung):

$$df := \nu := \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \cdot (n_1 - 1)} + \frac{s_2^4}{n_2^2 \cdot (n_2 - 1)}} = 25.669$$

- p-value: $p = 0.0006538$.
- KI bezüglich θ: $[-0.43; -0.13]$.
- Testentscheidung: Lehne H_0 ab, da $p < \alpha$, respektive 0 nicht im KI bezüglich θ enthalten ist.

Wilcoxons Tests

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression
Anhang: Wichtige Verteilungen

- Vorteil:
  - Annäherend „normalverteilte“ Daten werden **nicht** vorausgesetzt.
  - Tests sind sehr einfach durchzuführen
  - Auch „Rangmerkmale“ wie z.B. Schulnoten können „getestet“ werden.

Die Idee des Wilcoxon-Vorzeichenrangtests

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression
Anhang: Wichtige Verteilungen

- Dieser Test wird angewendet, falls es sich um paarige (abhängige) Messreihen handelt oder eine Hypothese über den Median einer Stichprobe überprüft werden soll

Beispiel:
eine Variable wird an einer Person zu verschiedenen Zeitpunkten gemessen (z.B. vor und nach medikamentöser Behandlung)

Beispiel:

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Für den Test werden 10 Personen nach ihrem Urteil über die Verträglichkeit zweier Behandlungen A und B mit Hilfe einer Bewertungsskala von 1 (sehr gut) bis 5 (sehr schlecht) gefragt.

Person Nr.	Produkt		Differenz	Rang (unabhängig vom Vorzeichen)
	A	B		
1	5	3	2	5.5
2	1	2	-1	$(-)2$
3	3	2	1	2
4	3	3	0	–
5	4	2	2	5.5
6	1	3	-2	$(-)5.5$
7	2	3	-1	$(-)2$
8	4	2	2	5.5
9	5	1	4	9
10	4	1	3	8

Summe der plus - Ränge: 35.5 und Summe der minus - Ränge: 9.5

Testentscheidung

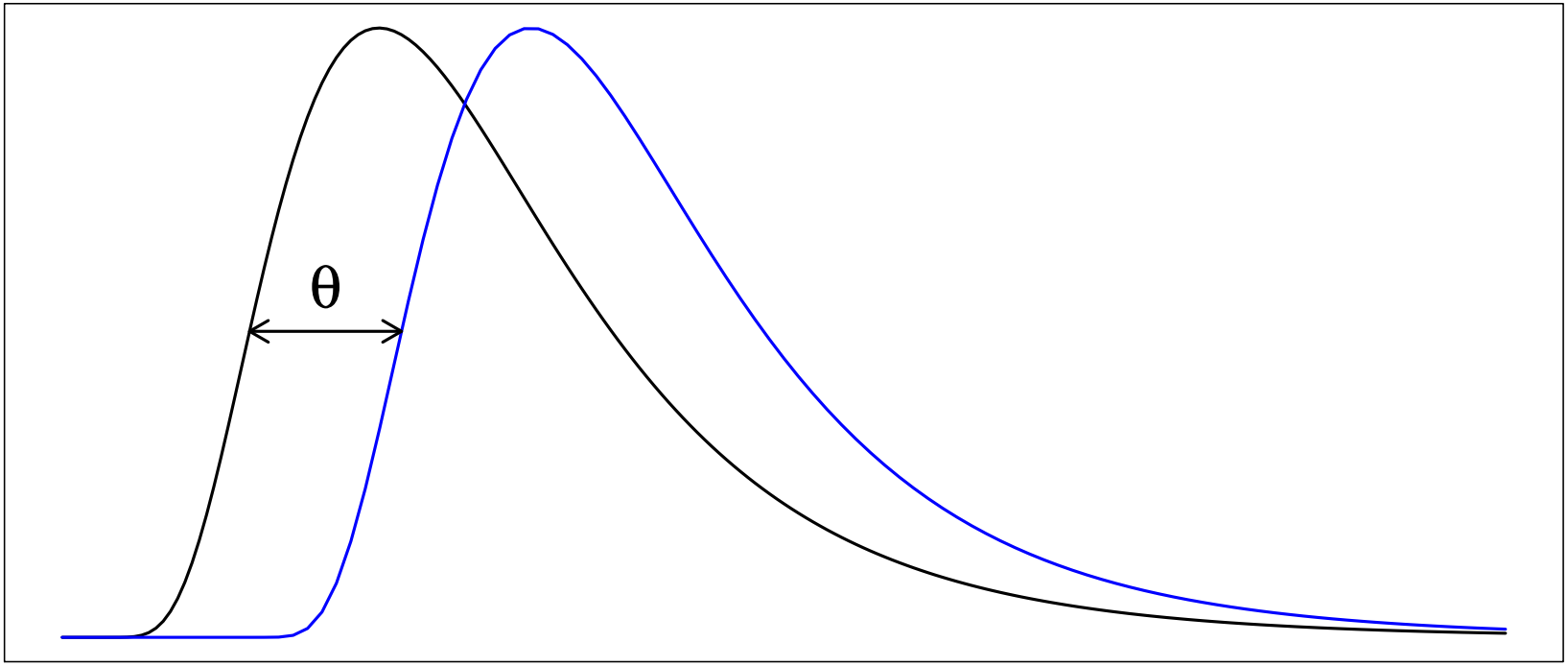
1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- je besser die Beurteilung von Produkt B:
  - desto größer die Summe der plus - Ränge
  - desto kleiner die minus - Summe der Ränge
- Verwerfung der Nullhypothese: Median der Differenzen von A und B = Null, wenn die kleinere der beiden Summen zu klein ist.
- der kritische Wert (Tabelle oder Software) ist $c = 6$ bei α = 5%
- die Summe der minus - Ränge (9.5) unterschreitet den kritischen Wert von 6 nicht,
- d.h. Produkt B wird nicht signifikant besser bewertet als Produkt A

Wilcoxon Rangsummentest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Annahme: Population ist nicht normalverteilt, Verteilung unterscheidet sich nur um einen „shift” und ist stetig für jede Gruppe.



- Für Hypothesen bzgl. θ kann ein nichtparametrischer Lagetest, z.B. der Wilcoxon Rangsummentest (Mann-Whitney’s U-Test) benutzt werden.

Wilcoxon Rangsummentest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Vergleich zweier unabh. Stichproben

- benötigt keine Normalverteilungsannahme, setzt aber stetig verteilte Daten voraus
- verwendet nur die Ränge der Beobachtungen der gemeinsamen Stichprobe:
- die kleinste Beobachtung erhält Rang 1; die größte Beobachtung erhält Rang $n_1 + n_2$
- Testidee basiert auf der Annahme: falls (hypothetisch) alle Daten der ersten Stichprobe kleinere Werte als die der zweiten Stichprobe habem, ist (logischwerweise) die Summe der Ränge in der ersten Stichprobe kleiner als die Summe der Ränge in der zweiten Stichprobe
- Teststatistik:

$$W = \sum \text{Ränge der ersten Stichprobe}$$

- Nullhypothese: $H_0 : Median_1 = Median_2$
- Alternativhypothese: $H_0 : Median_1 \neq Median_2$

Wilcoxon Rangsummentest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Effekt einer toxischen Substanz auf den Erwartungswert des Nierengewichts von weiblichen Laborratten.
- Zwei Zufallsstichproben von $n_1 = 13$ (Placebo-Gruppe) und $n_2 = 18$ (Verum-Gruppe).
- Annahme: Das Nierengewicht ist unabhängig in und über die Gruppen hinweg stetig verteilt. Beide Gruppen weisen die gleiche Verteilung auf, die sich lediglich um einen shift-Parameter unterscheidet.
- Hypothesen informal: H_0: Verum hat keinen Effekt vs. H_1 : Verum hat einen Effekt.
- Hypothesen formal: $H_0 : Median_1 = Median_2$ vs $H_0 : Median_1 \neq Median_2$.
- Irrtumswahrscheinlichkeit $\alpha = 0.05$.

Wilcoxon Rangsummentest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Hilfsmittel: Ränge, d.h.

Rang 1	$\equiv$	kleinste Beobachtung
Rang n	$\equiv$	größte Beobachtung

  Falls 2 oder mehr Beob. identisch sind wählt man mittlere Ränge.

- Wie entscheidet man?
  Intuitiv:
  - Je höher die Ränge in der „Verum“ Stichprobe, desto mehr spricht dies gegen H_0.
- Test: Zweiseitiger Wilcoxon Rangsummentest

$$W_{n_1,n_2} := \sum_{i=1}^{n_1} R(X_i) \overset{approx}{\sim} N\left(\frac{n_1\,(n_2+n_1+1)}{2}; \frac{n_2\,n_1\,(n_2+n_1+1)}{12}\right)$$

Wilcoxon Rangsummentest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Datentabelle

ID	Group	Weight	Rank	ID	Group	Weight	Rank
R01	Verum	2.15	23.5	R17	Verum	2.10	21
R02	Placebo	1.56	1	R18	Placebo	1.88	8.5
R03	Verum	1.96	13.5	R19	Verum	1.88	8.5
R04	Verum	2.49	31	R20	Placebo	1.93	12
R05	Placebo	2.30	27	R21	Placebo	1.96	13.5
R06	Verum	1.83	7	R22	Verum	2.19	25
R07	Placebo	2.01	18	R23	Verum	2.37	28
R08	Verum	2.12	22	R24	Verum	2.38	29
R09	Placebo	1.71	5	R25	Placebo	1.97	15
R10	Verum	1.89	10	R26	Verum	2.00	16.5
R11	Verum	2.03	19	R27	Verum	2.25	26
R12	Verum	2.43	30	R28	Verum	2.00	16.5
R13	Placebo	1.69	3.5	R29	Verum	2.05	20
R14	Placebo	1.92	11	R30	Placebo	1.63	2
R15	Placebo	1.69	3.5	R31	Verum	2.15	23.5
R16	Placebo	1.76	6				

Wilcoxon Rangsummentest

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Teststatistik: Summiere die Ränge der Verum-Gruppe

$$W_{n_1,n_2} := \sum_{i=1}^{n_1} R(X_i) = 370$$

- p-value: $p = 0.001098$.
- KI bezüglich θ: $[-0.44; -0.13]$.
- Testentscheidung: Lehne H_0 ab, da $p < \alpha$, respektive 0 nicht im KI bezüglich θ enthalten ist.

Varianzanalyse, ANOVA

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression
Anhang: Wichtige Verteilungen

Varianzanalyse (Vergleich von mehr als zwei Stichproben)

Parametrische Lösungen:

Die Idee des F-Tests

- p Stichproben mit möglicherweise unterschiedlichen Stichprobenumfängen $n_1, ..., n_p$
- die i-te Stichprobe folgt einer $N(\mu_i, \sigma^2)$-Verteilung:
  unbekannter Erwartungswert μ_i
  Varianz $\sigma^2 > 0$
- diese Parameter sollten in allen Stichproben annähernd gleich sein

F-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Test der Nullhypothese:

$$H_0 : \mu_1 = ... = \mu_p$$

gegen

$$H_1 : \text{mindestens zwei Erwartungswerte sind verschieden}$$

- Verwerfung von H_0, falls
  die Differenz zwischen den Stichprobenmitteln hinreichend groß ist
  die Streuung in den p Stichproben nicht zu groß ist
- Grundkomponente des Tests:

$$MST = \sum_{i=1}^{p} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(p-1)$$

$\bar{y}_{i.}$ ... Stichprobenmittel der i-ten Stichprobe
$\bar{y}_{..}$ ... Mittelwert aller Beobachtungen

F-Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- zweite Komponente:

$$MSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (N - p)$$

y_{ij} ... Beobachtungen
$\bar{y}_i$ ... Mittelwert der jeweiligen Stichprobe

- Ablehnung von H_0, falls

$$\frac{MST}{MSE} > F_{p-1,N-p;1-\alpha}$$

Nichtparametrische Anova

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Nichtparametrische Lösung:

Der Kruskal-Wallis-Test

- nichtparametrisches Gegenstück zur einfachen ANOVA

Idee des Tests:

- Verwerfung von H_0, falls sich die aus der Gesamtstichprobe berechneten Ränge der i-ten Stichprobe zu sehr unterscheiden

$$KW = \frac{1}{C}\left(\frac{12}{N(N+1)}\sum_{i=1}^{p}\frac{1}{n_i}\tau_{i.}^2 - 3(N+1)\right)$$

Kruskal-Wallis Test

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

wobei

$$C = 1 - \frac{1}{N^2(N-1)} \sum_{l=1}^{g} t_l^3 - t_l$$

eine Korrekturgröße ist
g ... Anzahl unterschiedlicher Werte mit Häufigkeiten τ_i und t_l

- Falls H_0 in diesem Test verworfen wird, gibt es signifikante Unterschiede zwischen den Stichproben.

Erweiterungen der einfachen Varianzanalyse: Zweifachklassifikation

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Einfachklassifikation: Gibt es Unterschiede zwischen den μ_i in I Populationen?

$$y_{ij} = \mu + \alpha_i + u_{ij} \text{ wobei } u_{ij} \sim N(0, \sigma^2), \alpha_i = \mu_i - \mu_0 \text{ und } \sum_{i=1}^{I} \alpha_i = 0.$$

Zweifachklassifikation: Ergänzung des Modells um weitere (qualitative) Faktoren β_j und Wechselwirkungen $(\alpha\beta)_{ij}$

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + u_{ijk}.$$

mit

$$\sum_{i=1}^{I} \alpha_i = 0, \sum_{j=1}^{J} \beta_j = 0 \text{ und } \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$$

und $i \in \{1, 2, \cdots, I\}, j \in \{1, 2, \cdots, J\}, k \in \{1, 2, \cdots, K\}$.

ANCOVA und repeated measurements

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Bei der Kovarianzanalyse wird in ein varianzanalytisches Modell zusätzlich ein quantitativer Faktor x_{ij}einbezogen:

$$y_{ij} = \mu_0 + \alpha_i + \gamma_j x_{ij} + u_{ij}.$$

x_{ij} besteht oftmals aus dem Baseline-Werten von y.
Falls eine Datenerhebung (an gleichen Untersuchungseinheiten) zu T Zeitpunkten erfolgt bietet sich ein rep. measurements Modell an:

$$y_{ijt} = \mu + \alpha_i + \theta_t + (\gamma\theta)_{it} + \pi_{j(i)} + u_{ijt}.$$

Die zeitlichen Abhängigkeiten werden über eine Kovarianzmatrix modelliert, die z.B. eine autorgressive Struktur 1. Ordnung haben kann, d.h. z.B. für $T = 4$:

$$\mathbf{Cov}(\mathbf{y_j}) = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_2\rho^2 & \sigma_1\sigma_2\rho^3 \\ & \sigma_2^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_2\rho^2 \\ & & \sigma_3^2 & \sigma_1\sigma_2\rho \\ & & & \sigma_4^2 \end{pmatrix}.$$

Multiplizität

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- **Problematik:**

  Werden mehrere Hypothesentests auf derselben Datenbasis durchgeführt, steigt die Wahrscheinlichkeit dafür, mindestens eine dieser Hypothesen fälschlicherweise abzulehnen, weit über die für den Fehler 1. Art festgesetzte Grenze.

- **Beispiel:**

  Angenommen, es werden 5 Mittelwerte durch paarweise t-Tests zum 5%-Niveau miteinander verglichen. Dann beträgt die Anzahl der Vergleiche $5(5-1)/2 = 10$. Unter der Voraussetzung, dass die Vergleiche unabhängig sind, lässt sich eine Obergrenze für die gesamte Fehlerrate des Experimentes ableiten:
  $1-(1-0.05)10 = 0.40$

- **Konsequenz:**

  a) Der kritische Wert, der für die Signifikanz überschritten muss, muss erhöht werden (multiple Vergleiche).
  b) Die Fehlerrate für die paarweisen Vergleiche muss gesenkt werden, um die Fehlerrate des gesamten Experimentes unter 5% zu halten (p-Wert Adjustierung).

Multiple Vergleiche

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Simultane (multiple) Vergleiche mehrerer Erwartungswerte

- Aufdeckung von Unterschieden in den Erwartungswerten bei mehr als zwei Stichproben

Scheffé- und Tukey-Test

- Berechnung der Differenzen der Mittelwerte

$$\bar{Y}_{i.} - \bar{Y}_{j.}$$

  für jedes Paar $(i, j), \quad i < j$
- der Unterschied ist signifikant, falls er eine Grenzdifferenz übersteigt

Multiple Vergleiche

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression
Anhang: Wichtige Verteilungen

Scheffé-Test

- setzt normalverteilte Daten voraus
- kann auch auf Stichproben unterschiedlichen Umfangs angewendet werden

- Grenzdifferenz:

$$FSD_\alpha = \sqrt{(p-1) \cdot MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right) \cdot F_{p-1,N-p;1-\alpha}}$$

Multiple Vergleiche

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Tukey-Test

- anwendbar auf:
  normalverteilte Daten
  Stichproben mit gleicher Länge

- Grenzdifferenz:

$$HSD_\alpha = \sqrt{\frac{MSE}{n}} \cdot q_{p,N-p;1-\alpha}$$

α ... Signifikanzniveau
$q_{p,N-p}$ ... Studentisierte Spannweite

Multiple Vergleiche

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Fisher's Least Significant Difference-Test (LSD)

- falls der F-Test in der ANOVA H_0 verwirft, Berechnung der Differenz:

$$t^{-1}_{N-p,1-\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

  MSE-Wert der entsprechenden ANOVA-Tabelle
- Entscheidung auf signifikante Unterschiede für jedes Paar, falls

$$|\bar{Y}_i. - \bar{Y}_j.| > t^{-1}_{N-p,1-\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Multiple Vergleiche

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Dunnett-Test

- Idee: Vergleich verschiedener Faktoren (Einflußgrößen) mit einem Standardfaktor μ_1
- Test der $(p-1)$ Differenzen $\mu_p - \mu_1$, $\mu_{p-1} - \mu_1, \cdots, \mu_2 - \mu_1$ auf Signifikanz

Vorgehensweise:

1. alle Stichproben haben den gleichen Umfang n:
   - Vergleich der Mittelwerte $|\hat{\mu}_i - \hat{\mu}_1|$ mit $DSD_{p-1,\alpha}$

$$DSD_{p-1,\alpha} = s \cdot \sqrt{\frac{2}{n}} \cdot |d|_{p-1,\nu;1-\alpha}$$

$|d|_{p-1,\nu;1-\alpha}$ aus entsprechenden Tabellen

Multiple Vergleiche

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

2. Standardbehandlung vom Umfang $m > n$:
   - Berechnung der neuen Grenzdifferenz $DSD^{\text{korr}}_{p-1,\alpha}$

$$DSD^{\text{korr}}_{p-1,\alpha} = s \cdot \sqrt{\frac{1}{m} + \frac{1}{n}} \cdot |d|_{p-1,\nu;1-\alpha}$$

$$\cdot \left[1 + \frac{w_{p-1,\nu;1-\alpha}}{100}\left(1 - \frac{n}{m}\right)\right]$$

$w_{p-1,\nu;1-\alpha}$ ... Korrekturgröße; $\nu = N - p$

p-Wert Adjustierung

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

Bonferroni-Holm-Prozedur

- Berechnung der p-Werte für alle möglichen paarweisen Vergleiche (Zwei-Stichproben-t-Tests oder Wilcoxon-Rangsummen-Test)
- Ordnung der p-Werte in aufsteigender Folge
- Entscheidung auf signifikante Unterschiede, falls der kleinste p-Wert kleiner ist als $\frac{\alpha}{c}$, der zweitkleinste kleiner als $\frac{\alpha}{c-1}$ etc.
  (c ...Anzahl der paarweisen Tests)
- dieser Test übersteigt das global Signifikanzniveau α nicht

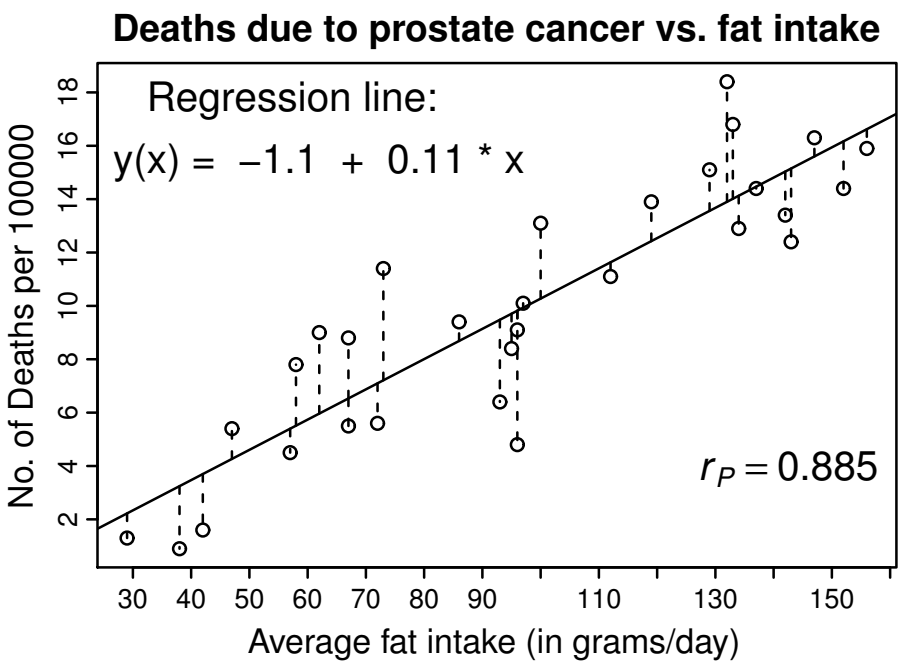
Bivariate Analyse: Regression

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
- Zufallsvariablen
- Verteilungen
- QQ-Plot
- Parameterschätzung
- Zentraler Grenzwertsatz
- Konfidenzintervalle
- Hypothesentests
- Stichprobenumfang
- Teststärke
- Wahl des Testverfahrens
- Anpassungstest
- Unabhängigkeitstest
- Parametrische Tests
- Nichtparametrische Tests
- Varianzanalyse
- Multiple Vergleiche
- Regression
- KQ-Residuen
- Modellgüte
- Standardfehler
- t-Tests
- Modellvarianten
- Box-Cox-Modelle
- Logistische Regression

- Situation: Metrische bivariate Datenstruktur mit nahezu linearem Zusammenhang ($r_p > 0.5$ und r_P signifikant von Null verschieden).
- Frage: Wie groß sind Steigung und Achsenabschnitt einer (optimalen) Geraden im Modell $y_t = \beta_1 + \beta_2 x_t + u_t$?
- Antwort: Geschätzte KQ-Regressionsfunktion $\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_t$

$\hat{\beta}_1$: Geschätzter erwarteter Achsenabschnitt. Erklärt die durchschnittliche Ausprägung von y falls $x = 0$.

$\hat{\beta}_2$: Geschätzte erwartete Steigung. Erklärt den durchschnittlichen Zuwachs in y, falls x um eine Einheit steigt.



KQ-Schätzer

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Lösung des Kleinste-Quadrate-Minimierungsproblems führt zu folgenden Schätzfunktionen:

$$\hat{\beta}_1(y_t, x_t, T) =: \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2(y_t, x_t, T) =: \hat{\beta}_2 = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

KQ-Residuen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Es ist nun möglich, den (unbekannten) Fehler u_t durch $\hat{u}_t$ (Residuum) folgendermaßen zu schätzen:

$$\hat{u}_t = y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t$$

- Die Varianz des Fehlers u_t ist generell unbekannt, kann aber geschätzt werden durch:

$$\widehat{\sigma_u^2} = \frac{1}{T-2} \sum_{t=1}^{T} \hat{u}_t^2$$

Streuungszerlegung und Bestimmtheitsmaß

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Man kann zeigen, dass im einfachen linearen Modell **mit Absolutglied** die sog. **Streuungszerlegung** gilt:

$$\underbrace{\sum_{t=1}^{T}(y_t-\bar{y})^2}_{\equiv TSS}=\underbrace{\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2}_{\equiv ESS}+\underbrace{\sum_{t=1}^{T}\hat{u}_t^2}_{\equiv RSS}$$

- Über die Streuungszerlegung gelangt man zum Bestimmtheitsmaß:

$$R^2=\frac{ESS}{TSS}=1-\frac{RSS}{TSS}$$

- Eigenschaften von R^2 im linearen Regressionsmodell **mit Absolutglied**:

$$R^2\in[0;1] \quad \text{und} \quad R^2=r_{yx}^2$$

t-Tests; Standardfehler

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- **Die Fehlervarianz σ_u^2 ist i.d.R. unbekannt: verwende daher zum Testen die geschätzten Standardfehler $se(\hat{\beta}_i)$:**

$$se(\hat{\beta}_1) = \sqrt{\frac{\widehat{\sigma_u^2} \sum_{t=1}^{T} x_t^2}{T \sum_{t=1}^{T} (x_t - \bar{x})^2}}, \quad se(\hat{\beta}_2) = \sqrt{\frac{\widehat{\sigma_u^2}}{\sum_{t=1}^{T} (x_t - \bar{x})^2}}$$

t-Tests für β_1 und β_2

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

1. Festlegung der Irrtumswahrscheinlichkeit α und der Hypothesen:

$$i) \quad H_0 : \beta_i = \beta_i^* \quad vs. \quad H_1 : \beta_i \neq \beta_i^* \text{ (zweiseitiger Test)}$$
$$ii) \quad H_0 : \beta_i \leq \beta_i^* \quad vs. \quad H_1 : \beta_i > \beta_i^* \text{ (einseitiger Test)}$$
$$iii) \quad H_0 : \beta_i \geq \beta_i^* \quad vs. \quad H_1 : \beta_i < \beta_i^* \text{ (einseitiger Test)}$$

3. Berechnung der Teststatistik:

$$\Rightarrow \quad t_{\beta_i} := \frac{\hat{\beta}_i - \beta_i^*}{se(\hat{\beta}_i)} \sim t_{(T-2)}$$

4. Ermittlung des kritischen Wertes $t_{T-2,1-\frac{\alpha}{2}}$.
5. Testentscheidung: sollte $|t_{\beta_i}| > t_{T-2,1-\frac{\alpha}{2}}$, kann die Nullhypothese abgelehnt werden, d.h. β_i ist signifikant von β_i^* verschieden.
6. Alternativ: Lehne H_0 ab, falls der p-Wert < 0.05

- Anmerkung: Die Standardwerte von β_i^* sind Null. Besonders wichtig ist der Test von $H_0 : \beta_2 = \beta_2^* = 0$

Modellvarianten

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Modellvarianten: Logarithmische Modelle:

$$\ln y_t = \beta_1 + \beta_2 x_t + u_t \qquad \text{(log-lin-Modell)}$$

$$y_t = \beta_1 + \beta_2 \ln x_t + u_t \qquad \text{(lin-log-Modell)}$$

$$\ln y_t = \beta_1 + \beta_2 \ln x_t + u_t \qquad \text{(log-log-Modell)}$$

Erweiterungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Power-Modell:

$$y_t = \beta_1 + \beta_2 x_t^p + u_t$$

wobei p bekannt sei.

$\hookrightarrow$ Je nach Definition von p, ergeben sich die folgenden Spezialfälle des Power-Modells:

$$y_t = \beta_1 + \beta_2 x_t^2 + u_t \qquad \text{(x-Quadrat-Modell)}$$

$$y_t = \beta_1 + \beta_2 \frac{1}{x_t} + u_t \qquad \text{(Reziprokes Modell)}$$

$$y_t = \beta_1 + \beta_2 \sqrt{x_t} + u_t \qquad \text{(Wurzelmodell)}$$

Box-Cox-Modelle

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression

- Box-Cox-Transformation:

$$y_t^{(p)} = \beta_1 + \beta_2 x_t^{(q)} + u_t$$

$$y_t^{(p)} := \begin{cases} \frac{y_t^p - 1}{p}, & \text{falls } p \neq 0, \\ \ln y_t, & \text{falls } p = 0. \end{cases}$$

$$x_t^{(q)} := \begin{cases} \frac{x_t^p - 1}{q}, & \text{falls } q \neq 0, \\ \ln x_t, & \text{falls } q = 0. \end{cases}$$

$\hookrightarrow$ Die Exponenten p, q können aus den Daten geschätzt werden und ermöglichen damit eine Auswahl zwischen linearen und nichtlinearen Modellen.

Logistische Regression

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
❖ Zufallsvariablen
❖ Verteilungen
❖ QQ-Plot
❖ Parameterschätzung
❖ Zentraler Grenzwertsatz
❖ Konfidenzintervalle
❖ Hypothesentests
❖ Stichprobenumfang
❖ Teststärke
❖ Wahl des Testverfahrens
❖ Anpassungstest
❖ Unabhängigkeitstest
❖ Parametrische Tests
❖ Nichtparametrische Tests
❖ Varianzanalyse
❖ Multiple Vergleiche
❖ Regression
❖ KQ-Residuen
❖ Modellgüte
❖ Standardfehler
❖ t-Tests
❖ Modellvarianten
❖ Box-Cox-Modelle
❖ Logistische Regression
Anhang: Wichtige Verteilungen

Das Logit-Modell

- Sei $y_t \in \{0, 1\}$ und x_t beliebig. Dann kann der Zusammenhang zwischen y_t und x_t dargestellt werden über

$$F(y_t) = \frac{1}{1 + \exp(-\beta_1 - \beta_2 x)} = \frac{1}{1 + e^{-\beta_1 - \beta_2 x}}$$

- Bei $F(y_t)$ handelt es sich um die Verteilungsfunktion der logistischen Verteilung, die eine stärkere Streuung als die Normalverteilung besitzt.
- Sei $\beta_1 + \beta_2 x_t$ abgekürzt durch z_t. Dann erhält man für $F(y_t)$ mit $p_t = \dfrac{1}{1 + \exp(-z_t)}$ ein lineares Modell der Form:

$$L_t = \ln\left(\frac{p_t}{1 - p_t}\right) = \beta_1 + \beta_2 x_t$$

- Die Größe $L_t = \dfrac{p_t}{1 - p_t}$ wird auch Logit genannt.

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

Anhang: Wichtige Verteilungen

Diskrete Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

Diskrete Verteilungen

- **Binomialverteilung**
- **Poisson-Verteilung**
- **Negative Binomialverteilung**

Diskrete Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

Binomialverteilung

- Wurde erstmals 1713 von Jakob Bernoulli in *Ars Conjectandi* veröffentlicht.

- Ergebnis von n voneinander unabhänigigen wiederholten Experimenten mit nur zwei möglichen Ausgängen.

- Entspricht einer „Urnenziehung mit Zurücklegen".

- Anwendungen: geheilt/nicht geheilt; UAW´s ja/nein

Diskrete Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- Wahrscheinlichkeitsfunktion :

$$P(X = x) = f_X(x|p; n) = \binom{n}{x} p^x (1-p)^{n-x} =: B(p; n)$$

$$mit \qquad \binom{n}{x} = \frac{n!}{x!(n-x)!} \qquad und \qquad x \in \{1; 2; ...; n\}$$

- Der Parameter p gibt die jeweilige Erfolgswahrscheinlichkeit an.

Diskrete Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

Sollte sich die Anzahl der Versuche auf $n = 1$ reduzieren, spricht man von einer

Bernoulliverteilung:

$$P(X = x) = f_X(x|p) = p^x(1-p)^{1-x} =: B(p)$$

- Verteilungsfunktion einer $B(p; n)$-Verteilten Zufallsvariablen :

$$P(X \leq x) = F_X(x|p) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

Diskrete Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- Erwartungswert und Varianz einer $B(p;n)$-Verteilten Zufallsvariablen X:

$$E(X) = np \qquad und \qquad Var(X) = np(1-p)$$

- Diese Verteilung besitzt die sog. Reproduktionseigenschaft. Es gilt:

$$B(p;n_1) + B(p;n_2) \xrightarrow{d} B(p;n_1+n_2)$$

- Approximation für $np(1-p) > 9$: $X \overset{iid}{\sim} N(np; np[1-p])$
- Approximation für $n \geq 50$, $p \leq 0,1$ und $np \leq 10$: $X \overset{iid}{\sim} Ps(np)$

Diskrete Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

Poisson-Verteilung

- Wurde von Poisson 1837 veröffentlicht. Die eigentliche Herleitung aus einer Sequenz von binomialverteilten ZV's wurde bereits von Moivre 1711 in „De Mensura Sortis" gegeben.

  Es gilt für $X \overset{iid}{\sim} B(p;n)$ und $w \subset \mathbb{N}_0$:

$$lim_{n\to\infty;np\to\lambda} \sum_w B_X(p;n) = \sum_w \frac{e^{-\lambda}\lambda^x}{x!} =: Ps(\lambda) = F_X(x|\lambda) = P(X \leq x)$$

- Wichtig: n wird „groß" und np bleibt endlich und hinreichend klein"!

- Daraus folgt: Anwendungen liegen bei der Erfassung von seltenen (λ klein) Ereignissen in großen Grundgesamtheiten (n groß).

Diskrete Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- Erste bekannte Anwendung (1898) „Tod durch Hufschlag bei der preußischen Armee" und „Selbstmord von Kindern unter zehn Jahren".

- Erwartungswert und Varianz einer $Ps(\lambda)$-Verteilten Zufallsvariablen X:

$$E(X) = \lambda \qquad und \qquad Var(X) = \lambda$$

- Diese Verteilung besitzt die sog. Reproduktionseigenschaft. Es gilt:

$$Ps(\lambda_1) + Ps(\lambda_2) \xrightarrow{d} Ps(\lambda_1 + \lambda_2)$$

- Approximation für $\lambda \geq 10$: $X \overset{iid}{\sim} N(\lambda; \lambda)$

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

Negative-Binomialverteilung

- Ergebnis von k voneinander unabhänigigen wiederholten Experimenten mit geometrischer Verteilung. Die Wahrscheinlichkeitsfunkion ist gegeben durch:

$$f_X(x|p;k) = P(X = x) = \binom{k+x-1}{k-1} p^x (1-p)^k$$

$$x \in \mathbb{N}_0 \qquad \eta, k > 0$$

- Fragestellung:

Wie hoch ist die Wahrscheinlichkeit nach $k-1$ Erfolgen im $k-1+x$-ten Versuch den k-ten Erfolg zu erzielen $(x = 1, 2, ...)$?

Diskrete Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- Erwartungswert und Varianz einer $nB(p,k)$-Verteilten Zufallsvariablen X:

$$E(X) = \frac{kp}{1-p} \qquad Var(X) = \frac{kp}{(1-p)^2} > \frac{kp}{1-p} = E(X)$$

Wird für Zähldatenmodelle verwendet
in denen die zu erwartende Varianz grösser als der Erwartungswert ist .

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

Stetige Verteilungen

- **Gleich-,Normal- und Lognormalverteilung**
- **Gamma-/χ^2- und Exponentialverteilung**
- **,t-, F-, Weibull- und Extremwertverteilung**

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **Gleichverteilung**

- Auch häufig unter dem Namen Rechteckverteilung zu finden.
- Dichtefunktion der gleichverteilten ZV X:

$$f_X(x|a,b) = \begin{cases} \frac{1}{b-a} & für \quad a \leq x \leq b \\ 0 & sonst \end{cases}$$

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- Erwartungswert und Varianz einer gleichverteilten ZV X:

$$E(X) = \frac{b+a}{2} \qquad und \qquad Var(X) = \frac{(b-a)^2}{12}$$

- Für $b = 1$ und $a = 0$ ist diese Verteilung ein Spezialfall der allgemeineren Betaverteilung:

$$f_X(x|a,b) = \frac{1}{\int_0^1 x^{a-1}(1-x)^{b-1}dx} x^{a-1}(1-x)^{b-1} \qquad 0 \leq x \leq 1$$

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **Normalverteilung**

- Faltet man die Gleichverteilung „oft genug" so erhält man im Limit die von C.F. Gauß hergeleitete Normalverteilung. Die Dichte hat folgende Darstellung:

$$f_X(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}; \qquad x \in I\!R; \qquad \mu \times \sigma^2 \in I\!R \times I\!R^+$$

- Erwartungswert und Varianz einer normalverteilten ZV X:

$$E(X) = \mu \qquad und \qquad Var(X) = \sigma^2$$

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **Lognormalverteilung**

Dichtefunktion:

$$f_X(x|\mu_*, \sigma_*^2) = \frac{1}{\sqrt{2\pi}\sigma_*}\frac{1}{x}exp\left\{-\frac{(lnx-\mu_*)^2}{2\sigma_*^2}\right\}; x \in I\!R^+; \mu_* \times \sigma_*^2 \in I\!R \times I\!R^+$$

- Erwartungswert und Varianz einer lognormalverteilten ZV X:

$$E(lnX) = \mu_* \qquad und \qquad Var(lnX) = \sigma_*^2$$

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **Gammaverteilung**

- Für die Modellierung von Warteschlangen in stetiger Zeit wird häufig auf die Gammaverteilung oder auf deren Spezialfälle unter bestimmten Parameterkonstellationen zurückgegriffen.

- Die Dichte ist durch folgenden Ausdruck gegeben:

$$f_X(x|\lambda, r) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} exp\{-\lambda x\} \qquad x \in I\!R^+; \qquad \lambda \times r \in I\!R^+ \times I\!R^+$$

- Wobei $\Gamma(r) = \int_0^\infty x^{r-1} exp\{-x\} dx$ die Gammafunktion mit Funktionsargument r ist.

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- Erwartungswert und Varianz einer gammaverteilten ZV X:

$$E(X) = \frac{r}{\lambda} \qquad und \qquad Var(X) = \frac{r}{\lambda^2}$$

- **χ^2-Verteilung**

- Für die Paramaterkonstellation $f_X(x|0.5; 0.5k)$ ergibt sich die χ^2-Verteilung:

$$f_X(x|k) = \frac{0.5^{k/2}}{\Gamma(k/2)} x^{k/2-1} exp\{-0.5x\} \qquad x \in I\!R^+; \qquad k \in I\!R^+$$

- Die Dichte einer χ^2-Verteilung ergibt sich auch aus der Summe von quadrierten $N(0; 1)$-verteilten Zufallsvariablen.

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **Exponentialverteilung**

- Für die Paramaterkonstellation $f_X(x|\lambda;1)$ ergibt sich die Exponentialverteilung:

$$f_X(x|\lambda) = \lambda exp\{-\lambda x\} \qquad x \in I\!R^+; \qquad \lambda \in I\!R^+$$

- Erwartungswert und Varianz einer exponentialverteilten ZV X verhalten sich analog zur Gammaverteilung mit $r = 1$, bzw. für eine χ^2-verteilte ZV X mit $r = 0.5k$ und $\lambda = 1$.

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **t-Verteilung (Student)**

- Sei $X \overset{iid}{\sim} N(0;1)$ und $Y \overset{iid}{\sim} \chi^2(k)$ voneinander unabhängige Zufallsvariablen, so folgt die unten angeführte Funktion einer zentralen t-Verteilung mit k Freiheitsgraden.

$$W = \frac{X}{\sqrt{Y/k}} \sim t_k$$

- Die Dichtefunktion ist gegeben durch:

$$f_W(w|k) = \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)} \cdot \frac{1}{\sqrt{k\pi}} \cdot \frac{1}{(1+w^2/k)^{(k+1)/2}} dw \qquad w \in I\!R; \qquad k \in I$$

- Erwartungswert und Varianz einer t-verteilten ZV W:

$$E(W) = 0 \quad für \quad k > 1 \qquad und \qquad Var(W) = \frac{k}{k-2} \quad für \quad k > 2$$

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **F-verteilung (Fisher)**

- Seien $X \overset{iid}{\sim} \chi^2(m)$ und $Y \overset{iid}{\sim} \chi^2(n)$ voneinander unabhängige Zufallsvariablen, so folgt die nachstehende Funktion einer zentralen F-Verteilung mit $m; n$ Freiheitsgraden.

$$R = \frac{X/m}{Y/n} \sim F_{(m;n)}$$

- Die Dichtefunktion ist gegeben durch:

$$f_R(r|m;n) = \frac{\Gamma[(m+n)/2]}{\Gamma[m/2]\Gamma[n/2]} \left(\frac{m}{n}\right)^{m/2} \cdot \frac{r^{(m-2)/2)}}{[1+(m/n)r]^{(m+n)/2}}$$

$$m \times n \in I\!N \times I\!N \qquad und \qquad r \in I\!R^+$$

- Erwartungswert und Varianz einer F-verteilten ZV R:

$$E(W) = \frac{n}{n-2} \quad für \quad n > 2 \qquad und \qquad Var(W) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **Weibullverteilung**

- Ein weiteres Instrument zur Verweildaueranalyse ist die Weibullverteilung. Die Dichte hat folgende Gestalt:

$$f_X(x|a;b;c) = \frac{c}{b}\left(\frac{x-a}{b}\right)^{c-1} exp\left\{-\left(\frac{x-a}{b}\right)^{c}\right\}$$

$$a \times b \times c \in I\!R \times I\!R^{+} \times I\!R^{+} \qquad und \qquad x \geq a$$

- Erwartungswert und Varianz einer weibullverteilten ZV X:

$$E(X) = a + b\Gamma\left(1 + \frac{1}{c}\right)$$

$$Var(X) = b^2\left\{\Gamma\left(1 + \frac{2}{c}\right) - \Gamma^2\left(1 + \frac{1}{c}\right)\right\}$$

Stetige Verteilungen

1. Basiswissen Statistik
2. Epidemiologie
3. Statistische Inferenz
Anhang: Wichtige Verteilungen
❖ Diskrete Verteilungen
❖ Stetige Verteilungen

- **Extremwertverteilung**

- Die Dichte hat folgende Gestalt:

$$f_X(x) = exp\left\{-exp\left\{-\frac{x-\alpha}{\beta}\right\}\right\}; \qquad \alpha \times \beta \in I\!R \times I\!R^+$$

- Erwartungswert und Varianz einer extremwertverteilten ZV X:

$$E(X) = a + b\gamma \qquad und \qquad Var(X) = \frac{\pi^2\beta^2}{6}$$

- Wobei $\gamma \approx 0.577216$ die Euler-Konstante darstellt.

Dieses Skript basiert anteilig auf einem englischsprachigen Kompendium. Wir danken Herrn Dr. Gerrit Eichner, Universität Giessen, für die freundliche Unterstützung.