Westfälische Wilhelms-Universität Münster

Datenanalyse

Dr. Ingolf Terveer

Data Science: Statistik und Optimierung

Sommersemester 2022

WIRTSCHAFTS INFORMATIK

1. Einführung

- Einordnung DA
- Aufbau der Vorlesung
- Organisation
- Literatur

Rückblick "Daten und Wahrscheinlichkeiten"

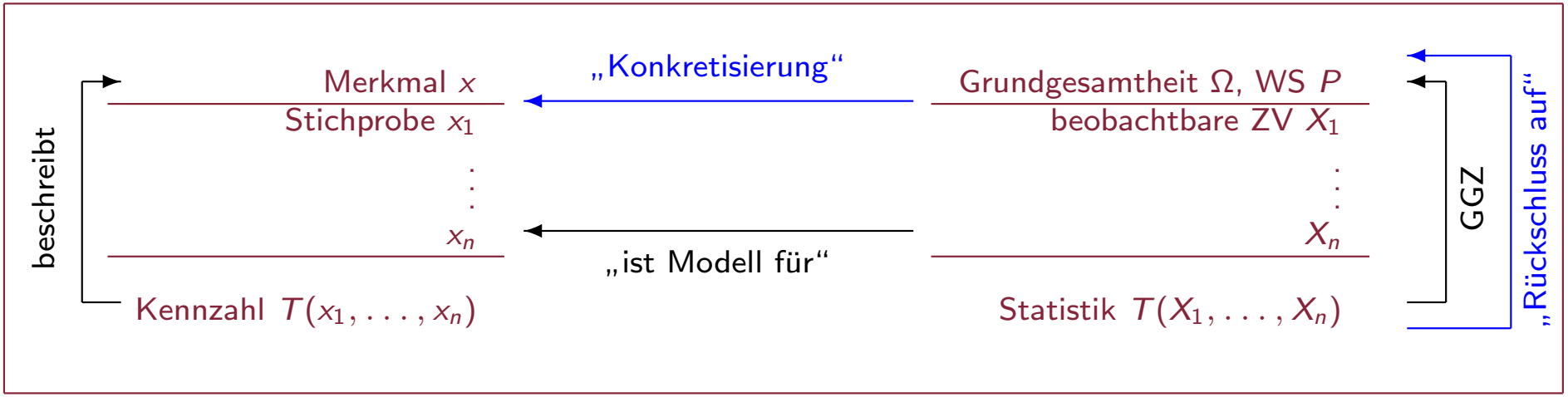
I. Deskriptive Statistik

- Analyse eines und zweier Merkmale
- Grafische Aufbereitung von Daten
- Lage, Streuungs- und Zusammenhangsmaße
- Clusteranalyse
- Statistische Software (R)

II. Wahrscheinlichkeitsrechnung

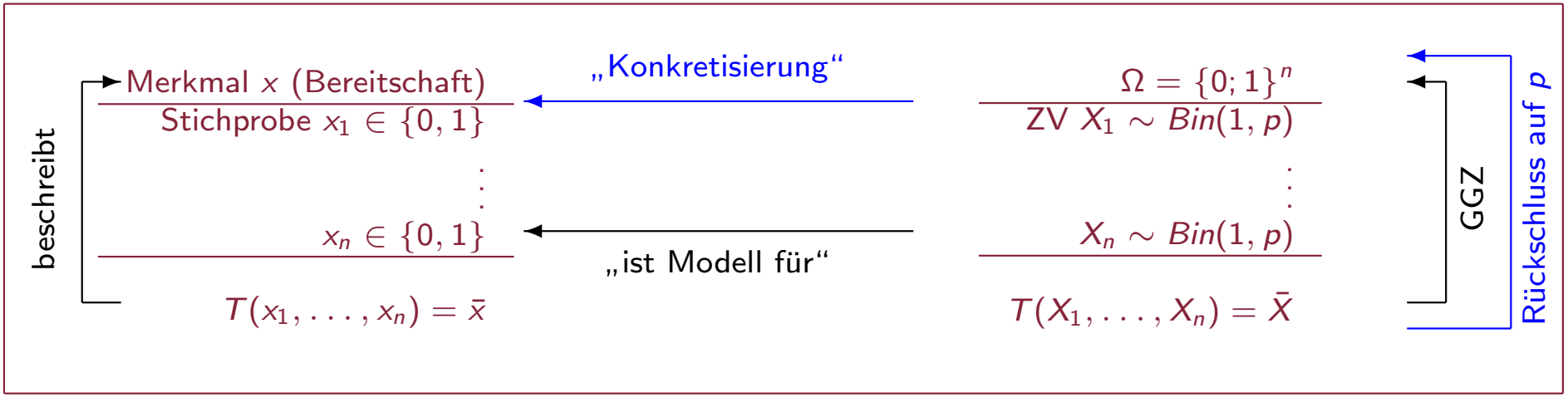
- Zufallsereignisse und Wahrscheinlichkeiten
- Zufallsvariablen
- Erwartungswert, (Ko-)varianz, Verteilungen von Zufallsvariablen
- Zufallsvektoren, Zusammenhangsmaße
- Gesetze großer Zahlen

Einordnung der Veranstaltungen DuW und EDA



- ↪ **Deskriptive Statistik:** Aggregation von Merkmalsausprägungen der einzelnen Merkmalsträger zu Aussagen über die Stichprobe.
- ↪ **Wahrscheinlichkeitsrechnung:** WS-Modellierung und -Kalkül.
- ↪ **Gesetze großer Zahlen:** Verknüpfung von Daten und WS-Modellen, Zusammenhang zwischen theoretischen und empirischen Kennzahlen.
- ↪ **Induktive Statistik:** Schluss von Stichprobe auf Grundgesamtheit
- ↪ **Konkretisierung:** Rückschluss von der Grundgesamtheit auf den Einzelfall

Beispiel: Zustimmung zu einer Umfrage ($n = 50$)



- 50 Personen gefragt: „Würden Sie an Umfrage zu Corona teilnehmen?“
- Modell: Bernoulli-Kette (n st.u. Einzelexperimente, $X_i \sim Bin(1,p)$), $nT(X) = X_1 + \cdots + X_n \sim Bin(n,p)$ (sog. **Stichprobenverteilung**)
- Gesetz großer Zahlen: $T(X) \overset{n\to\infty}{\longrightarrow} p$ fast sicher, aber: p unbekannt, n endlich.
- Induktive Statistik: Verwende $T(X)$ als Ersatz für p („Schätzer“). Quantifiziere Fehlentscheidungen mit Hilfe der Verteilung von $T(X)$.
- Konkretisierung („Plug-In“): Interpretation von $T(X)$ als Rücklaufquote in der späteren größeren Umfrage mit mehr Befragten.

$\hookrightarrow$ Angenommen k von n Befragten sind bereit zu der Umfrage.

$\hookrightarrow$ Die größere Umfrage sollte nur bei ausreichend hoher Rücklaufquote p durchgeführt werden, diese ist aber unbekannt.

$\hookrightarrow$ Typische, datenabhängig zu beantwortende Fragestellungen ($\alpha \in]0;1[$):

- □ Finde eine Größe, die p möglichst gut „ersetzt“ $\rightsquigarrow$ **Schätzung**
- □ Finde einen Bereich $[p_1(k); p_2(k)]$, in dem p mit $100(1-\alpha)\%$ Wahrscheinlichkeit liegt $\rightsquigarrow$ **Konfidenzintervall**
- □ Die größere Umfrage wird nur bei ausreichendem Interesse durchgeführt. Kann (z.B.) $p > 0,2$ aus der Vorumfrage mit $n = 50$ statistisch (d.h. zu einem Signifikanzniveau $100\alpha\%$) gesichert werden? $\rightsquigarrow$ **Hypothesentests**

Die Probleme hängen zusammen, werden mit der **Stichprobenverteilung**, d.h. der Verteilung von $T(X) = \bar{X}$ bzw. $nT(X) = X_1 + \cdots + X_n$ gelöst.

$\hookrightarrow$ Modellerweiterungen: Die Verteilungen von X_i hängen von weiteren beobachtbaren Merkmalen ab (z.B. sozialer Status), d.h. Modellierung von p abhängig von diesen Merkmalen $\rightsquigarrow$ **bedingte Verteilungen, Regression**

Aufbau der Vorlesung DA

I. Stichprobenverteilungen

II. Bedingte Erwartung

III. Schließende Statistik

- Punktschätzung
- Intervallschätzung
- statistische Tests

IV. Regression

Zusammenhang zwischen abhängiger Variable Y u. erklärenden Variablen X.

Vorlesung/Präsenzübung: I. Terveer

- ↪ Vorlesung: Videos (Folien), werden sukzessive bereitgestellt
- ↪ Übung: Zum Selbststudium enthalten die Folien Übungsteile, die für die folgende Präsenzübung vorbereitet werden müssen.
  Termine der Präsenzübung (Mi im J2, Di im Leo 1):

Mi	06.04.2022	10:00	12:00	V1: Einleitung
Mi	13.04.2022	10:00	12:00	V2: Stichprobenverteilungen
Mi	20.04.2022	10:00	12:00	V3: Bedingte Verteilung und bedingte Erwartung
Mi	27.04.2022	10:00	12:00	V4: ML-Schätzer und MM-Schätzer
Mi	04.05.2022	10:00	12:00	V5: Intervall- und Bereichsschätzer
Mi	11.05.2022	10:00	12:00	V6: Tests, Fehler 1./2. Art, Konstruktion. Einstichprobentests für EW
Mi	18.05.2022	10:00	12:00	V7: Anteilswerttests, Gütefunktion
Mi	25.05.2022	10:00	12:00	V8: Zweistichprobentests: Mittelwerttests, Wilcoxon-Test ( )
Mi	01.06.2022	10:00	12:00	V9: Verteilungstests: Chi-Quadrat-Unabh, Anp, KS
Mi	15.06.2022	10:00	12:00	V10: Regressionsanalyse theor. Lösung, Modell, Schätzansatz, einfache LR
Mi	22.06.2022	10:00	12:00	V11: KQ-Schätzer, Bestimmtheitsmaß, Hypothesentests
Mi	29.06.2022	10:00	12:00	V12: Parametertests, Bereichsschätzung, Residualanalyse
Di	05.07.2022	16:00	18:00	V13: Indikatorvariablen ( )
Mi	06.07.2022	10:00	12:00	V14: Fragestunde

- ↪ Bereitstellung von Lösungen: nach der Präsenzübung
- ↪ Notwendige Vorkenntnisse: MAWIWI (Differential- und Integralrechnung, Matrizenrechnung), DuW

Tutorium („Übungen zur Vorlesung Datenanalyse“)

- ↪ Wöchentlich: Übungszettel
  - □ u.a. Fortsetzung der R-Übungsaufgaben aus DuW
  - □ u.a. Aufgaben zur Klausurvorbereitung
  - □ Ausgabe nach VL-Übung,
  - □ Bearbeitung im Selbststudium
- ↪ In der Folgewoche:
  - □ Übungstermin, Di 16-18: Besprechung von Fragen
  - □ Bereitstellung von Musterlösungen (Ende der Woche)
  - □ Ggf. Besprechung von Fragen zu Übung der Vorwoche.

Weitere Informationen:

↪ DA im Learnweb (Materialien, Foren,Evaluation,. . .):

- https://sso.uni-muenster.de/LearnWeb/learnweb2/course/view.php?id=61044
- kein Einschreibeschlüssel bis Ende April, danach Einschreibung nur mit pM

↪ Informationen zur Prüfung:

- keine Anrechnung von Übungsaufgaben
- Klausur „Datenanalyse und Simulation“ 18.7.
  (120 min, voraussichtlich 80P DA, 40P Sim)

  Prüfungsformat: voraussichtlich Präsenzklausur.
- R ist Bestandteil der Klausur

Literatur

[1] Auer, B./Rottmann, H.: Statistik und Ökonometrie für Wirtschaftwissenschaftler, 2. Aufl. Gabler. 2011

[2] Durbin, J.: Distribution Theory for Tests Based on the Sample Distribution Function. Philadelpha: Society for Industrial and Applied Mathematics. 1973.

[3] Lilliefors, H.W.: On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. JASA, 62(318), 299-402. 1967

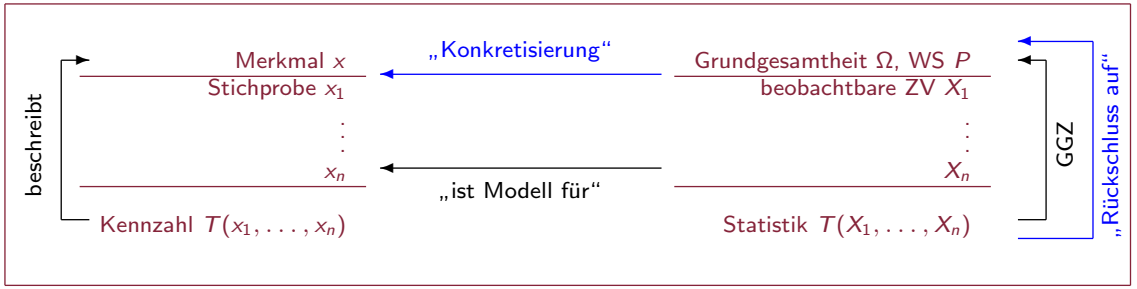
[4] Mood, A.M./Graybill, F.A./Boes, D.C.: Introduction to the theory of Statistics. McGraw-Hill, Singapore. 1974.

[5] Montgomery, D.C./Peck, E.A./Vining, G.G.: Introduction to Linear Regression Analysis, 5th ed. Wiley, Hoboken, New Jersey. 2012.

2. Stichprobenverteilungen

- Stichprobenstatistiken/Stichprobenverteilungen
- Transformationen von univariaten Zufallsvariablen
- Transformationsregeln für bivariate Verteilungen
- Funktionen zweier unabhängiger Zufallsvariablen
- Maximum/Minimum unabhängiger Zufallsvariablen
- Übersicht von Verteilungen und Transformationsregeln

2.1 Stichprobenstatistiken/Stichprobenverteilungen



- ↪ Beim Rückschluss von den Daten/Realisierungen auf das Modelll (Ω, P) wird Zufallsvariable (**Stichprobenstatistik**) $T(X_1, \ldots, X_n)$ verwendet.
- ↪ Der Rückschluss kann (bei endlichem n) nicht exakt sein.
  - □ Fehler/Ungenauigkeiten werden mit Wahrscheinlichkeiten bewertet.
  - □ Die Verteilung von T abhängig vom WS-Modell (Ω, P) wird **Stichprobenverteilung** genannt.
- ↪ Abgrenzung von der **Modellverteilung**: Verteilung(sannahme) der ZV $X_1, \ldots, X_n$ (ggf. mit Abhängigkeitsstruktur)

Beispiele für Modellverteilungen

- ↪ hier: Einstichprobenmodelle (meist) mit u.i.v. Zufallsvariablen $X_1, \ldots, X_n$
  - □ „u.“: $X_1, \ldots, X_n$ sind stochastisch unabhängig, d.h.
    $\forall$ Ereignisse A_i: $P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$
  - □ „i.“: Die $X_1, \ldots, X_n$ haben dieselbe („identische“) Verteilung, d.h.
    $\forall$ Ereignisse A und $\forall i, j$:$P(X_i \in A) = P(X_j \in A)$
- ↪ Beispiele diskreter Modelle
  - □ Bernoulli-Ketten: $X_i \sim Bin(1, p)$ (Binomialverteilung)
  - □ Klassierte Merkmale: $X_i \sim M(1, p_1, \ldots, p_K)$ (Multinomialverteilung)
  - □ Wartezeiten: $X_i \sim Geo(p)$ (geometrische Verteilung)
  - □ Ankunftszähler: $X_i \sim Pois(\lambda)$ (Poisson-Verteilung)
- ↪ Beispiele stetiger Modelle:
  - □ Wartezeiten: $X_i \sim Exp(\lambda)$ (Exponentialverteilung)
  - □ Messfehler: $X_i \sim \mathcal{N}(\mu, \sigma^2)$ (Normalverteilung)
  - □ Einkommen: $X_i \sim Par(\lambda, c)$ (Pareto-Verteilung)
- ↪ Diese Modellannahmen (inkl. „u.i.v.“) müss(t)en im Einzelfall geprüft werden.
  - □ Hierfür gibt es statistische Tests (Verteilungstests, s.u.)

Beispiele für Stichprobenstatistiken (Einstichprobenfall)

- $\hookrightarrow$ $T(X) = X_1 + \cdots + X_n$ bzw. $T(X) = \bar{X}$ bei Rückschluss auf Erwartungswert, z.B. bei Bernoulli-Ketten, klassierten Merkmale, Wartezeiten, Ankunftszähler
- $\hookrightarrow$ $T(X) = \max(X_1, \ldots, X_n)$ z.B. bei Einkommen
- $\hookrightarrow$ $T(X) = \sum(X_i - \bar{X})^2$ bzw. $T(X) = X_1^2 + \cdots + X_n^2$, z.B. bei Rückschluss auf die Varianz.

Bei Modell-Verteilungen mit mehreren Parametern müssen auch mehrere Stichprobenstatistiken gleichzeitig verwendet werden, z.B.:

- $\hookrightarrow$ Modellverteilung $\mathcal{N}(\mu, \sigma^2)$
  - $\square$ $T_1(X) = \bar{X}$
  - $\square$ $T_2(X) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- $\hookrightarrow$ Modellverteilung $\mathcal{R}(a, b)$ (Rechteckverteilung)
  - $\square$ $T_1(X) = \min(X_1, \ldots, X_n)$
  - $\square$ $T_2(X) = \max(X_1, \ldots, X_n)$
- $\hookrightarrow$ Für die Verfahren der schließenden Statistik wird dann manchmal die gemeinsame Verteilung von T_1, T_2 benötigt.

Ausgangssituation: Zugrunde liegt eine Stichprobe mit n (meist) unabhängigen Zufallsvariablen $X_1, \ldots, X_n$ oder eine ZV X

- $\hookrightarrow$ Wie ist $Y = X^2$ oder allgemeiner $Y = X^a$ verteilt? (Stichwort Datentransformation)
- $\hookrightarrow$ Wie ist $Y = X_1^2 + \ldots + X_n^2$ verteilt? (Stichwort χ^2-Verteilung)
- $\hookrightarrow$ Wie ist generell die Summe von Zufallsvariablen verteilt? (Faltung)
- $\hookrightarrow$ Wie ist das Maximum der Zufallsvariablen verteilt? (Ausfall von Netzwerken)
- $\hookrightarrow$ Wie ist generell die (gemeinsame) Verteilung der Zufallsvariable(n)

$$Y_1, \ldots, Y_k \text{ mit } Y_k = g_k(X_1, \ldots, X_n)?$$

Wir betrachten hier einige wichtige Spezialfälle:

- ☐ Univariate Verteilungen: $n = 1$ und $k = 1$
- ☐ Bivariate Verteilungen: $n = 2$ und $k = 2$
- ☐ Bivariate Verteilungen: $n = 2$ und $k = 1$

2.2 Transformationen von univariaten Zufallsvariablen

- Die Zufallsvariable X wird über eine Funktion $g(\cdot)$ in die Zufallsvariable Y überführt. Gesucht ist die Dichtefunktion von Y
- spezielle Vorgehensweisen, je nach Verteilungsart
  - □ diskrete Verteilungen: Bündelung von Wahrscheinlichkeiten, übertragbar auf Zufallsvektoren
  - □ stetige Verteilungen:
    Substitutionsregel für Integrale $\rightsquigarrow$ Dichtetransformationssatz

Diskreter Fall: Einleitendes Beispiel

Betrachte die diskrete, gleichverteilte ZV X mit Träger $\mathcal{T}_X = \{1, 2, 3, 4, 5\}$ und die transformierte ZV $Y = g(X) := (X - 2)^2$.

Es gilt $\mathcal{T}_Y = \{0, 1, 4, 9\}$ gilt, denn nur diese Werte werden von Y mit positiver W'keit angenommen.

Nun können wir einfach die Vorkommenswahrscheinlichkeiten aufsummieren:

$$
\begin{aligned}
f_Y(0) &= P(Y = 0) = P(X = 2) = 1/5, \\
f_Y(1) &= P(Y = 1) = P(X \in \{1, 3\}) = P(X = 1) + P(X = 3) = 2/5 \\
f_Y(4) &= P(Y = 4) = P(X = 4) = 1/5 \\
f_Y(9) &= P(Y = 9) = P(X = 5) = 1/5
\end{aligned}
$$

Diskreter Fall: Allgemein

Sei X <u>diskrete</u> ZV mit Träger $\mathcal{T}_X$ und Dichte $f_X(x) = P(X = x)$. Sei ferner $Y = g(X)$ univariate Transformation von X. Dann gilt:

$$f_y(y) = P(Y = y) = P(g(X) = y) = \sum_{\substack{x \in \mathcal{T}_X, \\ g(x)=y}} f_X(x).$$

Übung: Betrachte die ZV X mit Träger $\mathcal{T}_X = \{-2, -1, 0, 1, 2\}$ und Dichte $f_X(x) = 1/|\mathcal{T}_X|$. Bestimme die Verteilung von $Y = 2|X| + 1$.

Stetiger Fall: Idee

Wir gehen nun zum allgemeinen stetigen Fall über. Sei also X stetig verteilt mit Verteilungsfunktion $F_X(x)$. Sei ferner $Y = g(X)$ univariate Transformation von X.

Ansatz:

- $\hookrightarrow$ Bestimme die Verteilungsfunktion $F_Y(y)$ von Y mit Hilfe der bekannten Verteilungsfunktion von X.
- $\hookrightarrow$ Dazu muss die Ungleichung $g(X) \leq y$ so nach X aufgelöst werden, dass die Verteilungsfunktion F_X für die Berechnung der WS $P(g(X) \leq y)$ eingesetzt werden kann.
- $\hookrightarrow$ Die Dichte $f_Y(y)$ ist dann („fast sicher“) die Ableitung von $F_Y(y)$.

Stetiger Fall: Beispiel

Sei $X \sim \mathcal{N}(0,1)$-verteilt, d.h. mit Dichte φ und VF Φ

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$$

Betrachte die Transformation $Y = X^2$ und berechne die Verteilungsfunktion F_Y.

$\hookrightarrow$ Für $y < 0$ ist $F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = 0$ $\quad (X^2 \geq 0)$

$\hookrightarrow$ Für $y = 0$ ist $F_Y(0) = P(Y \leq 0) = P(X^2 \leq 0) = P(X = 0) = 0$ (X stetig v.)

$\hookrightarrow$ Für $y > 0$ ist:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= P(X \leq \sqrt{y}) - P(X < -\sqrt{y}) \\ &= P(X \leq \sqrt{y}) - P(X \leq -\sqrt{y}) \\ &= P(X \leq \sqrt{y}) - (1 - P(X \leq \sqrt{y})) \\ &= 2\Phi(\sqrt{y}) - 1 \end{aligned}$$

Die Dichte dazu ist dann

$\hookrightarrow$ $f_Y(y) = F_Y'(y) = 2\Phi'(\sqrt{y})\frac{1}{2\sqrt{y}} = \varphi(\sqrt{y})\frac{1}{\sqrt{y}} = \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}}$, $y > 0$.

Das ist die Dichte einer so genannten $\chi^2(1)$-Verteilung.

Übung: Es sei X eine stetig auf $[0;1]$ gleichverteilte Zufallsvariable. Berechnen Sie die Verteilung von $Y = -2\ln(X)$.

Alternative: Dichtetransformationssatz

Verwendung der Substitutionsregel anstelle der Kettenregel ergibt:

Dichtetransformationssatz

Sei X eine stetige Zufallsvariable mit Dichtefunktion $f_X(\cdot)$ und $\mathcal{T} = \{x : f_X(x) > 0\}$. Weiter sei $g : \mathcal{T} \to \Omega \subseteq \mathbb{R}$ eine Funktion mit den folgenden Eigenschaften

(i) g ist **bijektiv** (umkehrbar eindeutig) mit Umkehrfunktion $h = g^{-1}$,

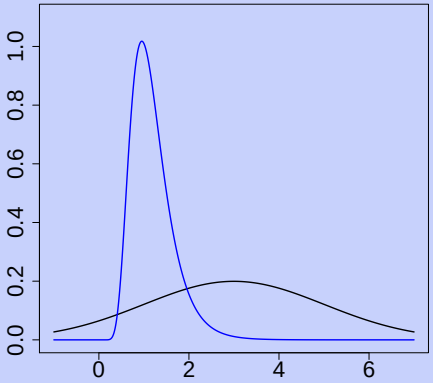
(ii) g <u>und</u> (dann auch!) h sind **stetig differenzierbar** mit $h'(y) \neq 0$ für alle $y \in \Omega$

Dann ist $Y = g(X)$ eine stetige Zufallsvariable mit der Dichte

$$f_Y(y) = f_X(h(y)) \cdot |h'(y)| \cdot \mathbb{1}_\Omega(y)$$

$X \sim \mathcal{N}(\mu, \sigma^2)$, gesucht ist Verteilung von $Y = g(X) = e^X$

$$h = g^{-1}(y) = \ln(y) \text{ sowie } h'(y) = \frac{1}{y}$$



Mit dem Dichtetransformationssatz folgt nun unmittelbar:

$$f_Y(y) = \underbrace{|h'(y)|}_{=1/y} f_X(h(y)) \cdot \mathbb{1}_\Omega(y)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{y} \cdot e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}} \mathbb{1}_{[0;\infty)}(y)$$

Das ist die Dichte einer Lognormalverteilung (Verwendung: Bedienungs-/Wartezeiten. Alternative zur Normalverteilung, wenn nur positive Werte auftreten können.)

Übung: Es sei X eine stetig auf $[0;1]$ gleichverteilte Zufallsvariable. Berechnen Sie die Verteilung von $Y = -2\ln(X)$ mit Hilfe des Dichtetransformationssatzes.

Übung: Sei $X \sim \mathcal{N}(\mu, \sigma^2)$. Zeige mit dem Dichtetransformationssatz, dass die lineare Transformation $Z = \frac{X-\mu}{\sigma}$ eine $\mathcal{N}(0,1)$-Verteilung besitzt.

2.3 Transformationsregeln für bivariate Verteilungen

- ↪ Gemeint sind Transformationen

$$(X, Y) \mapsto (U, V) = (g_1(X, Y), g_2(X, Y))$$

bei denen sich X, Y aus U, V „rekonstruieren lassen,z.B.

  - □ $(X, Y) \mapsto (X + Y, X - Y)$
  - □ $(X, Y) \mapsto (X \cdot Y, X/Y)$

- ↪ Zu den sich ergebenden bivariaten Verteilungen lassen sich jeweils die Randverteilungen berechnen, z.B.

  - □ $(X, Y) \mapsto (U, V) = (X + Y, X - Y) \quad \rightsquigarrow \quad X + Y$ (Summenverteilungen)
  - □ $(X, Y) \mapsto (U; V) = (X \cdot Y, X/Y) \quad \rightsquigarrow \quad X/Y$ (z.B. t-Verteilung)

- ↪ Dabei Beschränkung auf den stetigen Fall
- ↪ Hilfsmittel: Bivariater Dichtetransformationssatz.

Dichtetransformationssatz (bivariat), Verteilung von $(U, V) = g(X, Y)$

Es sei (X, Y) ein stetig verteilter Zufallsvektor mit gemeinsamer stetiger Dichte $f = f_{X,Y}$ und $\mathcal{T} = \{(x, y) \in \mathbb{R}^2 : f(x, y) > 0\}$

Weiter sei $g = (g_1, g_2) : \mathcal{T} \to \Omega \subseteq \mathbb{R}^2$ eine Funktion mit folgenden Eigenschaften:

(i) g ist stetig differenzierbar.

(ii) g ist bijektiv mit Umkehrfunktion $h = (h_1, h_2) : \Omega \to \mathcal{T}$, d.h. insbesondere

$$\left\{ \begin{array}{rcl} u & = & g_1(x, y) \\ v & = & g_2(x, y) \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{rcl} x & = & h_1(u, v) \\ y & = & h_2(u, v) \end{array} \right\} \qquad \forall (x, y) \in \mathcal{T}, \forall (u, v) \in \Omega$$

(iii) h ist differenzierbar, $\mathcal{J}_h(u, v) = \begin{pmatrix} \partial h_1/\partial u & \partial h_1/\partial v \\ \partial h_2/\partial u & \partial h_2/\partial v \end{pmatrix}$ ist invertierbar $\forall (u, v) \in \Omega$

Dann ist $(U, V) = g(X, Y)$ ein stetiger Zufallsvektor mit der Dichte

$$f_{U,V}(u, v) = f_{X,Y}(h(u, v)) \cdot |\det(\mathcal{J}_h(u, v))| \cdot 1_\Omega(u, v)$$

- Bezeichnung: $\mathcal{J}_h(u, v)$ heißt **Jacobi-Matrix** von h in (u, v).
- Transformationssatz übertragbar auf den Fall multivariater Verteilungen

Generisches Beispiel

Gegeben: Zufallsvektor (X, Y) mit stetiger WS-Dichte $f_{X,Y}$. Gesucht: die gemeinsame Verteilung von $(U, V) = (X + Y, X - Y)$. Dazu:

- $\begin{Bmatrix} u{=}x+y \\ v{=}x-y \end{Bmatrix}$ nach x, y auflösen: $\Leftrightarrow \begin{Bmatrix} u+v{=}2x \\ u-v{=}2y \end{Bmatrix} \Leftrightarrow \begin{Bmatrix} x{=}(u+v)/2 \\ y{=}(u-v)/2 \end{Bmatrix}$
- Schreibe Lösung als Funktion $h(u, v)$ und berechne die Jacobi-Matrix und den Betrag ihrer Determinante:
  $h(u, v) = (\frac{u+v}{2}, \frac{u-v}{2})$, $\mathcal{J}_h(u, v) = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}$ und $|\det \mathcal{J}_h(u, v)| = \frac{1}{2}$
- Wende den Dichtetransformationssatz an: $f_{U,V}(u, v) = f_{X,Y}(\frac{u+v}{2}, \frac{u-v}{2}) \cdot \frac{1}{2}$

Daraus (z.B.) die Randverteilungen:

- $f_{X+Y}(u) = \int\limits_{-\infty}^{\infty} f_{X,Y}(\frac{u+v}{2}, \frac{u-v}{2}) \cdot \frac{1}{2} dv \overset{t=\frac{u+v}{2}, dt=\frac{1}{2}dv, u-t=\frac{u-v}{2}}{=} \int\limits_{-\infty}^{\infty} f_{X,Y}(t, u-t) dt$
- $f_{X-Y}(v) = \int\limits_{-\infty}^{\infty} f_{X,Y}(\frac{u+v}{2}, \frac{u-v}{2}) \cdot \frac{1}{2} du \overset{t=\frac{u-v}{2}, dt=\frac{1}{2}du, v+t=\frac{u+v}{2}}{=} \int\limits_{-\infty}^{\infty} f_{X,Y}(v+t, t) dt$

Randverteilungen insb. wichtig bei st.u. X, Y, d.h. für $f_{X,Y}(x, y) = f_X(x) f_Y(y)$

Übung: Berechnen Sie in der obigen Situation eine Dichte der gemeinsamen Verteilung von $(U, V) = (X + Y, X)$ und daraus eine allgemeine Formel für die Dichte von $X + Y$ bei st.u. X, Y.

Ist das Ziel nur die Berechnung der Verteilung von $X + Y$, so ist dieser Rechenweg etwas einfacher als der Weg über $(X + Y, X - Y)$.

$f_{X,Y}(x,y) = e^{-(x+y)} = e^{-x} \cdot e^{-y}$, $x, y \geq 0$ (s.o., X, Y, st.u.)

$\hookrightarrow$ Gesucht: Dichte von $(U,V) = (X+Y, X-Y)$ bzw. $U = X+Y$

$$\hookrightarrow \quad f_{U,V}(u,v) = f_{X,Y}\left(\tfrac{u+v}{2}, \tfrac{u-v}{2}\right) \cdot \tfrac{1}{2}$$

$$= \tfrac{1}{2} \cdot e^{-\left(\frac{u+v}{2} + \frac{u-v}{2}\right)} \cdot 1_{[0;\infty[^2}\left(\tfrac{u+v}{2}, \tfrac{u-v}{2}\right)$$

$$= \tfrac{1}{2} \cdot e^{-u} \cdot 1_{[0;\infty[^2}\left(\tfrac{u+v}{2}, \tfrac{u-v}{2}\right)$$

Umformung der Bedingung $1_{[0;\infty]^2}(\ldots) = 1$:
$u+v \geq 0, u-v \geq 0 \Leftrightarrow v \geq -u, v \leq u \Leftrightarrow |v| \leq u \Leftrightarrow |v| \leq u, u \geq 0$

$$= \tfrac{1}{2} e^{-u} \cdot 1_{[0;\infty[}(u) \cdot 1_{[-u;u]}(v)$$

$\hookrightarrow$ Dichte von U ist

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_{U,V}(u,v)dv \\ &= \int_{-\infty}^{\infty} \tfrac{1}{2} e^{-u} \cdot 1_{[0;\infty[}(u) \cdot 1_{[-u;u]}(v)dv \\ &= ue^{-u} 1_{[0;\infty[}(u) \end{aligned}$$

Beachte: U ist Summe zweier st.u. $Exp(1)$-verteilter ZV. (Erlangverteilung)

2.4 Funktionen zweier unabhängiger Zufallsvariablen

- Im Folgenden: X, Y st.u. ZV und eine stetige Funktion $g : \mathbb{R}^2 \to \mathbb{R}$.
- Gesucht: Die Verteilung von $g(X, Y)$, z.B.
  - □ $g(X, Y) = X + Y$ (Faltung, engl.: Convolution)
  - □ $g(X, Y) = XY$
  - □ $g(X, Y) = X/Y$
  - □ $g(X, Y) = \max(X, Y)$
  - □ $g(X, Y) = \min(X, Y)$

$\hookrightarrow$ Diskreter Fall: X, Y seien st.u.

- X nehme Werte in $\mathcal{T}_1 = \{x_0, x_1, x_2, \dots\}$ an,
- Y nehme Werte in $\mathcal{T}_2 = \{y_0, y_1, y_2, \dots\}$ an.

$\hookrightarrow$ Betrachte eine Funktion $g : \mathcal{T}_1 \times \mathcal{T}_2 \to \mathbb{R}$. Die ZV $Z = g(X, Y)$ nimmt nur Werte der Form $g(x_i, y_j)$, d.h. Werte in der folgenden diskreten Menge an:

$$\mathcal{Z} := g(\mathcal{T}_1 \times \mathcal{T}_2) = \{z \in \mathbb{R} : \exists i, j \in \mathbb{N}_0 \; z = g(x_i, y_j)\}$$

$\hookrightarrow$ Dann hat Z eine diskrete Verteilung mit Dichte (für $z \in \mathcal{Z}$)

$$\begin{aligned} f_Z(z) &= P(Z = z) = P(g(X, Y) = z) = P(\{(X, Y) \in \mathcal{T}_1 \times \mathcal{T}_2 : g(X, Y) = z\}) \\ &= \sum_{(x,y) \in \mathcal{T}_1 \times \mathcal{T}_2} f_{X,Y}(x, y) \cdot 1_{\{z\}}(g(x, y)) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f_{X,Y}(x_i, y_j) \cdot 1_{\{z\}}(g(x_i, y_j)) \\ &\overset{st.u.}{=} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f_X(x_i) f_Y(y_j) \cdot 1_{\{z\}}(g(x_i, y_j)) \end{aligned}$$

$P(Z = z)$ ist die Summe all derjenigen Produkt-WS $f_X(x) f_Y(y)$ mit $g(x, y) = z$.

- Jetzt $g(x, y) = x + y$

Faltung von diskreten Zufallsvariablen

- Für st.u. reelle ZV X, Y mit Werten in $\mathcal{T}_1 = \{x_0, x_1, \ldots, \}$, $\mathcal{T}_2 = \{y_0, y_1, \ldots\}$ ist $Z = X + Y$ diskrete ZV mit Werten in $\mathcal{T} = \{z : \exists i, j \in \mathbb{N}\, z = x_i + y_j\}$ und Dichte

$$f_Z(z) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f_X(x_i) \cdot f_Y(y_j) \cdot 1_{\{z\}}(x_i + y_j). \quad \text{für } z \in \mathcal{T} \qquad (*)$$

- Spezialfall $\mathcal{T}_1 = \mathcal{T}_2 = \mathbb{N}_0$:

$$\begin{aligned} f_Z(z) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f_X(i) \cdot f_Y(j) \cdot 1_{\{z\}}(i + j) \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f_X(i) \cdot f_Y(j) \cdot 1_{\{j\}}(z - i) \text{ d.h.} \\ f_Z(z) &= \sum_{i=0}^{z} f_X(i) \cdot f_Y(z - i) \quad \text{für } z \in \mathbb{N}_0 \end{aligned}$$

Beispiel: Binomialverteilung

- X, Y st.u., $X \sim Bin(1,p)$, $Y \sim Bin(k,p)$, $\quad P(Y=j) = \binom{k}{j} p^j (1-p)^{k-j}$
- Für $z = 0$ gilt
  $P(X+Y=0) = P(X=0)P(Y=0) = (1-p)(1-p)^k = (1-p)^{k+1}$
- Für $z = k+1$ gilt $P(X+Y=k+1) = P(X=1, Y=k) = pp^k = p^{k+1}$
- Für $z \in \{1, \ldots, k\}$ gilt

$$\begin{aligned} P(X+Y=z) &= P(X=0)P(Y=z) + P(X=1)P(Y=z-1) \\ &= (1-p)\binom{k}{z} p^z (1-p)^{k-z} + p\binom{k}{z-1} p^{z-1} (1-p)^{k-(z-1)} \\ &= \left(\binom{k}{z} + \binom{k}{z-1}\right) p^z (1-p)^{(k+1)-z} \\ &= \binom{k+1}{z} p^z (1-p)^{(k+1)-z} \end{aligned}$$

- Insgesamt $X + Y \sim Bin(k+1, p)$

Faltung von Binomialverteilungen (Übungszettel!)

Falls X, Y st.u. mit $X \sim Bin(m,p)$, $Y \sim Bin(k,p)$, dann $X + Y \sim Bin(m+k, p)$

Übung: Es seien X, Y stochastisch unabhängig und $Poi(\lambda)$ bzw. $Poi(\mu)$-verteilt. Zeigen Sie zunächst für $n \in \mathbb{N}_0$: $P(X+Y=n) = \sum_{k=0}^{n} P(X=k)P(Y=n-k)$

Folgern Sie daraus $P(X+Y=n) = \frac{(\lambda+\mu)^n}{n!} e^{-(\lambda+\mu)}$, d.h. $X+Y \sim Poi(\lambda+\mu)$

Hilfsmittel: $P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$, binomische Formel: $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$

Summen von stetigen ZVen (Faltungssatz)

Seien X, Y st.u. stetig verteilte Zufallsvariablen und $Z = X + Y$. Dann gilt

$$f_Z(z) = f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy$$

Ergibt sich aus dem generischen Beispiel oben, z.B.

$$f_{X+Y}(z) = \int\limits_{-\infty}^{\infty} f_{X,Y}(x, z-x) dx \overset{st.u.}{=} \int\limits_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx$$

Beispiele

Seien X und Y unabhängige exponentialverteilte Zufallvariablen mit den Parametern λ_1 und λ_2, d.h. für $x, y > 0$: $f_X(x) = \lambda_1 e^{-\lambda_1 x}, \quad f_Y(y) = \lambda_2 e^{-\lambda_2 y}$

$$f_{X+Y}(z) = \int_0^z \lambda_1 e^{-\lambda_1 x} \lambda_2 e^{-\lambda_2 (z-x)} dx = \int_0^z \lambda_1 \lambda_2 e^{-\lambda_2 z} e^{-(\lambda_1 - \lambda_2)x} dx, \quad z > 0$$

Dann zwei mögliche Ergebnisse/Verteilungstypen:

- $\lambda_1 \neq \lambda_2$: Hypo-Exponentialverteilung: $f_{X+Y}(z) = \frac{\lambda_1 \lambda_2 (e^{-\lambda_1 z} - e^{-\lambda_2 z})}{\lambda_2 - \lambda_1} 1_{]0;\infty[}(z)$
- $\lambda_1 = \lambda_2 = \lambda$: Erlang(2)-Verteilung: $f_{X+Y}(z) = \lambda^2 z e^{-\lambda z} 1_{]0;\infty[}(z)$

Beide Verteilungen treten als Bedienzeitverteilungen auf (Simulation): Die Gesamt-Bedienzeit besteht dabei aus zwei unabhängigen exponentialverteilten Komponenten (Phasen). Verallgemeinerungen:

- Mehr Phasen, sequentiell (speziell: Erlang(n)-Verteilung)
- Mehr Phasen, Übergänge zwischen Phasen gemäß Matrix von Übergangs-WS.

Übung: Berechnen Sie die Dichte im obigen Beispiel für $\lambda_1 \neq \lambda_2$ und $z > 0$

Übung: Berechnen Sie die Dichte im obigen Beispiel für $\lambda_1 = \lambda_2 = \lambda$ und $z > 0$

- $*$-Schreibweise bei st.u. X, Y: Statt $\mathcal{L}(X+Y)$ schreibt man $\mathcal{L}(X) * \mathcal{L}(Y)$

Wichtige Faltungsformeln

- Diskret
  - $Bin(n,p) * Bin(k,p) = Bin(n+k,p)$
    Daraus: Sind $X_1, \ldots, X_n$ u.i.v $\sim Bin(1,p)$, dann $X_1 + \cdots + X_n \sim Bin(n,p)$.
  - $Poi(\lambda) * Poi(\mu) = Poi(\lambda + \mu)$
- Stetig
  - $\mathcal{N}(\mu_1, \sigma_1^2) * \mathcal{N}(\mu_2, \sigma_1^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
  - $Erlang(\lambda, n) * Erlang(\lambda, m) = Erlang(\lambda, n+m)$
    Daraus: Sind $X_1, \ldots, X_n$ u.i.v $\sim Exp(\lambda)$, dann $X_1 + \cdots + X_n \sim Erlang(\lambda, n)$
  - $\chi^2(n) * \chi^2(m) = \chi^2(n+m)$
    Daraus: Sind $X_1, \ldots, X_n$ u.i.v $\sim \mathcal{N}(0,1)$, dann $X_1^2 + \cdots + X_n^2 \sim \chi^2(n)$

Die letzten beiden Faltungsformeln sind Spezialfälle der Faltungsformel der Gamma-Verteilung.

2.5 Maximum/Minimum unabhängiger Zufallsvariablen

Sind $X_1, \ldots, X_n$ unabhängige und identisch stetig verteilte Zufallsvariablen mit Dichtefunktion $f_X(\cdot)$ und Verteilungsfunktion $F_X(\cdot)$, dann gilt für das Maximum Y_n und das Minimum Z_n der ZVen:

$$\begin{aligned} f_{Y_n}(y) &= n[F_X(y)]^{n-1} f_X(y), \\ f_{Z_n}(z) &= n[1 - F_X(z)]^{n-1} f_X(z). \end{aligned}$$

Dichten ergeben sich mit Kettenregel durch Ableiten der Verteilungsfunktionen:

- $F_{Y_n}(y) = P(\max(X_1, \ldots, X_n) \leq y) = P(X_1 \leq y, \ldots, X_n \leq y) \overset{uiv}{=} F_X(y)^n$
- $F_{Z_n}(z) = 1 - P(X_1 > z, \ldots, X_n > z) \overset{uiv}{=} 1 - (1 - F_X(z))^n$

Bsp. Systemausfall

System mit n Teilsystemen mit u.i.v. Ausfallzeitpunkten $X_1, \ldots, X_n \sim Exp(\lambda)$

Gesucht ist die Verteilung(sfunktion) des Ausfallzeitpunktes X des Gesamtsystems

a) bei Parallelschaltung, d.h. Ausfall in sämtlichen Teilsystemen. Für $x > 0$:

$$F_X(x) = P(X \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) \overset{uiv}{=} P(X_1 \leq x)^n = (1 - e^{-\lambda x})^n$$

Nach Ableiten: Dichte $f_X(x) = n\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{n-1} 1_{]0;\infty[}(x)$

b) Serienschaltung, d.h. Ausfall in einem Teilsystem. Für $x > 0$:

$$F_X(x) = P(\min(X_i) \leq x) \overset{uiv}{=} 1 - (1 - F_{X_1}(x))^n = 1 - (e^{-\lambda x})^n = 1 - e^{-n\lambda x}$$

ist Verteilungsfunktion einer $Exp(n\lambda)$-Verteilung.

2.6 Übersicht von Verteilungen und Transformationsregeln

Voraussetzung jeweils: $X_1, X_2, \ldots$ sind u.i.v. Zufallsvariablen.

	Verteilung	Träger	Dichte $f(x)$	Wenn $\mathcal{L}(X_i) =$	und $Y =$	dann $\mathcal{L}(Y) =$
diskret	Bernoulli $Bin(1,p)$	$\{0,1\}$	$p^x(1-p)^{1-x}$		$1_B(X_1)$	$Bin(1, P(X_1 \in B))$
	Binomial $Bin(n,p)$	$\{0,\ldots,n\}$	$\binom{n}{x}p^x(1-p)^{n-x}$	$Bin(k_i,p)$	X_1+X_2	$Bin(k_1+k_2,p)$
	Geometrisch $Geo(p)$	$\mathbb{N}$	$p(1-p)^{x-1}$	$Bin(1,p)$	$\inf\{i : X_i = 1\}$	$Geo(1,p)$
				$Exp(\lambda)$	$\lfloor X_1 \rfloor + 1$	$Geo(1-e^{-\lambda})$
	Neg.-Bin. $NBin(r,p)$	$\mathbb{N}_{\geq r}$	$(-1)^x\binom{-r}{x}p^r(1-p)^x$	$Bin(1,p)$	$\inf\{n : \sum_{i=1}^{n} X_i \geq r\}$	$NBin(r,p)$
				$NBin(r_i,p)$	X_1+X_2	$NBin(r_1+r_2,p)$
	Poisson $Poi(\lambda)$	$\mathbb{N}_0$	$\frac{\lambda^x}{x!}e^{-\lambda}$	$Poi(\lambda_i)$	X_1+X_2	$Poi(\lambda_1+\lambda_2)$
stetig	Rechteck $Re(a,b)$	$[a;b]$	$\frac{1}{b-a}$	$Re(a,b)$	$c+dX_1,\ d>0$	$Re(c+da, c+db)$
	Exponential $Exp(\lambda)$	$[0;\infty[$	$\lambda e^{-\lambda x}$	$Re(0,1)$	$-\frac{\ln(1-X_1)}{\lambda},\ \lambda>0$	$Exp(\lambda)$
				$Exp(\lambda_i)$	$\min(X_1,X_2)$	$Exp(\lambda_1+\lambda_2)$
	Normal $\mathcal{N}(\mu,\sigma^2)$	$]-\infty;\infty[$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mathcal{N}(\mu,\sigma^2)$	$aX_1+b, a\neq 0$	$\mathcal{N}(a\mu+b, a^2\sigma^2)$
				$\mathcal{N}(\mu_i,\sigma_i^2)$	X_1+X_2	$\mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$
	Lognormal $\mathcal{LN}(\mu,\sigma^2)$	$[0;\infty[$	$\frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}$	$\mathcal{N}(\mu,\sigma^2)$	e^{X_1}	$\mathcal{LN}(\mu,\sigma^2)$
	Gamma $\Gamma(\lambda,c)$	$[0;\infty[$	$\frac{\lambda^c}{\Gamma(c)}x^{c-1}e^{-\lambda x}$	$\Gamma(\lambda,c_i)$	X_1+X_2	$\Gamma(\lambda,c_1+c_2)$
	Pareto $Par(\lambda,c)$	$[\lambda;\infty[$	$\frac{c}{\lambda}\left(\frac{\lambda}{x}\right)^{c+1}$	$Exp(\mu)$	e^X	$Par(1,\mu)$
	Weibull $Wei(\lambda,c)$	$[0;\infty[$	$\frac{c}{\lambda}\left(\frac{x}{\lambda}\right)^{c-1}$	$Exp(\mu)$	$X^{1/c}$	$Wei(\mu^{1/c},c)$

3. Bedingte Verteilung, bedingte Erwartung

- Bedingte Verteilung
- Bedingter Erwartungswert

Erklärung und Prognose von Merkmalen

- ↪ Gegeben: Datensatz

Merkmal 1	...	Merkmal k	Merkmal $k+1$
x_{11}	...	x_{1k}	y_1
$\vdots$		$\vdots$	$\vdots$
x_{n1}	...	x_{nk}	y_n

,

  Modell n-fache (unabh.) Versuchswiederholung mit ZV $X_1, \ldots, X_k, Y$
- ↪ Aufgabe: Finde eine Funktion h mit $Y \approx h(X_1, \ldots, X_k)$, welche den Datensatz „gut“ beschreibt („Überwachtes Lernen“):
  - □ Erklärung von Y durch $X_1, \ldots, X_k$: Modell durch Festlegung von h an Daten anpassen.
  - □ Prognose von Y durch $X_1, \ldots, X_k$: Prognosefehler minimieren.
- ↪ Bedingte Sicht:
  - □ $X_1, \ldots, X_n$ sind Informationen, die zur Neubewertung der Modell-Informationen über Y führen
  - □ Aus der Wahrscheinlichkeitsrechnung bekanntes Konzept: Bedingte Wahrscheinlichkeiten von Ereignissen
  - □ Übertragung auf ZV: Bedingte Verteilung /bedingte Erwartung

3.1 Bedingte Verteilung

- ↪ in WS-Rechnung:
  - □ WS von Ereignissen $\rightsquigarrow$ Verteilungen von ZV $P(X \in B) = P(X^{-1}(B))$
  - □ Bedingte WS von Ereignissen: $P(B|A) = P(B \cap A)/P(A)$ für $P(A) > 0$
- ↪ Kombination: Bedingte WS von durch ZV X, Y induzierten Ereignissen
  - □ $$P(Y \in B|X \in A) = \frac{P(Y^{-1}(B) \cap X^{-1}(A))}{P(X^{-1}(A))} \quad \text{für } P(X \in A) > 0 \qquad (*)$$
- ↪ Dabei ist z.B. in Erklärungsmodellen ein konkreter beobachteter Wert $X = x$ gegeben, d.h. $A = \{X = x\}$. Die WS für Y wird unter $X = x$ neu bewertet. (**Bedingte Verteilung** von Y unter $X = x$, Notation: $\mathcal{L}(Y|X = x)$)
  - □ Gerade in stetigen WS-Modellen gilt aber $P(X = x) = 0$,
  - □ d.h. $(*)$ kann dann direkt nicht umgesetzt werden.
- ↪ Analog zur Bayes-Formel lässt sich für bivariate ZV (X, Y) mit gemeinsamer Dichte $f(x, y)$ „die“ Dichte einer bedingten Verteilung $\mathcal{L}(Y|X = x)$ erklären.
- ↪ Deren Kennzahlen lassen sich zur Prognose von Y gegeben $X = x$ einsetzen.

Übung: Es seien A, B Ereignisse mit $0 < P(A) < 1$ und $X = \mathbb{1}_A$, $Y = \mathbb{1}_B$. Berechnen Sie die bedingte Verteilung $\mathcal{L}(Y|X = x)$ für $x \in \{0, 1\}$.

Die bedingte Verteilung wird also durch die verschiedenen bedingten Wahrscheinlichkeiten auf Grundlage von A, A^c bzw B, B^c gegeben.

Beschränkung auf bivariaten Fall: Gegeben sei ein Zufallsvektor (X, Y) mit

- diskreter/stetiger Dichte $f_{X,Y}(x, y)$
- Randdichten $f_X(x)$, $f_Y(y)$.

Bedingte Verteilung $\mathcal{L}(Y|X = x)$

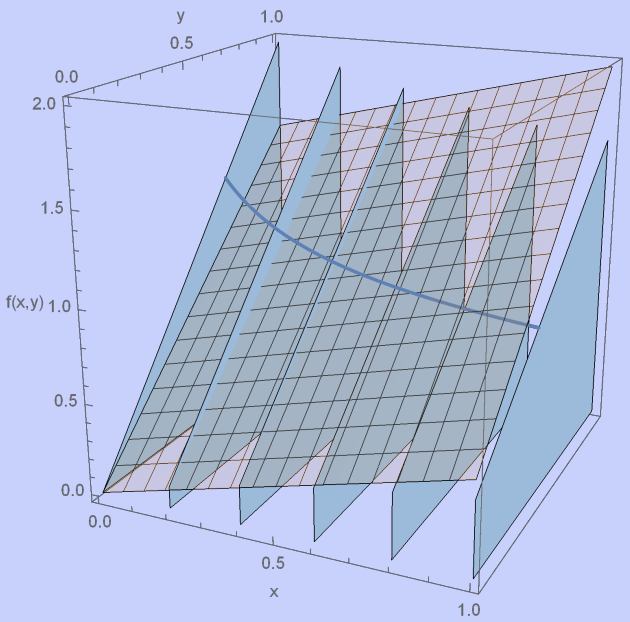
ist gegeben durch **bedingte Dichte**:

$$f_{Y|X=x}(y) = \begin{cases} f_{X,Y}(x,y)/f_X(x) & \text{falls Nenner} > 0 \\ f_Y(y) & \text{sonst} \end{cases} \qquad \text{(Bayes-Formel)}$$

„bedingte WS in y bilden bei festem x eine WS-Verteilung“.

Bed. Dichte zu stetigem Zufallsvektor (DuW)

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}x + \frac{3}{2}y & 0 \leq x, y \leq 1, \\ 0 & sonst \end{cases}, f_X(x) = \frac{1}{2}x + \frac{3}{4}, f_Y(y) = \frac{3}{2}y + \frac{1}{4}$$



bed. Dichte für $0 < x, y < 1$:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{2}x + \frac{3}{2}y}{\frac{1}{2}x + \frac{3}{4}}$$

Übung: Berechne $\mathcal{L}(Y|X=x)$ zu $f(x,y) = \dfrac{\mathbb{1}_{]0;1[}(x)\mathbb{1}_{]0;x[}(y)}{x}$:

1. $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \ldots$

2. $f_{Y|X=x}(y) = f(x,y)/f_X(x) = \ldots$

Diskretes Beispiel: $Y_1, \ldots, Y_m$ u.i.v. $\sim Bin(1,p)$, $X = Y_1 + \cdots + Y_m$

Bestimme bedingte Verteilung von $(Y_1, \ldots, Y_m)$ unter $X = x \in \{0, \ldots, m\}$.
Betrachte hierzu (ausschließlich) $y_1, \ldots, y_m \in \{0,1\}$ mit $y_1 + \cdots + y_m = x$:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_m = y_m | X = x) &= \frac{P(Y_1 = y_1, \ldots, Y_m = y_m, X = x)}{P(X = x)} \\ &= \frac{P(Y_1 = y_1, \ldots, Y_m = y_m)}{P(X = x)} \\ &= \frac{p^x (1-p)^{m-x}}{\binom{m}{x} p^x (1-p)^{m-x}} = 1/\binom{m}{x} \end{aligned}$$

$\hookrightarrow$ $\mathcal{L}(Y_1, \ldots, Y_m | X = x)$ ist also eine Gleichverteilung auf den $\binom{m}{x}$ m-Tupeln $(y_1, \ldots, y_m) \in \{0,1\}^m$ mit $y_1 + \cdots + y_m = x$.

$\hookrightarrow$ Die bedingte Verteilung ist unabhängig von p:

- Information über p ist schon vollständig in X enthalten.
- Schätzer, Tests, Konfidenzintervalle,... zu p dürfen Datenverdichtung X anstelle der Originaldaten $Y_1, \ldots, Y_m$ verwenden.
- Man sagt: $X = Y_1 + \cdots + Y_m$ ist **suffizient**.

3.2 Bedingter Erwartungswert

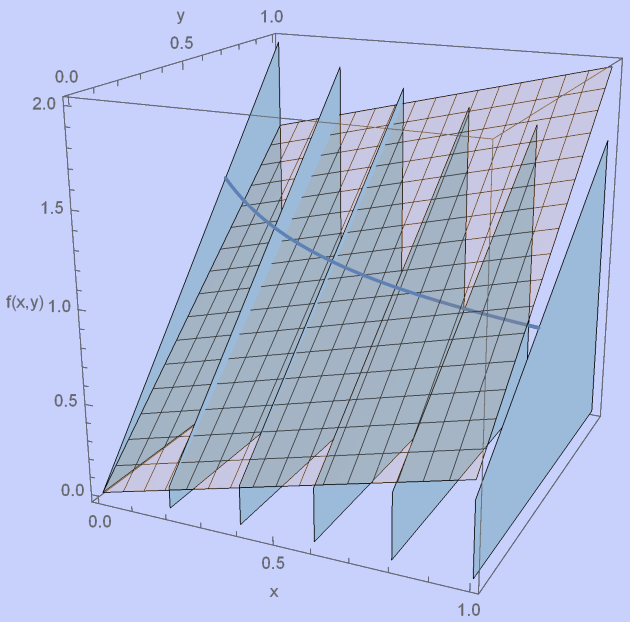
- $\hookrightarrow$ Erwartungswert der bedingten Verteilung $\mathcal{L}(Y|X=x)$.
- $\hookrightarrow$ Bei Vorliegen einer bedingten Dichte $f(y|x)$

$$h(x) := E(Y|X=x) = \begin{cases} \int\limits_{-\infty}^{\infty} y \cdot f(y|x)dy & \text{stetiger Fall} \\ \sum\limits_{y \in \mathcal{Y}} y \cdot f(y|x) & \text{diskreter Fall mit Träger } \mathcal{Y} \end{cases}$$

- $\hookrightarrow$ Schreibweise: $E(Y|X) = h(X)$
- $\hookrightarrow$ Der bed. EW löst das Problem $E((Y-h(X))^2) \stackrel{!}{=} \min$, d.h. gibt eine Funktion h an, welche den Ausdruck minimiert.
- $\hookrightarrow$ Konzept bedingter Verteilungen/Erwartungswerte übertragbar auf Zufallsvektoren:
  $\mathcal{L}(Y_1, \ldots, Y_m | X_1 = x_1, \ldots, X_k = x_k)$ bzw. $E(Y|X_1 = x_1, \ldots, X_n = x_n)$

Bed. Dichte zu stetigem Zufallsvektor (DuW)

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}x + \frac{3}{2}y & 0 \le x, y \le 1, \\ 0 & sonst \end{cases}, f_X(x) = \frac{1}{2}x + \frac{3}{4}, f_Y(y) = \frac{3}{2}y + \frac{1}{4}$$



bed. Dichte für $0 < x, y < 1$:

$$\begin{aligned} f_{Y|X=x}(y) &= \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{2}x + \frac{3}{2}y}{\frac{1}{2}x + \frac{3}{4}} \\ E(Y|X=x) &= \int_0^1 \frac{\frac{1}{2}xy + \frac{3}{2}y^2}{\frac{1}{2}x + \frac{3}{4}} dy \\ &= \left[\frac{\frac{1}{4}xy^2 + \frac{1}{2}y^3}{\frac{1}{2}x + \frac{3}{4}} \right]_{y=0}^{y=1} \\ &= \frac{\frac{1}{4}x + \frac{1}{2}}{\frac{1}{2}x + \frac{3}{4}} \end{aligned}$$

Übung: Berechne $E(Y|X=x)$ zu $f(x,y) = \dfrac{\mathbb{1}_{]0;1[}(x)\mathbb{1}_{]0;x[}(y)}{x}$. Bereits gerechnet:

1. $f_X(x) = \mathbb{1}_{]0;1[}(x)$, d.h. $\mathcal{L}(X)$ ist Standard-Rechteck-Verteilung
2. $f_{Y|X=x}(y) = f(x,y)/f_X(x) = \frac{1}{x}\mathbb{1}_{]0;x[}(y)$, d.h. $\mathcal{L}(Y|X=x) = Re(0,x)$

Daraus jetzt den bedingten Erwartungswert:

Regeln für bedingte Erwartungswerte

- ↪ Linearität: $E(a + bY_1 + cY_2|X) = a + bE(Y_1|X) + cE(Y_2|X)$
- ↪ Totale Wahrscheinlichkeit: $E(Y) = E(E(Y|X))$
- ↪ Faktorisierung: $E(Y \cdot g(X)|X) = g(X) \cdot E(Y|X)$
- ↪ Substituieren/Eliminieren (SE): Wenn X, Y st.u. sind, dann gilt:

(SE1) $P(h(X, Y) \in B|X = x)) = P(h(x, Y) \in B)$

(SE2) $E(h(X, Y)|X = x) = E(h(x, Y))$

Aussagen jeweils „fast sicher“ und unter Annahme existierender Erwartungswerte

Übung: Fortsetzung des Beispiels mit $f(x,y) = \mathbb{1}_{]0;1[}(x)\mathbb{1}_{]0;x[}(y)/x$.
Bereits berechnet wurden: $\mathcal{L}(X) = Re(0,1)$, $E(Y|X=x) = x/2$.

1. Berechne $E(Y)$ mit dem Satz von der totalen WS $E(Y) = E(E(Y|X))$.

2. Vergleiche mit dem direkten Rechenweg, d.h.
   - □ $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \ldots$ für $y \in [0;1]$
   - □ $E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y)dy = \ldots$

4. Punkt-Schätzer

- Punktschätzung
- Konstruktion von Schätzfunktionen
- Maximum-Likelihood-Methode
- Momentenmethode
- Gütekriterien für Punktschätzer
  - Erwartungstreue
  - Effizienz
  - Konsistenz
  - Asymptotische Effizienz
- Zusammenfassung

4 Punktschätzer

Statistik wird häufig eingesetzt, um Informationen über bestimmte Charakteristika einer Grundgesamtheit zu beschaffen.

z.B.: Durchschnittseinkommen aller Bachelorstudenten, durchschnittliche erwartete Verspätung einer Zuglinie, Defektrate im Fertigungsprozess, Basisreproduktionszahl bei einer Epidemie etc.

- $\hookrightarrow$ Vollerhebung der Grundgesamtheit i.A. nicht möglich.
- $\hookrightarrow$ Idee: Ziehe aus einer „repräsentativen“ Stichprobe Rückschlüsse auf die Grundgesamtheit.
- $\hookrightarrow$ Dabei wichtig: homogene Grundgesamtheit muss vorliegen (Partitionierung von Grundgesamtheit bzw. Stichprobe durch Clusterverfahren)

1) Punktschätzverfahren

- ↪ liefern einen einzelnen Schätzwert für einen unbekannten Parameter einer Grundgesamtheit/Verteilung.
- ↪ Stichprobenwerte werden mit Hilfe einer Schätzfunktion aggregiert. Die dabei verwendeten Datenkennzahlen (gemäß DuW) korrespondieren oft zum unbekannten Parameter (z.B. aufgrund der Gesetze großer Zahlen).

2) Intervallschätzverfahren

- ↪ Schätzergebnis ist ein Intervall, in das der unbekannte Parameter der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit fällt.
- ↪ Oft ist ein „plausibler“ Punktschätzer die Intervallmitte, und die Intervallbreite reflektiert die Unsicherheit über die Genauigkeit des Schätzers.

3) Statistische Tests

- ↪ Überpüfung einer bzgl. der Grundgesamtheit erstellten Hypothese
- ↪ Konstruktion derart, dass eine korrekte Hypothese nur mit einer sehr geringen Wahrscheinlichkeit verworfen wird.

4.1 Punktschätzung

Möglichst genaue Annäherung eines unbekannten Grundgesamtheitsparameters.

Kennwerte einer beliebigen, unbekannten Verteilung

- $\hookrightarrow$ Erwartungswert bzw. Varianz einer Zufallsvariablen
- $\hookrightarrow$ Median oder allg. Quantil einer Zufallsvariablen
- $\hookrightarrow$ Korrelation zweier Zufallsvariablen ...

Spezifische Parameter eines zugrundegelegten Verteilungsmodells

- $\hookrightarrow$ z.B. X: Anzahl Schadensmeldungen innerhalb eines Monats, $X \sim \mathit{Poisson}(\lambda)$, interessierender Parameter λ
- $\hookrightarrow$ z.B. X: Durchmesser von produzierten Schrauben, $X \sim \mathcal{N}(\mu, \sigma^2)$, interessierende Parameter μ und σ^2
- $\hookrightarrow$ z.B. X: Lebensdauer von Glühbirnen, $X \sim \mathit{Exp}(\lambda)$, interessierender Parameter λ

Ausgangspunkt:

$\hookrightarrow$ n Stichprobenziehungen/Zufallsexperimente, repräsentiert durch ZV $X_1, \ldots, X_n$

$\hookrightarrow$ $X_1, \ldots, X_n$ heißen auch Stichprobenvariablen

$\hookrightarrow$ häufig: Stichprobenvariablen sind unabhängige Wiederholungen von X:

- Experimente, die den ZVen $X_1, \ldots, X_n$ zugrundeliegen, sind unabhängig,
- jedes Mal wird dasselbe Zufallsexperiment durchgeführt

Für Konsistenzuntersuchung angenommen: u.i.v.-**Folge** $X_1, \ldots, X_n, \ldots$

$\hookrightarrow$ auf Basis der Realisierungen $x_1, \ldots, x_n$ soll auf θ geschlossen werden.

Eine **Punktschätzung** für θ ist eine Funktion $t = g(x_1, \ldots, x_n)$.

$\hookrightarrow$ z.B. $g(x_1, \ldots, x_n) = \sum_{i=1}^{n} \frac{x_i}{n}$ ist (sinnvolle) Punktschätzung für $\theta = E(X)$.

$\hookrightarrow$ die Stichproben sind Realisationen von Zufallsvariablen $X_1, \ldots, X_n$

$\hookrightarrow$ die Punktschätzung ist als Funktion von ZVen auch eine ZVe

$\hookrightarrow$ Variabilität wird durch die Variabilität der ZVen $X_1, \ldots, X_n$ bestimmt

Eine **Schätzfunktion** oder **Schätzstatistik** für den Grundgesamtheitsparameter θ ist eine Funktion der Stichprobenvariablen $X_1, \ldots, X_n$:

$$T = g(X_1, \ldots, X_n)$$

Der **Schätzwert** ergibt sich aus den Realisationen $x_1, \ldots, x_n$:

$$\hat{\theta} = g(x_1, \ldots, x_n)$$

Beispiele

$$\bar{X} = g(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{(Schätzfkt. für } E(X)\text{)}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad X_i \in \{0,1\} \quad \text{(Schätzfkt. für Anteilswert } \pi = P(X=1)\text{)}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{(Schätzfkt. für } \sigma^2 = Var(X)\text{)}$$

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{(Schätzfkt. für } \sigma^2 = Var(X)\text{)}$$

Beispiele

X : Verspätung der Rückgabe eines Mathebuches in der Bibliothek in Tagen.

X_i : X bei i-ter Messung.

Seien $(x_1, \ldots, x_{10}) = (2, 14, 10, 0, 9, 20, 8, 2, 3, 2)$. $\mu = E(X) =$???

Annahme: $X \sim F$ mit unbekannter Verteilung $F \Rightarrow X_1, \ldots, X_n \overset{iiv}{\sim} F$

Mögliche Schätzungen:

$$\begin{aligned} \hat{\mu}_1 &= \bar{x} = 7 \\ \hat{\mu}_2 &= x_1 = 2 \\ \hat{\mu}_3 &= 3 \cdot x_8 = 6 \\ \hat{\mu}_4 &= \tfrac{1}{9} \textstyle\sum_{i=1}^{9} x_i = 7.56 \end{aligned}$$

Welche ist die "Beste"?

Antwort allgemein abhängig vom verwendeten WS-Modell und Gütekriterium.

Wie findet man geeignete Schätzungen für unbekannte Parameter?

Idee: Suche für Realisationen $x_1, \ldots, x_n$ denjenigen Parameter $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$, der die plausibelste Erklärung für die Beobachtung dieser Realisationen liefert.

- $\hookrightarrow$ **ML-Schätzer** $\hat{\theta}_{ML}$: Durch Maximierung der gemeinsamen WS-Dichte der Stichprobenvariablen als Funktion des Parameters.
- $\hookrightarrow$ **MM-Schätzer** $\hat{\theta}_{MM}$: Durch Gleichsetzung von theoretischen und Stichproben-Kennzahlen und Auflösen dieser Gleichung(en) nach θ.
- $\hookrightarrow$ **KQ-Schätzer (Regression)**: „bestmögliche Erklärung des Zusammenhangs zwischen abhängiger und unabhängiger Variable durch Minimierung der „Residuenquadratsumme“der abhängigen Variable
- $\hookrightarrow$ **MD-Schätzer**: Gegenüberstellung von Stichprobenverteilung und theoretischer Verteilung (z.B. deren Verteilungsfunktion, Quantilfunktion) und Minimierung der „Distanzen“ dazwischen. Hier nicht behandelt.
- $\hookrightarrow$ **Bayes-Schätzer**: Gewichtetes Mittel aus datenabhängiger Schätzung und „Vorinformation “ des Schätzers (Idee: von der a-priori-Verteilung zur a-posteriori-Verteilung). Hier nicht behandelt.

4.3 Maximum-Likelihood-Methode (ML)

- ↪ Betrachtung der „Plausibilität“ innerhalb einer parametrischen Verteilungsklasse anhand einer konkreten Stichprobe $x_1, \ldots, x_n$.
- ↪ geht v.a. auf Sir Ronald Aylmer Fisher zurück (Anfang des 20. Jahrhunderts)

- ↪ Genetiker, Evolutionstheoretiker, Eugeniker, Statistiker, $*$ 17.02.1890 in London; $+$ 29.07.1962 in Adelaide
- ↪ wichtige Beiträge zur Statistik: ML-Methode, Varianzanalyse, F-Test, Versuchsplanung, Suffizienz, Verteilungsfreiheit, Fisher-Information, Diskriminanzanalyse

Maximum-Likelihood-Prinzip

Seien $X_1, \ldots, X_n$ unabhängige und identische Wiederholungen eines Experimentes.

$\hookrightarrow$ **Likelihood-Funktion**: Fasse die gemeinsame Dichte als Funktion des unbekannten Parameters θ bei festen Realisationen $x_1, \ldots, x_n$ auf:

$$L(\theta) = L(\theta|x_1, \ldots, x_n) := f(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta)$$

$\hookrightarrow$ **Maximum-Likelihood-Schätzung**

$\hat{\theta}_{ML} = \hat{\theta}_{ML}(x_1, \ldots, x_n)$ ist erklärt als Maximalstelle der Funktion $\theta \mapsto L(\theta)$

$\hookrightarrow$ **Log-Likelihood**: Statt L wird meist $\ln(L(\theta)) = \sum_{i=1}^{n} \ln(f(x_i|\theta))$ maximiert.

- **Maximum-Likelihood-Schätzer**: $g(X_1, \ldots, X_n) = \hat{\theta}(X_1, \ldots, X_n)$ (ist eine Zufallsvariable, die ML-Schätz<u>ung</u> ist deren Realisation)
- Durch die Logarithmierung wird das Maximierungsproblem i.A. einfacher. Das maximierende θ selbst ist identisch, da der Logarithmus eine streng monotone Transformation ist.

Bsp: Elfmeter (aus *DuW*)

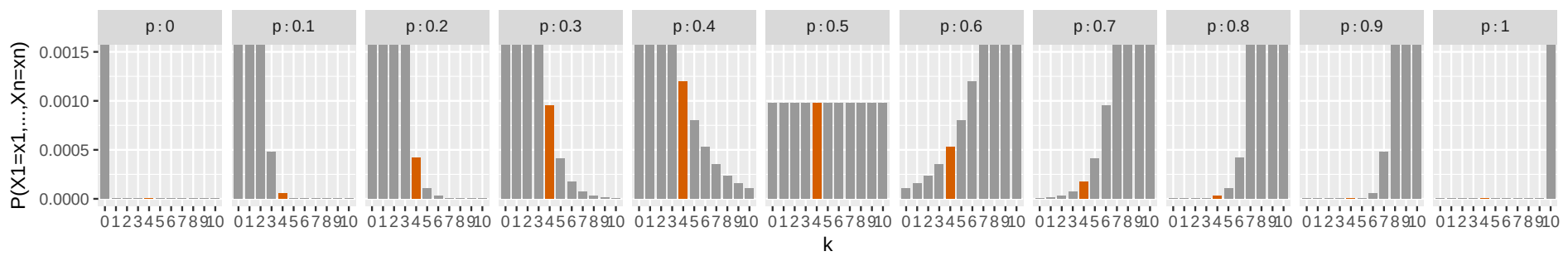
- ↪ Ergebnis werde als Bernoulli-Experiment mit Treffer-WS p angesehen (problematisch: u.i.v-Annahme)
- ↪ Bei $n = 10$ Schüssen folgende Ergebnisse: 1,1,0,0,1,0,0,0,1,0 ($k = 4$ Treffer)

Ziel: Schätzung von p mit ML-Methode.

Likelihood=Gemeinsame Dichte, hier allgemein (mit $k = x_1 + \cdots x_n$)

$$\begin{aligned} f_p(x_1, \ldots, x_n) &= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \quad = \quad p^k(1-p)^{n-k} \end{aligned}$$

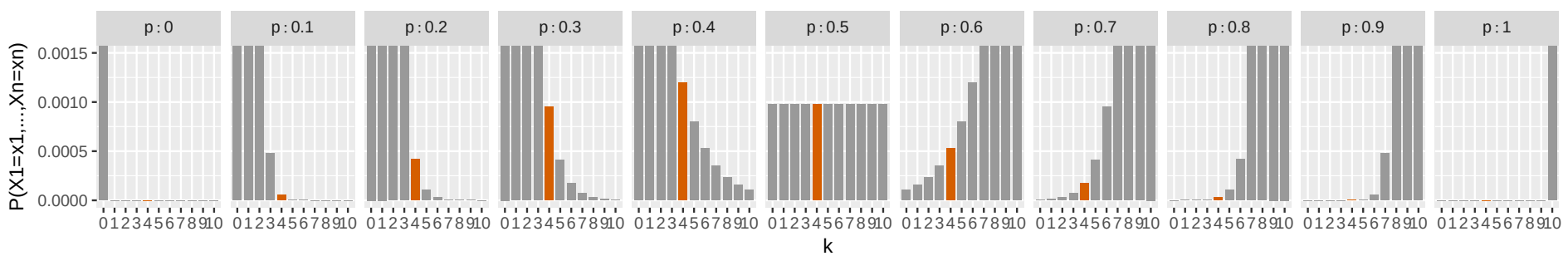
Plot der WS ($n = 10$, diverse p) für $k = 0, \ldots, 10$



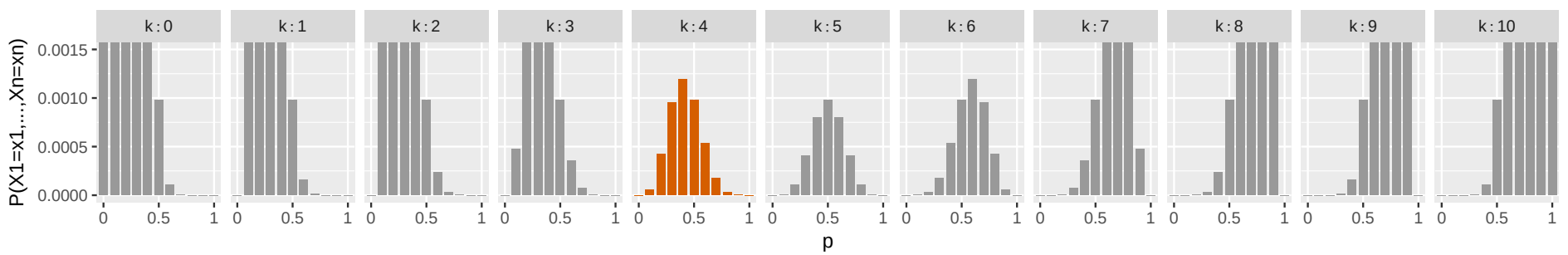
Hervorgehoben: jeweils $k = 4$

Likelihoodfunktion der konkreten Stichprobe ($k = x_1 + \cdots + x_{10} = 4$):

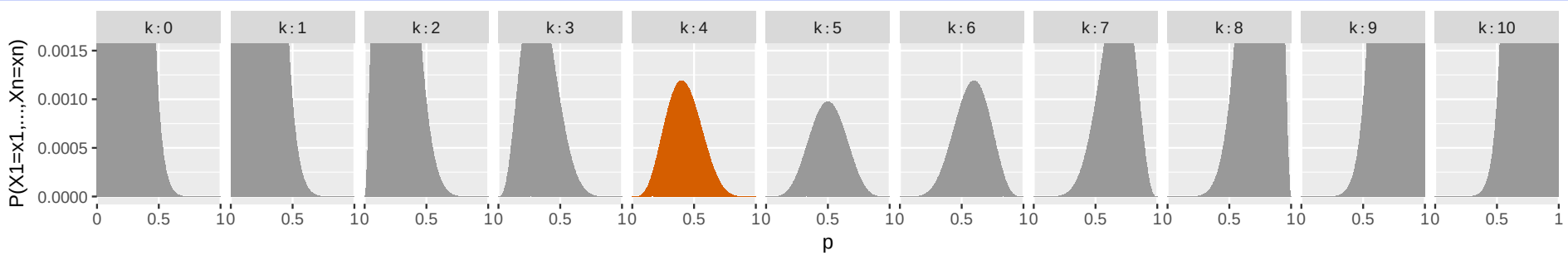
Dichte (s.o.)



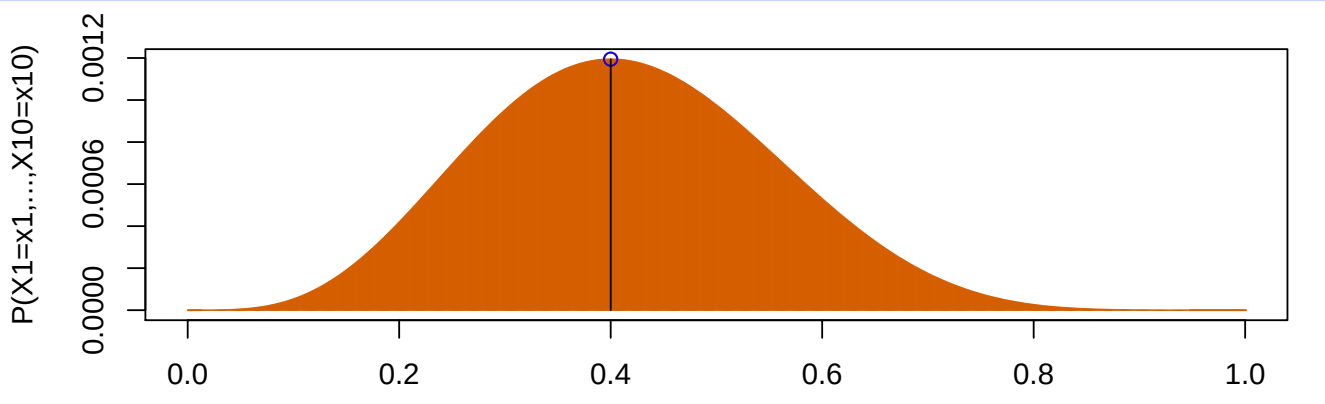
Likelihood: Stelle Werte der Dichte gebündelt nach k dar:



vollständige Graphen der Likelihood (Polynome 10. Grades):



Bsp (Elfmeter): Likelihoodfunktion der konkreten Stichprobe ($k = 4$)



- $\hookrightarrow$ Likelihood mit $k = \sum x_i$ rechnerisch: $L(p) = L(x_1, \ldots, x_n, p) = p^k(1-p)^{n-k}$
- $\hookrightarrow$ Log-Likelihood: $\ln(L(p)) = \ln(p^k(1-p)^{n-k}) = k\ln(p) + (n-k)\ln(1-p)$
- $\hookrightarrow$ Maximierung der Log-Likelihood: $\frac{\partial \ln(L(p))}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p}$
  - □ Notwendig: $0 \stackrel{!}{=} \frac{k}{p} - \frac{n-k}{1-p} \Leftrightarrow \frac{k}{p} = \frac{n-k}{1-p} \Leftrightarrow k - kp = np - kp \Leftrightarrow p = \frac{k}{n}$
  - □ Hinreichend: Zielfunktion ist konkav: $\frac{\partial^2 \ln(L(p))}{\partial^2 p} = -\frac{k}{p^2} - \frac{n-k}{(1-p)^2} \leq 0$

  Maximal ist die (Log-)Likelihood für $p = \frac{k}{n} = \frac{x_1 + \cdots + x_n}{n} = \bar{x}$

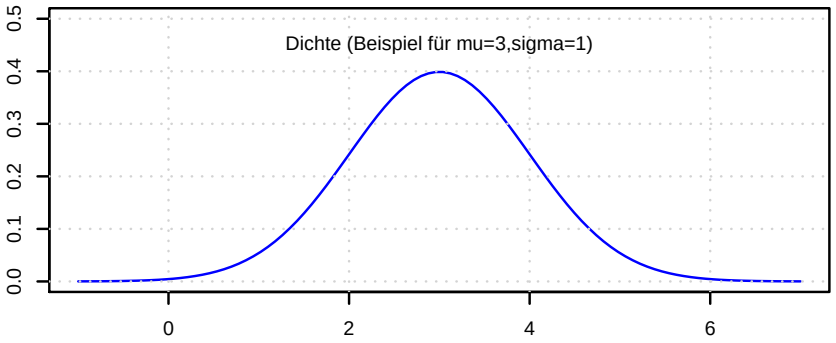
Der ML-Schätzer ist $g(X_1, \ldots, X_n) = \hat{p}_{ML}(X_1, \ldots, X_n) = \hat{p}_{ML} = \bar{X}$

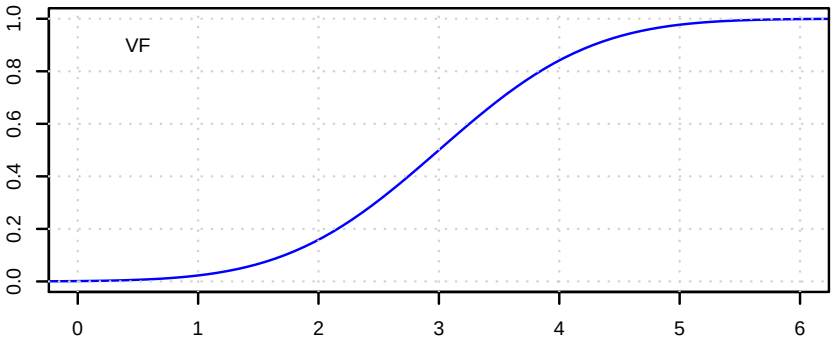
Bei Ergebnisfolge 1,1,0,0,1,0,0,0,1,0 ($k = 4$ Treffer) ML-Schätzung $\hat{p}_{ML} = 0.4$

Übung: In einem Fertigungsprozess werden 5x wöchentlich je 4 Items auf Fehler überprüft. Die Defektzahlen $X_1, \ldots, X_5$ seien st.u., $\mathcal{L}(X_i) = Bin(4, p)$.
In einer bestimmten Woche werden die Defektzahlen 2,0,0,1,0 beobachtet.
Berechnen Sie eine ML-Schätzung für die Defektwahrscheinlichkeit p eines Items.

Memo (DuW) Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ mit $\mu \in \mathbb{R}, \sigma > 0$

- ↪ Träger: $\mathcal{X} = \mathbb{R}$
- ↪ Dichte: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- ↪ VF $F(x) = \Phi(\frac{x-\mu}{\sigma})$
- ↪ $E(X) = med(X) = \mu$
- ↪ $var(X) = \sigma^2$
- ↪ R: dnorm,pnorm,qnorm,rnorm
- ↪ Speziell: VF, Dichte zu $\mathcal{N}(0,1)$:
  $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt = 1 - \Phi(-x)$,
  $\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$
  Schreibweise: $u_\alpha := \Phi^{-1}(\alpha)$





- ↪ Als Modellverteilung stetiger Merkmale verwendet (z.B. Regressionsanalyse), aber oft ungerechtfertigt – fehlende Symmetrie, begrenzter Träger der Daten
- ↪ Approximative Stichprobenverteilung standardisierter Summen (Zentraler Grenzwertsatz), für derart kumulierte Daten dann auch Modellverteilung.

Beispiel: Statistico (DuW)

- ↪ Der Spielehersteller „R-Games“ benötigt zum Testen seines Spiels „Statistico“ eine Gruppe von Testspielern , deren IQ demjenigen der Grundgesamtheit entspricht (Annahme: normalverteilt, $\mu = 100$, $\sigma = 15$).
- ↪ Von 20 möglichen Testpersonen liegen die folgenden IQ-Ergebnisse vor: 119, 95, 120, 119, 106, 77, 86, 96, 100, 136, 111, 88, 83, 96, 96, 94, 104, 87, 107, 81
- ↪ Kennzahlen der Stichprobe: $\bar{x} = 100.05$, $\tilde{s}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = 221.8475$
- ↪ Kann man bei dieser Stichprobe davon ausgehen, dass sie die Grundgesamtheit ausreichend genau repräsentiert?
- ↪ Unter der Annahme, dass die IQ-Werte über die Grundgesamtheit normalverteilt sind, sind zunächst (ML-)Schätzungen für μ und σ gesucht.

Ziel: ML-Schätzung von $\theta = (\mu, \sigma)$ in u.i.v. Stichprobe $X_1, \ldots, X_n \overset{u.i.v.}{\sim} \mathcal{N}(\mu, \sigma^2)$

1.) Likelihood: Schreibe $L(\mu, \sigma)$ für $L(\mu, \sigma, x_1, \ldots, x_n)$

$$L(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right) \cdots \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

2.) Log-Likelihood:

$$\ln L(\mu, \sigma) = \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right] = \sum_{i=1}^{n} \left[-\ln\left(\sqrt{2\pi}\right) - \ln(\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

3.) Maximierung (hier nur notwendige Bedingungen)

- $\frac{\partial \ln L(\mu,\sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} \overset{!}{=} 0$ d.h. $\sum_{i=1}^{n} x_i - n \cdot \mu = 0 \Leftrightarrow \boxed{\hat{\mu} = \bar{x}}$
- $\frac{\partial \ln L(\mu,\sigma)}{\partial \sigma} = \sum_{i=1}^{n} \left(-\frac{1}{\sigma} + \frac{2(x_i-\mu)^2}{2\sigma^3}\right) \overset{!}{=} 0,$

  d.h. $-\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0 \Leftrightarrow \sum_{i=1}^{n} (x_i - \mu)^2 = n \cdot \sigma^2,$

  d.h. $\hat{\sigma} := \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

ML-Schätzer für (μ, σ) sind die Schätzstatistiken $(\bar{X}, \tilde{S})$.

Statistico-Stichprobe: $\hat{\mu} = \bar{x} = 100.05$, $\hat{\sigma} = \tilde{s} \approx 14.89$

$\hookrightarrow$ ML-Schätzer sind oft nicht (nur) mittels Differentialrechnung zu bestimmen:

Doppelexponentialverteilung

$\hookrightarrow$ $X_1, \ldots, X_n$ u.i.v. mit Dichte (!) $f(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}$, $\mu \in \mathbb{R}$, $\sigma > 0$.

$\hookrightarrow$ Verwendung anstelle Normalverteilung (höhere WS extremer Ereignisse, Verteilungen mit „schweren Flanken“, z.B. bei Finanzdaten)

$\hookrightarrow$ Log-Likelihood: $\ln(L(\mu, \sigma)) = -n \ln(2\sigma) - \frac{1}{\sigma} \sum_{i=1}^{n} |x_i - \mu|$

- Maximierung in μ bei festem σ:

  L maximal, wenn $\sum_{i=1}^{n} |x_i - \mu|$ minimal, d.h. (DuW) für $\hat{\mu} = med(x)$

  (dieser Schritt zwangsläufig ohne Differentialrechnung)
- Maximierung in σ bei $\mu = \hat{\mu}$:

  $\ln(L(\hat{\mu}, \sigma)) = -n \ln(2\sigma) - \frac{n \cdot MA(x)}{\sigma}$ mit $MA(x) = \frac{1}{n} \sum_{i=1}^{n} |x_i - med(x)|$

  Notwendig: $\frac{\partial L(\hat{\mu}, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{n \cdot MA(x)}{\sigma^2} \stackrel{!}{=} 0 \Rightarrow \hat{\sigma} = MA(x)$

MA-Schätzer sind also $\hat{\mu}_{ML} = med(X)$ und $\hat{\sigma}_{ML} = MA(X)$

Übung: Berechnen Sie im vorigen Beispiel mit $f(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}$ die angegebene Log-Likelihood $\ln(L(\mu, \sigma)) = -n \ln(2\sigma) - \frac{1}{\sigma} \sum_{i=1}^{n} |x_i - \mu|$

Fritzbrötchen

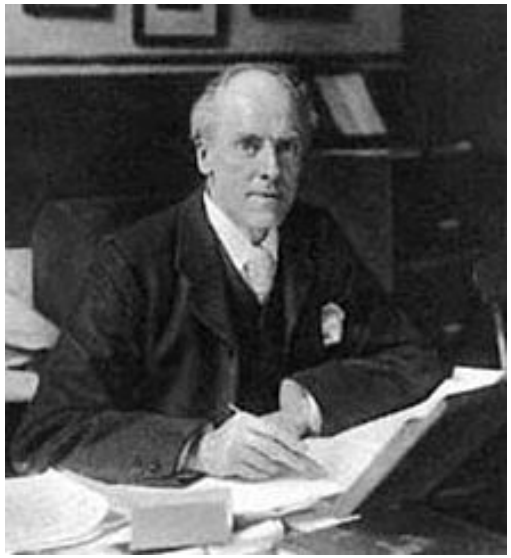
Bäcker Kalkoves neueste Kreation ist das Fritzbrötchen, welches er in der Bäckerei neben dem Hörsaal verkaufen möchte. Unter 10 Studierenden hat er folgende Preisbereitschaften (PB) erfragt (in EuroCent): 20, 70, 85, 50, 75, 0, 40, 90, 95, 60.

Übung: Die Likelihood-Funktion für (mit $\theta > 0$) u.i.v. $\mathrm{Re}(0, \theta)$-verteilte PB lautet $L(\theta) = \frac{1}{\theta^n} \mathbb{1}_{[v;\infty[}(\theta)$ mit $v = \max(x_1, \ldots, x_n)$

Übung: Begründen Sie $\hat{\theta}_{ML} = \max(X_1, \ldots, X_n)$ mit einer Skizze von L.

4.4 Momentenmethode (MM)

- ↪ Vergleich der Momente der Stichprobenvariablen $X_1, \ldots, X_n$ mit den empirischen Momenten der konkreten Stichprobe $x_1, \ldots, x_n$.
- ↪ geht auf Karl Pearson zurück (Ende des 19. Jahrhunderts)



- ↪ britischer Mathematiker, $*$ 27.03.1857 in London; $+$ 27.04.1936 in Coldharbour
- ↪ wissenschaftliche Schwerpunkte: Mathematik, Physik, Deutsche Literatur, Recht, Theologie und Geschichte, Statistik, Psychologie
- ↪ Gründung der weltweit ersten Statistik-Fakultät 1911, University College of London
- ↪ wichtige Beiträge: Korrelationskoeffizient, Momentenmethode, statistische Testtheorie, ...

Modellannahmen: Wir wollen $\theta = (\theta_1, \ldots, \theta_m)$ schätzen. Betrachte hierzu eine Stichprobe $x_1, \ldots, x_n$ zu den Stichprobenvariablen $X_1, \ldots, X_n \overset{u.i.v.}{\sim} F$.

Vorgehen (Kochrezept)

1.) Bestimme k-te empirische Momente $\hat{m}_k(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i^k \quad \forall\ k$

2.) Gleichsetzen empirischer und theoretischer Momente:

$$\hat{m}_k(x_1, \ldots, x_n) = g_k(\theta) = E(X_1^k) \quad \forall\ k \qquad (\star)$$

Beachte: $E(X_1^k)$ hängt von $\theta_1, \ldots, \theta_m$ ab.

3.) Bestimmung der Lösung $\hat{\theta}(x_1, \ldots, x_n)$ des Gleichungssystems. (vorausgesetzt werden die eindeutige Lösbarkeit sowie die „Messbarkeit“ von $\hat{\theta} : \mathbb{R}^n \to \Theta \subset \mathbb{R}^m$ als Stichprobenfunktion.

Momentenschätzer

Der durch $(\star)$ gegebene Zufallsvektor $\hat{\theta}_{MM}(X_1, \ldots, X_n)$ heißt **Momentenschätzer** des Parametervektors θ.

Beispiel: Normalverteilte Stichprobenvariablen

$X_1, \ldots, X_n \overset{u.i.v.}{\sim} \mathcal{N}(\mu, \sigma^2)$, μ und σ^2 unbekannt
$\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times (0, \infty)$

1.) Bestimmung der 1. und 2. empirischen Momente der Stichprobe:

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \hat{m}_2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

2.) Gleichsetzen der empirischen und theoretischen Momente:

$$\begin{aligned} g_1(\mu, \sigma^2) &= E(X_1) = \mu = \frac{1}{n} \sum_{i=1}^{n} x_i, \\ g_2(\mu, \sigma^2) &= E(X_1^2) = \sigma^2 + \mu^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Beispiel: Normalverteilte Stichprobenvariablen

3.) Auflösen des Gleichungssystems:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2$$

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

Beispiel: Normalverteilte Stichprobenvariablen

4.) Konstruktion der Momentenschätzer

$$\begin{aligned}\hat{\mu}_{MM}(X_1, \ldots, X_n) &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ (\hat{\sigma}^2)_{MM}(X_1, \ldots, X_n) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\end{aligned}$$

$\hookrightarrow$ $(\hat{\sigma}^2)_{MM} \triangleq \tilde{S}^2$

Beispiel: Binomialverteilte Stichprobenvariablen

$X_1, \ldots, X_n \overset{u.i.v.}{\sim} Bin(k,p)$, $\theta = (\theta_1, \theta_2) = (k,p)$, $\Theta = \mathbb{N} \times (0,1)$

1.) Bestimmung des 1./2. emp. Moments: $\hat{m}_1 = \frac{1}{n}\sum_{i=1}^n x_i, \quad \hat{m}_2 = \frac{1}{n}\sum_{i=1}^n x_i^2$

2.) Gleichsetzen der empirischen und theoretischen Momente:

- $g_1(k,p) = E(X_1) = k \cdot p \overset{!}{=} \frac{1}{n}\sum_{i=1}^n x_i$
- $g_2(k,p) = E(X_1^2) = \text{var}(X_1) + E(X_1)^2 = k \cdot p \cdot (1-p) + (k \cdot p)^2 \overset{!}{=} \frac{1}{n}\sum_{i=1}^n x_i^2$

3.) Auflösen des Gleichungssystems:

- $\hat{k} = \frac{1}{n \cdot \hat{p}}\sum_{i=1}^n x_i = \frac{1}{\hat{p}}\bar{x}$
- $$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n x_i^2 &= kp(1-p) + (kp)^2 \\ \Leftrightarrow \quad & \frac{1}{n}\sum_{i=1}^n x_i^2 &= \bar{x}(1-p) + \bar{x}^2 \\ \Leftrightarrow \quad & \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2 &= \bar{x}(1-p) \\ \Leftrightarrow \quad & \hat{p} = 1 - \frac{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}{\bar{x}} &= \frac{\bar{x} - \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}{\bar{x}} \end{aligned}$$

Beispiel: Binomialverteilte Stichprobenvariablen

4.) Konstruktion der Momentenschätzer

$$\hat{k}(X_1, \ldots, X_n) = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{p}(X_1, \ldots, X_n) = \frac{\bar{X}}{\hat{k}(X_1, \ldots, X_n)}$$

für $(X_1, \ldots, X_n) \neq (0, \ldots, 0)$.

Fritzbrötchen

Bäcker Kalkoves neueste Kreation ist das Fritzbrötchen, welches er in der Bäckerei neben dem Hörsaal verkaufen möchte. Unter 10 Studierenden hat er folgende Preisbereitschaften (PB) erfragt (in EuroCent): 20, 70, 85, 50, 75, 0, 40, 90, 95, 60.

Übung: Berechnen Sie $E(X_i)$, $var(X_i)$ für (mit $\theta > 0$) u.i.v. Re$(0, \theta)$-verteilte PB

Übung: Berechnen Sie einen MM-Schätzer für θ und vergleichen Sie mit $\hat{\theta}_{ML}$.

4.5 Gütekriterien für Punktschätzer

Nachfolgend betrachten wir folgende Problemstellungen:

Fragestellung 1

Gegeben: Punktschätzer T für μ_X
Frage: Wie „gut“ ist T?

Beispiel: Gegeben Körpergrößen $X_1, \ldots, X_n$ in cm. Wir schätzen die mittere Körpergröße konstant durch $T = 150$. Gut oder schlecht?

Fragestellung 2

Gegeben: Mehrere Punktschätzer $T_1, \ldots, T_k$ für μ_X
Frage: Welcher Schätzer ist „der beste“? Bzw. wie kann man überhaupt Schätzer vergleichen?

Beispiel: Betrachte erneut Körpergrößen $X_1, \ldots, X_n$ und die zwei Schätzer $T_1 = 150$ und $T_2 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Welcher ist besser?

Zu Fragestellung 1

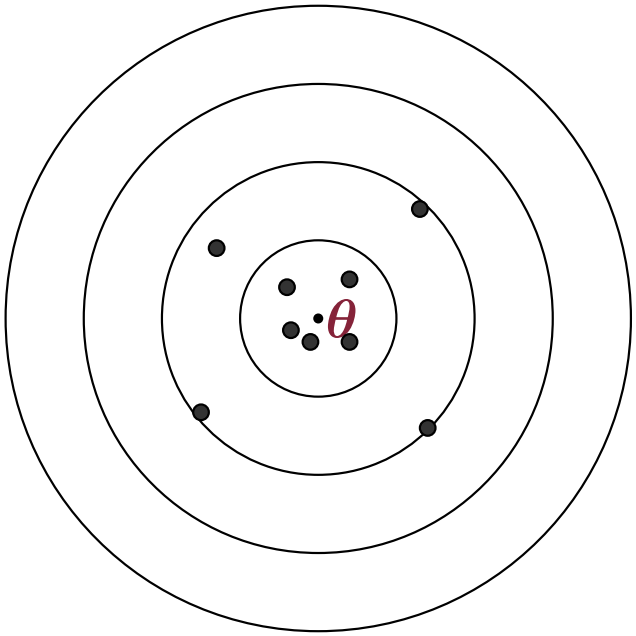
Betrachten wir erneut Körpergrößen $X_1, \ldots, X_n$. Die Schätzung des Erwartungswertes μ_X mit Hilfe des Schätzers

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

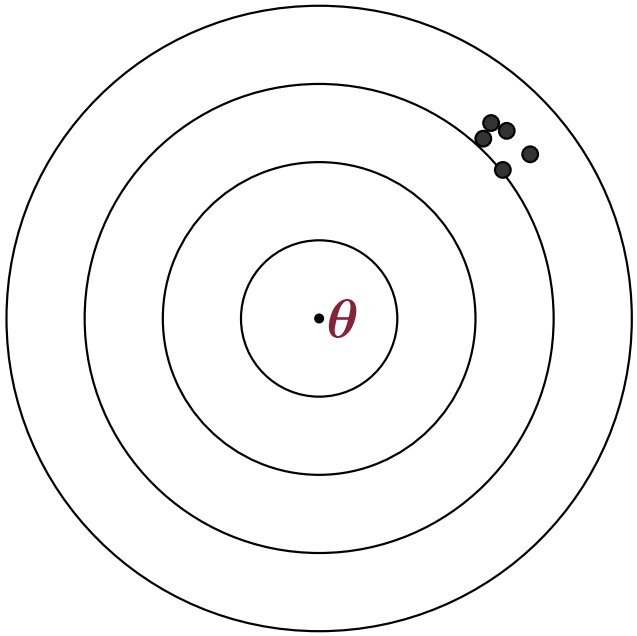
„macht intuitiv Sinn“!
Wir werden sehen, dass $\bar{X}$ tatsächlich ein „guter“ Schätzer (für μ_X) ist. Aber was heißt denn nun überhaupt gut?

Eine Analogie: Darts

Wir stellen uns eine Stichprobe eines Schätzers T für den Parameter θ als Pfeilwurf auf eine Dartscheibe mit **Bullseye** θ vor:



Im Mittel richtig



Systematisch daneben

Zur Erinnerung

Eine **Schätzfunktion** oder **Schätzstatistik** für einen Parameter θ ist eine **Funktion der Stichprobenvariablen** $X_1, \ldots, X_n$:

$$T = g(X_1, \ldots, X_n)$$

D. h. der Schätzer ist selbst eine Zufallsvariable!
D. h. wiederum, dass wir etwa Erwartungswert und Varianz von T betrachten können!

Erwartungstreue

Eine Statistik $T = g(X_1, \ldots, X_n)$ heißt **erwartungstreu** oder unverzerrt, wenn gilt

$$E_\theta(T) = \theta$$

Erwartungstreue des Stichprobenmittels einer u.i.v.-Stichprobe

$\bar{X}$ ist erwartungstreue Schätzstatistik für den Erwartungswert $\mu = E(X)$:

$$E_\mu(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} E_\mu(X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu$$

Beispiel Buchrückgabe: $\mu_1 = \bar{X}$

$$E(\hat{\mu}_1) = \frac{1}{10}\sum_{i=1}^{n} E(X_i) = \frac{1}{10}\cdot 10 \cdot \mu = \mu \quad \text{erwartungstreu}$$

Übung: Sind $\hat{\mu}_2 = X_1$, $\hat{\mu}_3 = 3 \cdot X_8$ bzw. $\hat{\mu}_4 = \frac{1}{9}\sum_{i=1}^{9} X_i$ erwartungstreu für μ?

$\rightsquigarrow$ "Erwartungstreue" als einziges Kriterium ist nicht ausreichend - und auch nicht immer sinnvoll.

Weitere Beispiele (u.i.v.-Wiederholungen)

$$E_p(\bar{X}) = p \quad \text{mit } X_i \in \{0,1\}, p = P(X_i = 1) \qquad \text{erwartungstreu}$$

$$E_{\sigma^2}(S^2) = E_{\sigma^2}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \sigma^2 \qquad \text{erwartungstreu}$$

$$E_{\sigma^2}(\tilde{S}^2) = E_{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$$

$$= \frac{n-1}{n} \cdot E_{\sigma^2}(S^2) = \frac{n-1}{n} \cdot \sigma^2 \qquad \text{nicht erwartungstreu}$$

(dabei $\sigma^2 = var(X_i)$)

Beachten Sie: Wie in diesen Beispielen entstehen Schätzfunktionen oft aus deskriptiven Lage- und Streuungskennzahlen, denen eine Zufallstichprobe mit geeigneten Zufallsvariablen (d.h. WS-Modell) zugrundeliegt.

Übung: Zeigen Sie in verschiedenen Schritten $E_{\sigma^2}(S^2) = \sigma^2$

$$E(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2) = \frac{1}{n-1}\sum_{i=1}^{n}E(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(E(X_i^2 - 2E(X_i\bar{X}) + E(\bar{X}^2))$$

1. Zeigen Sie zunächst: $E(X_i^2) = \sigma^2 + \mu^2$ (mit $\mu = E(X_i)$)

2. Stellen Sie entsprechend den zweiten Erwartungswert dar

3. Verfahren Sie ebenso mit dem dritten Erwartungswert.

4. Zeigen Sie $E_{\sigma^2}(S^2) = \sigma^2$ mit den Ergebnissen aus 1. bis 3.

Beispiel Buchrückgabe ($(x_1, \ldots, x_{10}) = (2, 14, 10, 0, 9, 20, 8, 2, 3, 2)$)

X_i: Verspätung der Buchabgabe bei i-ter Messung ($i = 1, \ldots, 10$),

$X_i \sim F$ (unbekannt)

Erwartungstreue Schätzung der Varianz σ^2 durch $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

$$\begin{aligned} S^2 &= \frac{1}{9} \left[(2-7)^2 + (14-7)^2 + (10-7)^2 + (0-7)^2 + (9-7)^2 \right. \\ &\quad \left. + (20-7)^2 + (8-7)^2 + (2-7)^2 + (3-7)^2 + (2-7)^2 \right] \\ &= \frac{1}{9} (25 + 49 + 9 + 49 + 4 + 169 + 1 + 25 + 16 + 25) \\ &= \frac{1}{9} \cdot 372 = 41.\bar{3} \end{aligned}$$

```
> x = c(2, 14, 10, 0, 9, 20, 8, 2, 3, 2)
> var(x)
[1] 41.33333
```

Bias / Verzerrung

Eine Statistik $T = g(X_1, \ldots, X_n)$ heißt **erwartungstreu** , wenn gilt $E_\theta(T) = \theta$.

Der *systematische* Fehler einer Schätzstatistik heißt **Verzerrung** oder **Bias**:

$$Bias_\theta(T) = E_\theta(T) - \theta$$

Verzerrung eines nicht erwartungstreuen Schätzers

$$E_{\sigma^2}(\tilde{S}^2) = E_{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{n-1}{n} \cdot \sigma^2 \quad \text{nicht erwartungstreu}$$

$$\rightsquigarrow \quad Bias_{\sigma^2}(\tilde{S}^2) = E_{\sigma^2}(\tilde{S}^2) - \sigma^2 = \left(\tfrac{n}{n} \cdot \sigma^2 - \tfrac{1}{n} \cdot \sigma^2\right) - \sigma^2 = -\tfrac{1}{n} \cdot \sigma^2$$

Beispiel Buchrückgabe: Annahme: $E(X_i) = \mu$

$$E(\hat{\mu}_3) \quad = \quad E(3 \cdot X_8) = 3 \cdot \mu \neq \mu \quad \text{nicht erwartungstreu}$$

Übung: Bestimmen Sie den Bias des Schätzers $\hat{\mu}_3$.

Asymptotische Erwartungstreue

Eine Schätzstatistik heißt **asymptotisch erwartungstreu** für θ, wenn gilt:

$$\lim_{n\to\infty} E_\theta(T) = \theta$$

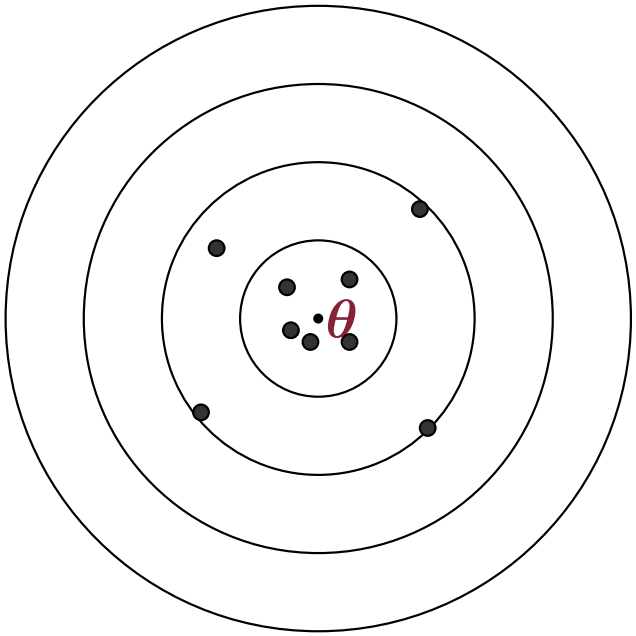
$\hookrightarrow$ $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ ist asymptotisch erwartungstreu, da

$$\lim_{n\to\infty} E_{\sigma^2}(\tilde{S}^2) = \lim_{n\to\infty} \frac{n-1}{n} \cdot \sigma^2 = \sigma^2$$

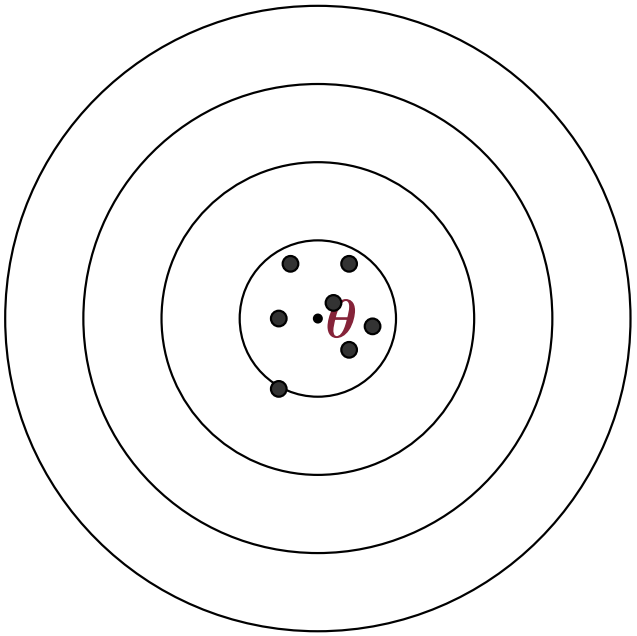
.

$\hookrightarrow$ aber: für kleines n kann der Bias erheblich sein.

Eine Analogie: Darts



Im Mittel richtig, aber ungenau



im Mittel richtig und ziemlich genau

Standardfehler und Varianz einer Schätzstatistik

Die **Schätzervarianz** einer Schätzstatistik $T = g(X_1, \ldots, X_n)$ ist

$$var(T) = E\left([T - E(T)]^2\right)$$

Der **Standardfehler** einer (meist erwartungstreuen) Schätzstatistik ist

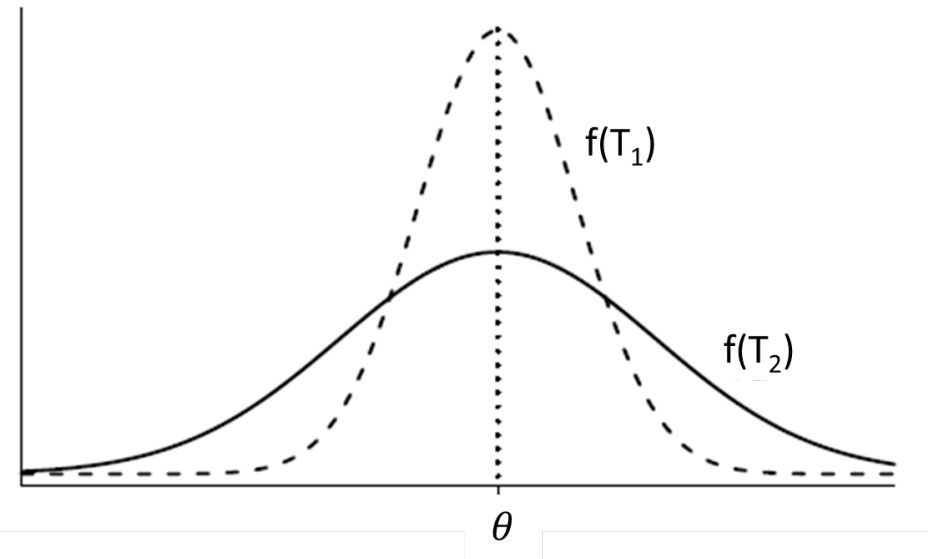
$$\sigma_T = \sqrt{var(T)}.$$

Beispiel: Standardfehler des Gruppenmittels

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{Var(X)}{n}}.$$

- $\hookrightarrow$ Gütekriterium für Schätzer: erwartungstreu mit minimaler Schätzervarianz (nur in Ausnahmefällen realisierbar).
- $\hookrightarrow$ Der (theoretische) Standardfehler ist unbekannt. Für ihn werden in den Anwendungen Schätzungen benötigt, welche selber als (empirische) Standardfehler bezeichnet werden (z.B. für Intervallschätzer benötigt).

Effiziente Schätzstatistiken



Die Verteilung eines „guten“ Schätzers sollte:

- keine oder nur geringe systematische Abweichung nach oben und unten zum unbekannten Parameter aufweisen
- eine geringe Streuung besitzen.

Beide Kriterien lassen sich im MSE vereinen.

Fehler eines Schätzers (MSE)

Die **erwartete mittlere quadratische Abweichung** (MSE: *mean squared error*) bestimmt sich durch

$$
\begin{aligned}
MSE(T) &= E\Big([T-\theta]^2\Big) = E\Big([T - E(T) + E(T) - \theta]^2\Big) \\
&= E\Big([T-E(T)]^2 + 2\cdot[T-E(T)]\cdot[E(T)-\theta] + [E(T)-\theta]^2\Big) \\
&= E\Big([T-E(T)]^2\Big) + 2\cdot E\Big([T-E(T)]\cdot[E(T)-\theta]\Big) \\
&\quad + E\Big([E(T)-\theta]^2\Big) \qquad \text{beachte: } [E(T)-\theta] \text{ ist konstant} \\
&= E\Big([T-E(T)]^2\Big) + [E(T)-\theta]\cdot 2\cdot \underbrace{E\Big([T-E(T)]\Big)}_{=E(T)-E(T)=0} + [E(T)-\theta]^2 \\
&= E\Big([T-E(T)]^2\Big) + [E(T)-\theta]^2 = Var(T) + [Bias(T)]^2
\end{aligned}
$$

$\hookrightarrow$ führt Schätzervarianz und Bias in einem Gütekriterium zusammen.

$\hookrightarrow$ meist nur in „linearen“ Ausnahmefällen (Existenz erwartungstreuer Schätzer) verwendbar, wird asymptotisch (Stichprobenumfang $\to \infty$) nützlich.

Eigenschaften erwartungstreuer Schätzer

Für einen zum Parameter θ erwartungstreuen Schätzer T gilt:

a) $E_\theta(T) = \theta$

b) $Bias_\theta(T) = E_\theta(T) - \theta \overset{a)}{=} 0$

c) $MSE_\theta(T) = Var_\theta(T) + [Bias_\theta(T)]^2 \overset{b)}{=} Var_\theta(T)$.

Schätzervergleich durch MSE

Im Vergleich zweier Schätzstatistiken T_1 und T_2 heißt T_1 **MSE-effizienter**, wenn für alle zugelassenen Verteilungen gilt:

$$MSE(T_1) \leq MSE(T_2)$$

Eine Schätzstatistik heißt **MSE-effizient**, wenn ihr MSE den kleinsten möglichen Wert für alle zugelassenen Schätzer annimmt.

- $\hookrightarrow$ Für erwartungstreue Statistiken reduziert sich der Vergleich auf die entsprechenden Varianzen der Schätzstatistiken.
- $\hookrightarrow$ ansonsten alleiniger Varianzvergleich sinnlos (z.B. konstante Schätzstatistiken)

Beispiel: 200-facher Wurf einer verbogenen Münze

Seien $X_1, \ldots, X_{200}$ u.i.v. mit $X_i \sim \mathcal{B}(1, p)$ (1 ≡Z, 0 ≡W)

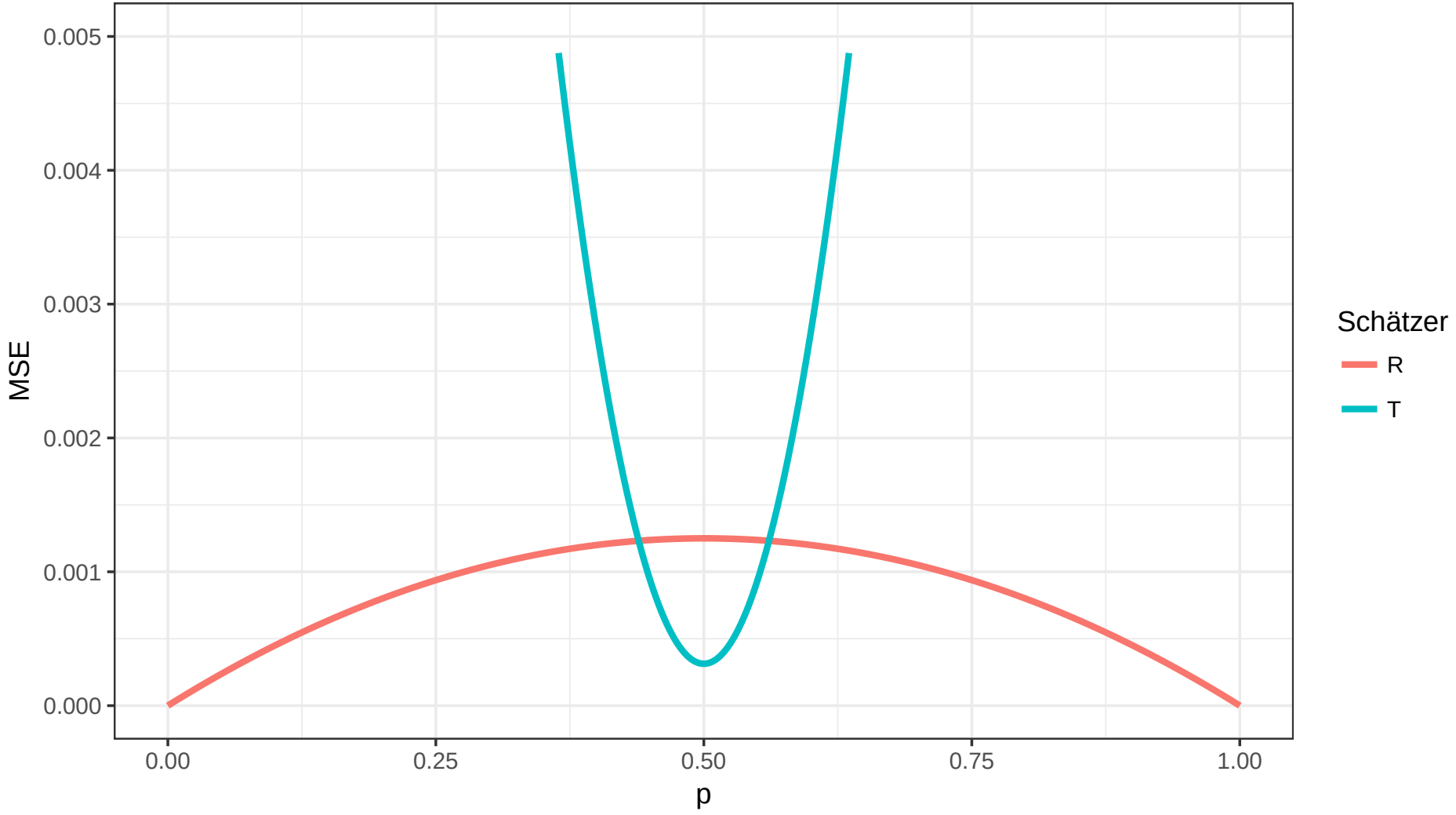
Mögliche Schätzer für p sind bspw.:

- $T = \bar{X}_{200}$ (hier: $t = 0.45$) (e-treu)

  $MSE(T) = var(T) = \frac{p \cdot (1-p)}{200}$ (minimal unter e-treuen)

- $R = \frac{1}{2} T + \frac{1}{4}$ (hier: $r = 0.475$) (nicht e-treu)

$$
\begin{aligned}
MSE(R) &= var(R) + [E(R) - p]^2 \\
&= var\left(\tfrac{1}{2} \cdot T + \tfrac{1}{4}\right) + \left[E\left(\tfrac{1}{2} \cdot T + \tfrac{1}{4}\right) - p\right]^2 \\
&= \tfrac{1}{4} \cdot var\left(T\right) + \left(\tfrac{1}{4} - \tfrac{p}{2}\right)^2 \\
&= \tfrac{1}{4} \cdot \tfrac{p \cdot (1-p)}{200} + \tfrac{1}{4} \cdot \left(p - \tfrac{1}{2}\right)^2
\end{aligned}
$$



$\rightsquigarrow$ Beispiel für die Unvergleichbarkeit von Schätzern

Übung: Vergleichen Sie die drei erwartungstreuen Schätzer $\hat{\mu}_1 = \frac{1}{10}\sum_{i=1}^{10} X_i$, $\hat{\mu}_2 = X_1$ und $\hat{\mu}_4 = \frac{1}{9}\sum_{i=1}^{9} X_i$ im Buch-Beispiel hinsichtlich der MSE-Effizienz.

Verallgemeinerung des MSE: Verlustfunktion, Risikofunktion

↪ Es bezeichne:

- Θ die Menge aller möglichen Parameter,
- $\mathcal{T}$ den Wertebereich der (zur Verfügung stehenden) Schätzer.

↪ Eine **Verlustfunktion** ist eine Funktion $L : \Theta \times \mathcal{T} \to [0; \infty]$.

↪ Das **Risiko** einer Schätzfunktion $T = T(X)$ bei vorliegendem Parameter θ ist

$$R(\theta, T) = E_\theta(L(\theta, T(X)).$$

Die **Risikofunktion** von $T = T(X)$ ist die Funktion $R : \theta \mapsto R(\theta, T(X))$.

Beispiele von Verlustfunktionen

↪ $L(\theta, t) = (t - \theta)^2$ (= MSE)

↪ $L(\theta, t) = |\theta - t|^r$ mit $r > 0$

↪ Auch Fehlerwahrscheinlichkeiten 1./2. Art bei statistischen Tests lassen sich als Risiken darstellen (0/1-wertige Verlustfunktion)

- ↪ Konsistenz eines Schätzers bedeutet, dass dieser mit wachsendem Stichprobenumfang dem unbekannten Parameter immer näher kommt.
- ↪ Entsprechend den verschiedenen Konvergenzarten (s.o. Gesetze großer Zahlen) gibt es verschiedene Konzepte für Konsistenz:
  - ↪ Konsistenz im quadratischen Mittel
  - ↪ schwache Konsistenz
  - ↪ starke Konsistenz

Konsistenz im quadratischen Mittel (für einen Schätzer)

Eine Schätzstatistik T_n heißt **konsistent im quadratischen Mittel**, wenn gilt

$$MSE(T_n) = Var(T_n) + [Bias(T_n)]^2 \overset{n\to\infty}{\longrightarrow} 0.$$

- $\hookrightarrow$ sowohl Verzerrung als auch Varianz verschwinden für wachsende Stichprobenumfänge
- $\hookrightarrow$ asymptotische Eigenschaft, kann für endliches n erhebliche Bias- und Varianzanteile aufweisen

Schwache Konsistenz eines Schätzers

Die Schätzstatistik $T_n = g(X_1, \ldots, X_n)$ heißt **schwach konsistent**, wenn zu beliebigem $\varepsilon > 0$ gilt:

$$\lim_{n\to\infty} P(|T_n - \theta| < \varepsilon) = 1 \quad \text{bzw.}$$

$$\lim_{n\to\infty} P(|T_n - \theta| \geq \varepsilon) = 0$$

D.h. für wachsendes n konvergiert die Wahrscheinlichkeit, mit der die Schätzstatistik T_n höchstens um ε vom wahren Wert θ abweicht, gegen Null (*Konvergenz in Wahrscheinlichkeit*).

Schwache Konsistenz ist oft eine Folge des schwachen Gesetzes großer Zahlen bzw. der Tschebytscheff-Ungleichung (s.o.)

T_n ist konsistent im quadratischen Mittel $\Rightarrow$ T_n ist schwach konsistent.

Beweis: mit der Markoff-Ungleichung (mit $g(x) = x^2$) gilt für $\varepsilon > 0$

$$P(|T_n - \theta| \geq \varepsilon) \leq \frac{E\left([T_n - \theta]^2\right)}{\varepsilon^2} = \frac{1}{\varepsilon^2} \cdot MSE(T_n) \overset{n\to\infty}{\longrightarrow} 0$$

Folglich ist auch eine Schätzstatistik schwach konsistent, wenn sie erwartungstreu ist und $Var(T_n) \to 0$ für $n \to \infty$ gilt.

Beweis: Wegen obiger Abschätzung (inkl. Markoff-Ungleichung) und $E(T_n) = \theta$ gilt

$$P(|T_n - \theta| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \cdot \underbrace{MSE(T_n)}_{=Var(T_n)+[E(T_n)-\theta]^2} = \frac{1}{\varepsilon^2} \cdot Var(T_n) \overset{n\to\infty}{\longrightarrow} 0.$$

Beispiel

Sei $X \sim \mathcal{N}(\mu, \sigma^2)$. Für unabhängige Wiederholungen $X_1, \ldots, X_n$ erhält man

$E(\bar{X}) = \mu,\ Var(\bar{X}) = \frac{\sigma^2}{n}$, d.h. $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

$\hookrightarrow$ $\bar{X}$ ist erwartungstreu, abnehmende Varianz für wachsendes n
$\Rightarrow$ konsistent im quadratischen Mittel

$\hookrightarrow$ Überprüfung der schwachen Konsistenz:

$$\begin{aligned} P(|\bar{X} - \mu| \leq \varepsilon) &= P\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \leq \frac{\varepsilon}{\sigma/\sqrt{n}}\right) \\ &= P\left(-\frac{\varepsilon}{\sigma} \cdot \sqrt{n} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{\varepsilon}{\sigma} \cdot \sqrt{n}\right) \\ &= \Phi\left(\frac{\varepsilon}{\sigma} \cdot \sqrt{n}\right) - \Phi\left(-\frac{\varepsilon}{\sigma} \cdot \sqrt{n}\right) = 2 \cdot \Phi\left(\frac{\varepsilon}{\sigma} \cdot \sqrt{n}\right) - 1 \overset{n\to\infty}{\longrightarrow} 1 \end{aligned}$$

$\Rightarrow$ $\bar{X}$ ist schwach konsistent.

Übung: Ist folgende Aussage richtig? Eine asymptotisch erwartungstreue Schätzstatistik T_n ist schwach konsistent, wenn $var(T_n) \to 0$ für $n \to \infty$ gilt.

Starke Konsistenz eines Schätzers

Die Schätzstatistik $T_n = g(X_1, \ldots, X_n)$ heißt **stark konsistent**, falls:

$$P_\theta \left(\lim_{n \to \infty} |T_n - \theta| = 0 \right) \quad = \quad 1 \qquad \forall\, \theta$$

$\hookrightarrow$ *fast sichere Konvergenz* $(T_n \xrightarrow{f.s.} \theta)$

- $\hookrightarrow$ T ist stark konsistent $\Rightarrow$ T ist schwach konsistent.
- $\hookrightarrow$ Unter bestimmten Zusatzannahmen gilt auch:
  T ist stark konsistent $\Rightarrow$ T ist konsistent im quadratischen Mittel.
- $\hookrightarrow$ Die starke Konsistenz eines Schätzers ergibt sich fast immer aus dem starken Gesetz großer Zahlen.

- ↪ Viele Schätzer T_n lassen sich bei festem n nicht vergleichen bzw. es existieren keine optimalen Schätzer bei festem n.
- ↪ Um die Schätzer zu „selektieren“, wird oft die Konsistenz von $T_n(X)$ für θ in einem geeigneten Sinne gefordert. Das reicht aber oft auch noch nicht aus.
- ↪ Zusätzlich zur Konsistenz wird dann die asymptotische Normalität gefordert:

$$\lim_{n\to\infty} P(\sqrt{n}\frac{T_n - \theta}{\sigma(\theta)} \leq t) = \Phi(t)$$

  Diese ist oft aufgrund des zentralen Grenzwertsatzes erfüllt.
- ↪ $\sigma^2(\theta)$ wird als **asymptotische Schätzervarianz** bezeichnet. Je kleiner $\sigma^2(\theta)$, desto (asymptotisch) besser ist der Schätzer. Ein Schätzer mit minimalem $\sigma^2(\theta)$ heißt **asymptotisch effizient**.
- ↪ Untergrenze für $\sigma^2(\theta)$ ist oft die Cramer-Rao-Schranke $1/I(\theta)$ mit der Fisher-Information $I(\theta) = var_\theta(S_\theta)$ und Score-Funktion $S_\theta = \frac{\partial}{\partial\theta}\ln(f(x,\theta))$.
- ↪ Viele ML-Schätzer sind asymptotisch effizient, d.h. (stark) konsistent und ihre asymptotische Schätzervarianz ist $1/I(\theta)$

Elfmeter (Bernoullikette, $X_1, X_2, \cdots \sim Bin(1,p)$, u.i.v.-Folge)

- ↪ Score-Funktion: $S_p(x) = \frac{\partial}{\partial p} \ln(p^x(1-p)^{1-x}) = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)}$
- ↪ Fisher-Information: $I(p) = var(S_p(X_1)) = var(\frac{X_1-p}{p(1-p)}) = \frac{var(X_1)}{p^2(1-p)^2} = \frac{1}{p(1-p)}$
- ↪ Cramer-Rao-Schranke: $1/I(p) = p(1-p)$
- ↪ Asymptotische Schätzervarianz des ML-Schätzers $T_n = \bar{X}_n$:
  - □ T_n ist stark konsistent wegen des SGGZ: $\bar{X} \to p$ f.s.
  - □ T_n ist asymptotisch normal wegen des ZGS: $\sqrt{n}\frac{\bar{X}_n-p}{\sqrt{p(1-p)}} \to \mathcal{N}(0,1)$
  - □ Die asymptotische Schätzervarianz ist dann $\sigma^2(p) = p(1-p)$
- ↪ Der ML-Schätzer hat die Cramer-Rao-Schranke als asymptotische Schätzervarianz.

Übung: $X_1, X_2, \ldots$ sei eine u.i.v.-Folge von Poi(λ)-verteilten ZV (z.B. die Anzahl der Patienten in einer Notaufnahme an aufeinander folgenden Tagen). Bestimmen Sie zunächst einen ML-Schätzer für λ auf Basis von $X_1, \ldots, X_n$ und prüfen Sie, ob der Schätzer asymptotisch effizient ist. Hinweis: Dichte: $f_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda}, x \in \mathbb{N}_0$

4.6 Zusammenfassung

Schätzer sind Zufallsvariablen $T(X_1, \ldots, X_n)$ mit denen (z.B.) Parameter der Modellverteilung ermittelt werden sollen.

- $\hookrightarrow$ Konstruktionsprinzipien: ML, MM,...
- $\hookrightarrow$ Nicht jeder Schätzer ist gut. Es gilt Vor- und Nachteile abzuwägen.
- $\hookrightarrow$ Erwartungstreue als eine grundlegende Eigenschaft (zumindest asymptotisch)
- $\hookrightarrow$ Unter den erwartungstreuen Schätzern sind die mit der kleinsten Varianz zu bevorzugen.
- $\hookrightarrow$ Allgemein werden Schätzer anhand ihres Risikos verglichen.
- $\hookrightarrow$ Schätzervergleich oft nur für wachsendes n durchführbar. Hier sind konsistente Schätzer vorzuziehen. Unter den konsistenten Schätzern sind diejenigen zu bevorzugen, welche asymptotisch effizient sind.

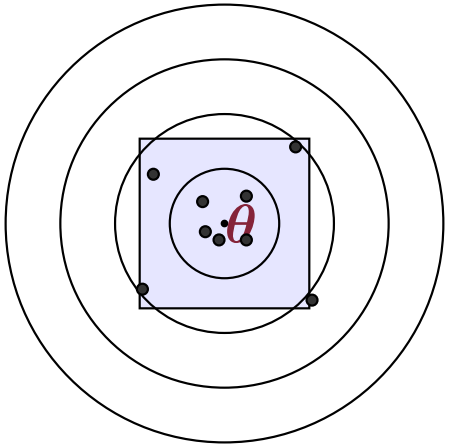
Ausblick: Weil Schätzer zufällig sind, können sie den „richtigen“ Wert eines Parameters nicht angeben. Statt dessen Bestimmung von Bereichen, welche den Parameter mit einer bestimmten (Mindest-)-WS „überdecken“.

5. Intervall- und Bereichsschätzer

- Intervallschätzung
- Konstruktionsprinzip für Konfidenzintervalle
- Konfidenzintervalle für Erwartungswert
- Konfidenzintervall für die Varianz
- Zusammenfassung

5.1 Intervallschätzung

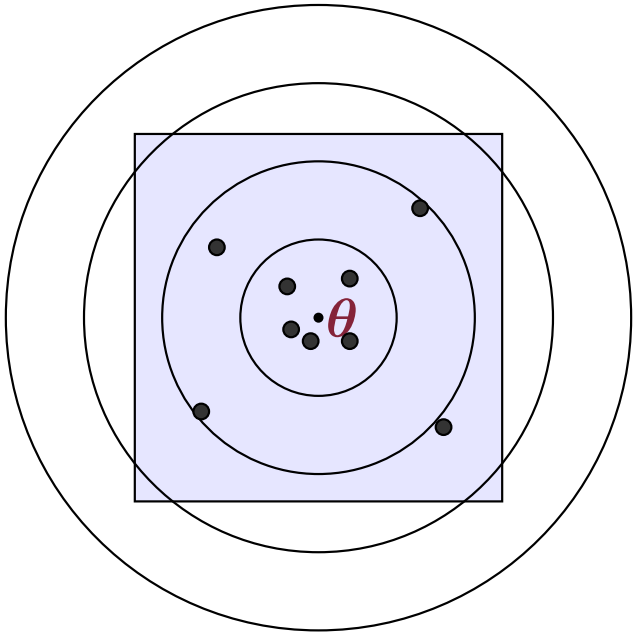
Bisher: Verwendung von Stichprobenergebnissen zur **Punkt**schätzung $\hat{\theta}$ eines unbekannten Parameters θ (bzw. einer unbek. Kennzahl) der Grundgesamtheit.



- $\hookrightarrow$ **Folgerung:** Schätzwert $\hat{\theta}$ und geschätzter Parameter θ stimmen (so gut wie) nie überein!
- $\hookrightarrow$ **Jetzt:** Angabe eines **Schätzintervals** $[G_u, G_o]$ in dem sich der unbekannte Parameter θ mit (kontrollierbarer) *Überdeckungswahrscheinlichkeit* $(1-\alpha)$, $\alpha \in (0,1)$ befindet.[1]

Mit dem Parameter $\alpha \in (0, 1)$ – der sogenannten Irrtumswahrscheinlichkeit – kontrollieren wir die Größe des Intervals. Intuitiv:

- $\hookrightarrow$ Je größer α (je kleiner die vorgegebene Überdeckungswahrscheinlichkeit) ist, desto kleiner darf das Intervall sein.
- $\hookrightarrow$ Je kleiner α (je größer die vorgegebene Überdeckungswahrscheinlichkeit) ist, desto größer muss das Intervall sein.



Zweiseitiges $(1-\alpha)$-Konfidenzintervall

Ein (zweiseitiges) $(1-\alpha)$-Konfidenzintervall (KI) $[G_u, G_o]$ zur Irrtumswahrscheinlichkeit $\alpha \in (0,1)$, wird gebildet aus Schätzstatistiken

$$G_u = g_u(X_1, \ldots, X_n) \text{ und } G_o = g_o(X_1, \ldots, X_n)$$

wenn folgende Voraussetzungen erfüllt sind

$$P(G_u \leq G_o) = 1 \text{ und } P(\theta \in [G_u; G_o]) \geq 1-\alpha.$$

Einseitige $(1-\alpha)$-Konfidenzintervalle

Ein **einseitiges** $(1-\alpha)$-**Konfidenzintervall** $]-\infty; G_o]$ liegt vor, wenn

$$P(\theta \leq G_o) \geq 1-\alpha$$

Ein **einseitiges** $(1-\alpha)$-**Konfidenzintervall** $[G_u; \infty[$ liegt vor, wenn

$$P(\theta \geq G_u) \geq 1-\alpha$$

Ein paar Anmerkungen:

- ↪ Die Intervallgrenzen sind Zufallsvariablen $G_{u/o} = g_{u/o}(X_1, \ldots, X_n)$, die von den Stichprobenvariablen abhängen.
- ↪ Die Schreibweisen der Ereignisse suggerieren, dass θ die ZV ist, zu der eine WS berechnet wird. Tatsächlich handelt es sich um folgende Ereignisse:
  - □ $\{\theta \in [G_u, G_o]\} = \{\omega \in \Omega : G_u(\omega) \leq \theta, G_o(\omega) \geq \theta\}$
  - □ $\{\theta \leq G_o\} = \{\omega \in \Omega : G_o(\omega) \geq \theta\}$
  - □ $\{\theta \geq G_u\} = \{\omega \in \Omega : G_u(\omega) \leq \theta\}$
- ↪ Die Überdeckungs-WS $1 - \alpha$ ist eine (vorgegebene) Eigenschaft des konkreten Schätzverfahrens. Mit Mindest-WS $1 - \alpha$ wird ein Intervall (z.B. $[g_u(x_1, \ldots, x_n),\ g_o(x_1, \ldots, x_n)]$) generiert, das den wahren Wert θ überdeckt.
- ↪ Im stetigen Modellen für die Grundgesamtheit wird das KI so angesetzt, dass die Mindest-WS $1 - \alpha$ exakt eingehalten wird.
- ↪ Zweiseitige KI werden oftmals symmetrisch angesetzt, d.h. $P(\theta > G_o) \leq \frac{\alpha}{2}$, $P(\theta < G_u) \leq \frac{\alpha}{2}$ (im stetigen Fall: „$= \frac{\alpha}{2}$“).

5.2 Konstruktionsprinzip für Konfidenzintervalle

1. Bestimme einen **Punktschätzer** $T_n = T(X_1, \ldots, X_n)$ für θ (z. B. mittels ML, MM, ...).
2. Lege das (zweiseitige) Konfidenzintervall fest mit

$$G_u = T_n - t \cdot A_n \text{ und } G_o = T_n + t \cdot A_n$$

   wobei $t > 0$ und $A_n = a(X_1, \ldots, X_n) > 0$ eine geeignete Statistik ist.

   Es gilt $\theta \in [T_n - t \cdot A_n \,;\; T_n + t \cdot A_n] \quad \Leftrightarrow \quad -t \leq \dfrac{T_n - \theta}{A_n} \leq t$
3. Wähle $t > 0$ derart, dass gilt $P_\theta \left(-t \leq \frac{T_n - \theta}{A_n} \leq t \right) \overset{!}{=} 1 - \alpha \quad (\text{bzw. } \overset{!}{\approx} 1 - \alpha)$

Typische Wahl für A_n: Schätzer für den Standardfehler von T_n, so dass der Bruch (zumindest approximativ) eine Standardisierung von T_n darstellt.

Übung: Zeigen Sie, dass gilt:

$$\theta \in [T_n - t \cdot A_n \; ; \; T_n + t \cdot A_n] \quad \Leftrightarrow \quad -t \leq \frac{T_n - \theta}{A_n} \leq t$$

Festlegung der Größen t, A_n im KI $[T_n - t \cdot A_n\ ,\ \ T_n + t \cdot A_n]$

- $\hookrightarrow$ $t > 0$ ist (meist) $(1 - \frac{\alpha}{2})$-Quantil einer symmetrischen „Standardverteilung“ (oft approximiert durch Anwendung des zentralen Grenzwertsatzes).
- $\hookrightarrow$ A_n ist (meist konsistenter) Schätzer für den Standardfehler $\sqrt{var(T_n)}$ (wird als empirischer Standardfehler bzw. ebenfalls als Standardfehler bezeichnet).
- $\hookrightarrow$ In Ausnahmefällen ist A_n zufallsunabhängig (KI mit fester Breite).

Analog bei einseitigen KI oder unsymmetrischen Intervallgrenzen.

- □ Fall „linksseitig unbeschränkt“: $]-\infty; G_o]$ mit $G_o = T_n + t \cdot A_n$
- □ Erläuterung: Die Schätzung T_n für θ muss nach oben etwas ausgeweitet werden, um θ mit ausreichender WS zu „überdecken“.
- □ $P(G_o \geq \theta) = 1 - \alpha$, d.h. $P(T_n + t \cdot A_n \geq \theta) = 1 - \alpha$
- □ Es gilt $T_n + t \cdot A_n \geq \theta \Leftrightarrow T_n - \theta \geq -t \cdot A_n \Leftrightarrow \frac{T_n - \theta}{A_n} \geq -t$.
- □ Also wird gefordert: $P(\frac{T_n - \theta}{A_n} \geq -t) \stackrel{!}{=} 1 - \alpha$
- □ Bei stetigen Modellen ist das i.d.R. gleichwertig zu $P(\frac{T_n - \theta}{A_n} \leq -t) \stackrel{!}{=} \alpha$
- □ $-t$ ist also das α-Quantil der Stichprobenverteilung von $\frac{T_n - \theta}{A_n}$, welche meist approximativ nicht mehr von θ abhängt (ZGS).

Übung: Wie lautet der entsprechende Ansatz für ein rechtsseitig unbeschränktes $(1-\alpha)$-Konfidenzintervall $[G_n; \infty[$

5.3 Konfidenzintervalle für den Erwartungswert

Seien $X_1, \ldots, X_n \overset{u.i.v}{\sim} \mathcal{N}(\mu, \sigma^2)$. Gesucht ist ein (zweiseitiges) Konfidenzintervall für den Erwartungswert μ.

Erster Fall: σ ist bekannt

Ein bekannter Punktschätzer für μ ist das arithmetische Mittel $T_n = \bar{X}$. Es gilt:

$$X_1 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \bar{X} = \frac{X_1 + \cdots + X_n}{n} \sim \mathcal{N}(\mu, \sigma^2/n)$$

$$\Rightarrow \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \quad (\star)$$

$(\star)$ enthält den unbekannten Parameter μ; die Verteilung von $(\star)$ ist unabhängig von μ.

Das KI kann in der Form $[T_n - t \cdot A_n; T_n + t \cdot A_n]$ mit $T_n = \bar{X}$ und $A_n = \sigma/\sqrt{n}$ gewählt werden. Seine Breite ist dann zufallsunabhängig $2t \cdot \sigma/\sqrt{n}$.

Zur Festlegung von t in $[T_n - t \cdot A_n; T_n + t \cdot A_n]$:

$\hookrightarrow$ Es gilt

$$1 - \alpha \stackrel{!}{=} P(\bar{X} - t\sigma/\sqrt{n} \leq \mu \leq \bar{X} + t\sigma/\sqrt{n}) \quad = \quad P\left(-t \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq t\right)$$

$$= \Phi(t) - \Phi(-t) = 2\Phi(t) - 1$$

$\hookrightarrow$ Dabei gilt $2\Phi(t) - 1 = 1 - \alpha \Leftrightarrow \Phi(t) = 1 - \alpha/2$

Zweiseitiges KI für μ, wenn σ^2 bekannt

$$[G_u; G_o] = \left[\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\ ;\ \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

mit $z_{1-\alpha/2}$, d.h. dem $(1 - \alpha/2)$-Quantil der Standardnormalverteilung.

z.B. mit R und $\alpha = 0.01 \rightarrow z_{1-\alpha/2} = z_{0.995} = 2.575 \quad \rightarrow$ `qnorm(0.995)`

$\hookrightarrow$ Bestimme die Breite $b_{\text{KI}} = G_o - G_u$ des Konfidenzintervalls

$$[G_u\ ;\ G_o] = \left[\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\ ;\ \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

$\hookrightarrow$ Was passiert, wenn wir α klein oder groß machen? Was passiert, wenn wir den Stichprobenumfang n gegen unendlich laufen lassen?

$$\begin{aligned} b_{\text{KI}} = G_o - G_u &= \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} - \left(\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \\ &= 2 \cdot z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}. \end{aligned}$$

$\hookrightarrow$ Effekt von α (bei festem n) .

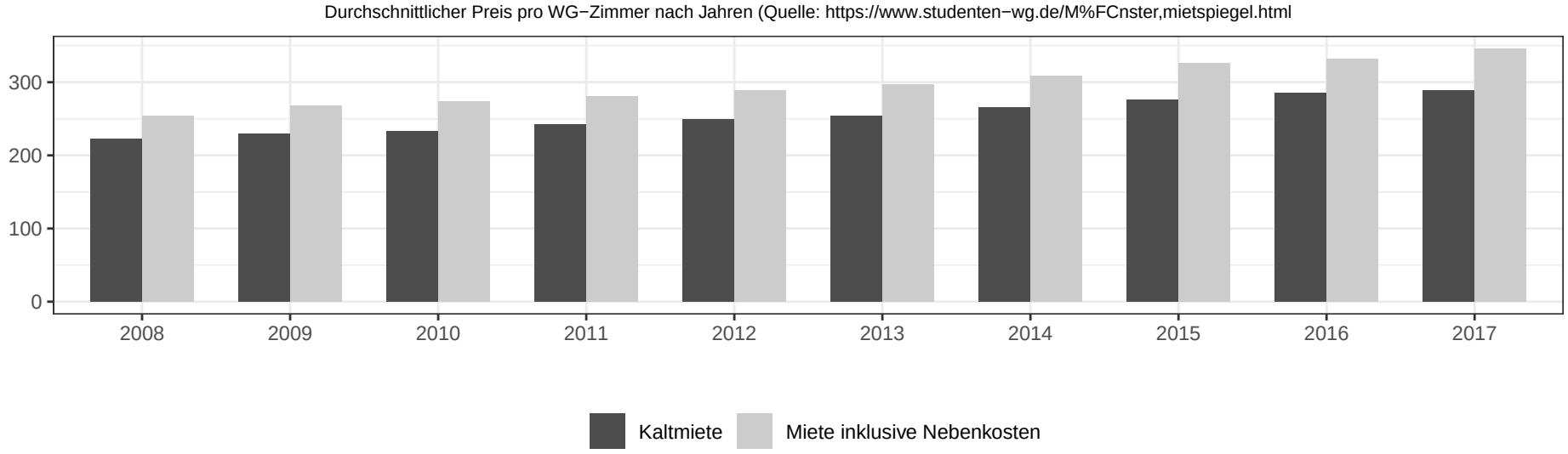
- $\square$ α klein $\Rightarrow$ Breite nimmt zu
- $\square$ α groß $\Rightarrow$ Breite nimmt ab.

$\hookrightarrow$ Bei konstantem α nimmt die Breite für wachsendes n ab, d.h.

$$\lim_{n\to\infty} 2 \cdot z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 2 \cdot z_{1-\alpha/2} \cdot \sigma \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0.$$

Fiktives Zahlenbeispiel zur Monatsmiete von Studenten in MS

- ↪ n=144, $\bar{X}$=300
- ↪ Annahme: $X_1, \ldots, X_n \overset{u.i.v}{\sim} \mathcal{N}(\mu, \sigma^2)$. Annahme: $\sigma = 100$ ist bekannt (?).
- ↪ ges.: Zweiseitiges Konfidenzintervall für μ, $\alpha = 0.05$



Übung: Bestimmen Sie das zweiseitige Konfidenzintervall für μ ($\alpha = 0.05$) gegeben $n = 144$, $\bar{X} = 300$, $\sigma = 100$.

Übung: Wie groß müsste n mindestens sein, um zu einem Signifikanzniveau $\alpha = 0.05$ ein zweiseitiges Konfidenzintervall für μ max. Breite 20 zu erhalten?

Übung: Verifizieren Sie Ihr Ergebnis, indem Sie zunächst das entsprechende Konfidenzintervall und anschließend dessen Breite ermitteln.

Einseitiges Konfidenzintervall $[G_u; \infty[$ für μ bei bekanntem σ^2

$\hookrightarrow$ Analog zum Vorgehen im zweiseitigen Fall:

$$1 - \alpha = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \ \leq \ z_{1-\alpha}\right) = P\left(\mu \ \geq \ \bar{X} - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$\hookrightarrow$ Also $G_u = \bar{X} - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$

$\hookrightarrow$ Analog mit oberer Grenze: $\quad KI = \left]-\infty \ ; \ \bar{X} + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right]$

Beispiel: Monatsmieten von Studenten in Münster

- ↪ n=144, $\bar{X}$=300, Annahme: $X_1, \ldots, X_n \overset{u.i.v}{\sim} \mathcal{N}(\mu, 100^2)$.
- ↪ ges.: Einseitiges oberes Konfidenzintervall für μ, $\alpha \in \{0.01, 0.05, 0.10\}$

- ↪ Für allgemeines α: $-\infty < \mu \leq \bar{X} + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = 300 + z_{1-\alpha} \cdot \frac{100}{\sqrt{144}}$
- ↪ $z_{0.99} \approx 2.326$, $z_{0.05} \approx 1.645$, $z_{0.9} \approx 1.282$
- ↪ Für $\alpha = 0.01$: $-\infty < \mu \leq 300 + 2.326 \cdot \frac{100}{\sqrt{144}} = 319.38$
- ↪ Für $\alpha = 0.05$: $-\infty < \mu \leq 300 + 1.645 \cdot \frac{100}{\sqrt{144}} = 313.71$
- ↪ Für $\alpha = 0.1$: $-\infty < \mu \leq 300 + 1.282 \cdot \frac{100}{\sqrt{144}} = 310.68$
- ↪ Beachte wiederum: Je größer α, desto kleiner die vorgegebene Überdeckungs-WS, desto kleiner das KI.

$X_1, \ldots, X_n \overset{u.i.v}{\sim} \mathcal{N}(\mu, \sigma^2)$. Wir suchen erneut Konfidenzintervallschätzung für μ.

Zweiter Fall: σ ist unbekannt

Konstruktion einer Zufallsvariablen, die den wahren Parameter enthält, mit bekannter Verteilung, die unabhängig ist vom unbekannten Parameter.

$$(\star\star) \quad \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n} \sim t_{n-1} \text{ mit } S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

- ↪ **Beachte**: Der Ausdruck unterscheidet sich von jenem bei bekannter Varianz σ^2 nur durch die Ersetzung eben jener Varianz durch die Schätzung S^2.
- ↪ Das Konfidenzintervall ist dann von der Form

$$[G_u \,;\, G_o] \quad = \quad [T_n - t \cdot S/\sqrt{n} \,;\, T_n + t \cdot S/\sqrt{n}]$$

  mit einem geeigneten Wert $t > 0$ und zufallsabhängiger Breite $B = 2tS/\sqrt{n}$.
- ↪ Für die Bestimmung von t muss die Verteilung $(\star\star)$ verwendet werden, dazu zwei Exkurse:

Exkurs 1: χ^2-Verteilung

Seien $X_1, \ldots, X_n \overset{u.i.v.}{\sim} \mathcal{N}(0,1)$. Dann heißt die Verteilung der Zufallsvariablen

$$Z := X_1^2 + \ldots + X_n^2$$

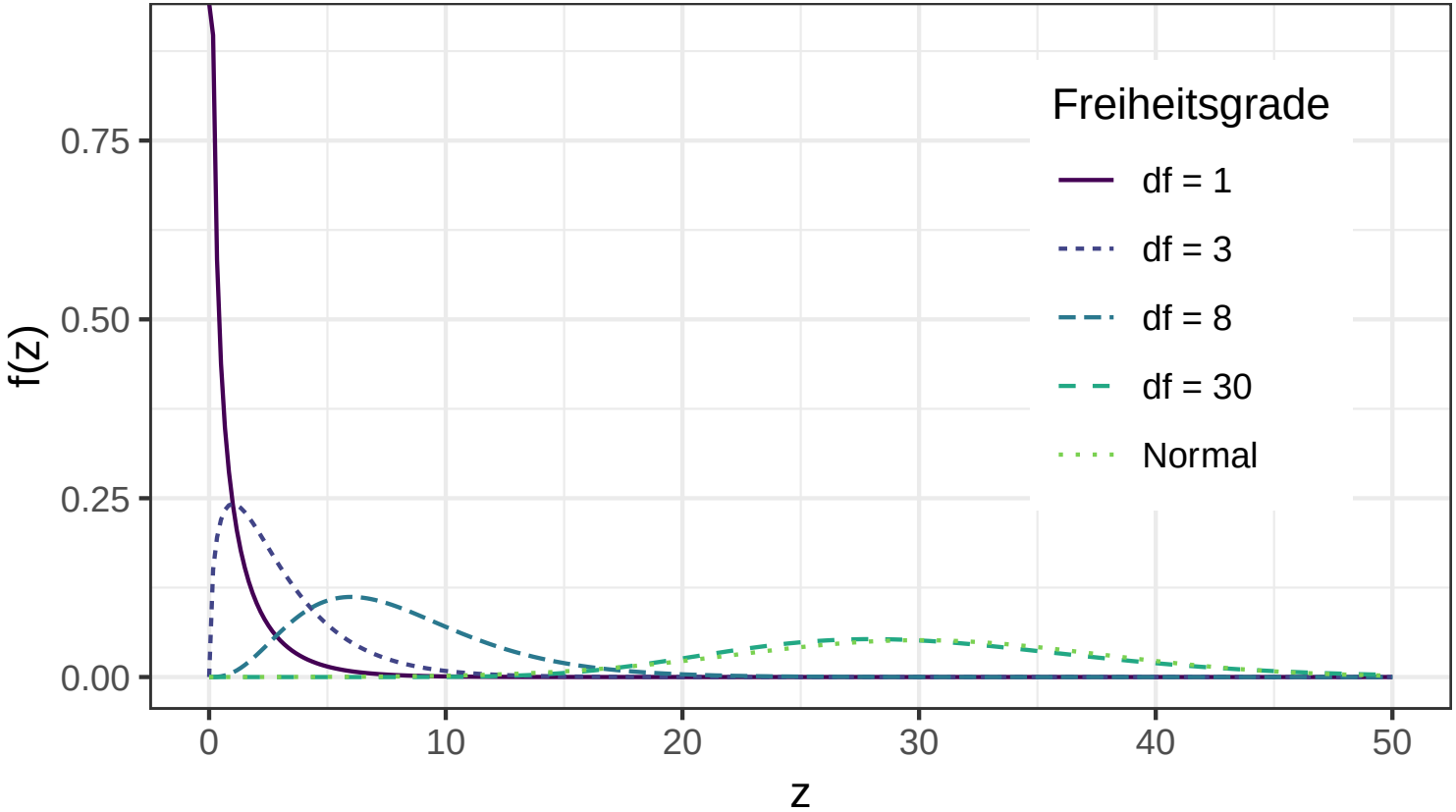
Chi-Quadrat Verteilung mit n Freiheitsgraden, Symbol χ_n^2.

- $\hookrightarrow$ Ihre Dichte lässt sich durch Faltung bestimmen
  $$f_n(z) = \begin{cases} \frac{z^{n/2-1} \cdot e^{-z/2}}{2^{n/2} \cdot \Gamma(n/2)} & \text{, falls } z > 0 \\ 0 & \text{, falls } z \leq 0 \end{cases} \quad \text{mit } \Gamma(t) = \int_0^\infty x^{t-1} \cdot e^{-x} dx$$
- $\hookrightarrow$ Momente: $E(Z) = n$ und $var(Z) = 2n$
- $\hookrightarrow$ Dichte, Verteilungs-, Quantilfunktion mit R: `dchisq()`, `pchisq()`, `qchisq()`

Für $X_1, \ldots, X_n \overset{u.i.v.}{\sim} \mathcal{N}(\mu, \sigma^2)$ gilt: $T = \sum_{i=1}^n (X_i - \bar{X})^2 / \sigma^2$ ist χ_{n-1}^2-verteilt.

Heuristik für $\mu = 0, \sigma = 1$: $Z = \underbrace{\sum_{i=1}^n X_i^2}_{\sim \chi_n^2} = \underbrace{\sum_{i=1}^n (X_i - \bar{X})^2}_{\sim \chi_{n-1}^2} + \underbrace{n\bar{X}^2}_{\sim \chi_1^2} = T + (\underbrace{\sqrt{n}\bar{X}}_{\sim \mathcal{N}(0,1)})^2$

Dichte der χ^2-Verteilung - Visualisierung



$\hookrightarrow$ für kleine n linkssteil

$\hookrightarrow$ für $n > 100$ Normalverteilungsapproximation: $Z \overset{approx.}{\sim} \mathcal{N}(n, 2n)$

Exkurs 2: t-Verteilung

Seien X und Z unabhängig, $X \sim \mathcal{N}(0,1), Z \sim \chi^2_n$. Die Verteilung der ZV

$$T := \frac{X}{\sqrt{Z/n}}$$

heißt t-**Verteilung mit** n **Freiheitsgraden**, Symbol: t_n

$\hookrightarrow$ Die Dichte berechnet sich z.B. mit dem Dichtetransformationssatz aus der gemeinsamen Dichte der Zufallsvariablen $X\sqrt{Z}$ und $X/\sqrt{Z}$

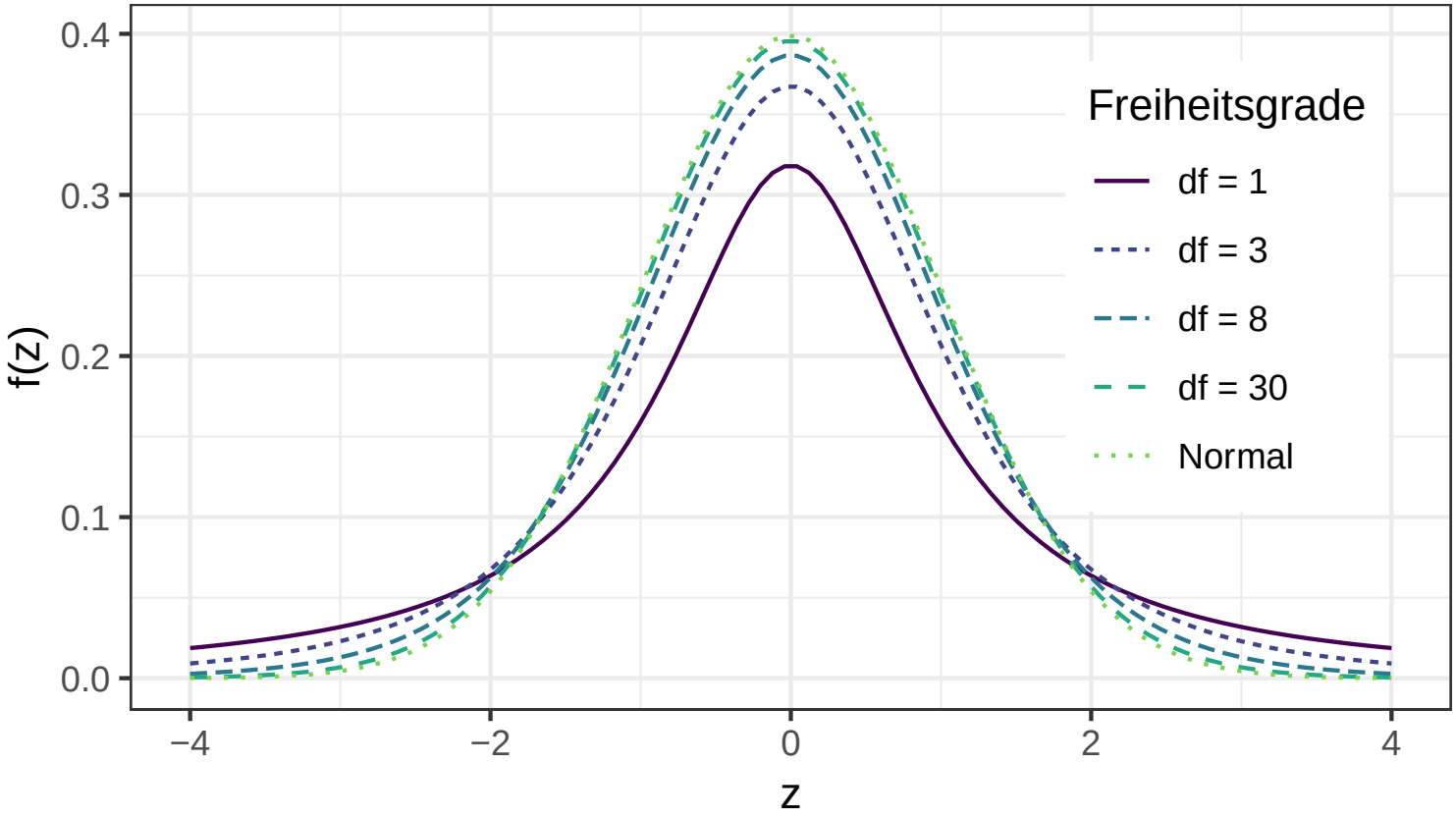
$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n \cdot \pi} \cdot \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

$\hookrightarrow$ $E(T) = 0$ und $var(T) = \frac{n}{n-2}$ (für $n \geq 3$)

$\hookrightarrow$ Dichte, Verteilungs- und Quantilfunktion mit R: `dt( ...)`, `pt(...)`, `qt(...)`

Für $X_1, \ldots, X_n \overset{u.i.v.}{\sim} \mathcal{N}(\mu, \sigma^2)$ sind $\bar{X}$ und S st.u. und es gilt $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Dichte der t-Verteilung - Visualisierung



$\hookrightarrow$ schwerere Ränder im Vergleich zur Normalverteilung

$\hookrightarrow$ für $n \to \infty$ Konvergenz zur Normalverteilung, ab $n = 30$ Approximation gut

$X_1, \ldots, X_n \overset{u.i.v}{\sim} \mathcal{N}(\mu, \sigma^2)$. Wir suchen erneut Konfidenzintervallschätzung für μ, diesmal aber für unbekanntes σ
Aufgrund der Exkurse gilt:

$$\begin{aligned}
1-\alpha &= P\left(-t_{1-\alpha/2;\,n-1} \;\leq\; \frac{\bar{X}-\mu}{S/\sqrt{n}} \;\leq\; t_{1-\alpha/2;\,n-1}\right) \\
&= P\left(-t_{1-\alpha/2;\,n-1}\cdot\frac{S}{\sqrt{n}} \;\leq\; \bar{X}-\mu \;\leq\; t_{1-\alpha/2;\,n-1}\cdot\frac{S}{\sqrt{n}}\right) \\
&= P\left(\bar{X}-t_{1-\alpha/2;\,n-1}\cdot\frac{S}{\sqrt{n}} \;\leq\; \mu \;\leq\; \bar{X}+t_{1-\alpha/2;\,n-1}\cdot\frac{S}{\sqrt{n}}\right)
\end{aligned}$$

Somit lautet das zugehörige Konfidenzintervall:

$$\left[\bar{X}-t_{1-\alpha/2;\,n-1}\cdot\frac{S}{\sqrt{n}}, \bar{X}+t_{1-\alpha/2;\,n-1}\cdot\frac{S}{\sqrt{n}}\right]$$

mit $b_{\mathrm{KI}} = 2\cdot t_{1-\alpha/2;\,n-1}\cdot\frac{S}{\sqrt{n}}$

Approximative Konfidenzintervalle für μ

Bei **beliebiger Verteilung der Grundgesamtheit** lassen sich die Verteilungen von

$$(\star) \quad \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} \text{ und } (\star\star) \quad \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

für großes n („Faustregel“ $n > 30$) durch die Standardnormalverteilung approximieren!
Wir erhalten für bekanntes σ^2

$$\left[\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

und für unbekanntes σ^2

$$\left[\bar{X} - z_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}}\right]$$

Übung: (Anteilswertschätzung) $X_1, \ldots, X_n$ seien u.i.v. $Bin(1, p)$-verteilt. Bestimmen Sie ein approximatives zweiseitiges $(1-\alpha)$-Konfidenzintervall für p auf Grundlage des Schätzers $\bar{X}$.

1. Stellen Sie das KI mit dem Ansatz ($\star\star$) der vorigen Folie dar.

2. Führen Sie S (bzw. zunächst $(n-1)S^2$) auf $\bar{X}$ zurück und vereinfachen Sie die Ausdrücke im KI so weit wie möglich. Hinweis: $X_i = X_i^2$ (warum?)

5.4 Konfidenzintervall für die Varianz

- ↪ Seien $X_1, \ldots, X_n \overset{u.i.v}{\sim} \mathcal{N}(\mu, \sigma^2)$. Wir suchen jetzt eine Konfidenzintervallschätzung für σ^2, dabei sei μ unbekannt.
- ↪ Nach Exkurs 1 gilt: $\frac{n-1}{\sigma^2} \cdot S^2 \sim \chi^2_{n-1}$, d.h. mit dem α-Quantil q_α zu χ^2_{n-1}

$$
\begin{aligned}
1-\alpha &= P\left(q_{\alpha/2} \leq \frac{n-1}{\sigma^2} \cdot S^2 \leq q_{1-\alpha/2}\right) \\
&= P\left(\frac{q_{\alpha/2}}{(n-1)\cdot S^2} \leq \frac{1}{\sigma^2} \leq \frac{q_{1-\alpha/2}}{(n-1)\cdot S^2}\right) \\
&= P\left(\frac{(n-1)\cdot S^2}{q_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)\cdot S^2}{q_{\alpha/2}}\right)
\end{aligned}
$$

Ein zweiseitiges $(1-\alpha)$-KI für σ^2 ist $\left[\frac{(n-1)\cdot S^2}{q_{1-\alpha/2}} \; ; \; \frac{(n-1)\cdot S^2}{q_{\alpha/2}}\right]$

5.5 Zusammenfassung

- ↪ Schätzer T_n für eine unbekannte Kennzahl θ einer Grundgesamtheit treffen (fast) nie den tatsächlichen Wert von θ, allenfalls approximativ oder unverfälscht mit geringem Standardfehler.
- ↪ $(1-\alpha)$-Konfidenzintervalle für θ überdecken den tatsächlichen Wert von θ mit (approximativer) Wahrscheinlichkeit $1-\alpha$, geben also eine „statistische“ Genauigkeit der (ursprünglichen) Schätzung.
- ↪ Zweiseitige Konfidenzintervalle für θ ergeben sich oft aus dem Ansatz

$$P(-t \leq (T_n - \theta)/A_n \leq t) \stackrel{!}{=} 1-\alpha$$

  wobei t approximativ Quantil einer Standardverteilung ist, und Auflösung der beiden Ungleichungen nach θ.
- ↪ Die Breite eines $(1-\alpha)$-KI ist abhängig von α und n, i.d.R. ist sie auch stichprobenabhängig.
- ↪ Ausblick: Es bestehen enge Zusammenhänge zwischen KI und statistischen Tests (siehe nächstes Kapitel).

6. Hypothesentests

- Testen von Hypothesen
  - Nullhypothese vs. Alternativhypothese
  - Fehler und Signifikanzniveau eines Tests
  - Teststatistik V und ihre Verteilung F_V
  - Testentscheidung
- Einstichproben-Erwartungswert-Tests
- Der p-value
- Einstichproben-Anteilswert-Tests
- Einstichproben-Varianztests
- Gütefunktionen von Parametertests

$\hookrightarrow$ Bisher behandelte statistische Verfahren

- Punktschätzer: Aus Stichprobe wird „gute“ Schätzung für den Wert einer Kennzahl / eines Parameters einer Grundgesamtheit gewonnen.

Covid-19: theor. Reproduktionsfaktor ρ in Bevölkerung unbekannt, bestimme eine Schätzung R anhand der Infektionszahlen (zufallsabhängig, tagweise Änderung)

- Intervallschätzer: Zu „gutem“ Schätzer und Überdeckungs-WS $1-\alpha$ werden „Toleranzen nach unten und oben“ angegeben.

Covid-19: Angabe eines „möglichst schmalen“ KI $[R-\Delta, R+\Delta]$ mit 99% Überdeckungswahrscheinlichkeit für ρ.

$\hookrightarrow$ Jetzt Fragestellung: Erfüllt die Grundgesamtheit eine bestimmte Hypothese?

Covid-19: Liegt der (theor.) Reproduktionsfaktor unter 1 (unter 0,75,...)?

Als Verfahren zur Beantwortung solcher Fragen werden statistische Tests eingesetzt. Die Entscheidungen können - mit Bezug auf den wahren Parameter - falsch sein, wobei zwei mögliche Fehler (1./2. Art) auftreten können. Ein Test sollte wenigstens den Fehler 1. Art „unter Kontrolle haben“.

6.1 Testen von Hypothesen

- ↪ Anhand Stichproben: Prüfung von Annahmen über eine Grundgesamtheit
- ↪ Formulierung der Annahmen mit Bezug auf
  - □ Verteilungsparameter
  - □ Verteilungskennzahlen (Erwartungswert, ...)

  als komplementäres Hypothesenpaar: Nullhypothese H_0/Gegenhypothese H_1.
- ↪ H_0 wird verworfen (statistisch widerlegt), wenn das Stichprobenergebnis „deutlich gegen H_0 spricht“.
  - □ Verwerfung bedeutet nicht, dass H_0 falsch bzw. H_1 wahr ist.
  - □ Nichtverwerfung bedeutet nicht, dass H_0 wahr bzw. H_1 falsch ist.
- ↪ Verfahren werden **Hypothesentests** oder **statistische Tests** genannt.
- ↪ Unterschieden werden Parametertests und Verteilungstests.
- ↪ Tests sind nach einem einheitlichen Schema aufgebaut.

Bsp.: Verkauf von Kartoffeln in 10kg-Säcken. Wird das Sollgewicht eingehalten?
Stichprobe: $\bar{x} = 9.85$ ($n = 100$). Modell: $X_1, \ldots, X_{100}$ u.i.v. $\sim \mathcal{N}(\mu, 4)$, μ unbek.

Generelles Testschema

1. Festlegung von
   - ☐ **Nullhypothese** H_0 und **Gegen- bzw. Alternativhypothese** H_1

     $H_0 : \mu = 10$ vs. $H_1 : \mu \neq 10$
   - ☐ **Signifikanzniveau** α (= WS H_0 irrtümlich abzulehnen)

     $\alpha = 0.05$ (H_0 wird mit WS 5% irrtümlich verworfen)
2. Herleitung einer **Teststatistik** $V = V(X_1, \ldots, X_n)$ und Verteilung F_V für H_0

   $V = \sqrt{100} \cdot \frac{\bar{X}-10}{\sqrt{4}} \quad \sim \quad F_V = \mathcal{N}(0,1)$ für $\mu = 10$
3. Aufstellen einer **Entscheidungsregel** anhand der statistischen Verteilung F_V und Festlegung des **kritischen Bereiches** K mit Ablehnung von H_0 für $V \in K$.

   H_0 verwerfen, wenn $|V| > z_{1-\alpha/2} = 1.96$ (0.975-Quantil von $\mathcal{N}(0,1)$)
4. **Entscheidung** anhand der berechneten Teststatistik $V = v$

   Aus $\bar{x} = 9.85$ folgt $|v| = \left|10 \cdot \frac{9.85-10}{2}\right| = 0.75$

   $|v| \leq 1.96 \Rightarrow H_0$ kann nicht verworfen werden.

Hypothesen (über die Grundgesamtheit)

↪ **Verteilungshypothesen:** Annahme eines bestimmten Verteilungstyps (oder einer speziellen Verteilung), auch: einer bestimmten Verteilungseigenschaft.

Bsp.: Normalverteilungsannahme $H_0 : F = \mathcal{N}$ vs. $H_1 : F \neq \mathcal{N}$

↪ **Parameterhypothesen:** beziehen sich auf einen bestimmten Parameter

$H_0 : \theta = \theta_0$	vs.	$H_1 : \theta \neq \theta_0$	Punkthypothese, zweiseitig
$H_0 : \theta \geq \theta_0$	vs.	$H_1 : \theta < \theta_0$	Bereichshypothese, rechtsseitig
$H_0 : \theta \leq \theta_0$	vs.	$H_1 : \theta > \theta_0$	Bereichshypothese, linksseitig

Beispiele für θ: Anteilswert, Erwartungswert, Varianz, ...

↪ **Beachte:**

- □ Sprechen in den Daten „genügend Indizien“ gegen H_0 , so wird H_0 verworfen und man entscheidet sich aktiv für H_1 („H_0 wird verworfen“)
- □ H_0 kann *nie (aktiv) akzeptiert*, nur mangels genügend Informationen nicht wiederlegt werden („H_0 wird beibehalten bzw. nicht verworfen“)
- □ Formulierung des „intendierten“ Sachverhaltes (sofern möglich) als H_1

Fehler und Signifikanzniveau eines Tests

↪ **Fehler 1. und 2. Art:**

	Testentscheidung	
Wahrer Zustand	Nichtablehnung von H_0	Ablehnung von H_0
H_0 korrekt	richtig	α-Fehler (Fehler 1. Art)
H_0 falsch	β-Fehler (Fehler 2. Art)	richtig

↪ **Signifikanzniveau α:**

- vorgegebener Maximalwert für die Wahrscheinlichkeit des Fehlers 1. Art (d.h. W.-keit H_0 abzulehnen, obwohl H_0 korrekt ist)
- muss vor der Anwendung des Tests festgelegt werden
- üblich sind $\alpha = 0.01$, $\alpha = 0.05$, $\alpha = 0.1$

Testbestimmung: Konfliktäre Ziele

- ↪ Idealerweise würde man die WS des α- und β-Fehler simultan minimieren
- ↪ Problem: $\alpha \downarrow \quad \Rightarrow \quad \beta \uparrow$
- ↪ Wie groß sind die WS der Fehler 1./2. Art, wenn H_0 stets abgelehnt wird?
- ↪ Wie groß sind die WS der Fehler 1./2. Art, wenn H_0 stets angenommen wird?
- ↪ Ausweg: Wähle für fixes α den Test mit niedrigstem β (mächtigster Test).

Teststatistik V und ihre Verteilung F_V

- ↪ Aggregation der Stichprobenwerte zu einer Teststatistik $V = V(X)$, anhand derer die Testentscheidung (H_0 beibehalten oder verwerfen) getroffen wird
- ↪ V ist charakteristisches Merkmal eines statistischen Tests und ergibt sich oft
  - aus einer Schätzfunktion bei Parameterhypothesen
  - aus Gegenüberstellung von empirischen und theoretischen Kennzahlen.
- ↪ Als Aggregation von Zufallsvariablen ist die Teststatistik ebenfalls eine Zufallsvariable (und folgt einer Verteilung F_V)

Beispiel

- Nullhypothese H_0: $\mu = \mu_0$
- Stichprobe: $X_1, \ldots, X_n$ mit $X_i \sim \mathcal{N}(\mu_0, \sigma^2)$
- Teststatistik: $V = \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\sigma} \sim \mathcal{N}(0, 1)$

  $\Rightarrow$ zugehörige Verteilung der Teststatistik V: $F_V = \mathcal{N}(0, 1)$

Teststatistik V und ihre Verteilung F_V

- $\hookrightarrow$ Über die statistische Verteilung F_V wird die Einhaltung des Signifikanzniveaus α geprüft.
- $\hookrightarrow$ Verteilung F_V hängt von der Verteilung der Grundgesamtheit, insbesondere von der Korrektheit von H_0, ab
- $\hookrightarrow$ Bei Korrektheit von H_0 muss F_V (mindestens approximativ) bestimmbar sein.

Entscheidungsregel (I)

- $\hookrightarrow$ anhand einer konkreten Stichprobe x wird die Realisation $v = v(x)$ bestimmt
- $\hookrightarrow$ Bestimmung eines **kritischen Bereiches** K: Ablehnung von H_0 für $v \in K$

  H_0 wird abgelehnt, falls $|v| > z_{1-\alpha/2}$, d.h. $K = (-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$

  - $\hookrightarrow$ Sicherstellung, dass $P(\text{Fehler 1. Art}) \leq \alpha$.
  - $\hookrightarrow$ endgültige Festlegung von K so, dass $P(\text{Fehler 2. Art})$ minimal ist.
  - $\hookrightarrow$ oft (z.B. stetiges F_V) realisiert durch $P(\text{Fehler 1. Art}) = \alpha$ für „ungünstigsten Fall der Hypothese“.

Faustregel zur Festlegung des Hypothesenpaars

Statistische Tests können nur zur Widerlegung von H_0 verwendet werden – nicht zur Bestätigung.

$\rightsquigarrow$ Die zu bestätigende Hypothese muss als Gegenhypothese formuliert werden.

Entscheidungsregel (II)

$v \in K$:

$\hookrightarrow$ Stichprobe steht in signifikantem Widerspruch zu H_0

$\hookrightarrow$ H_0 **kann abgelehnt werden**

$\hookrightarrow$ H_1 ist statistisch bestätigt mit Signifikanzniveau α.

$v \notin K$:

$\hookrightarrow$ Stichprobe steht nicht in signifikantem Widerspruch zu H_0

$\hookrightarrow$ H_0 **kann nicht abgelehnt werden** anhand der konkreten Stichprobe

$\hookrightarrow$ H_0 ist dadurch nicht bestätigt!

Klassifizierungen von Tests anhand der Anzahl einbezogener Stichproben

Einstichprobentest

- ↪ Grundlage bildet eine u.i.v. Stichprobe $X_1, \ldots, X_n$ zu einer Grundgesamtheit
- ↪ Bsp: Tests für den Anteilswert p oder den Erwartungswert μ

Zweistichprobentest

- ↪ für den Vergleich verschiedener Grundgesamtheiten

  Unterscheidet sich das mittlere Einkommen der 20- bis 30-Jährigen signifikant vom mittleren Einkommen der 30- bis 40-Jährigen?
- ↪ Stichproben $X_{11}, \ldots, X_{1n_1}$ und $X_{21}, \ldots, X_{2n_2}$, gezogen aus 2 Grundgesamtheiten
- ↪ Annahme: unabhängige Stichproben
- ↪ Bsp: Tests für die Differenz zweier Erwartungswerte oder zweier Anteilswerte

Spezialfall: Zweistichprobentest für verbundene Stichproben

- ↪ Ziehen einer zweidimensionalen Stichprobe aus einer einzelnen Grundgesamtheit.
- ↪ pro Merkmalsträger werden mehrere Merkmale erhoben
- ↪ Bsp: Differenzentests für das arithmetische Mittel oder Korrelations- oder Kontingenztest

I) Besteht eine Abhängigkeit in den Leistungen der Studierenden in den Fächern Mathematik und Physik?

II) Blutdruck von n Patienten vor und nach der Einnahme eines blutdrucksenkenden Medikamentes. Hat das Medikament einen Einfluss?

III) Einkommen von n Erwerbslosen vor und nach einer Arbeitsbeschaffungsmaßnahme. Hat diese einen Einfluss?

Gemeinsame Struktur aller behandelten Tests

Abhängig von Hypothese (und Signifikanzniveau α) haben sämtliche Tests folgende Struktur mit spezifischer Teststatistik $V = V(X)$:

	Nullhypothese H_0	**kritischer Bereich**	p-**value**
a)	zweiseitig: $\theta = \theta_0$	$V \notin \left[q_{\alpha/2}; q_{1-\alpha/2}\right]$	$2 \cdot \min(F(v), 1 - F(v))$
	F symmetrisch	$\lvert V \rvert > q_{1-\alpha/2}$	$2 \cdot (1 - F(\lvert v \rvert)$
b)	rechtsseitig: $\theta \geq \theta_0$	$V < q_\alpha$	$F(v)$
c)	linksseitig: $\theta \leq \theta_0$	$V > q_{1-\alpha}$	$1 - F(v)$

$\hookrightarrow$ Dabei ist q_α das α-Quantil der Stichprobenverteilung $F = F_V$ von V

$\hookrightarrow$ F und q_α werden i.d.R. auch bei b) und c) für $\theta = \theta_0$ bestimmt, weil die WS eines Fehlers 1. Art dann besonders hoch ist.

$\hookrightarrow$ p-value: für alternative Darstellung der Test-Entscheidung (s.u.)

6.2 Einstichproben-Erwartungswert-Tests

Konstruktionsschema (s.o.)

Zunächst werden die Modellannahmen weiter spezifiziert, danach:

1. Aufstellen der **Hypothesen** H_0 und H_1, sowie des Signifikanzniveaus α.
2. Bestimmung der **Teststatistik** $V = V(X_1, \ldots, X_n)$ mitsamt zugehöriger Stichprobenverteilung F_V
3. Aufstellen der **Entscheidungsregel** bzgl. der Ablehnung von H_0

Anhand der konkreten Stichprobe $x_1, \ldots, x_n$ wird dann mit dem berechneten Wert $v(x_1, \ldots, x_n)$ der Teststatistik $V(X_1, \ldots, X_n)$ die Entscheidung (H_0 beibehalten oder verwerfen) getroffen.

Einstichprobentests für den Erwartungswert

- ↪ Stichprobe $X_1, \ldots, X_n$ von u.i.v. ZVen mit existierendem (unbekanntem) Erwartungswert μ und existierender (bekannter/unbekannter) Varianz σ^2
- ↪ Die Vorgehensweise ist jeweils ähnlich, unterscheidet sich im Detail nach
  - □ Verteilung der Grundgesamtheit: $X_i \sim \mathcal{N}(\mu, \sigma^2)$ bzw. $\mathcal{L}(X)$ unspezifiziert
  - □ Varianz: bekannt bzw. (in der Praxis meist) unbekannt

1. Hypothesen aufstellen

Wähle aus den folgenden Möglichkeiten das zum Fall passende Hypothesenpaar:

a) $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$

b) $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$

c) $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$

Einkommen von niedergelassenen Ärzten

Köln – *Niedergelassene Ärzte haben die Erträge ihrer Praxen zwischen 2007 und 2011 deutlich steigern können. Das geht aus Erhebungen des Statistischen Bundesamts hervor. Danach stieg der Reinertrag je Arztpraxis, der dem Überschuss nach Abzug der Praxisaufwendungen vor Steuern und sonstigen Abgaben entspricht, in diesen vier Jahren um 21 Prozent auf durchschnittlich 234.000 Euro. Nach Angaben des Statistischen Bundesamts gab es bei den Arzteinkommen je nach Fachrichtung große Unterschiede. Allgemeinarztpraxen verbuchten 2011 Reinerträge von durchschnittlich 181.000 Euro, Kinderarztpraxen brachten es auf 191.000 Euro, Orthopädie-Praxen auf 293.000 Euro und Praxen für Haut- und Geschlechtskrankheiten auf 249.000 Euro.* *Quelle:* `https://www.aerzteblatt.de/nachrichten/55525/`

Über alle Sparten hinweg beträgt demnach das mittlere Einkommen[a] 181 T€ oder mehr. Gilt dies auch für Ärzte in unserer Region? D.h. Hypothesenpaar:

$$H_0 : \mu \geq 181 \quad \text{vs.} \quad H_1 : \mu < 181$$

[a] Dass die Angaben (näherungsweise) Grundgesamtheitswerte sind, ist allerdings kritisch, da scheinbar nur freiwillige Angaben von ca. 4000 (6%) der deutschen Arztpraxen zugrunde lagen.

Festlegung: „Grenzfall“ $\mu = \mu_0$ (hier $\mu = 181$) gehört zu H_0, nicht zu H_1.

Fiktives Datenbeispiel: In einer Stichprobe von $n = 200$ Arztpraxen unserer Region wurde ein Durchschnittseinkommen von $\bar{X} = 168$ T€ ermittelt.

Durchschnittliche Praktikumsgehälter, Clevis Praktikantenspiegel 2016

Öffentlicher Sektor	1194	Pharma	1280
TK, IT und Internet	1216	Baugewerbe, -industrie	1321
Transport, Logistik	1128	Personaldienstleistungen	1334
Medien, Unterhaltung	1231	Konsum-, Gebrauchsgüter	1380
Finanzen, Banken, Versicherungen	1260	Consulting, WP & Recht	1383

Werte gerundet, Quelle: `http://www.clevis.de/CLEVIS_Praktikantenspiegel/2016.pdf`

Das Studienergebnis soll für Praktikanten in der **Versicherungsbranche** geprüft werden, d.h. Hypothesen[a]:

$$H_0 : \mu = 1260 \quad \text{vs.} \quad H_1 : \mu \neq 1260$$

[a] Auch hier ist die Annahme, dass $\mu = 1260$ eine Angabe zur Grundgesamtheit ist, kritisch zu beurteilen.

Fiktives Datenbeispiel: Von $n = 40$ Studierenden, die 2015 ein Praktikum in der Versicherungsbranche absolvierten, wurde das Praktikumsgehalt erfragt:
1346,1187,1353,1347,1261,1066,1127,1191,1219,1460,1296,1140,1105,1191,1190,
1179,1245,1131,1264,1096,1198,1258,1233,1300,1214,1270,1329,1151,1092,1225,
1196,1166,1177,1155,1293,1335,1319,1177,1344,1192 Mittelwert: 1225,45 €

2. Bestimmung der Teststatistik V und zugehöriger Verteilung F_V

V und F_V hängen davon ab, ob σ^2 bekannt ist und welches WS-Modell für die u.i.v.-Stichprobe vorliegt.

Fall	σ^2	V	Modell für X_i	Verteilung von V	Testbezeichnung
i)	bekannt	$\sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\sigma}$	Normalvtlg.	$\mathcal{N}(0,1)$	Gaußtest
ii)			beliebig	approx. $\mathcal{N}(0,1)$	approx. Gaußtest
iii)	unbekannt	$\sqrt{n} \cdot \frac{\bar{X} - \mu_0}{S}$	Normalvtlg.	t_{n-1}	t-Test
iv)			beliebig	approx. $\mathcal{N}(0,1)$	approx. Gaußtest

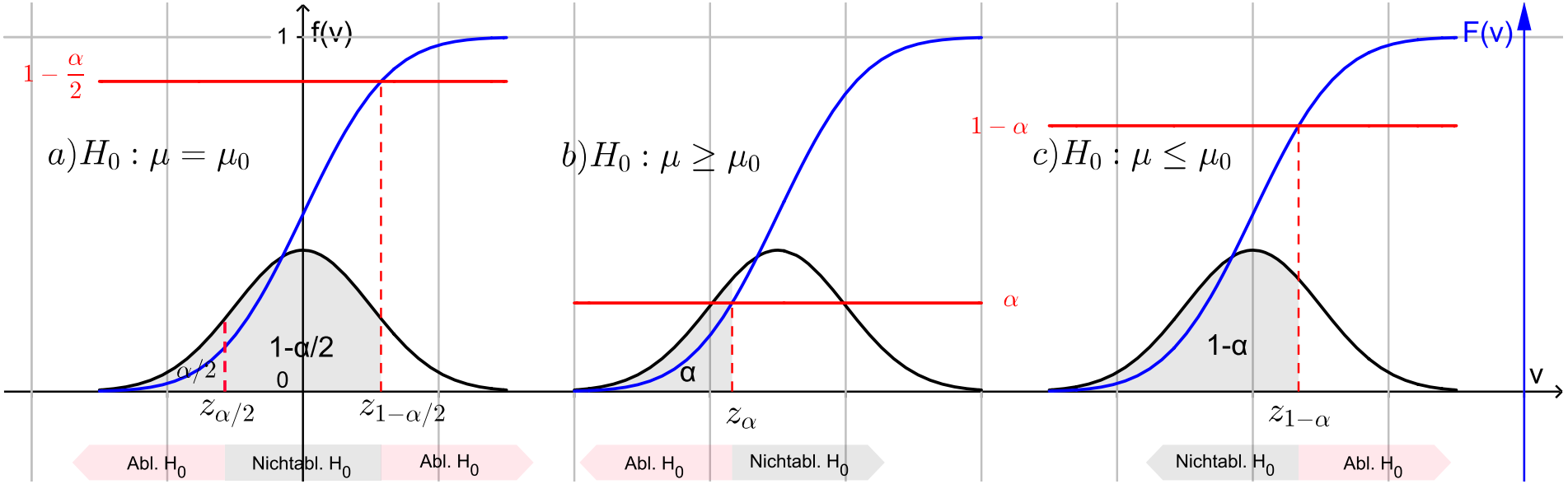
$\hookrightarrow$ Dabei $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$\hookrightarrow$ in R: Gaußtest: `compositions::Gauss.test(...)`, t-Test: `ttest(...)`

Ärzteeinkommen ($n = 200, \mu_0 = 181, \bar{X} = 168$): Bestimmen Sie den Wert v der Teststatistik $V = \sqrt{n} \cdot \frac{\bar{X}-\mu_0}{\sigma}$ die Verteilung von V für $\mu = \mu_0$ ($\sigma = 100$ bekannt)

Praktikumsbeispiel ($n = 40, \mu_0 = 1260, \bar{X} = 1225,45$):
Bestimmen Sie $V = \sqrt{n} \cdot \frac{\bar{X}-\mu_0}{\sigma}$ (Ann.: $\sigma = 120$ bekannt) bzw. $V = \sqrt{n} \cdot \frac{\bar{X}-\mu_0}{S}$ (Ann.: σ unbekannt) und jeweils die Verteilung von V unter $\mu = \mu_0$.

Bezeichne z_α das α-Quantil von $\mathcal{N}(0,1)$. In R: `qnorm(alpha)`



https://ggbm.at/u3kbgjby

3. Entscheidungsregel für die Fälle i), ii) und iv)

a) $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$: Ablehnung von H_0, falls $|v| > z_{1-\alpha/2}$

b) $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$: Ablehnung von H_0, falls $v < z_\alpha$

c) $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$: Ablehnung von H_0, falls $v > z_{1-\alpha}$

Durchschnittseinkommen von niedergelassenen Ärzten

Stichprobe von $n = 200$ Arztpraxen, $\bar{X} = 168$T€. Laut einer statistischen Bundesamt liegt das „wahre Durchschnittseinkommen“ bei mindestens 181T€.
Wird das „wahre“ Durchschnittseinkommen (bei einem Signifikanzniveau von 5%) durch die Stichprobe gestützt?
Annahme: $X \sim \mathcal{N}(\mu, \sigma^2), \sigma = 100$ bekannt.

- $\hookrightarrow$ Hypothesen: $H_0 : \mu \geq 181$, vs. $H_1 : \mu < 181$
- $\hookrightarrow$ Teststatistik: $v = \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{\sigma} \approx -1.83$

Ermitteln Sie die Testentscheidung.

Tabelle der α-Quantile der Standardnormalverteilung:

α	0.001	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995	0.999
z_α	-3.090	-2.576	-2.326	-1.960	-1.645	-1.282	1.282	1.645	1.960	2.326	2.576	3.090

Durchschnittseinkommen von Praktikanten

Stichprobe von $n = 40$ Studenten mit durchschnittlichem Einkommen $\bar{X} = 1225.45$.
Laut einer Studie liegt das „wahre Durchschnittseinkommen“ bei 1260 €.
Wird die Angabe der Studie (bei einem Signifikanzniveau von $\alpha = 5\%$) durch die Stichprobe gestützt?

Annahme: $X \sim \mathcal{N}(\mu, \sigma^2), \sigma = 120$ als bekannt angenommen.

Wie lautet, ausgehend von Ihren bisherigen Berechnungen, Ihre Testentscheidung?

Tabelle der α-Quantile der Standardnormalverteilung:

α	0.001	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995	0.999
z_α	-3.090	-2.576	-2.326	-1.960	-1.645	-1.282	1.282	1.645	1.960	2.326	2.576	3.090

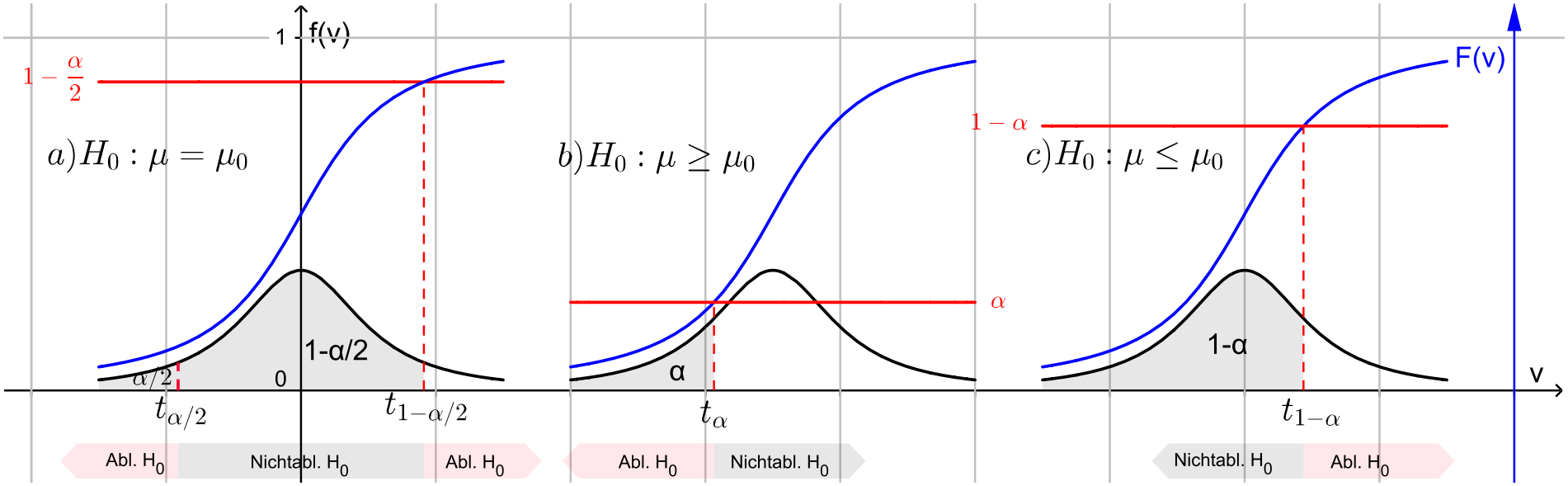
Bezeichne $t_\alpha = t_{\alpha;\, n-1}$ das α-Quantil von t_{n-1}. In R: `qt(alpha, df = n - 1)`

3. Entscheidungsregel bei iii), d.h. Normalvtlg. mit unbekanntem σ^2

a) $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$: Ablehnung von H_0, falls $|v| > t_{1-\alpha/2}$

b) $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$: Ablehnung von H_0, falls $v < t_\alpha$

c) $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$: Ablehnung von H_0, falls $v > t_{1-\alpha}$



Für $n > 30$: Approximation der t-Verteilung durch die Standardnormalverteilung.

Durchschnittseinkommen von Praktikanten

Stichprobe von $n = 40$ Studenten mit durchschnittlichem Einkommen $\bar{X} = 1225.45$.
Laut einer Studie liegt das „wahre Durchschnittseinkommen“ bei 1260 €.
Wird die Angabe der Studie (bei einem Signifikanzniveau von $\alpha = 5\%$) durch die Stichprobe gestützt?

Annahme: $X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 unbekannt (wird durch $S^2 = 87.23354^2$ geschätzt).

Wie lautet, ausgehend von Ihren bisherigen Berechnungen, Ihre Testentscheidung?

(Tabelle der t-Quantile: siehe nächste Folie)

Tabelle: Quantile von Standardnormal- und t-Verteilung

	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
∞	1.28	1.64	1.96	2.33	2.58	3.09	3.29
1	3.08	6.31	12.71	31.82	63.66	318.31	636.62
2	1.89	2.92	4.30	6.96	9.92	22.33	31.60
3	1.64	2.35	3.18	4.54	5.84	10.21	12.92
4	1.53	2.13	2.78	3.75	4.60	7.17	8.61
5	1.48	2.02	2.57	3.36	4.03	5.89	6.87
6	1.44	1.94	2.45	3.14	3.71	5.21	5.96
7	1.41	1.89	2.36	3.00	3.50	4.79	5.41
8	1.40	1.86	2.31	2.90	3.36	4.50	5.04
9	1.38	1.83	2.26	2.82	3.25	4.30	4.78
10	1.37	1.81	2.23	2.76	3.17	4.14	4.59
11	1.36	1.80	2.20	2.72	3.11	4.02	4.44
12	1.36	1.78	2.18	2.68	3.05	3.93	4.32
13	1.35	1.77	2.16	2.65	3.01	3.85	4.22
14	1.35	1.76	2.14	2.62	2.98	3.79	4.14
15	1.34	1.75	2.13	2.60	2.95	3.73	4.07
16	1.34	1.75	2.12	2.58	2.92	3.69	4.01
17	1.33	1.74	2.11	2.57	2.90	3.65	3.97
18	1.33	1.73	2.10	2.55	2.88	3.61	3.92
19	1.33	1.73	2.09	2.54	2.86	3.58	3.88
20	1.33	1.72	2.09	2.53	2.85	3.55	3.85
21	1.32	1.72	2.08	2.52	2.83	3.53	3.82
22	1.32	1.72	2.07	2.51	2.82	3.50	3.79
23	1.32	1.71	2.07	2.50	2.81	3.48	3.77
24	1.32	1.71	2.06	2.49	2.80	3.47	3.75
25	1.32	1.71	2.06	2.49	2.79	3.45	3.73
26	1.31	1.71	2.06	2.48	2.78	3.43	3.71
27	1.31	1.70	2.05	2.47	2.77	3.42	3.69
28	1.31	1.70	2.05	2.47	2.76	3.41	3.67
29	1.31	1.70	2.05	2.46	2.76	3.40	3.66
30	1.31	1.70	2.04	2.46	2.75	3.39	3.65
31	1.31	1.70	2.04	2.45	2.74	3.37	3.63
32	1.31	1.69	2.04	2.45	2.74	3.37	3.62
33	1.31	1.69	2.03	2.44	2.73	3.36	3.61
34	1.31	1.69	2.03	2.44	2.73	3.35	3.60
35	1.31	1.69	2.03	2.44	2.72	3.34	3.59
36	1.31	1.69	2.03	2.43	2.72	3.33	3.58
37	1.30	1.69	2.03	2.43	2.72	3.33	3.57
38	1.30	1.69	2.02	2.43	2.71	3.32	3.57
39	1.30	1.68	2.02	2.43	2.71	3.31	3.56
40	1.30	1.68	2.02	2.42	2.70	3.31	3.55

	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
41	1.30	1.68	2.02	2.42	2.70	3.30	3.54
43	1.30	1.68	2.02	2.42	2.70	3.29	3.53
44	1.30	1.68	2.02	2.41	2.69	3.29	3.53
45	1.30	1.68	2.01	2.41	2.69	3.28	3.52
46	1.30	1.68	2.01	2.41	2.69	3.28	3.51
47	1.30	1.68	2.01	2.41	2.68	3.27	3.51
49	1.30	1.68	2.01	2.40	2.68	3.27	3.50
50	1.30	1.68	2.01	2.40	2.68	3.26	3.50
51	1.30	1.68	2.01	2.40	2.68	3.26	3.49
52	1.30	1.67	2.01	2.40	2.67	3.25	3.49
53	1.30	1.67	2.01	2.40	2.67	3.25	3.48
54	1.30	1.67	2.00	2.40	2.67	3.25	3.48
56	1.30	1.67	2.00	2.39	2.67	3.24	3.47
57	1.30	1.67	2.00	2.39	2.66	3.24	3.47
59	1.30	1.67	2.00	2.39	2.66	3.23	3.46
62	1.30	1.67	2.00	2.39	2.66	3.23	3.45
63	1.30	1.67	2.00	2.39	2.66	3.22	3.45
64	1.29	1.67	2.00	2.39	2.65	3.22	3.45
66	1.29	1.67	2.00	2.38	2.65	3.22	3.44
68	1.29	1.67	2.00	2.38	2.65	3.21	3.44
69	1.29	1.67	1.99	2.38	2.65	3.21	3.44
71	1.29	1.67	1.99	2.38	2.65	3.21	3.43
73	1.29	1.67	1.99	2.38	2.64	3.21	3.43
74	1.29	1.67	1.99	2.38	2.64	3.20	3.43
76	1.29	1.67	1.99	2.38	2.64	3.20	3.42
77	1.29	1.66	1.99	2.38	2.64	3.20	3.42
79	1.29	1.66	1.99	2.37	2.64	3.20	3.42
81	1.29	1.66	1.99	2.37	2.64	3.19	3.41
85	1.29	1.66	1.99	2.37	2.63	3.19	3.41
88	1.29	1.66	1.99	2.37	2.63	3.19	3.40
89	1.29	1.66	1.99	2.37	2.63	3.18	3.40
96	1.29	1.66	1.98	2.37	2.63	3.18	3.39
99	1.29	1.66	1.98	2.36	2.63	3.17	3.39
102	1.29	1.66	1.98	2.36	2.62	3.17	3.39
106	1.29	1.66	1.98	2.36	2.62	3.17	3.38
112	1.29	1.66	1.98	2.36	2.62	3.16	3.38
118	1.29	1.66	1.98	2.36	2.62	3.16	3.37
128	1.29	1.66	1.98	2.36	2.61	3.16	3.37
129	1.29	1.66	1.98	2.36	2.61	3.15	3.37
132	1.29	1.66	1.98	2.35	2.61	3.15	3.37

	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
134	1.29	1.66	1.98	2.35	2.61	3.15	3.36
152	1.29	1.65	1.98	2.35	2.61	3.14	3.36
154	1.29	1.65	1.98	2.35	2.61	3.14	3.35
159	1.29	1.65	1.97	2.35	2.61	3.14	3.35
171	1.29	1.65	1.97	2.35	2.60	3.14	3.35
182	1.29	1.65	1.97	2.35	2.60	3.14	3.34
185	1.29	1.65	1.97	2.35	2.60	3.13	3.34
202	1.29	1.65	1.97	2.34	2.60	3.13	3.34
222	1.29	1.65	1.97	2.34	2.60	3.13	3.33
237	1.29	1.65	1.97	2.34	2.60	3.12	3.33
247	1.28	1.65	1.97	2.34	2.60	3.12	3.33
259	1.28	1.65	1.97	2.34	2.59	3.12	3.33
285	1.28	1.65	1.97	2.34	2.59	3.12	3.32
332	1.28	1.65	1.97	2.34	2.59	3.11	3.32
401	1.28	1.65	1.97	2.34	2.59	3.11	3.31
433	1.28	1.65	1.97	2.33	2.59	3.11	3.31
473	1.28	1.65	1.96	2.33	2.59	3.11	3.31
538	1.28	1.65	1.96	2.33	2.58	3.11	3.31
555	1.28	1.65	1.96	2.33	2.58	3.10	3.31
675	1.28	1.65	1.96	2.33	2.58	3.10	3.30
1712	1.28	1.65	1.96	2.33	2.58	3.09	3.30
∞	1.28	1.64	1.96	2.33	2.58	3.09	3.29

Beispiele und Nutzungshinweise:

- $t_{0.95}(15) \approx 1.75$
- Für nicht aufgeführte n nächstkleineres gelistetes n nutzen, z.B. $t_{0.9}(250) \approx t_{0.9}(247)$
- $z_{0.975} = t_{0.975}(\infty) = 1.96$
- Andere α:
  - $t_\alpha(n) = -t_{1-\alpha}(n)$
  - $z_\alpha = -z_{1-\alpha}$
- Für $n > 2000$ gilt $t_\alpha(n) \approx t_\alpha(\infty)$

6.3 Der p-value

- ↪ für praktische Durchführung: Bringe den Ablehnungsbereich auf eine einheitliche Form:

$$v \in K \Leftrightarrow p^* < \alpha$$

  Dabei ist $p^* = p^*(v)$ eine (geeignete) Transformation der Teststatistik.
- ↪ p^* wird als p-**value** (p-Wert, Signifikanz) bezeichnet.
- ↪ Der p-Wert entspricht der geringsten vorgegebenen Irrtumswahrscheinlichkeit, zu der die Nullhypothese bei vorliegender Stichprobe abgelehnt werden kann. aber (!): Der p-Wert <u>ist keine Wahrscheinlichkeit</u>, sondern zufallsabhängig.
- ↪ Vorteile:
  - □ Einheitlichkeit der Test-Darstellung,
  - □ Nutzung in statistischer Software.
  - □ Nur noch p^* wird berechnet, α kommt vom Anwender.
- ↪ Nachteile:
  - □ p^* muss i.d.R. mit Software berechnet werden.
  - □ Verschiedene Hypothesenpaare $\rightsquigarrow$ verschiedene Formeln für p^*.

- ↪ Klassische Testdarstellung: Teststatistik $V = v$, Schwellenwert q_α (Quantil)
- ↪ Für Umschreibung auf p-value: Transformation mittels Verteilungsfunktion:

Memo DuW: Verteilungsfunktion versus Quantilfunktion

- ↪ Die Verteilungsfunktion einer ZV ist die Funktion $x \mapsto F(x) = P(X \leq x)$
  Symmetrische Verteilung: $1 - F(-x) = F(x) \quad \forall x$ (z.B. $\mathcal{N}(0,1)$, t_n)
- ↪ Die Quantilfunktion einer ZV X mit VF F ist die Funktion

$$\alpha \mapsto F^{-1}(\alpha) := q_\alpha := \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$$

- ↪ Für alle $x \in \mathbb{R}, \alpha \in]0;1[$ gilt:

$$F(x) \geq \alpha \Leftrightarrow x \geq F^{-1}(\alpha)$$

- ↪ Falls F stetig, streng monoton, so ist F^{-1} Umkehrfunktion von F,d.h. es gilt

$$F(q_\alpha) = \alpha \quad \text{für alle } \alpha \in]0;1[\qquad (*)$$

- ↪ Die Quantilfunktion ist Ersatz für die Umkehrfunktion, wenn F nicht umkehrbar ist. Daher auch Bezeichung: „verallgemeinerte Umkehrfunktion“.

$p-$ *value* beim Mittelwerttest (σ bekannt) für (b) $H_0 : \mu \geq \mu_0$, $H_1 : \mu < \mu_0$

$\hookrightarrow$ Ablehnung von H_0 für $\sqrt{n}\frac{\bar{x}-\mu_0}{\sigma} = v < z_\alpha$ $\quad (z_\alpha = \Phi^{-1}(\alpha))$.

$\hookrightarrow$ Äquivalent $\Phi(v) < \Phi(z_\alpha) \overset{(*)}{=} \alpha$

$\hookrightarrow$ d.h. der p-value ist $p^*(v) = F(v)$ mit $F = \Phi$ $\quad$ (Φ VF zu $\mathcal{N}(0,1)$)

Durchschnittseinkommen von niedergelassenen Ärzten

Stichprobe von $n = 200$ Arztpraxen, $\bar{X} = 168$T€. Laut einer statistischen Bundesamt liegt das „wahre Durchschnittseinkommen“ bei mindestens 181T€.
Annahme: $X \sim \mathcal{N}(\mu, \sigma^2), \sigma = 100$ bekannt.

$\hookrightarrow$ Hypothesen: $H_0 : \mu \geq 181$, vs. $H_1 : \mu < 181$, $v = \sqrt{n} \cdot \frac{\bar{x}-\mu_0}{\sigma} \approx -1.83$

Ermitteln Sie den p-value und die Testentscheidung ($\alpha = 0.05$).

Tabelle der VF zu $\mathcal{N}(0,1)$ auf folgender Folie:

	000	005	010	015	020	025	030	035	040	045	050	055	060	065	070	075	080	085	090	095
0.0	500	502	504	506	508	510	512	514	516	518	520	522	524	526	528	530	532	534	536	538
0.1	540	542	544	546	548	550	552	554	556	558	560	562	564	566	567	569	571	573	575	577
0.2	579	581	583	585	587	589	591	593	595	597	599	601	603	604	606	608	610	612	614	616
0.3	618	620	622	624	626	627	629	631	633	635	637	639	641	642	644	646	648	650	652	654
0.4	655	657	659	661	663	665	666	668	670	672	674	675	677	679	681	683	684	686	688	690
0.5	691	693	695	697	698	700	702	704	705	707	709	711	712	714	716	717	719	721	722	724
0.6	726	727	729	731	732	734	736	737	739	741	742	744	745	747	749	750	752	753	755	756
0.7	758	760	761	763	764	766	767	769	770	772	773	775	776	778	779	781	782	784	785	787
0.8	788	790	791	792	794	795	797	798	800	801	802	804	805	806	808	809	811	812	813	815
0.9	816	817	819	820	821	823	824	825	826	828	829	830	831	833	834	835	836	838	839	840
1.0	841	843	844	845	846	847	848	850	851	852	853	854	855	857	858	859	860	861	862	863
1.1	864	865	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883	884
1.2	885	886	887	888	889	890	891	892	893	893	894	895	896	897	898	899	900	901	901	902
1.3	903	904	905	906	907	907	908	909	910	911	911	912	913	914	915	915	916	917	918	918
1.4	919	920	921	921	922	923	924	924	925	926	926	927	928	929	929	930	931	931	932	933
1.5	933	934	934	935	936	936	937	938	938	939	939	940	941	941	942	942	943	944	944	945
1.6	945	946	946	947	947	948	948	949	949	950	951	951	952	952	953	953	954	954	954	955
1.7	955	956	956	957	957	958	958	959	959	960	960	960	961	961	962	962	962	963	963	964
1.8	964	964	965	965	966	966	966	967	967	967	968	968	969	969	969	970	970	970	971	971
1.9	971	972	972	972	973	973	973	974	974	974	974	975	975	975	976	976	976	976	977	977
2.0	977	978	978	978	978	979	979	979	979	980	980	980	980	981	981	981	981	981	982	982
2.1	982	982	983	983	983	983	983	984	984	984	984	984	985	985	985	985	985	986	986	986
2.2	986	986	986	987	987	987	987	987	987	988	988	988	988	988	988	989	989	989	989	989
2.3	989	989	990	990	990	990	990	990	990	990	991	991	991	991	991	991	991	991	992	992
2.4	992	992	992	992	992	992	992	993	993	993	993	993	993	993	993	993	993	994	994	994
2.5	994	994	994	994	994	994	994	994	994	995	995	995	995	995	995	995	995	995	995	995
2.6	995	995	995	996	996	996	996	996	996	996	996	996	996	996	996	996	996	996	996	996
2.7	997	997	997	997	997	997	997	997	997	997	997	997	997	997	997	997	997	997	997	997
2.8	997	997	998	998	998	998	998	998	998	998	998	998	998	998	998	998	998	998	998	998
2.9	998	998	998	998	998	998	998	998	998	998	998	998	998	998	999	999	999	999	999	999
3.0	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
3.1	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
3.2	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	≈ 1

Beispiel:
$\Phi(1.240) = \Phi(1.3 + .040) \approx 0.893.$
Für nicht aufgeführte x:

- $x \geq 3.2$:
  $\Phi(x) \approx 1$ für $x \geq 3.2$.
- Für $x < 0$:
  $\Phi(x) = 1 - \Phi(-x)$
- Interpolation:
  für vertafelte x, y und
  $t = \lambda x + (1 - \lambda)y)$, $\lambda \in]0; 1[$
  $\Phi(t) \approx \lambda\Phi(x) + (1 - \lambda)\Phi(y)$

Berechnung des p-values in den anderen Fällen (σ bekannt, mit $F = \Phi$):

c) $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$: $\quad v \in K \Leftrightarrow v > z_{1-\alpha}$

$p^* = 1 - \Phi(v)$

denn $v > z_{1-\alpha} \quad \Leftrightarrow \quad F(v) > F(z_{1-\alpha}) \overset{(*)}{=} 1 - \alpha \quad \Leftrightarrow \quad 1 - F(v) < \alpha$

a) $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$: $\quad v \in K \Leftrightarrow |v| > z_{1-\alpha/2}$

$p^* = 2\min(\Phi(v), 1 - \Phi(v)) = 2\min(1 - \Phi(-v), 1 - \Phi(v)) = 2(1 - \Phi(|v|)$

$$
\begin{aligned}
|v| > z_{1-\alpha/2} \quad \Leftrightarrow \quad & v < -z_{1-\alpha/2} \quad \text{oder} \quad v > z_{1-\alpha/2} \\
\Leftrightarrow \quad & -v > z_{1-\alpha/2} \quad \text{oder} \quad v > z_{1-\alpha/2} \\
\Leftrightarrow \quad & F(-v) > 1 - \tfrac{\alpha}{2} \quad \text{oder} \quad F(v) > 1 - \tfrac{\alpha}{2} \\
\Leftrightarrow \quad & 1 - F(-v) < \tfrac{\alpha}{2} \quad \text{oder} \quad 1 - F(v) < \tfrac{\alpha}{2} \\
\Leftrightarrow \quad & F(v) < \tfrac{\alpha}{2} \quad \text{oder} \quad 1 - F(v) < \tfrac{\alpha}{2} \\
\Leftrightarrow \quad & \min(F(v), 1 - F(v)) < \tfrac{\alpha}{2} \quad \Leftrightarrow \quad 2\min(F(v), 1 - F(v)) < \alpha
\end{aligned}
$$

Durchschnittseinkommen von Praktikanten

Stichprobe von $n = 40$ Studenten mit durchschnittlichem Einkommen $\bar{X} = 1225.45$.
Laut einer Studie liegt das „wahre Durchschnittseinkommen“ bei 1260 €.
Wird die Angabe der Studie (bei einem Signifikanzniveau von $\alpha = 5\%$) durch die Stichprobe gestützt? Berechnen Sie den p-value.

Annahme: $X \sim \mathcal{N}(\mu, \sigma^2), \sigma = 120$ als bekannt angenommen.

Wie lautet, ausgehend von Ihren bisherigen Berechnungen, Ihre Testentscheidung?

Berechnung des p-Wertes

a) bei zweiseitiger Hypothese $H_0 : \mu = \mu_0$ (zweiseitigem Test):

$$p^* = 2 \cdot \min(P(V \leq v)\,,\ 1 - P(V \leq v))$$

b) bei rechtssseitiger Hypothese $H_0 : \mu \geq \mu_0$ (rechtsseitigem Test):

$$p^* = P(V \leq v)$$

c) bei linksseitiger Hypothese $H_0 : \mu \leq \mu_0$ (linkssseitigem Test):

$$p^* = 1 - P(V \leq v)$$

$\hookrightarrow$ In allen Fällen wird p^* jeweils mit $\mu = \mu_0$ berechnet.

p-Werte p^* bei Mittelwert-Tests (Fälle i),ii),iv), d.h. Normalverteilung mit σ^2 bekannt, bzw. approx. Normalverteilung s.o.)

$$
\begin{array}{llcl}
a) & H_0 : \mu = \mu_0 & p^* & = & 2 \cdot (1 - P(V \leq |v|)) = 2 \cdot (1 - \Phi(|v|)) \\
b) & H_0 : \mu \geq \mu_0 & p^* & = & P(V \leq v) = \Phi(v) \\
c) & H_0 : \mu \leq \mu_0 & p^* & = & (1 - P(V \leq v)) = 1 - \Phi(v)
\end{array}
$$

Fall iv): Für Normalverteilung und unbekanntes σ^2 ist die Nullverteilung die t_{n-1}-Verteilung, dann analog:

$$
\begin{array}{llcl}
a) & H_0 : \mu = \mu_0 & p^* & = & 2 \cdot (1 - P(V \leq |v|)) = 2 \cdot (1 - F_{t_{n-1}}(|v|)) \\
b) & H_0 : \mu \geq \mu_0 & p^* & = & P(V \leq v) = F_{t_{n-1}}(v) \\
c) & H_0 : \mu \leq \mu_0 & p^* & = & (1 - P(V \leq v)) = 1 - F_{t_{n-1}}(v)
\end{array}
$$

Berechnung mittels Software (R) oder (für $n > 30$ approximativ) durch $F_{t_{n-1}} \approx \Phi$

Durchschnittseinkommen von Praktikanten

Stichprobe von $n = 40$ Studenten mit durchschnittlichem Einkommen $\bar{X} = 1225.45$.
Laut einer Studie liegt das „wahre Durchschnittseinkommen“ bei 1260 €.
Wird die Angabe der Studie (bei einem Signifikanzniveau von $\alpha = 5\%$) durch die Stichprobe gestützt? Berechnen Sie den p-value.
Annahme: $X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 unbekannt (wird durch $S^2 = 87.23354^2$ geschätzt).

$\hookrightarrow$ Formeln für p-Werte auf viele stetige(!) Modelle übertragbar:

- $H : \theta = \theta_0$: $p^* = 2 \cdot \min(F(v), 1 - F(v))$
- $H : \theta \geq \theta_0$: $p^* = F(v)$
- $H : \theta \leq \theta_0$: $p^* = 1 - F(v)$

Geeignet bei unimodaler Nullverteilung F.

$\hookrightarrow$ Bei diskreten Modellen Formeln nicht einfach übertragbar, weil die VF nicht invertierbar ist. Grundsätzlich aber weiterhin:

$p^* \cong$ min. Signifikanzniveau, zu dem H_0 bei $V = v$ abgelehnt wird.

Vor allem bei zweiseitigen Hypothesen dann diverse Ansätze, z.B.:

„Principle of minimum likelihood“ (z.T. in R verwendet)

$p^* \cong$ Summe aller WS $P(V = k)$ mit $P(V = k) \leq P(V = v)$.

Vgl.: *Gibbons/Pratt: P-values: Interpretation and Methodology, Amer. Statist. 29 (1975), 20-25.*

6.4 Einstichproben-Anteilswert-Tests

Modell: $X_i \sim \mathcal{B}(1, p)$

1. Hypothesensysteme

a) $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$

b) $H_0 : p \geq p_0$ vs. $H_1 : p < p_0$

c) $H_0 : p \leq p_0$ vs. $H_1 : p > p_0$

Beispiele:

a) Der Anteil iPad-Besitzer unter den Studenten in Münster liegt bei p_0.
Eine (verbogene) Münze ist fair ($p = \frac{1}{2}$).

b) Der Anteil Kunden, die einem Vertreterbesuch zustimmen, ist mindestens p_0.

c) Der Anteil fehlerhafter Artikel einer Tagesproduktion ist höchstens p_0.

Anteilswert-Tests sind Mittelwert-Tests für dichotome X_i. ($E(X_i) = p$). Sie lassen sich exakt oder approximativ ausführen.

Exakter Test

$\hookrightarrow$ Teststatistik: $V = X_1 + \cdots + X_n \sim Bin(n, p_0)$ für $p = p_0$

a) $H_0 : p = p_0$: Ablehnung wenn $v \notin [q_{\alpha/2}; q_{1-\alpha/2}]$, p^* s.u.

b) $H_0 : p \geq p_0$: Ablehnung wenn $v < q_\alpha$, $p^* = F(v)$.

c) $H_0 : p \leq p_0$: Ablehnung, wenn $v > q_{1-\alpha}$, $p^* = 1 - F(v-1)$.

$p^* \cong$ min. Sig.-niveau, mit dem H_0 abgelehnt wird.
Davon abweichend in R: „minimum likelihood principle“

Approx. Test für $np_0(1-p_0) \geq 9$ (Fall iv) der Mittelwerttests)

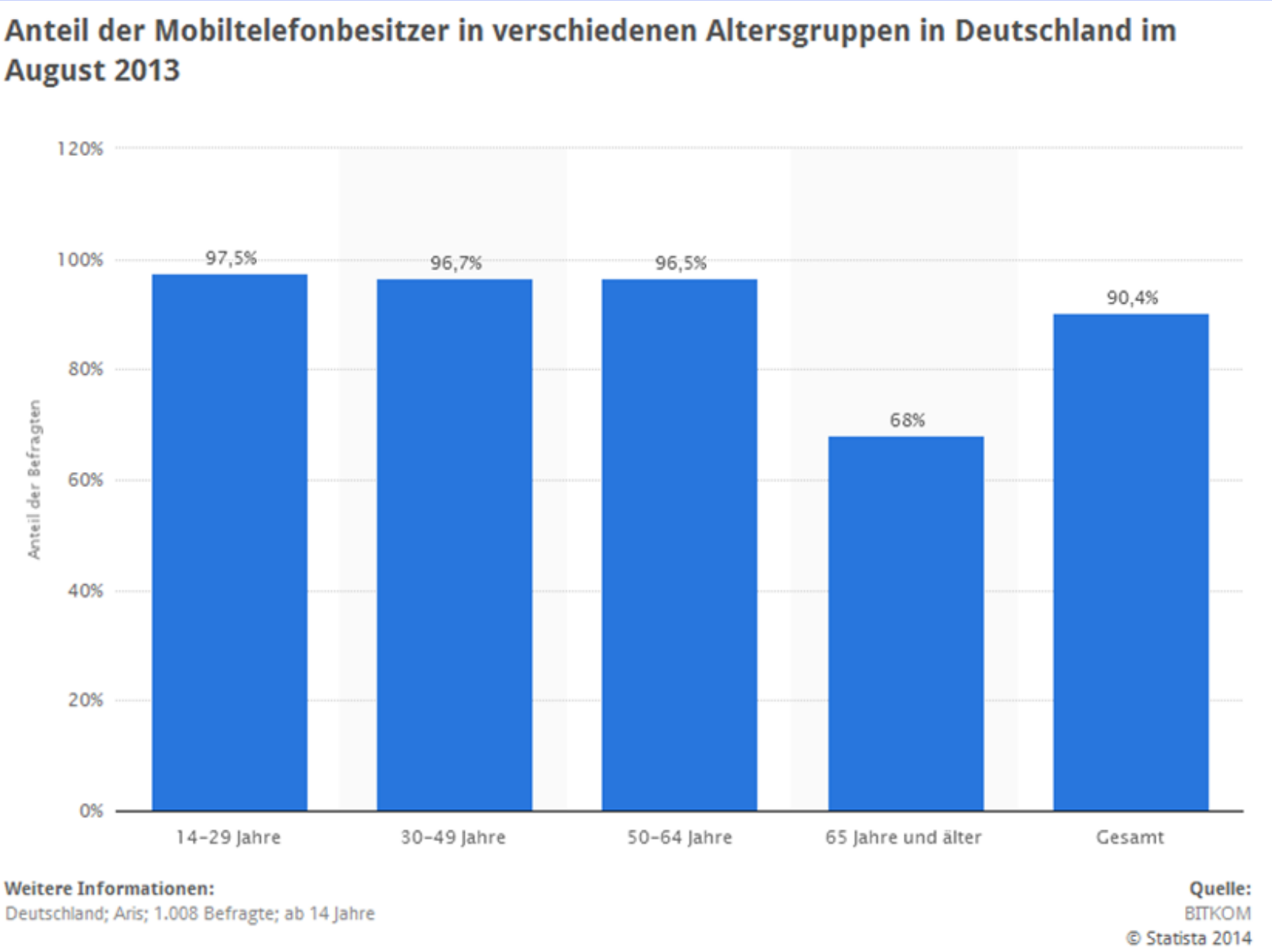
$\hookrightarrow$ Teststatistik $V = \sqrt{n}\frac{\bar{X}-p_0}{\sqrt{p_0(1-p_0)}}$, näherungsweise $\sim \mathcal{N}(0,1)$ für $p = p_0$.

a) $H_0 : p = p_0$: Ablehnung wenn $|v| > z_{1-\alpha/2}$, $p^* = 2(1 - \Phi(|v|))$.

b) $H_0 : p \geq p_0$: Ablehnung wenn $v < z_\alpha$, $p^* = \Phi(v)$.

c) $H_0 : p \leq p_0$: Ablehnung, wenn $v > z_{1-\alpha}$, $p^* = 1 - \Phi(v)$.

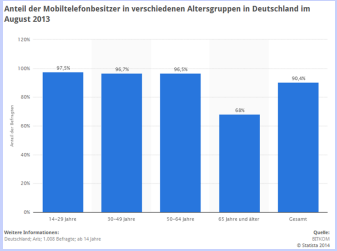
Exakter und approx. Test ergeben nicht immer die gleiche Entscheidung (Übung!)

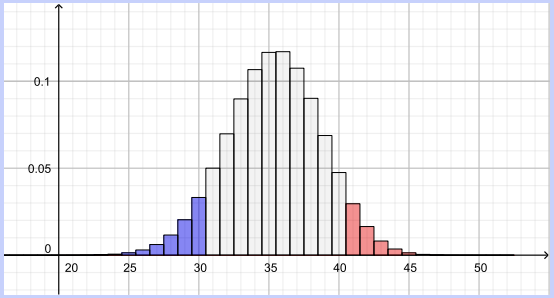

Anteil der Mobiltelefonbesitzer in verschiedenen Altersgruppen in Deutschland im August 2013
120%
100%
80%
60%
40%
20%
0%
Anteil der Befragten
97,5%
96,7%
96,5%
68%
90,4%
14-29 Jahre
30-49 Jahre
50-64 Jahre
65 Jahre und älter
Gesamt
Weitere Informationen:
Deutschland; Aris; 1.008 Befragte; ab 14 Jahre
Quelle:
BITKOM
© Statista 2014

Beispiel: Handynutzung unter Rentnern

Laut einer Studie besitzen 68% der Rentner ein Handy. Wird dies durch eine Stichprobe von $n = 52$ Rentnern, von denen 30 ein Handy besitzen, gestützt ($\alpha = 0.05$)?

Hypothesen: $H_0 : p = p_0 = 0.68$, $H_1 : p \neq 0.68$





https://ggbm.at/rqrpwked

$\hookrightarrow$ exakter Test:

- $V = v = 30$
- $q_{0.025} = 29 \leq 30 \leq q_{0.975} = 42$:
  H_0 wird nicht verworfen.
- $p^* = P(V \leq 30) + P(V \geq 41) \approx 0.136$

in R: binom.test(30,52,0.68)$p.value („minimum likelihood principle“)

$\hookrightarrow$ approximativer Test:

- $V = v = \sqrt{n}\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)}} = \sqrt{52}\frac{30/52-0.68}{\sqrt{0.68 \cdot 0.32}} = -1.59$
- $|v| = 1.59 < z_{0.975} = 1.96$: H_0 wird nicht verworfen.
- $p^* = 2 \cdot (1 - \Phi(1.59)) = 0.11 > 0.05$

Gültigkeit der Normalapproximation: $52 \cdot 0.68 \cdot 0.32 = 11.32 > 9$

Anteil der Tablet-Nutzer in Deutschland in den Jahren 2012 bis 2017
Anteil an der Bevölkerung
60%
50%
40%
30%
20%
10%
0%
13%
15%
28%
40%
41%
47%
2012
2013
2014
2015
2016
2017*
Quelle
Bitkom
© Statista 2019
Weitere Informationen:
Deutschland; Bitkom Research; ab 14 Jahre
statista

Umfrage unter Studierenden

Von $n = 40$ Befragten teilten 25 mit, dass sie im Jahr 2018 Tablet-Nutzer gewesen seien. Lässt sich hieraus statistisch schließen, dass der Anteil der Tablet-Nutzer in 2018 auf über 47% gestiegen ist ($\alpha = 5\%$)?

Anmerkung: Eine verallgemeinernde Aussage ist aufgrund der speziellen Auswahl der Befragten eher nicht möglich („selection bias“).

Zusammenhang von Test und Konfidenzintervall

$H_0 : p = p_0$ wird nicht abgelehnt, falls

$$
\begin{aligned}
|V| &= \left| \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} \right| \leq z_{1-\alpha/2} \\
\Leftrightarrow \quad -z_{1-\alpha/2} &\leq \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} \leq z_{1-\alpha/2} \\
\Leftrightarrow \quad \hat{p} - z_{1-\alpha/2}\sigma_{\hat{p}} &\leq p_0 \leq \hat{p} + z_{1-\alpha/2}\sigma_{\hat{p}}
\end{aligned}
$$

- $\hookrightarrow$ entspricht dem (approximativen) Konfidenzintervall für den Anteilswert (s.o.)
- $\hookrightarrow$ H_0 **wird abgelehnt, wenn der hypothetische Wert** p_0 **außerhalb der Realisation des Konfidenzintervalles liegt.**
- $\hookrightarrow$ übertragbar auf einseitige Tests/KI und (z.B.) Mittelwert-Tests

6.5 Chi-Quadrat-Test für die Varianz bei $X_i \sim \mathcal{N}(\mu, \sigma^2)$

1. Hypothesensystem

a) $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$

b) $H_0 : \sigma^2 \geq \sigma_0^2$ vs. $H_1 : \sigma^2 < \sigma_0^2$

c) $H_0 : \sigma^2 \leq \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$

2. Teststatistik und ihre Verteilung

Für $\sigma = \sigma_0$ gilt (s.o.) mit $\hat{\sigma}^2 = S^2$:

$$V = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1) \quad (\mu \text{ unbekannt}),$$

$$V = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \mu)^2 \sim \chi^2(n) \quad (\mu \text{ bekannt})$$

3. Entscheidungsregel

Für unbekanntes μ wird H_0 abgelehnt, wenn:

a) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$ $\quad v < \chi^2_{\alpha/2}(n-1)$ oder $v > \chi^2_{1-\alpha/2}(n-1)$

b) $H_0 : \sigma^2 \geq \sigma_0^2, H_1 : \sigma^2 < \sigma_0^2$ $\quad 0 < v < \chi^2_{\alpha}(n-1)$

c) $H_0 : \sigma^2 \leq \sigma_0^2, H_1 : \sigma^2 > \sigma_0^2$ $\quad v > \chi^2_{1-\alpha}(n-1)$

$\hookrightarrow$ die χ^2-Verteilung ist nicht symmetrisch, deshalb unsymmetrische Festlegung bei zweiseitiger Hypothese.

p-Werte

a) $H_0, \sigma^2 = \sigma_0^2$: $\quad p^* = 2 \cdot \min(F_{\chi^2(n-1)}(v), 1 - F_{\chi^2(n-1)}(v))$

b) $H_0, \sigma^2 \geq \sigma_0^2$: $\quad p^* = F_{\chi^2(n-1)}(v)$

c) $H_0, \sigma^2 \leq \sigma_0^2$: $\quad p^* = 1 - F_{\chi^2(n-1)}(v)$

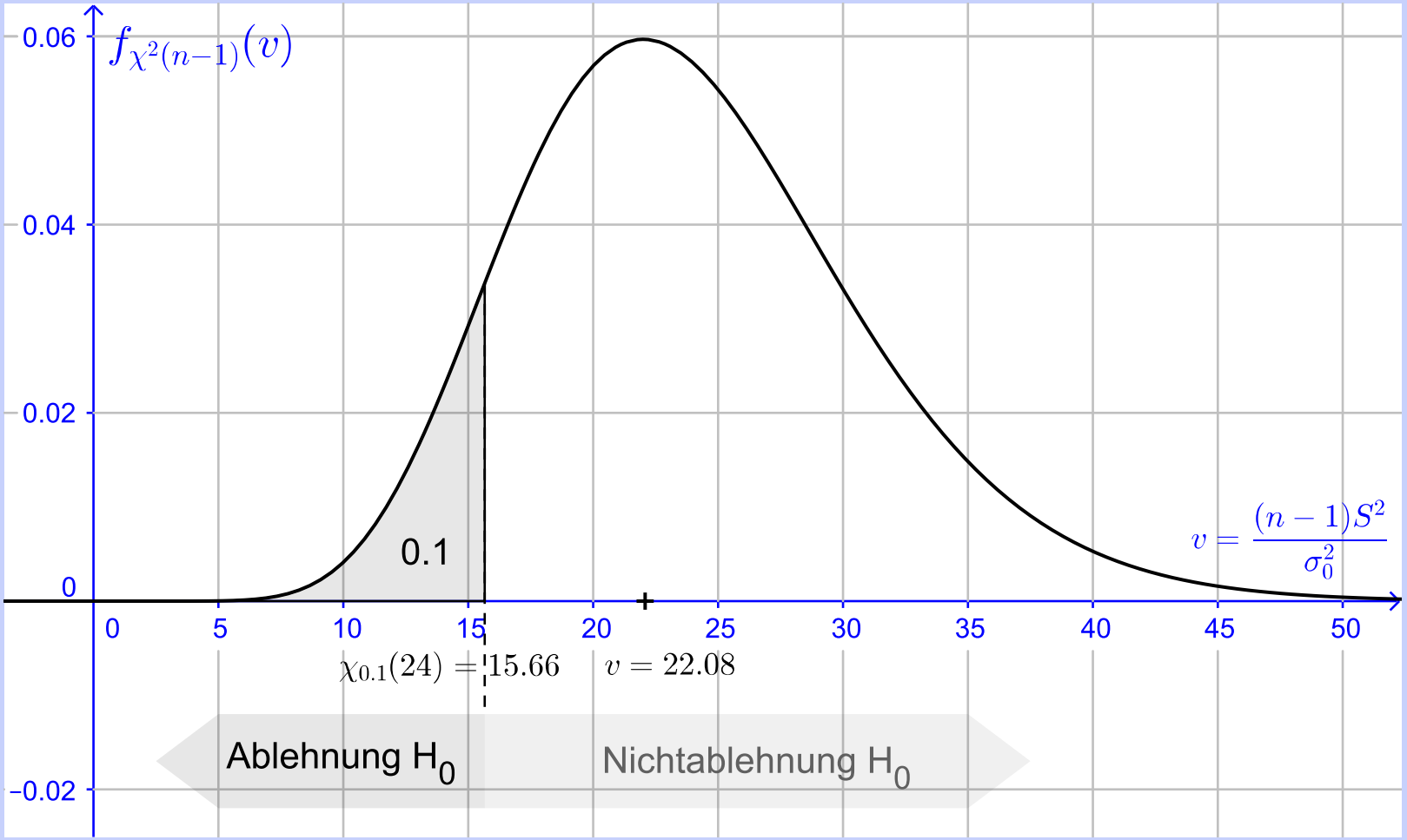
$\hookrightarrow$ Für bekanntes μ werden Quantile und Verteilungsfunktion der χ^2-Verteilung mit n Freiheitsgraden verwendet (aber eher unrealistisch).

Lebensdauer eines Verschleißteils

Annahme: Lebensdauer X eines Verschleißteils sei normalverteilt. In der Vergangenheit galt $\sigma^2 = 2.5$ Jahre^2. Nach Umstellung des Produktionsverfahrens soll geprüft werden, ob die Variabilität zurückgegangen ist ($\alpha = 0.1$). Eine Stichprobe bei unbekanntem μ ergab: $n = 25, \hat{\sigma}^2 = 2.3$.

In R: `qchisq(0.1,df=c(24,25))` ergibt c(15.65868,16.47341)

Illustration



6.6 Gütefunktionen von Parametertests

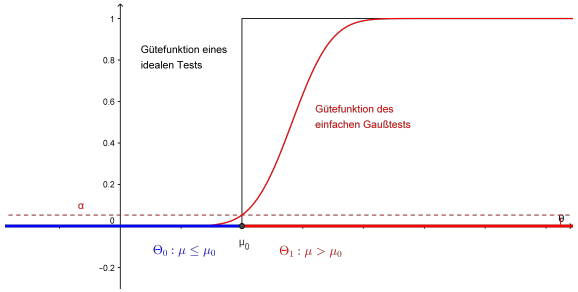
- ↪ Partition des Parameterraums gemäß Hypothesenpaar H_0, H_1 in $\Theta = \Theta_0 \cup \Theta_1$.
- ↪ Bestimmung der Teststatistik $v = V = V(X)$. Damit Testentscheidungen:
  - □ $d(v) = d_1$ (Ablehnung von H_0) für $v \in K$
  - □ $d(v) = d_0$ (Nichtablehnung von H_0) für $v \notin K$
- ↪ Risiko[2]: $R(\theta, V) = \begin{cases} P_\theta(V \in K) & \theta \in \Theta_0 \text{ (Fehler 1. Art)} \\ P_\theta(V \notin K) & \theta \in \Theta_1 \text{ (Fehler 2. Art)} \end{cases}$
- ↪ Tests sollten ein möglichst kleines Risiko haben ⇝ 2 konfliktäre Ziele:
  - □ Für $\theta \in \Theta_0$: $P_\theta(V \in K)$ möglichst klein
  - □ Für $\theta \in \Theta_1$: $P_\theta(V \notin K)$ möglichst klein, d.h. $P_\theta(V \in K)$ möglichst groß

Gütefunktion

$$g : \Theta \to [0; \infty], g(\theta) \quad = \quad P_\theta(\text{„Ablehnung von } H_0\text{“}) = P_\theta(V \in K)$$

[2] $R(\theta, V) = E_\theta(L(\theta, d(V)))$ m. Verlustfunktion $L(\theta, d) = \begin{cases} 1 & \text{für } \theta \in \Theta_i, d = d_j,\ i \neq j \\ 0 & \text{sonst} \end{cases}$

Beispiel: linksseitiger Gaußtest : $\Theta_0 :]-\infty; \mu_0]$ $\Theta_1 :]\mu_0; \infty[$



https://ggbm.at/cedn5dkr

$$
\begin{aligned}
g(\mu) &= P_\mu\left(\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} > z_{1-\alpha}\right) = P_\mu\left(\sqrt{n}\frac{\bar{X}-\mu+\mu-\mu_0}{\sigma} > z_{1-\alpha}\right) \\
&= P_\mu\left(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} + \sqrt{n}\frac{\mu-\mu_0}{\sigma} > z_{1-\alpha}\right) \\
&= P_\mu\left(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} > z_{1-\alpha} - \sqrt{n}\frac{\mu-\mu_0}{\sigma}\right) = 1 - \Phi\left(z_{1-\alpha} - \sqrt{n}\frac{\mu-\mu_0}{\sigma}\right) \\
&= \Phi\left(\sqrt{n}\frac{\mu-\mu_0}{\sigma} - z_{1-\alpha}\right) = \Phi\left(\frac{\sqrt{n}}{\sigma}\cdot\left(\mu - \left(\mu_0 + \frac{\sigma}{\sqrt{n}}z_{1-\alpha}\right)\right)\right) = \Phi(b(\mu - a))
\end{aligned}
$$

Die Gütefunktion entsteht aus der VF Φ zu $\mathcal{N}(0,1)$ durch lineare Transformation:

- Skalenfaktor $b = \frac{\sqrt{n}}{\sigma}$ legt „Steilheit“ fest. Je größer n/kleiner σ, desto steiler.
- $a = \mu_0 + \frac{\sigma}{\sqrt{n}}z_{1-\alpha}$ ist Wendestelle (Übung).

Ein Test mit steiler verlaufender Gütefunktion stellt einen besseren Kompromiss zwischen Fehler-WS dar.

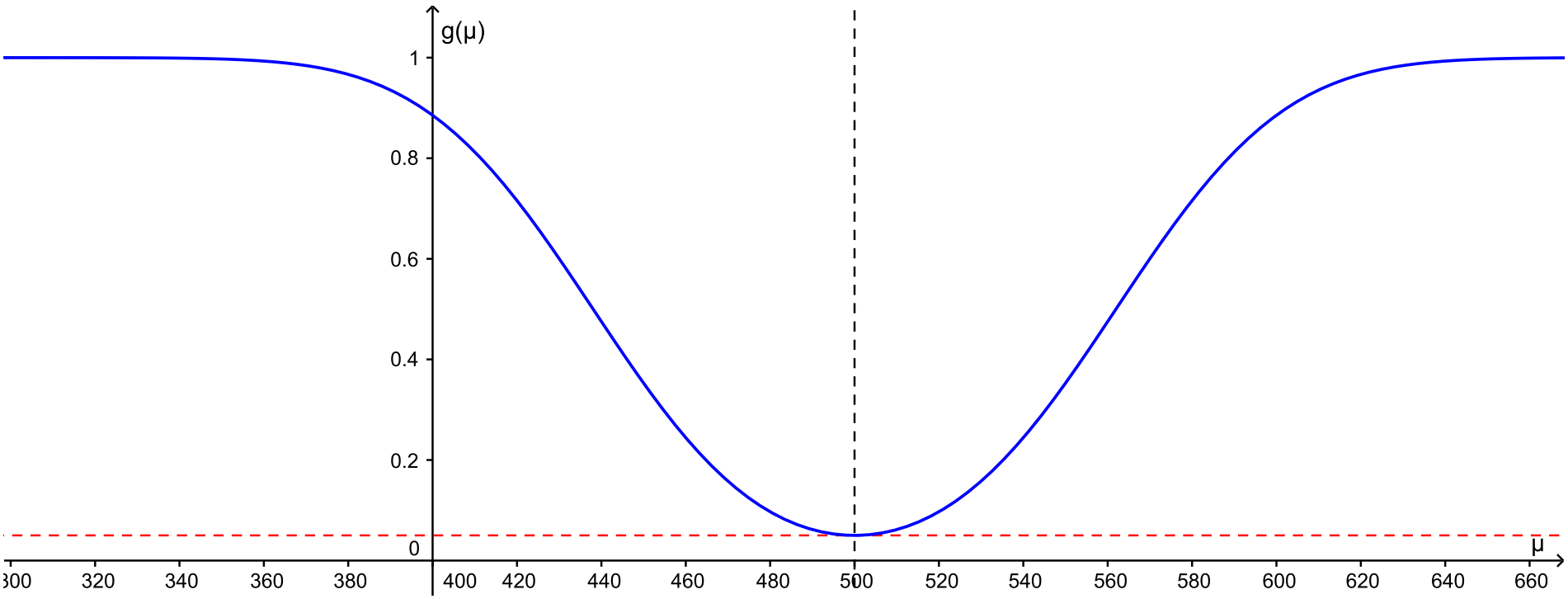
Einfluss des Stichprobenumfangs n auf die Testgüte

Durch Erhöhung von n können Vorgaben an beide Fehlertypen erfüllt werden.

Konkret: Wie groß muss n sein, damit die Gütefunktion an der Stelle μ^* einen vorgegebenen Wert β überschreitet?

Gütefunktion im zweiseitigen Testproblem

$$g(\mu) = \Phi\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right) + \Phi\left(-\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right)$$



Rechnung dazu: Übung

7. Zweistichprobentests

- Gauß-Test (Normalverteilung mit bekannten Varianzen)
- t-Test (Normalverteilungen mit unbekannter Varianz)
- Welch-Test
- Approximativer Gauß-Test
- t-Test bei verbundenen Stichproben
- Wilcoxon-Rangsummentest

Zweistichprobentests

- ↪ Ziehung zweier Stichproben, Analyse möglicher Unterschiede
- ↪ Herkunft der zwei Stichproben
  - □ aus verschiedenen Grundgesamtheiten, unabh. Stichproben
    z.B. Vergleich des Durchschnittseinkommens von Ärzten und Anwälten
  - □ zwei unabh. Stichproben aus einer Grundgesamtheit
    z.B. Aufteilung von n Studenten zufällig auf zwei Gruppen mit verschiedenem Übungskonzept, Vergleich der Leistungen beider Gruppen
  - □ zwei gebundene Stichproben aus einer Grundgesamtheit
    z.B. WI-Studierende: Abiturnote vs. Bachelornote

Grundsätzliche Vorgehensweise bei Mittelwerttests:

$\hookrightarrow$ Ziehen zweier Stichproben der Umfänge n_1 und n_2 mit $\bar{x}_1$ und $\bar{x}_2$

$\hookrightarrow$ Überprüfung von Hypothesen bzgl. der Differenz $\mu_1 - \mu_2$ anhand von $d = \bar{x}_1 - \bar{x}_2$

$$a)\ H_0 : (\mu_1 - \mu_2) = \delta_0 \qquad \text{vs.} \qquad H_1 : (\mu_1 - \mu_2) \neq \delta_0$$

$$b)\ H_0 : (\mu_1 - \mu_2) \geq \delta_0 \qquad \text{vs.} \qquad H_1 : (\mu_1 - \mu_2) < \delta_0$$

$$c)\ H_0 : (\mu_1 - \mu_2) \leq \delta_0 \qquad \text{vs.} \qquad H_1 : (\mu_1 - \mu_2) > \delta_0$$

7.1 Gauß-Test (Normalverteilung mit bekannten Varianzen)

Annahme: $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, σ_i^2 bekannt.

1. Hypothesen

a) $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. $H_1 : \mu_1 - \mu_2 \neq \delta_0$

b) $H_0 : \mu_1 - \mu_2 \geq \delta_0$ vs. $H_1 : \mu_1 - \mu_2 < \delta_0$

c) $H_0 : \mu_1 - \mu_2 \leq \delta_0$ vs. $H_1 : \mu_1 - \mu_2 > \delta_0$

2. Teststatistik V und Stichprobenverteilung F_V

$$V = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1)$$

3. Entscheidungsregel – Gauß-Test

Ablehnung von H_0, falls	a) $\lvert v\rvert > z_{1-\alpha/2}$	b) $v < z_\alpha$	c) $v > z_{1-\alpha}$
p-Wert $p^* =$	$2(1-\Phi(\lvert v\rvert))$	$\Phi(v)$	$1-\Phi(v)$

Durchschnittseinkommen von Praktikanten

Stichproben von $n_1 = 35$ Studenten in Stadt A, $\bar{x}_1 = 600$
$n_2 = 40$ Studenten in Stadt B, $\bar{x}_2 = 520$

Annahme: $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2), \sigma_1^2 = 8000, \sigma_2^2 = 5000$ (bekannt).

Kann man aus den Daten folgern, dass die Einkommen von Praktikanten in Stadt A höher als in Stadt B sind ($\alpha = 0.01$)?

$\hookrightarrow$ Hypothesen $H_0 : \mu_1 \leq \mu_2$ vs. $H_1 : \mu_1 > \mu_2$ (Fall c) mit $\delta_0 = 0$)

$\hookrightarrow$ Teststatistik $V = \dfrac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{600-520-0}{\sqrt{8000/35+5000/40}} = \frac{80}{10.177} = 7.86$

$\hookrightarrow$ Entscheidung: $v = 7.86 > z_{1-0.01} = z_{0.99} = 2.33$ ($\alpha = 0.05$),
also wird H_0 abgelehnt.

$\hookrightarrow$ $p^* = (1 - \Phi(7.86)) \approx 0 < 0.01$

Bemerkung: Die Annahme, dass die Varianz bekannt ist (wenn gleichzeitig der Erwartungswert unbekannt ist), ist eher unrealistisch. Für $n_1, n_2 > 30$ kann die Varianz durch die Stichprobenvarianz ersetzt werden (approx. Gaußtest)

7.2 t-Test (Normalverteilungen mit unbekannter Varianz)

Annahme: $X_1 \sim \mathcal{N}(\mu_1, \sigma^2), X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, σ^2 unbekannt.

1. Hypothesen

a) $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. $H_1 : \mu_1 - \mu_2 \neq \delta_0$

b) $H_0 : \mu_1 - \mu_2 \geq \delta_0$ vs. $H_1 : \mu_1 - \mu_2 < \delta_0$

c) $H_0 : \mu_1 - \mu_2 \leq \delta_0$ vs. $H_1 : \mu_1 - \mu_2 > \delta_0$

2. Teststatistik V und Stichprobenverteilung F für $\mu_1 - \mu_2 = \delta_0$

$$V = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{(1/n_1 + 1/n_2)\,\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}} \sim t(n) \text{ mit } n = n_1 + n_2 - 2$$

3. Entscheidungsregel – t-Test

Abl. von H_0, falls	a) $\lvert v\rvert > t_{1-\alpha/2}(n)$	b) $v < t_\alpha(n)$	c) $v > t_{1-\alpha}(n)$
p-Wert $p^* =$	$2(1 - F(\lvert v\rvert))$	$F(v)$	$1 - F(v)$

Durchschnittseinkommen von Praktikanten

Stichproben von $n_1 = 35$ Studenten in Stadt A, $\bar{x}_1 = 600$, $s_1 = 100$
$n_2 = 40$ Studenten in Stadt B, $\bar{x}_2 = 520, s_2 = 70$

Annahme: $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2), \sigma_1 = \sigma_2$ unbekannt.

Sind die Durchschnittseinkommen der Studenten beider Städte gleich($\alpha = 0.05$)?

7.3 Welch-Test

Annahme: $X_j \sim F_j$, $n_1, n_2 > 30$, oder $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, σ_j^2 unbekannt.

1. Hypothesen

a) $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. $H_1 : \ldots \neq \delta_0$

b) $H_0 : \mu_1 - \mu_2 \geq \delta_0$ vs. $H_1 : \cdots < \delta_0$

c) $H_0 : \mu_1 - \mu_2 \leq \delta_0$ vs. $H_1 : \cdots > \delta_0$

2. Teststatistik V und Stichprobenverteilung F für $\mu_1 - \mu_2 = \delta_0$

$$V = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \overset{appr}{\sim} t(k),\ k = \left\lfloor \frac{S_1^2/n_1 + S_2^2/n_2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2} \right\rfloor,\ S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

3. Entscheidungsregel – Welch-Test

Abl. von H_0, falls	a) $\lvert v\rvert > t_{1-\alpha/2}(k)$	b) $v < t_\alpha(k)$	c) $v > t_{1-\alpha}(k)$
p-Wert $p^* =$	$2(1 - F(\lvert v\rvert))$	$F(v)$	$1 - F(v)$

Alternativ für $n_1, n_2 > 30$: Gauß-Test mit obigem $V \overset{approx.}{\sim} \mathcal{N}(0, 1)$

7.4 Approximativer Gauß-Test

Annahme: $X_{ji} \sim F_j$ für $i = 1, \ldots, n_j$, $j_{=}1, 2$. Dabei $n_1, n_2 > 30$

1. Hypothesen

a) $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. $H_1 : \ldots \neq \delta_0$

b) $H_0 : \mu_1 - \mu_2 \geq \delta_0$ vs. $H_1 : \cdots < \delta_0$

c) $H_0 : \mu_1 - \mu_2 \leq \delta_0$ vs. $H_1 : \cdots > \delta_0$

2. Teststatistik V und Stichprobenverteilung F für $\mu_1 - \mu_2 = \delta_0$

$$V = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \overset{approx}{\sim} \mathcal{N}(0,1), \text{ mit } S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

3. Entscheidungsregel – approx. Gauß-Test

Ablehnung von H_0, falls	a) $\lvert v \rvert > z_{1-\alpha/2}$	b) $v < z_\alpha$	c) $v > z_{1-\alpha}$
p-Wert $p^* =$	$2(1 - F(\lvert v \rvert))$	$F(v)$	$1 - F(v)$

Alternativ: Welch-Test, nichtparametrische Verfahren (z.B. Wilcoxon-Test, s.u.)

7.5 t-Test (verb. Stichpr.) $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2), \delta = \mu_2 - \mu_1$

- ↪ Erhebung $X_{11}, \ldots, X_{1n}, X_{21}, \ldots, X_{2n}$ z.B. an denselben Merkmalsträgern
- ↪ **Differenzentest**: Einstichproben-t-Test auf Differenzen $D_i = X_{2i} - X_{1i}$

1. Hypothesen

a) $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$

b) $H_0 : \delta \geq 0$ vs. $H_1 : \delta < 0$

c) $H_0 : \delta \leq 0$ vs. $H_1 : \delta > 0$

2. Prüfgröße V

$V = \sqrt{n}\bar{D}/\hat{\sigma}_D \sim t(n-1)$ für $\delta = 0$

mit $\sigma_D^2 = var(X_{i2} - X_{i1})$,

$\hat{\sigma}_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2$

3. Entscheidung: H_0 wird abgelehnt, wenn

a) $H_0 : \delta = 0$: $|v| > t_{1-\alpha/2}(n-1)$ $\quad p^* = 2(1 - F(|v|))$

b) $H_0 : \delta \geq 0$: $v < t_{\alpha}(n-1)$ $\quad p^* = F(v)$

c) $H_0 : \delta \leq 0$: $v > t_{1-\alpha}(n-1)$ $\quad p^* = 1 - F(v)$

- ↪ Bei beliebig verteilter Grundgesamtheit: approx. Gaußtest ($V \overset{approx.}{\sim} \mathcal{N}(0,1)$)
- ↪ Auch verfügbar: Wilcoxon-Test bei gebundenen Stichproben

Bearbeitungsdauer zweier Programme P1,P2 auf einem Maschinentyp

Maschine i	1	2	3	4	5	6	7	8
Dauer P1 (X_{1i})	5.9	5.1	3.1	6.5	5.6	4.9	4.6	4.2
Dauer P2 (X_{2i})	5.0	4.0	5.3	6.2	5.1	4.6	7.7	3.9
$D_i = X_{2i} - X_{1i}$	-0.9	-1.1	2.2	-0.3	-0.5	-0.3	3.1	-0.3

Ist Bearbeitungsdauer bei P1 signifikant höher als bei P2, d.h. $\mu_1 > \mu_2$ ($\alpha = 0.05$)?

7.6 Wilcoxon-Rangsummentest

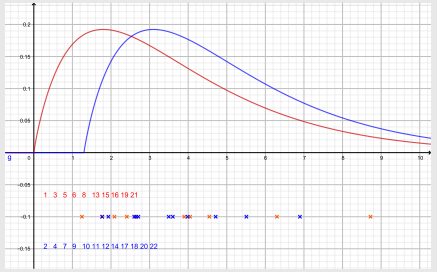
- ↪ die vorangegangenen Tests setzen entweder die Normalverteilung oder einen hinreichend großen Stichprobenumfang voraus
- ↪ nichtparametrische Alternative basierend auf Rängen
- ↪ In R: wilcox.test (...)

Annahmen:

$$X = X_1, \ldots, X_{n_1} \overset{u.i.v.}{\sim} F, \text{ (stetig)}$$

$$Y = Y_1, \ldots, Y_{n_2} \overset{u.i.v.}{\sim} G \text{ (stetig)}$$

X und Y sind st.u., $F_Y(x) = F_X(x - a)$



https://www.geogebra.org/m/khj3n2by

Idee:

- ↪ Verteilungsfunktionen von der Form her gleich, u.U. Verschiebung einer VF
- ↪ Gleichheit der Mediane entspricht Gleichheit der Verteilungen. Keine Existenz von Erwartungswerten vorausgesetzt.

1. Hypothese: H_0

a) $med(X) = med(Y)$ b) $med(X) \geq med(Y)$ c) $med(X) \leq med(Y)$

Unter der Annahme $F_Y(x) = F_X(x - a)$ ist dies gleichwertig zu

a) $a = 0$ b) $a \leq 0$ c) $a \geq 0$

2. Teststatistik:

$$W = \sum_{i=1}^{n_1} rg(X_i),$$

in R: $V = \sum_{i,j} 1_{\{X_i \leq Y_j\}} \stackrel{!}{=} W - \frac{n_1(n_1+1)}{2}$

- ↪ Teststatistik basiert auf X-Rängen der gepoolten Stichprobe $Z = (X, Y)$.
- ↪ Bei Gleichheit der Gruppenmediane wird eine gute Durchmischung von Z erwartet. Nullverteilung von W ist dann:
  - □ für $n_i \leq 25$: Wilcoxon-Verteilung (Quantile w_α) vertafelt bzw. mit R
  - □ für $\max(n_1, n_2) > 25$ approximativ: $\stackrel{approx.}{\sim} \mathcal{N}\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 \cdot n_2(n_1+n_2+1)}{12}\right)$
- ↪ Wichtig bei der Anwendung: Bindungsfreie Daten (Stetige Verteilungen!)

3. Testentscheidung: H_0 wird abgelehnt, falls

a) $H_0 : med(X) = med(Y)$: $w > w_{1-\alpha/2}(n_1, n_2)$ oder $w < w_{\alpha/2}(n_1, n_2)$

b) $H_0 : med(X) \geq med(Y)$: $w < w_{\alpha}(n_1, n_2)$

c) $H_0 : med(X) \leq med(Y)$: $w > w_{1-\alpha}(n_1, n_2)$

p-values für $\max(n_1, n_2) \leq 25$ (Berechnung von F z.B. mit R, pwilcox (...)):

a) $p^* = 2\min(F(w), 1 - F(w-1))$

b) $p^* = F(w)$

c) $p^* = 1 - F(w-1)$

p-values für $\max(n_1, n_2) > 25$ mit Normalapproximation: $\tilde{w} = \frac{w - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}}$

a) $p^* = 2(1 - \Phi(|\tilde{w}|)$

b) $p^* = \Phi(\tilde{w})$

c) $p^* = 1 - \Phi(\tilde{w})$

Tabelle der α-Quantile w_α für $n_1, n_2 \leq 25$ (mit R berechnet, `qwilcox(...)`):

.05	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	3	3	3	4	4	4	5	5	5	5	6	6	6	7	7	7	8	8	8	9	9	9	10	10
3	6	6	7	8	9	9	10	10	11	12	12	13	14	14	15	16	16	17	18	18	19	20	20	21
4	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	28	29	30	31	32	33	34
5	16	17	18	20	21	22	24	25	27	28	29	31	32	34	35	36	38	39	41	42	44	45	46	48
6	22	24	25	27	29	30	32	34	36	38	39	41	43	45	47	48	50	52	54	56	58	59	61	63
7	29	31	33	35	37	40	42	44	46	48	50	53	55	57	59	62	64	66	68	70	73	75	77	79
8	38	40	42	45	47	50	52	55	57	60	63	65	68	70	73	76	78	81	84	86	89	91	94	97
9	47	49	52	55	58	61	64	67	70	73	76	79	82	85	88	91	94	97	100	103	106	109	112	115
10	57	60	63	67	70	73	76	80	83	87	90	93	97	100	104	107	111	114	118	121	124	128	131	135
11	68	72	75	79	83	86	90	94	98	101	105	109	113	117	121	124	128	132	136	140	144	148	152	156
12	81	84	88	92	96	100	105	109	113	117	121	126	130	134	139	143	147	151	156	160	164	169	173	177
13	94	98	102	107	111	116	120	125	129	134	139	143	148	153	157	162	167	172	176	181	186	190	195	200
14	108	113	117	122	127	132	137	142	147	152	157	162	167	172	177	183	188	193	198	203	208	213	219	224
15	124	128	133	139	144	149	154	160	165	171	176	182	187	193	198	204	209	215	221	226	232	237	243	249
16	140	145	151	156	162	167	173	179	185	191	197	202	208	214	220	226	232	238	244	250	256	262	268	274
17	157	163	169	174	180	187	193	199	205	211	218	224	231	237	243	250	256	263	269	275	282	288	295	301
18	176	181	188	194	200	207	213	220	227	233	240	247	254	260	267	274	281	288	295	302	308	315	322	329
19	195	201	208	214	221	228	235	242	249	256	263	271	278	285	292	300	307	314	321	329	336	343	351	358
20	215	222	229	236	243	250	258	265	273	280	288	295	303	311	318	326	334	341	349	357	365	372	380	388
21	237	243	251	258	266	273	281	289	297	305	313	321	329	337	345	353	362	370	378	386	394	402	411	419
22	259	266	274	282	290	298	306	314	322	331	339	348	356	365	373	382	390	399	408	416	425	433	442	451
23	282	290	298	306	314	323	331	340	349	358	367	375	384	393	402	411	420	429	438	447	456	466	475	484
24	307	314	323	331	340	349	358	367	376	386	395	404	414	423	432	442	451	461	470	480	489	499	508	518
25	332	340	349	358	367	376	386	395	405	415	424	434	444	454	463	473	483	493	503	513	523	533	543	553

.025	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	3	3	3	3	3	3	4	4	4	4	5	5	5	5	5	6	6	6	6	7	7	7	7	7
3	6	6	6	7	8	8	9	9	10	10	11	11	12	12	13	13	14	14	15	15	16	16	17	17
4	10	10	11	12	13	14	15	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	28	29
5	15	16	17	18	19	21	22	23	24	25	27	28	29	30	31	33	34	35	36	38	39	40	41	43
6	21	23	24	25	27	28	30	32	33	35	36	38	39	41	43	44	46	47	49	51	52	54	55	57
7	28	30	32	34	35	37	39	41	43	45	47	49	51	53	55	57	59	61	63	65	67	69	71	73
8	37	39	41	43	45	47	50	52	54	56	59	61	63	66	68	71	73	75	78	80	82	85	87	90
9	46	48	50	53	56	58	61	63	66	69	72	74	77	80	83	85	88	91	94	96	99	102	105	108
10	56	59	61	64	67	70	73	76	79	82	85	89	92	95	98	101	104	108	111	114	117	120	123	127
11	67	70	73	76	80	83	86	90	93	97	100	104	107	111	114	118	122	125	129	132	136	140	143	147
12	80	83	86	90	93	97	101	105	108	112	116	120	124	128	132	136	140	144	148	152	156	160	164	168
13	93	96	100	104	108	112	116	120	125	129	133	137	142	146	151	155	159	164	168	172	177	181	186	190
14	107	111	115	119	123	128	132	137	142	146	151	156	161	165	170	175	180	184	189	194	199	204	208	213
15	122	126	131	135	140	145	150	155	160	165	170	175	180	185	191	196	201	206	211	217	222	227	232	238
16	138	143	148	152	158	163	168	174	179	184	190	196	201	207	212	218	223	229	235	240	246	252	257	263
17	156	160	165	171	176	182	188	193	199	205	211	217	223	229	235	241	247	253	259	265	271	277	283	289
18	174	179	184	190	196	202	208	214	220	227	233	239	246	252	258	265	271	278	284	291	297	304	310	317
19	193	198	204	210	216	223	229	236	243	249	256	263	269	276	283	290	297	304	310	317	324	331	338	345
20	213	219	225	231	238	245	252	259	266	273	280	287	294	301	309	316	323	330	338	345	352	360	367	374
21	235	240	247	254	261	268	275	282	290	297	305	312	320	328	335	343	351	358	366	374	382	389	397	405
22	257	263	270	277	284	292	299	307	315	323	331	339	347	355	363	371	379	387	395	404	412	420	428	436
23	280	286	294	301	309	317	325	333	341	350	358	366	375	383	392	400	409	417	426	434	443	452	460	469
24	304	311	318	326	334	343	351	360	368	377	386	395	403	412	421	430	439	448	457	466	475	484	493	502
25	329	336	344	353	361	370	379	388	397	406	415	424	433	443	452	461	471	480	489	499	508	518	527	537

.01	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	5
3	6	6	6	6	6	7	7	8	8	8	9	9	9	10	10	11	11	11	12	12	12	13	13	14
4	10	10	10	11	12	12	13	14	14	15	16	16	17	18	18	19	20	20	21	22	22	23	24	24
5	15	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
6	21	21	23	24	25	26	28	29	30	31	33	34	35	37	38	40	41	42	44	45	46	48	49	51
7	28	29	30	32	33	35	36	38	40	41	43	45	46	48	50	52	53	55	57	59	60	62	64	65
8	36	37	39	41	43	44	46	48	50	52	54	57	59	61	63	65	67	69	71	73	75	77	79	82
9	45	47	49	51	53	55	57	60	62	64	67	69	72	74	77	79	82	84	86	89	91	94	96	99
10	55	57	59	62	64	67	69	72	75	78	80	83	86	89	92	94	97	100	103	106	109	111	114	117
11	66	68	71	74	76	79	82	85	89	92	95	98	101	104	108	111	114	117	120	124	127	130	133	137
12	78	81	84	87	90	93	96	100	103	107	110	114	117	121	125	128	132	135	139	143	146	150	154	157
13	92	94	97	101	104	108	112	115	119	123	127	131	135	139	143	147	151	155	159	163	167	171	175	179
14	106	108	112	116	119	123	128	132	136	140	144	149	153	157	162	166	171	175	179	184	188	193	197	201
15	121	124	128	132	136	140	145	149	154	158	163	168	172	177	182	187	191	196	201	206	211	215	220	225
16	137	140	144	149	153	158	163	168	173	178	183	188	193	198	203	208	213	219	224	229	234	239	245	250
17	154	158	162	167	172	177	182	187	192	198	203	209	214	220	225	231	236	242	247	253	259	264	270	276
18	172	176	181	186	191	196	202	208	213	219	225	231	237	242	248	254	260	266	272	278	284	290	296	302
19	192	195	200	206	211	217	223	229	235	241	247	254	260	266	273	279	285	292	298	304	311	317	324	330
20	212	216	221	227	233	239	245	251	258	264	271	278	284	291	298	304	311	318	325	332	338	345	352	359
21	233	237	243	249	255	262	268	275	282	289	296	303	310	317	324	331	338	345	353	360	367	374	382	389
22	255	259	265	272	278	285	292	299	307	314	321	329	336	344	351	359	366	374	381	389	397	404	412	420
23	278	283	289	296	303	310	317	325	332	340	348	356	364	371	379	387	395	403	411	419	427	435	444	452
24	302	307	314	321	328	336	343	351	359	367	376	384	392	400	409	417	425	434	442	451	459	468	476	485
25	327	333	339	347	355	362	371	379	387	396	404	413	421	430	439	448	456	465	474	483	492	501	510	518

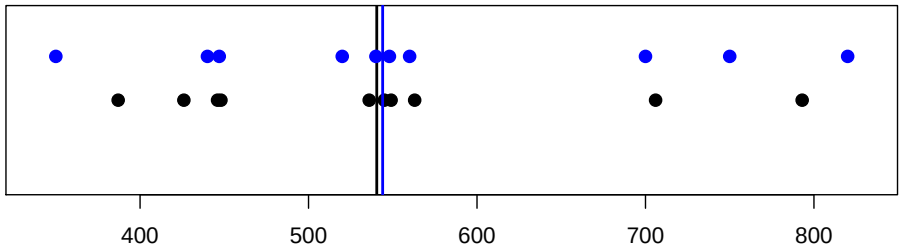
.005	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4
3	6	6	6	6	6	6	6	7	7	7	8	8	8	9	9	9	9	10	10	10	11	11	11	12
4	10	10	10	10	11	11	12	12	13	13	14	14	15	16	16	17	17	18	19	19	20	20	21	21
5	15	15	15	16	17	17	18	19	20	21	22	23	23	24	25	26	27	28	29	30	30	31	32	33
6	21	21	22	23	24	25	26	27	28	29	31	32	33	34	35	37	38	39	40	41	43	44	45	46
7	28	28	29	30	32	33	35	36	38	39	41	42	44	45	47	48	50	51	53	54	56	58	59	61
8	36	36	38	39	41	43	44	46	48	50	52	54	55	57	59	61	63	65	67	69	71	72	74	76
9	45	46	47	49	51	53	55	57	59	62	64	66	68	70	73	75	77	79	82	84	86	89	91	93
10	55	56	58	60	62	65	67	69	72	74	77	80	82	85	87	90	93	95	98	100	103	106	108	111
11	66	67	69	72	74	77	80	83	85	88	91	94	97	100	103	106	109	112	115	118	121	124	127	130
12	78	80	82	85	88	91	94	97	100	103	106	110	113	116	120	123	126	130	133	137	140	143	147	150
13	91	93	95	99	102	105	109	112	116	119	123	126	130	134	137	141	145	149	152	156	160	164	167	171
14	105	107	110	113	117	121	124	128	132	136	140	144	148	152	156	160	164	169	173	177	181	185	189	193
15	120	123	126	129	133	137	141	145	150	154	158	163	167	172	176	181	185	190	194	199	203	208	212	217
16	136	139	142	146	150	155	159	164	168	173	178	182	187	192	197	202	207	211	216	221	226	231	236	241
17	153	156	160	164	169	173	178	183	188	193	198	203	208	214	219	224	229	235	240	245	250	256	261	266
18	171	174	178	183	188	193	198	203	209	214	219	225	230	236	242	247	253	259	264	270	276	281	287	293
19	191	194	198	203	208	213	219	224	230	236	242	248	254	260	265	272	278	284	290	296	302	308	314	320
20	211	214	219	224	229	235	241	247	253	259	265	271	278	284	290	297	303	310	316	323	329	336	342	349
21	232	235	240	246	251	257	264	270	276	283	290	296	303	310	316	323	330	337	344	350	357	364	371	378
22	254	258	263	268	275	281	288	294	301	308	315	322	329	336	343	350	358	365	372	379	387	394	401	409
23	277	281	286	292	299	306	312	320	327	334	341	349	356	364	371	379	386	394	402	409	417	425	432	440
24	301	305	311	317	324	331	338	346	353	361	369	376	384	392	400	408	416	424	432	440	448	456	465	473
25	326	331	336	343	350	358	365	373	381	389	397	405	413	422	430	438	447	455	464	472	481	489	498	506

Dabei: $w_{1-\alpha}(n_1 = n, n_2 = m) = n_1(n_1 + n_2 + 1) - w_\alpha(n_1, n_2)$

Durchschnittsgehälter von Praktikanten in Stadt 1 und Stadt 2:

$$\begin{aligned} X &= (387, 426, 446, 448, 536, 545, 549, 563, 706) \\ Y &= (350, 440, 447, 520, 540, 548, 560, 700, 750) \end{aligned}$$

Unterscheiden sich die Gehälter ($\alpha = 0,05$)?



$rg(X) = (2, 3, 5, 7, 9, 11, 13, 15, 17)$, $rg(Y) = (1, 4, 6, 8, 10, 12, 14, 16, 18)$

$$H_0: \quad med(X) = med(Y)$$

$$w = \sum_{i=1}^{9} rg(X_i) = 82$$

$$w_{0.025}(9,9) = 63, \; w_{1-0.025}(9,9) = 9 \cdot (9 + 9 + 1) - 63 = 108$$

$\Rightarrow$ H_0 kann nicht abgelehnt werden ($\alpha = 0.05$).

Durchschnittsgehälter von Praktikanten in Stadt 1 und Stadt 2:

$$\begin{aligned} X &= (325, 350, 380, 448, 536, 545, 559, 620, 806) \\ Y &= (450, 470, 547, 550, 640, 648, 760, 800, 830) \end{aligned}$$

Sind Absolventengehälter in Stadt 1 geringer als in Stadt 2 ($\alpha = 0,05$)?

8. Verteilungstests

- χ^2-Unabhängigkeitstest
- χ^2-Anpassungstest
- Kolmogorov-Smirnov-Test

Verteilungstests

Bezeichnung für Tests, die einen Verteilungstyp oder eine Verteilungseigenschaft ohne parametrische Annahme prüfen. Hier besprochen:

- ↪ χ^2-Unabhängigkeitstest (Kontingenztest),
  - Test mit zwei gebundenen Stichproben
  - Nullhypothese: Stichproben zu zwei „unabhängigen Grundgesamtheiten“
- ↪ χ^2-Anpassungstest
  - Einstichprobentest
  - Vergleicht diskretes (bzw. diskretisiertes) Merkmal mit einer vorgegebenen endlich-diskreten Verteilung.
- ↪ Kolmogorov-Smirnov-Anpassungstest
  - Einstichprobentest
  - Prüft, ob Daten zu einer vorgegebenen Verteilung passen
  - Vergleicht empirische und theoretische Verteilungsfunktion.

8.1 χ^2-Unabhängigkeitstest (Kontingenztest)

- ↪ Datenbasis ist verbundene u.i.v.-Stichprobe $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- ↪ Es soll geprüft werden, ob X_ℓ, Y_ℓ st.u. sind
- ↪ Kann in dieser Allgemeinheit nicht getestet werden.
- ↪ Vereinfachende Annahmen (falls nicht erfüllt, durch Klassierung erzwingen):
  - □ $X_1, \ldots, X_n$ sind diskrete ZV mit Werten $A_1, \ldots, A_K$
  - □ $Y_1, \ldots, Y_n$ sind diskrete ZV mit Werten $B_1, \ldots, B_L$

1. Hypothese

H_0 : X und Y sind st. unabhängig vs. H_1 : X und Y sind st. abhängig

- ↪ Unter H_0 gilt also

$$P(X_\ell = A_i, Y_\ell = B_j) = P(X_\ell = A_i)P(Y_\ell = B_j) \quad \forall i, j \qquad (*)$$

- ↪ Idee des χ^2-Tests: Stelle den theoretischen WS geeignete empirische Kennzahlen gegenüber und formuliere $(*)$ approximativ mit diesen Kennzahlen.

$\hookrightarrow$ Erstelle **Kontingenztafel** (vgl. DuW)

	B_1	B_2	$\dots$	B_L	Σ
A_1	H_{11}	H_{12}	$\dots$	H_{1L}	$H_{1\bullet}$
A_2	H_{21}	H_{22}	$\dots$	H_{2L}	$H_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_K	H_{K1}	H_{K2}	$\dots$	H_{KL}	$H_{K\bullet}$
Σ	$H_{\bullet 1}$	$H_{\bullet 2}$	$\dots$	$H_{\bullet L}$	n

d.h. $H_{ij} = \sum_{\ell=1}^{n} 1_{\{X_\ell = A_i\}} 1_{\{Y_\ell = B_j\}}, \quad H_{i\bullet} = \sum_{j=1}^{L} H_{ij}, \quad H_{\bullet j} = \sum_{i=1}^{K} H_{ij}$

Beispiel (vgl. DuW) Sichtung von 84 Bewerbungsmappen:

Ab	Zus	Ab	Zus	Ab	Zus	Ab	Zus	Ab	Zus	Ab	Zus	Ab	Zus
Gym	vllt	Real	nein	Gym	vllt	Gym	nein	Real	nein	Ges	vllt	Real	vllt
Real	ja	Gym	vllt	Real	vllt	Gym	nein	Real	ja	Gym	nein	Gym	nein
Real	nein	Ges	ja	Real	vllt	Gym	ja	Gym	nein	Gym	ja	Gym	nein
Gym	vllt	Gym	vllt	Gym	nein	Ges	vllt	Gym	ja	Gym	nein	Gym	vllt
Real	vllt	Real	ja	Real	nein	Gym	vllt	Gym	vllt	Real	vllt	Ges	nein
Gym	vllt	Real	nein	Gym	ja	Gym	nein	Gym	nein	Gym	ja	Gym	nein
Real	vllt	Real	vllt	Real	nein	Real	nein	Ges	vllt	Ges	vllt	Ges	nein
Gym	nein	Gym	vllt	Gym	vllt	Gym	vllt	Gym	nein	Gym	ja	Real	ja
Gym	vllt	Gym	vllt	Ges	ja	Gym	ja	Gym	vllt	Real	vllt	Gym	vllt
Gym	vllt	Gym	ja	Ges	nein	Gym	vllt	Real	vllt	Real	vllt	Real	vllt
Gym	vllt	Ges	vllt	Ges	vllt	Gym	vllt	Real	vllt	Real	nein	Ges	nein
Gym	nein	Gym	nein	Gym	vllt	Gym	vllt	Gym	vllt	Gym	nein	Gym	vllt

	Zusage			
Ab	ja	vllt	nein	$H_{i\bullet}$
Real	4	12	8	24
Gym	8	24	16	48
Ges	2	6	4	12
$H_{\bullet j}$	14	42	28	84

$\hookrightarrow$ Ansatz für eine Teststatistik:

- Unter H_0 gilt: $P(X = A_i, Y = B_j) = P(X = A_i)P(Y = B_j)$
- ML-Schätzungen für die genannten WS sind $\frac{H_{ij}}{n}$, $\frac{H_{i\bullet}}{n}$ und $\frac{H_{\bullet j}}{n}$
- Unter H_0 sollte deshalb für alle Häufigkeiten gelten:

$$\frac{H_{ij}}{n} \approx \frac{H_{i\bullet}}{n} \cdot \frac{H_{\bullet j}}{n} \qquad \Leftrightarrow \qquad H_{ij} \approx \frac{H_{i\bullet} \cdot H_{\bullet j}}{n} := \tilde{H}_{ij}$$

  $\tilde{H}_{ij}$ sind die unter Unabhängigkeit erwarteten Häufigkeiten.
- Abweichungen zwischen H_{ij} und $\tilde{H}_{ij}$ sprechen gegen die Unabhängigkeit.
- Eine Teststatistik sollte die Abweichungen $(H_{ij} - \tilde{H}_{ij})^2$ verwenden.

2. Teststatistik V und Verteilung unter H_0

$$V = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{(H_{ij} - \tilde{H}_{ij})^2}{\tilde{H}_{ij}} = n \cdot \left(\sum_{i=1}^{K} \sum_{j=1}^{L} \frac{H_{ij}^2}{H_{i\bullet} \cdot H_{\bullet j}} - 1 \right) \overset{approx.}{\sim} \chi^2_{(K-1)\cdot(L-1)}$$

$\hookrightarrow$ Faustregel: Die Approximation ist ausreichend genau, wenn $\min(\tilde{H}_{ij}) \geq 5$ (d.h. wenn jede Zelle der Kontingenztafel mindestens 5 Einträge enthält)

Ab	Zusage ja	vllt	nein	$H_{i\bullet}$
Real	4	12	8	24
Gym	8	24	16	48
Ges	2	6	4	12
$H_{\bullet j}$	14	42	28	84

, erwartet:

Ab	Zusage ja	vllt	nein	$H_{i\bullet}$
Real	4	12	8	24
Gym	8	24	16	48
Ges	2	6	4	12
$H_{\bullet j}$	14	42	28	

Approximation hier nicht ausreichend.

$$v = \sum_{i=1}^{K}\sum_{j=1}^{L}\frac{(H_{ij}-\tilde{H}_{ij})^2}{\tilde{H}_{ij}} = \frac{(4-4)^2}{4} + \frac{(12-12)^2}{12} + \cdots + \frac{(4-4)^2}{4} = 0$$

Rechenweg 2:

$$\begin{aligned} v &= n \cdot \left(\sum_{i=1}^{K}\sum_{j=1}^{L}\frac{H_{ij}^2}{H_{i\bullet}\cdot H_{\bullet j}} - 1 \right) \\ &= 84 \cdot \left(\frac{4^2}{24\cdot 14} + \frac{12^2}{24\cdot 42} + \ldots + \frac{4^2}{12\cdot 28} - 1 \right) = 84\cdot(1-1) = 0 \end{aligned}$$

3. Entscheidungsregel (und p-Wert)

Entscheidungsregel:

$$H_0 \text{ wird abgelehnt} \quad \Longleftrightarrow \quad v > \chi^2_{(1-\alpha);(K-1)\cdot(L-1)}$$

p-Wert:

$$p^* = 1 - F_{\chi^2_{(K-1)\cdot(L-1)}}(v)$$

↪ In R: Quantil und p-Wert mit `qchisq(alpha, df = ...)` und `pchisq(v, df = ...)`

- ☐ geg.: zwei Merkmale X und Y, bestehend aus $K = 3$ und $L = 5$ Klassen
- ☐ hierzu ergibt sich der kritische Wert zum Signifikanzniveau $\alpha = 5\%$ durch: `qchisq(1 - 0.05, df = (3 - 1)*(5 - 1))` (Ergebnis: `15.50731`)
- ☐ entsprechend ergibt sich der p-Wert einer realisierten Teststatistik $v = 13$ zu `1 - pchisq(13, df = (3 - 1)*(5 - 1))` (Ergebnis: `0.1118496`)

↪ Durchführung in R: `chisq.test(...)` (auch für Anpassungstest, s.u.)

↪ Test hat i.d.R. geringe Güte.

Beispiel Bewerbungsmappe, $\alpha = 0.05$

	Zusage			
Ab	ja	vllt	nein	$H_{i\bullet}$
Real	4	12	8	24
Gym	8	24	16	48
Ges	2	6	4	12
$H_{\bullet j}$	14	42	28	84

, erwartet:

	Zusage			
Ab	ja	vllt	nein	$H_{i\bullet}$
Real	4	12	8	24
Gym	8	28	16	48
Ges	2	6	4	12
$H_{\bullet j}$	14	42	28	

$$\begin{aligned} v &= 0 \\ \chi^2_{(1-\alpha);(K-1)\cdot(L-1)} &= \chi^2_{0.95;2\cdot 2} = 9.48 > v \end{aligned}$$

Also kann H_0 nicht abgelehnt werden.

Extrembeispiel: Merkmale sind empirisch unabhängig (s. DuW)

Modifiziertes Beispiel „Bewerbungsmappe“

Ab	Zusage ja	vllt	nein	$H_{i\bullet}$
Real	4	12	8	24
Gym	20	24	4	48
Ges	2	6	4	12
$H_{\bullet j}$	26	42	16	84

$$
\begin{aligned}
v &= n \cdot \left(\sum_{i=1}^{K} \sum_{j=1}^{L} \frac{H_{ij}^2}{H_{i\bullet} \cdot H_{\bullet j}} - 1 \right) \\
&= 84 \cdot \left(\frac{4^2}{24 \cdot 26} + \frac{12^2}{24 \cdot 42} + \frac{8^2}{24 \cdot 16} \cdots + \frac{4^2}{12 \cdot 16} - 1 \right) = 84 \cdot (0.13) = 10.9
\end{aligned}
$$

$\hookrightarrow$ $\alpha = 0.05$: $\chi^2_{(1-\alpha);(K-1)\cdot(L-1)} = \chi^2_{0.95;\,4} = 9.48 < v$ $\Rightarrow$ H_0 wird abgelehnt.

$\hookrightarrow$ $\alpha = 0.01$: $\chi^2_{0.99\;4} = 13.48 > v$ $\Rightarrow$ H_0 wird nicht abgelehnt.

$\hookrightarrow$ p-Wert: $p = 1 - F_{\chi^2_4}(10.9) = 0.027$ mit R: `1 - pchisq(10.9, df = 4)`

Für den Zusammenhang zwischen Geschlecht und Note bei einer Mathematikklausur ergibt sich folgende Häufigkeitstabelle:

	Note				
Geschlecht	1	2	3	4	5
m	8	55	87	40	122
w	4	13	42	34	66

Führen Sie den χ^2-Test auf Unabhängigkeit per Hand aus (Signifikanzniveau $\alpha = 0.01$). Verwenden Sie die Quantilstabelle der nachfolgenden Folie.

Quantile der χ^2_n-Verteilung zum Signifikanzniveau α:

	α																				
n	0.005	0.01	0.02	0.025	0.03	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.97	0.975	0.98	0.99	0.995
1	0.000	0.000	0.001	0.001	0.001	0.004	0.016	0.064	0.148	0.275	0.455	0.708	1.074	1.642	2.706	3.841	4.709	5.024	5.412	6.635	7.879
2	0.010	0.020	0.040	0.051	0.061	0.103	0.211	0.446	0.713	1.022	1.386	1.833	2.408	3.219	4.605	5.991	7.013	7.378	7.824	9.210	10.597
3	0.072	0.115	0.185	0.216	0.245	0.352	0.584	1.005	1.424	1.869	2.366	2.946	3.665	4.642	6.251	7.815	8.947	9.348	9.837	11.345	12.838
4	0.207	0.297	0.429	0.484	0.535	0.711	1.064	1.649	2.195	2.753	3.357	4.045	4.878	5.989	7.779	9.488	10.712	11.143	11.668	13.277	14.860
5	0.412	0.554	0.752	0.831	0.903	1.145	1.610	2.343	3.000	3.655	4.351	5.132	6.064	7.289	9.236	11.070	12.375	12.833	13.388	15.086	16.750
6	0.676	0.872	1.134	1.237	1.330	1.635	2.204	3.070	3.828	4.570	5.348	6.211	7.231	8.558	10.645	12.592	13.968	14.449	15.033	16.812	18.548
7	0.989	1.239	1.564	1.690	1.802	2.167	2.833	3.822	4.671	5.493	6.346	7.283	8.383	9.803	12.017	14.067	15.509	16.013	16.622	18.475	20.278
8	1.344	1.646	2.032	2.180	2.310	2.733	3.490	4.594	5.527	6.423	7.344	8.351	9.524	11.030	13.362	15.507	17.010	17.535	18.168	20.090	21.955
9	1.735	2.088	2.532	2.700	2.848	3.325	4.168	5.380	6.393	7.357	8.343	9.414	10.656	12.242	14.684	16.919	18.480	19.023	19.679	21.666	23.589
10	2.156	2.558	3.059	3.247	3.412	3.940	4.865	6.179	7.267	8.295	9.342	10.473	11.781	13.442	15.987	18.307	19.922	20.483	21.161	23.209	25.188
11	2.603	3.053	3.609	3.816	3.997	4.575	5.578	6.989	8.148	9.237	10.341	11.530	12.899	14.631	17.275	19.675	21.342	21.920	22.618	24.725	26.757
12	3.074	3.571	4.178	4.404	4.601	5.226	6.304	7.807	9.034	10.182	11.340	12.584	14.011	15.812	18.549	21.026	22.742	23.337	24.054	26.217	28.300
13	3.565	4.107	4.765	5.009	5.221	5.892	7.042	8.634	9.926	11.129	12.340	13.636	15.119	16.985	19.812	22.362	24.125	24.736	25.472	27.688	29.819
14	4.075	4.660	5.368	5.629	5.856	6.571	7.790	9.467	10.821	12.078	13.339	14.685	16.222	18.151	21.064	23.685	25.493	26.119	26.873	29.141	31.319
15	4.601	5.229	5.985	6.262	6.503	7.261	8.547	10.307	11.721	13.030	14.339	15.733	17.322	19.311	22.307	24.996	26.848	27.488	28.259	30.578	32.801
16	5.142	5.812	6.614	6.908	7.163	7.962	9.312	11.152	12.624	13.983	15.338	16.780	18.418	20.465	23.542	26.296	28.191	28.845	29.633	32.000	34.267
17	5.697	6.408	7.255	7.564	7.832	8.672	10.085	12.002	13.531	14.937	16.338	17.824	19.511	21.615	24.769	27.587	29.523	30.191	30.995	33.409	35.718
18	6.265	7.015	7.906	8.231	8.512	9.390	10.865	12.857	14.440	15.893	17.338	18.868	20.601	22.760	25.989	28.869	30.845	31.526	32.346	34.805	37.156
19	6.844	7.633	8.567	8.907	9.200	10.117	11.651	13.716	15.352	16.850	18.338	19.910	21.689	23.900	27.204	30.144	32.158	32.852	33.687	36.191	38.582
20	7.434	8.260	9.237	9.591	9.897	10.851	12.443	14.578	16.266	17.809	19.337	20.951	22.775	25.038	28.412	31.410	33.462	34.170	35.020	37.566	39.997

8.2 χ^2-Anpassungstest

- ↪ Prüft die Hypothese:
  Ist F_0 Stichprobenverteilung einer u.i.v.-Stichprobe $X_1, \ldots, X_n$?
- ↪ z.B. Tests von Standardzufallszahlen.
- ↪ Test prüft nicht die Ausgangshypothese, sondern eine Vergröberung durch...
- ↪ ...Klassierung des Stichprobenraums, z.B. Intervalle $A_1, \ldots, A_K$,
  Konkretisierung anhand der unter F_0 gegebenen Klassen-W'keiten $p_1, \ldots, p_K$.

1. Hypothese:

$$H_0 : P(X = j) = p_j, \quad \forall j \in \{1, \ldots, K\}$$

- ↪ Vergleich erwarteter Häufigkeiten und beobachteter Häufigkeiten $H(A_j)$

2. Teststatistik und Nullverteilung

$$V = \sum_{j=1}^{K} \frac{(H(A_j) - n \cdot p_j)^2}{n \cdot p_j} \overset{approx}{\sim} \chi^2_{K-1} \text{ für } \min(n \cdot p_1, \ldots, n \cdot p_K) \geq 5$$

χ^2-Anpassungstest

3. Entscheidungsregel (und p-Wert)

Entscheidungsregel (Signifikanzniveau α):

$$H_0 \text{ wird abgelehnt} \quad \Longleftrightarrow \quad v > \chi^2_{(1-\alpha);\, K-1}$$

p-Wert:

$$p^* = 1 - F_{\chi^2_{K-1}}(v)$$

Prüfung d. Kassendaten einer Eisdiele durch Steuerfahndung ($n = 180$)

Für die führende Ziffer[a] wird bei einwandfreier Kassenführung die **Benford-Verteilung** angenommen: $p_j = \log_{10}(j+1) - \log_{10}(j),\ j = 1, 2, \ldots, 9.$

Die Kassendaten einer Eisdiele sollen darauf geprüft werden, ob sie einer Benford-Verteilung genügen:

	j	1	2	3	4	5	6	7	8	9	Σ
Daten	$H(A_j)$	45	36	26	20	17	11	13	6	6	180
theo-	p_j	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	1
retisch	$n \cdot p_j$	54.19	31.70	22.49	17.44	14.25	12.05	10.44	9.21	8.24	180
Test-	$(H(A_j) - n \cdot p_j)^2$	84.37	18.52	12.33	6.53	7.55	1.10	6.56	10.29	5.00	
statistik	$\frac{(H(A_j)-n\cdot p_j)^2}{n\cdot p_j}$	1.56	0.58	0.55	0.37	0.53	0.09	0.63	1.12	0.61	$v = 6.038$

$\hookrightarrow$ Zum Signifikanzniveau $\alpha = 0.05$ gilt $v = 6.038 \leq \chi^2_{0.95;\,8} = 15.50731$,

$\hookrightarrow$ d.h. H_0: „einwandfreie Kassenführung“ kann nicht abgelehnt werden.

$\hookrightarrow$ Der p-Wert beträgt $p^* = 1 - F_{\chi^2_8}(6.038) \approx 0.643 > \alpha$

[a] bspw. hat 13,80 € die führende Ziffer $x = 1$ und -0,00057 die führende Ziffer $x = 5$

Übung: Schreiben Sie eine R-Funktion `leading.digit(x)`, welche aus einer beliebigen Zahl $x \neq 0$ die führende Ziffer der Mantisse berechnet.

Übung: Der Pseudozufallszahlen-Generator des TI-59 war der lineare Kongruenzgenerator $y_{i+1} \equiv (24298 \cdot y_i + 99991) \mod 199017$. Schreiben Sie ein R-Programm, welches mit Startwert $y_0 = 0$ die Pseudozufallszahlen $u_i = \frac{y_i}{199017}$, $i = 1, \ldots, 200$ erzeugt. Prüfen Sie, ob die Annahme einer Gleichverteilung nicht abgelehnt werden kann.

8.3 Kolomogorov-Smirnov Anpassungstest

- $\hookrightarrow$ Anwendung: Tests von Standardzufallszahlen („Diehard-Testbatterie“)
- $\hookrightarrow$ bei kleiner Stichprobe Alternative zum χ^2-Anpassungstest (höhere Güte)
- $\hookrightarrow$ keine Kategorisierung der Daten, dafür nur Vergleich mit stetiger Verteilung(sfunktion)

Gegeben eine u.i.v.-Stichprobe $X_1, \ldots, X_n$ mit (stetiger) Verteilungsfunktion F und angeordneter Stichprobe $X_{(1)} < \cdots < X_{(n)}$ (d.h. keine Bindungen)

1. Hypothese

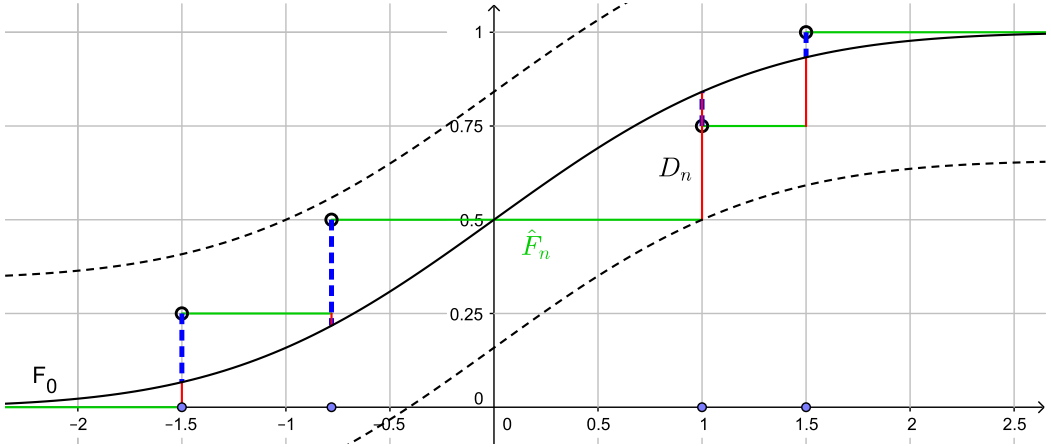
H_0: $F(x) = F_0(x)\ \forall x$ mit vorgegebener VF F_0.

2.1 Teststatistik

- ↪ basiert auf der empirischen VF $\hat{F}_n(x)$ der Realisationen $x_1, \ldots, x_n$
  $\hat{F}_n(x)$ ist der Anteil der Beobachtungen, die höchstens den Wert x aufweisen
- ↪ die Teststatistik misst den größten vertikalen Abstand zwischen hypothetischer und empirischer Verteilungsfunktion
  - □ $D_n = \sup_{x \in \mathbb{R}} |F_0(x) - \hat{F}_n(x)| = \max_{i \in \{0,1\},\, j \in \{1,\ldots,n\}} \left\{ \left| F_0(X_{(j)}) - \frac{j-i}{n} \right| \right\}$
  - □ $V = \sqrt{n} \cdot D_n$

Illustration $F_0(x) = \Phi(x)$

j	x_j	$F_0(x_j)$	$\left\|F_0(x_j) - \frac{j-1}{4}\right\|$	$\left\|F_0(x_j) - \frac{j}{4}\right\|$
1	-1.50	0.07	0.07	0.18
2	-0.78	0.22	0.03	0.28
3	1.00	0.84	$D_4 = 0.34$	0.09
4	1.50	0.93	0.18	0.07



2.2 Verteilung F_V von V unter $H_0 : F = F_0$

- $\hookrightarrow$ für n klein: numerische Berechnung von F_V nach Durbin
- $\hookrightarrow$ Approximativ ($n > 40$): $$P(V \leq t) = 1 + 2 \cdot \sum_{j=1}^{\infty} (-1)^j \cdot \exp(-2 \cdot j^2 \cdot t^2)$$
- $\hookrightarrow$ Quantile $d_{n,\alpha}$ durch numerische Inversion der VF.

Tabelle der Quantile (bei fehlendem Eintrag nächstgrößerer Stichprobenumfang):

α	n												
	5	6	7	8	9	10	15	20	25	30	35	40	> 40
0.90	1.14	1.15	1.15	1.16	1.16	1.17	1.18	1.18	1.19	1.19	1.19	1.20	1.22
0.95	1.26	1.27	1.28	1.28	1.29	1.29	1.31	1.32	1.32	1.32	1.33	1.33	1.36
0.99	1.49	1.51	1.52	1.53	1.54	1.55	1.57	1.58	1.58	1.59	1.59	1.59	1.63

(nach Durbin)

3. Entscheidungsregel

Entscheidungsregel (Signifikanzniveau α):

$$H_0 \text{ wird abgelehnt} \quad \Longleftrightarrow \quad v \geq d_{n,1-\alpha}$$

in R: `ks.test(x, y = "p<distr>",...)`

dabei ist `p<distr>` die zu testende VF

```
ks.test(x = rnorm(50), y = "pgamma", shape = 3, rate = 2)
```

testet, ob ein Vektor bestehend aus 50 standardnormalverteilten Beobachtungen einer Gamma-Verteilung mit Parametern `shape = 3` und `rate = 2` folgt

Bsp: Test auf Gleichverteilung ($\alpha = 0.05$, $n = 10$)

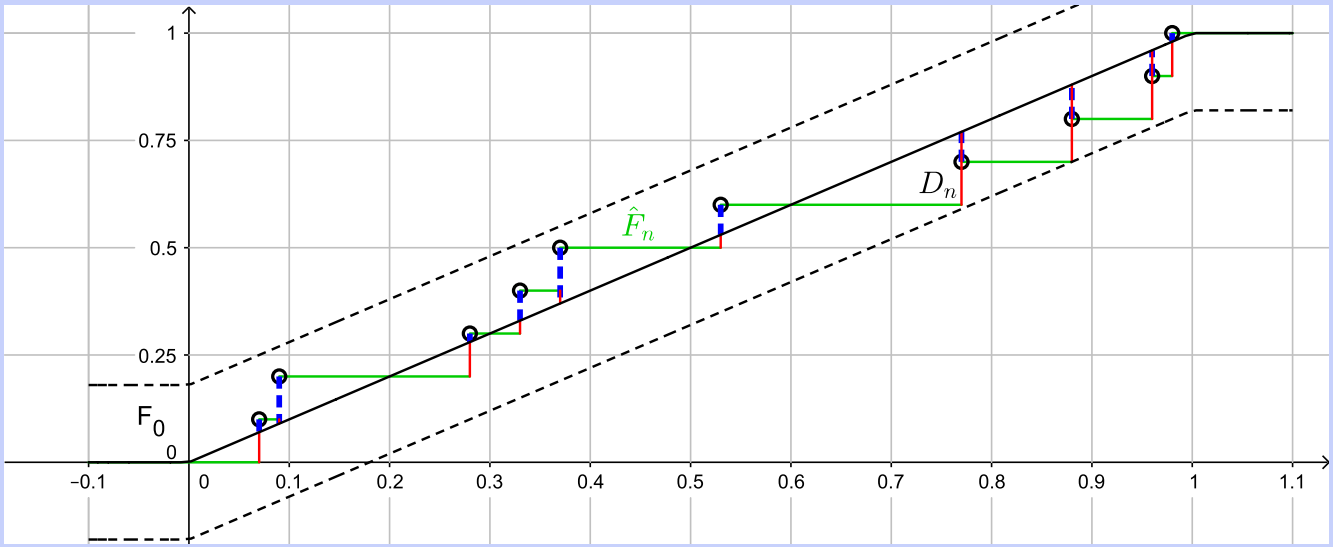
$H_0 : F_0 = Re(0,1)$ vs. $H_1 : F_0 \neq Re(0,1)$ (unter H_0 gilt $F_0(x) = x$)

j	1	2	3	4	5	6	7	8	9	10
$\left\lvert F_0(x_j) - \frac{j-1}{n} \right\rvert$	0.07	0.01	0.08	0.03	0.03	0.03	0.17	0.18	0.16	0.08
$\frac{j-1}{n}$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
x_j	0.07	0.09	0.28	0.33	0.37	0.53	0.77	0.88	0.96	0.98
$\frac{j}{n}$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$\left\lvert F_0(x_j) - \frac{j}{n} \right\rvert$	0.03	0.11	0.02	0.07	0.13	0.07	0.07	0.08	0.06	0.02

in R:

```
x = c(0.07, 0.09, 0.28, 0.33, 0.37 0.53, 0.77, 0.88, 0.96, 0.98)
ks.test(x, "punif", min = 0, max = 1, alternative = "two.sided")
```

j	1	2	3	4	5	6	7	8	9	10
$\left\lvert F_0(x_j) - \frac{j-1}{n} \right\rvert$	0.07	0.01	0.08	0.03	0.03	0.03	0.17	0.18	0.16	0.08
$\frac{j-1}{n}$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
x_j	0.07	0.09	0.28	0.33	0.37	0.53	0.77	0.88	0.96	0.98
$\frac{j}{n}$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$\left\lvert F_0(x_j) - \frac{j}{n} \right\rvert$	0.03	0.11	0.02	0.07	0.13	0.07	0.07	0.08	0.06	0.02



$\sqrt{10} \cdot D_n = \sqrt{10} \cdot 0.18 = 0.5691 < 1.29 = d_{10,\ 0.95}$

$\Rightarrow$ Nichtablehnung von H_0

Gegeben sind die folgenden 5 Beobachtungen: -1.0, -0.2, 0.45, 1.05, 1.69.
Testen Sie, ob die Daten einer Normalverteilung mit $\mu = 0.5$ und $\sigma = 1$ folgen!

Nutzen Sie die folgende Tabelle zur Bestimmung von $F_0(x)$
(wählen Sie μ geeignet; $\sigma = 1$ ist bereits vorgegeben):

	μ								
x	-1.000	-0.950	-0.500	-0.050	0.000	0.050	0.500	0.950	1.000
-1.00	0.500	0.480	0.309	0.171	0.159	0.147	0.067	0.026	0.023
-0.20	0.788	0.773	0.618	0.440	0.421	0.401	0.242	0.125	0.115
0.45	0.926	0.919	0.829	0.691	0.674	0.655	0.480	0.309	0.291
1.05	0.980	0.977	0.939	0.864	0.853	0.841	0.709	0.540	0.520
1.69	0.996	0.996	0.986	0.959	0.954	0.949	0.883	0.770	0.755

Gegeben sind die folgenden 5 Beobachtungen: -1.0, -0.2, 0.45, 1.05, 1.69.

	μ								
x	-1.000	-0.950	-0.500	-0.050	0.000	0.050	0.500	0.950	1.000
-1.00	0.500	0.480	0.309	0.171	0.159	0.147	0.067	0.026	0.023
-0.20	0.788	0.773	0.618	0.440	0.421	0.401	0.242	0.125	0.115
0.45	0.926	0.919	0.829	0.691	0.674	0.655	0.480	0.309	0.291
1.05	0.980	0.977	0.939	0.864	0.853	0.841	0.709	0.540	0.520
1.69	0.996	0.996	0.986	0.959	0.954	0.949	0.883	0.770	0.755

Bestimmen Sie zunächst die Realisation der zugehörigen Teststatistik!

Gegeben sind die folgenden 5 Beobachtungen: -1.0, -0.2, 0.45, 1.05, 1.69.

	μ								
x	-1.000	-0.950	-0.500	-0.050	0.000	0.050	0.500	0.950	1.000
-1.00	0.500	0.480	0.309	0.171	0.159	0.147	0.067	0.026	0.023
-0.20	0.788	0.773	0.618	0.440	0.421	0.401	0.242	0.125	0.115
0.45	0.926	0.919	0.829	0.691	0.674	0.655	0.480	0.309	0.291
1.05	0.980	0.977	0.939	0.864	0.853	0.841	0.709	0.540	0.520
1.69	0.996	0.996	0.986	0.959	0.954	0.949	0.883	0.770	0.755

Entscheiden Sie mit Hilfe der kritischen Werte nach Durbin, ob H_0 zu einem Signifikanzniveau von $\alpha = 5\%$ verworfen werden kann!

Spezialfall für den Test auf Normalverteilung:

- ↪ für fixe Parameter μ_0, σ_0^2: $F_0(x) = \Phi\left(\frac{x-\mu_0}{\sigma_0}\right)$
- ↪ für nicht spezifierte Parameter ist der KS-Test nicht direkt anwendbar (hält das Signifikanzniveau strikt ein, lehnt H_0 zu selten ab: „zu konservativ“)
- ↪ stattdessen Vorgehensweise mit modifizierten kritischen Werten:
  - □ die Parameter μ, σ werden aus der Stichprobe geschätzt; die Verteilung hängt nicht mehr von den unbekannten Parametern ab, unterscheidet sich jedoch von der ursprünglichen Verteilung von D_n.

$$\sqrt{n} \cdot D_n^N = \sqrt{n} \cdot \sup_x \left| \hat{F}_n(x) - \Phi\left(\frac{x-\bar{x}}{s}\right) \right|$$

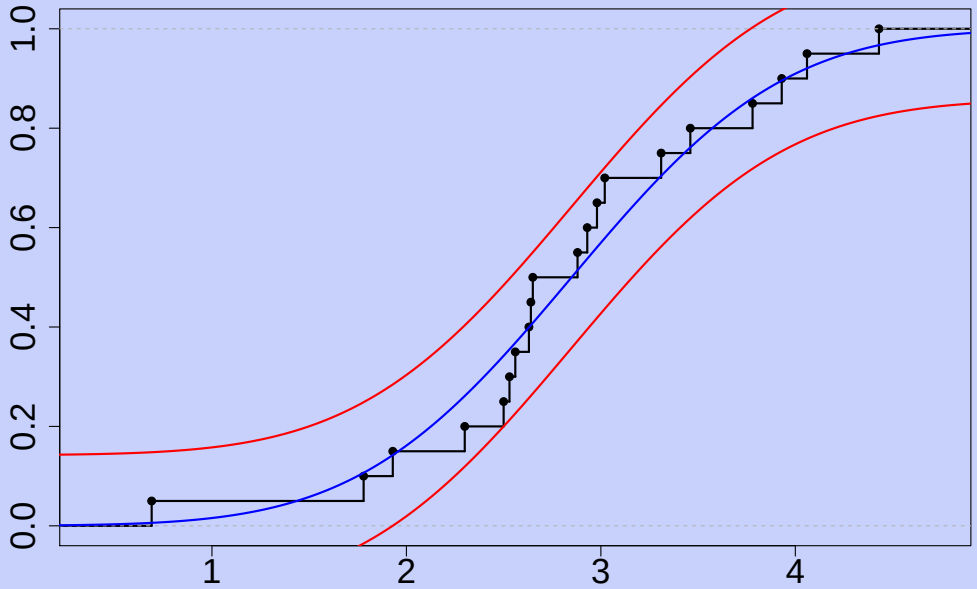
  - □ H_0 wird abgelehnt, falls $\sqrt{n} \cdot D_n^N \geq d_{n,(1-\alpha)}^N$
  - □ kritische Werte $d_{n,\alpha}^N$ aus Simulationsstudien (Lilliefors):

	n					
α	5	8	10	20	30	> 30
0.90	0.72	0.74	0.76	0.79	0.80	0.81
0.95	0.76	0.81	0.82	0.85	0.88	0.89
0.99	0.91	0.94	0.94	1.03	1.03	1.04

Bsp: Test auf Normalverteilung

$x = (0.69, 1.78, 1.93, 2.30, 2.50, 2.53, 2.56, 2.63, 2.64, 2.65, 2.88, 2.93, 2.98, 3.02, 3.31, 3.46, 3.78, 3.93, 4.06, 4.43)$

$$H_0 : F_0 = \mathcal{N}(\bar{x}, \sigma_x^2) \quad \text{vs.} \quad H_1 : F_0 \neq \mathcal{N}(\bar{x}, \sigma_x^2)$$



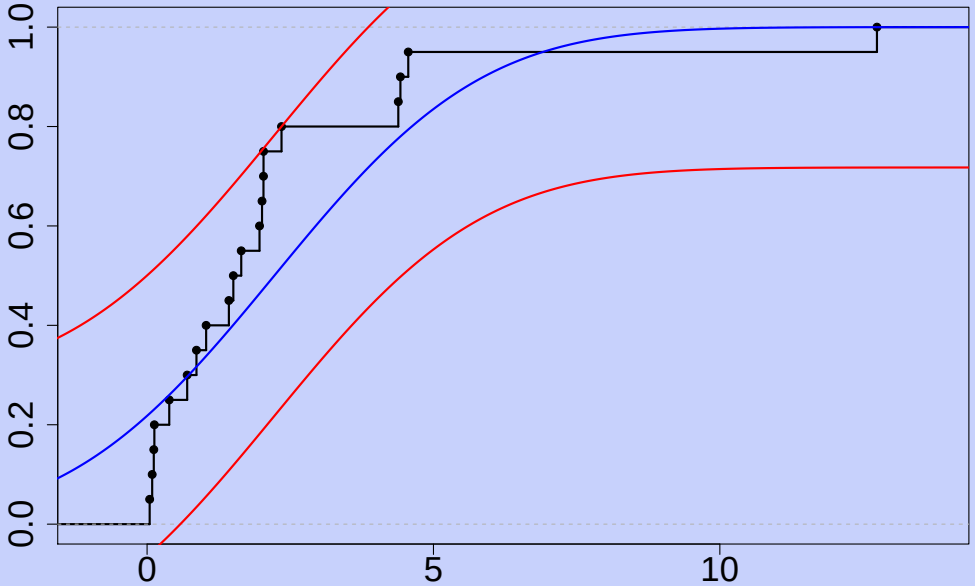
$\sqrt{20} \cdot D_n^N = \sqrt{20} \cdot 0.142 = 0.636 < 0.85 = d_{20,0.95}^N$

$\Rightarrow H_0$ kann nicht abgelehnt werden.

Bsp: Test auf Normalverteilung (II)

$x = (0.045, 0.089, 0.117, 0.127, 0.387, 0.700, 0.862, 1.031, 1.428, 1.506, 1.644, 1.963, 2.007, 2.032, 2.033, 2.347, 4.386, 4.423, 4.560, 12.743)$

$$H_0 : F_0 = \mathcal{N}(\bar{x}, \sigma_x^2) \quad \text{vs.} \quad H_1 : F_0 \neq \mathcal{N}(\bar{x}, \sigma_x^2)$$



$\sqrt{20} \cdot D_n^N = \sqrt{20} \cdot 0.2824 = 1.263 > d_{20,0.95}^N = 0.85$

$\Rightarrow H_0$ wird abgelehnt.

9. Regressionsanalyse

- Einfache lineare Regression
- Multiple lineare Regression
- Hypothesentests in der linearen Regression
  - Varianzzerlegung und Bestimmtheitsmaß
  - F-Test auf Modellgüte
  - F-Test auf Einfluss von Parametergruppen
  - Tests für die Regressionskoeffizienten
  - Konfidenz- und Prognoseintervalle
- Überprüfung der Modellannahmen
  - Motivation
  - Residuenanalyse
- Indikatorvariablen

9 Regressionsanalyse [1,4,5]

Ziel: Erklärung/Prognose von Variablen durch andere Variablen

- **Regressand** Y: zu erklärende, abhängige Zufallsvariable
- **Regressoren** $X_1, \ldots, X_k$: erklärende, unabhängige Zufallsvariablen

Beispiele

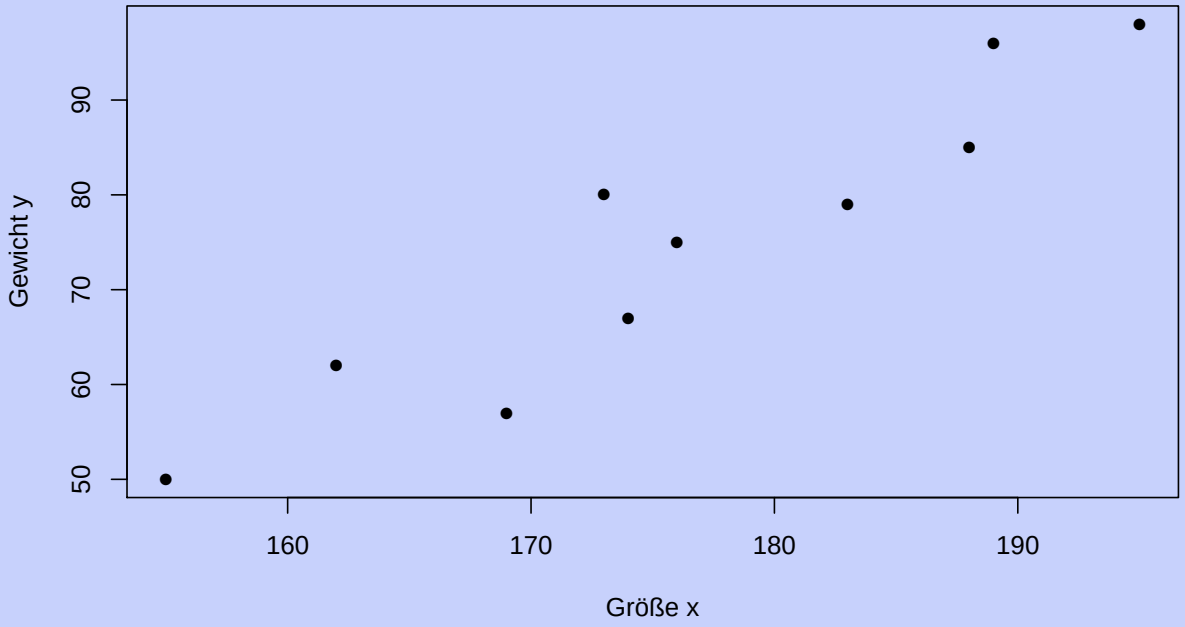
- ↪ Mietspiegel: Y=qm-Kaltmiete, X_1=Wohnfläche, X_2= Lage, X_3=Baujahr,...
- ↪ Gebrauchtwagen: Y =Marktwert, X_1 =Alter, X_2 =TKM, X_3 =Zustand,...
- ↪ Paket-Service: Y =Tagesfahrtzeit eines Wagens, X_1=Zahl der Anlieferungen, X_2=Routenlänge, X_3=Zielregion...
- ↪ Kunden-Umsätze: Y=Wert der Aufträge, X_1=Einkommen, X_2=Familienstand, X_3=Berufsgruppe, X_4=Anzahl der Aufträge,...

In der Praxis liegen Datensätze für diese Variablen/Merkmale vor.

Beispiel 1: Größe und Gewicht

Von 10 Personen wurden Körpergröße x und Körpergewicht y gemessen:

	A	B	C
1	Person	x	y
2	1	162	62
3	2	173	80
4	3	169	57
5	4	189	96
6	5	176	75
7	6	188	85
8	7	155	50
9	8	174	67
10	9	183	79
11	10	195	98



Lässt sich ein rechnerischer Zusammenhang zwischen x,y finden, der die Daten „gut“ beschreibt? Wie lässt sich das Gewicht einer Person „vorhersagen“, welche 181 cm groß ist?

Beispiel 2: Lieferzeitenanalyse (aus [4])

Ein Unternehmen füllt Getränke-/Süßigkeitenautomaten nach. In der Vergangenheit wurden einige Liefervorgänge protokolliert:

Y: Lieferzeit (Minuten), `delTime`

X_1: Anzahl nachzufüllender Produkte, `n.prod`

X_2: Fußweg (ft), `distance`

	n.prod	distance	delTime		n.prod	distance	delTime		n.prod	distance	delTime
1	7	560	16.68	9	30	1460	79.24	17	6	200	15.35
2	3	220	11.50	10	5	605	21.50	18	7	132	19.00
3	3	340	12.03	11	16	688	40.33	19	3	36	9.50
4	4	80	14.88	12	10	215	21.00	20	17	770	35.10
5	6	150	13.75	13	4	255	13.50	21	10	140	17.90
6	7	330	18.11	14	6	462	19.75	22	26	810	52.32
7	2	110	8.00	15	9	448	24.00	23	9	450	18.75
8	7	210	17.83	16	10	776	29.00	24	8	635	19.83
								25	4	150	10.75

Für die Tourenplanung wird eine Formel gesucht, mit der die Lieferzeit möglichst gut anhand eines konkreten Auftrags prognostiziert werden kann.

Datenquelle in R: `library(robustbase),data(delivery)`

Beispiel 3: Lebensdauer eines Werkzeugs (aus [5])

Für 20 Werkzeuge wurden gemessen:

- Y: Lebensdauer des Schneidwerkzeuges
- X_1: Umdrehungen pro Minute
- X_2: Werkzeugtyp

	y	x1	x2		y	x1	x2
1	18.73	610	A	11	30.16	670	B
2	14.52	950	A	12	27.09	770	B
3	17.43	720	A	13	25.40	880	B
4	14.54	840	A	14	26.05	1000	B
5	13.44	980	A	15	33.49	760	B
6	24.39	530	A	16	35.62	590	B
7	13.34	680	A	17	26.07	910	B
8	22.71	540	A	18	36.78	650	B
9	12.68	890	A	19	34.95	810	B
10	19.32	730	A	20	43.67	500	B

Wie wirken sich Drehzahl und Werkzeugtyp auf die Lebensdauer aus?

Ziel: Erklärung/Prognose von Variablen durch andere Variablen

- **Regressand** Y: zu erklärende, abhängige Zufallsvariable
- **Regressoren** $X_1, \ldots, X_k$: erklärende, unabhängige Zufallsvariablen

$\hookrightarrow$ Kausalmodell $Y = f(X_1, \ldots, X_n)$:

- Dabei liegt f in vorgegebener „passender“ Funktionsklasse $\mathcal{F}$, z.B. $f(x_1, \ldots, x_k) = a_0 + a_1 x_1 + \cdots + a_k x_k$
- Dieser Idealfall ist bei stochastischen Daten unrealistisch.

$\hookrightarrow$ Annahme: „gestörtes“ Kausalmodell $Y = f(X_1, \ldots, X_k) + \epsilon$ (f, ϵ unbekannt).

$\hookrightarrow$ Ziel: Finde „optimales“ f in Funktionsklasse $\mathcal{F}$.

$\hookrightarrow$ Lösung mit der KQ-Methode:

- theoretisch: $E((Y - f(X_1, \ldots, X_k))^2) \stackrel{!}{=} \min\limits_{f \in \mathcal{F}}$
- empirisch: $\sum\limits_{i=1}^{n} (y_i - f(x_{i1}, \ldots, x_{ik}))^2 \stackrel{!}{=} \min\limits_{f \in \mathcal{F}}$ mit Daten x_{ij}, y_i

Memo: (Regeln für Erwartungswerte, Varianzen und Kovarianzen)

$$
\begin{aligned}
(*) \quad & E(X^2) = var(X) + (E(X))^2 \\
(**) \quad & var(X + c) = var(X) \text{ für } c \in \mathbb{R} \\
(***) \quad & var(X + Y) = var(X) + 2cov(X, Y) + var(Y)
\end{aligned}
$$

Beispiel: Minimierungsproblem für lineare Funktionen $f(x) = ax + b$

Für bekannte $\mu_1 = E(X), \mu_2 = E(Y), \sigma_1^2 = var(X), \sigma_2^2 = var(Y), \rho = cor(X, Y)$:

$$
\begin{aligned}
E(Y - (aX + b))^2) &= var(Y - (aX + b)) + (E(Y - (aX + b))^2 \\
&= var(Y - aX) + (E(Y) - aE(Y) - b))^2 \\
&= \underbrace{\sigma_2^2 - 2a\,cov(X, Y) + a^2\sigma_1^2}_{\text{min. für } a = \frac{cov(X,Y)}{\sigma_1^2} = \frac{\sigma_2}{\sigma_1}\rho} + \underbrace{(\mu_2 - a\mu_1 - b)^2}_{\text{min. für } b = \mu_2 - a\mu_1}
\end{aligned}
$$

d.h. beste lin. Vorhersage ist $f(X) = \frac{\sigma_2}{\sigma_1}\rho X + (\mu_2 - \frac{\sigma_2}{\sigma_1}\rho\mu_1) = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho(X - \mu_1)$

Ohne Vorgaben an f: Problem wird durch bedingten Erwartungswert gelöst.

Übung: Löse das Minimierungsproblem $E((Y - f(X))^2) \stackrel{!}{=} \min$ für

1. $f(x) = a$, mit $a \in \mathbb{R}$. Minimiere hierzu $E((Y - a)^2) = \ldots$ in a.

2. $f(x) = ax$ mit $a \in \mathbb{R}$. Minimiere hierzu $E((Y - aX)^2) = \ldots$ in a

Memo: Bedingte Verteilung/Bedingter Erwartungswert

- ↪ Hier bivariater Fall: Zufallsvektor (X, Y) mit
  - □ diskreter/stetiger Dichte $f_{X,Y}(x, y)$
  - □ Randdichten $f_X(x)$, $f_Y(y)$.
- ↪ **Bedingte Verteilung** $\mathcal{L}(Y|X = x)$: ist gegeben durch **bedingte Dichte**:

$$f_{Y|X=x}(y) = \begin{cases} f_{X,Y}(x, y)/f_X(x) & \text{falls Nenner} > 0 \\ f_Y(y) & \text{sonst} \end{cases}$$

  (Bayes-Formel) „bedingte WS in y bilden bei festem x eine WS-Verteilung“.

- ↪ **Bedingter Erwartungswert**: $f(x) = E(Y|X = x)$:
  - □ Erwartungswert der bedingten Verteilung $\mathcal{L}(Y|X = x)$.
  - □ Schreibweise für $f(X)$: $E(Y|X) = f(X)$
  - □ Der bed. EW löst das (obige) Problem $E((Y - f(X))^2) \stackrel{!}{=} \min$.
- ↪ Konzept übertragbar auf Zufallsvektoren:
  $\mathcal{L}(Y_1, \ldots, Y_m|X_1 = x_1, \ldots, X_k = x_k)$ bzw. $E(Y|X_1 = x_1, \ldots, X_n = x_n)$

Beispiel (DuW): Bivariate Normalverteilung

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_1}{\sigma_1})^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + (\frac{y-\mu_2}{\sigma_2})^2\right)}$$

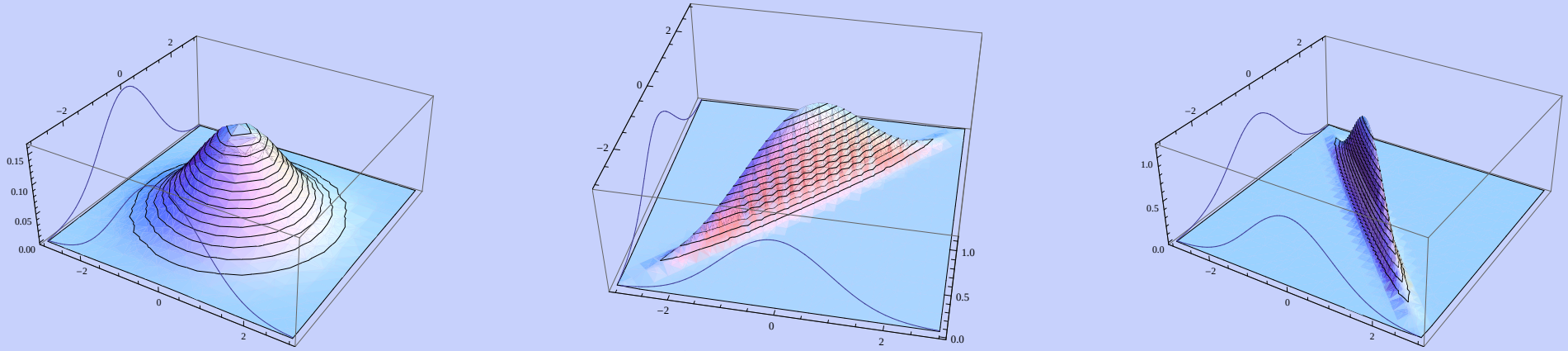
mit $\mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 > 0$, $-1 < \rho < 1$.

$\hookrightarrow$ Randverteilungen sind $\mathcal{L}(X) = \mathcal{N}(\mu_1, \sigma_1^2)$ und $\mathcal{L}(Y) = \mathcal{N}(\mu_2, \sigma_2^2)$

$\hookrightarrow$ ρ ist Pearson-Korrelation von X, Y.

Illustration: `https://www.geogebra.org/m/whgyqhmf`

Für $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$ und $\rho = 0/\rho = 0.99/\rho = -0.99$



Jeweils: $f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)})$ und $\mathcal{N}(0,1)$-Randverteilungen.

$\hookrightarrow$ Betrachte speziell $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$.Bedingte Dichte ist

$$\begin{aligned}
f_{Y|X=x}(y) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right) \Big/ \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \\
&= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right) - \frac{x^2}{2}\right) \\
&= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2 - (1-\rho^2)x^2\right)\right) \\
&= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(y^2 - 2\rho yx + \rho^2 x^2\right)\right) \\
&= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right)
\end{aligned}$$

d.h. $\mathcal{L}(Y|X = x) = \mathcal{N}(\rho x, 1 - \rho^2)$ und $E(Y|X = x) = \rho x$

$\hookrightarrow$ Für allgemeine μ_i, σ_i^2:

- $\mathcal{L}(Y|X = x) = \mathcal{N}(\mu_2 + \frac{\sigma_2}{\sigma_1} \cdot \rho(x - \mu_1), \sigma_2^2(1 - \rho^2))$
- $E(Y|X = x) = \mu_2 + \frac{\sigma_2}{\sigma_1} \cdot \rho(x - \mu_1)$

„In (multivariaten) Normalverteilungsmodellen ist die beste Prognose linear.“

Memo: Regeln für bedingte Erwartungswerte

- ↪ Linearität: $E(a + bY_1 + cY_2|X) = a + bE(Y_1|X) + cE(Y_2|X)$
- ↪ Totale Wahrscheinlichkeit: $E(Y) = E(E(Y|X))$
- ↪ Faktorisierung: $E(Y \cdot g(X)|X) = g(X) \cdot E(Y|X)$
- ↪ Substituieren/Eliminieren (SE): Wenn X, Y st.u. sind, dann gilt:

  (SE1) $P(h(X, Y) \in B|X = x)) = P(h(x, Y) \in B)$

  (SE2) $E(h(X, Y)|X = x) = E(h(x, Y))$
- ↪ Varianzzerlegung: Mit $var(Y|X = x) = E((Y - E(Y|X = x))^2|X = x) = E(Y^2|X = x) - (E(Y|X = x))^2$ gilt

$$var(Y) = var(E(Y|X)) + E(var(Y|X))$$

  - □ linker Summand: Variabilität in Y erklärt durch X
  - □ rechter Summand: Nicht durch X erklärte Variabilität.

Aussagen jeweils „fast sicher“ und unter Annahme existierender Erwartungswerte

Allgemeines Regressionsmodell mit Fehlerterm

- $\hookrightarrow$ $Y = f(X_1, \ldots, X_k) + \sigma V$, $f \in \mathcal{F}$ (vorgegebene Klasse von Funktionen)
- $\hookrightarrow$ σV: nicht beobachtbarer - Fehler mit $E(V) = 0$, $var(V) = 1$, σ unbekannt.
- $\hookrightarrow$ Grundannahme an den Fehler: $cor(f(X_1, \ldots, X_k), V) = 0$. Damit lassen sich KQ-Schätzer für f, σ gewinnen, aber (noch) keine Bereichsschätzer/Tests.
- $\hookrightarrow$ Stärkere Grundannahme $(X_1, \ldots, X_k)$ und V sind st.u.
  Dann wird $\mathcal{L}(Y|X_1, \ldots, X_k)$ durch $f \in \mathcal{F}$, $\mathcal{L}(V)$ und σ bestimmt:
  $$\begin{aligned} P(Y \leq y | X_1 = x_1, \ldots, X_k = x_k) &\overset{\text{(SE1)}}{=} P(f(x_1, \ldots, x_k) + \sigma V \leq y) \\ &= P(V \leq \frac{y - f(x_1, \ldots, x_k)}{\sigma}) \end{aligned}$$
- $\hookrightarrow$ $\mathcal{L}(X_1, \ldots, X_k)$ und damit $\mathcal{L}(Y, X_1, \ldots, X_k)$ bleiben i.d.R. unspezifiziert. Aussagen (Tests, Schätzer, Bereichsschätzer) werden fast ausschließlich bedingt an den Werten $X_1 = x_1, \ldots, X_k = x_k$ getroffen.
- $\hookrightarrow$ Fast ausschließlich verwendet: klassisches Regressionsmodell: $\mathcal{L}(V) = \mathcal{N}(0, 1)$.

Regressionsmodell zu einer Stichprobe

- ↪ Es liegt eine u.i.v.-Stichprobe $(Y_i, X_{i1}, \ldots, X_{ik})$, $i = 1, \ldots, n$, vor.
- ↪ $Y_i = f(X_{i1}, \ldots, X_{ik}) + \sigma V_i$ mit unbekannter Funktion $f \in \mathcal{F}$, d.h.
  - □ bekannter „Typ“ der Regressionsfunktion, keine vollständige Spezifikation.
  - □ nicht beobachtbarer Fehler $\epsilon_i = \sigma V_i$
- ↪ $E(\epsilon_i)$=0, $var(\epsilon_i) = \sigma^2$ (Homoskedastizität)
- ↪ ggf. Normalverteilungsannahme für ϵ_i.

Statistische Aufgaben im Regressionsmodell

- ↪ Schätzung von f, σ^2, bei parametrischem Typ auch Konfidenzintervalle.
- ↪ Schätzung durch KQ/OLS-Methode: $\sum_{i=1}^{n}(Y_i - f(X_{i1}, \ldots, X_{ik}))^2 \stackrel{!}{=} \min_{f \in \mathcal{F}}$
- ↪ Unter Normalverteilungsannahme:
  - □ Schätzung durch ML-Methode
  - □ Prüfung von Hypothesen zu f.

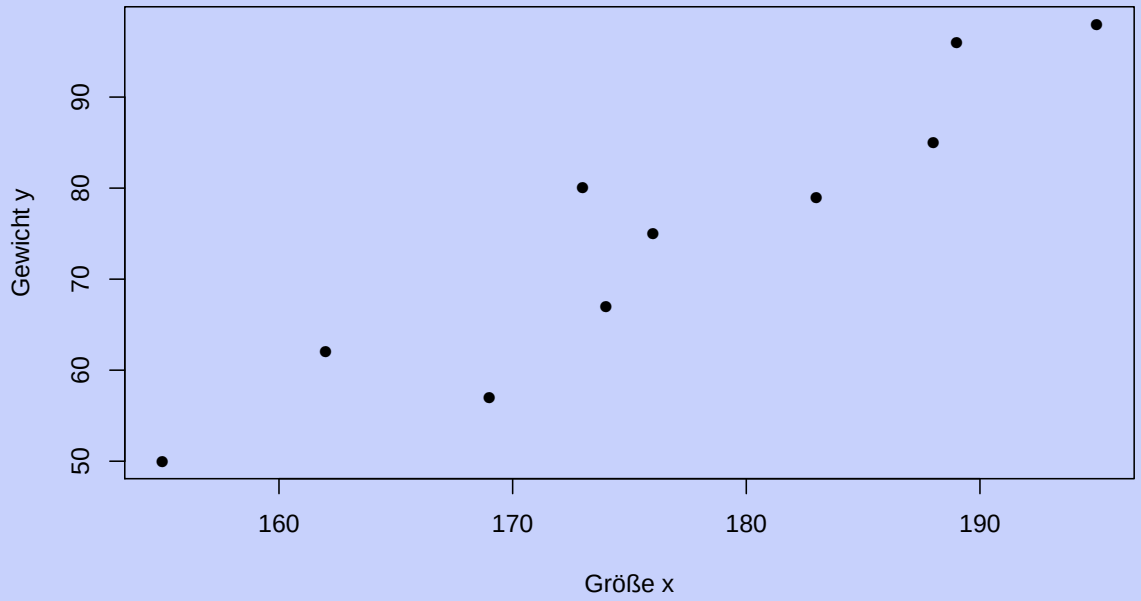
9.1 Einfache lineare Regression

Ein Regressor, f linear, $f(x) = \beta_0 + \beta_1 x$ mit $\beta_0, \beta_1 \in \mathbb{R}$

$\hookrightarrow$ $Y = \beta_0 + \beta_1 X + \epsilon$ mit st.u. ϵ, X

- $E(Y|X = x) = \beta_0 + \beta_1 x$,
- $var(Y|X = x) = var(\beta_0 + \beta_1 x + \epsilon) = var(\epsilon) = \sigma^2$

Größe und Gewicht



	A	B	C	D	E
1	beta	30	0,1		
2	Person	const	x	y	ydach
3	1	1	162	62	46,2
4	2	1	173	80	47,3
5	3	1	169	57	46,9
6	4	1	189	96	48,9
7	5	1	176	75	47,6
8	6	1	188	85	48,8
9	7	1	155	50	45,5
10	8	1	174	67	47,4
11	9	1	183	79	48,3
12	10	1	195	98	49,5

Gesucht: Eine Funktion $f(x) = \beta_0 + \beta_1 x$, welche die Daten möglichst gut beschreibt.

$\hookrightarrow$ Schätzung von β_0, β_1 anhand u.i.v.-Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$. Dabei mindestens zwei verschiedene x_i angenommen.

$\hookrightarrow$ **KQ-Methode**: $K(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \stackrel{!}{=} \min_{\beta_0, \beta_1}$. Notwendig: $\nabla K = \bar{0}$

$$\left.\begin{aligned} &\text{(I)} \quad \frac{\partial K}{\partial \beta_0} = -2n(\bar{y} - \beta_0 - \beta_1 \bar{x}) \stackrel{!}{=} 0 \\ &\text{(II)} \quad \frac{\partial K}{\partial \beta_1} = -2\Big(\sum_{i=1}^{n} x_i y_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2\Big) \stackrel{!}{=} 0 \end{aligned}\right\} \quad \textbf{(Normalgleichungen)}$$

$\hookrightarrow$ Auflösen von (I) nach β_0 und Einsetzen in (II)

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_i y_i - n(\bar{y} - \beta_1 \bar{x})\bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - \beta_1 \Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\Big) \end{aligned} \quad \Rightarrow \quad \begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{cov(x,y)}{s_x^2} = \frac{s_y}{s_x} r_{xy} \end{aligned}$$

$\hookrightarrow$ Prognostizierte Werte: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hookrightarrow$ $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, $i = 1, \ldots, n$ heißen **Residuen**. Nutzen:

- $\square$ Erwartungstreue KQ-Schätzung von σ^2 durch $\frac{1}{n-2} \sum e_i^2$
- $\square$ Überprüfung der Modellgüte durch $R^2 = 1 - \sum e_i^2 / \sum (y_i - \bar{y})^2$.

	x	y	x^2	xy	y^2	$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$	$e = y - \hat{y}$	e^2
1	162	62	26244	10044	3844	57.74	4.26	18.13
2	173	80	29929	13840	6400	70.85	9.15	83.74
3	169	57	28561	9633	3249	66.08	-9.08	82.50
4	189	96	35721	18144	9216	89.91	6.09	37.05
5	176	75	30976	13200	5625	74.42	0.58	0.33
6	188	85	35344	15980	7225	88.72	-3.72	13.85
7	155	50	24025	7750	2500	49.40	0.60	0.36
8	174	67	30276	11658	4489	72.04	-5.04	25.41
9	183	79	33489	14457	6241	82.76	-3.76	14.17
10	195	98	38025	19110	9604	97.06	0.94	0.88
Σ	1764	749	312590	133816	58393	749.00	0.00	$SS_{Res} = 276.41$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{133816 - 10 \times 176.4 \times 74.9}{312590 - 10 \times 176.4^2} = \frac{1692.4}{1420.4} \approx 1.1915$$

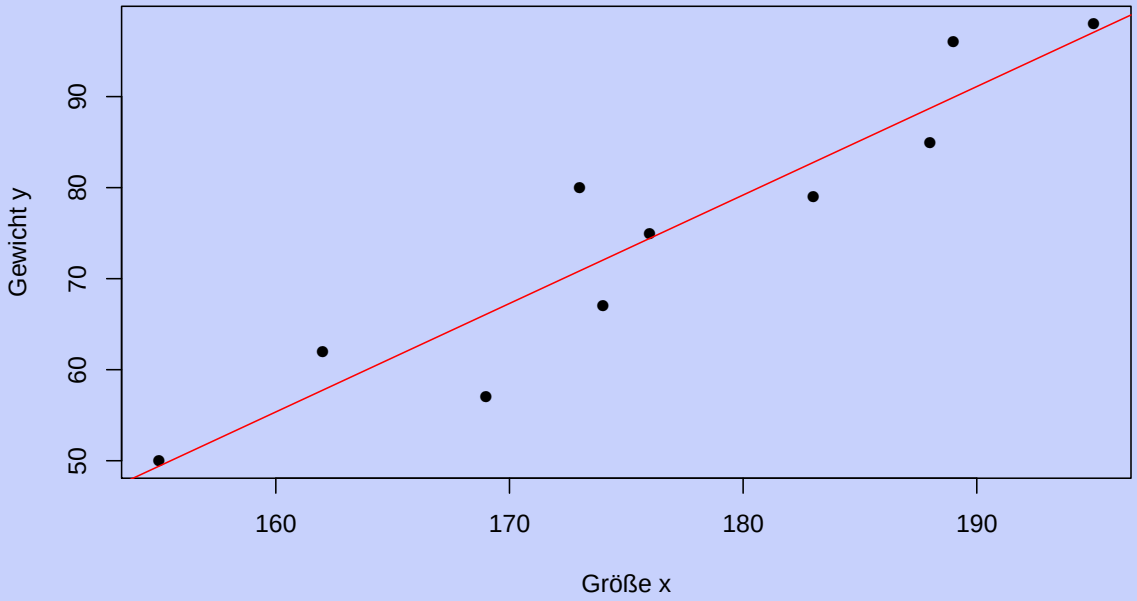
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 74.9 - 1.1915 \cdot 176.4 \approx -135.2798$$

$$SS_T = \Sigma y_i^2 - n\bar{y}^2 = 58393 - 10 \times 74.9^2 = 2292.9$$

$$R^2 = 1 - SS_{Res}/SS_T = 1 - \frac{276.41}{2292.9} \approx 0.879$$

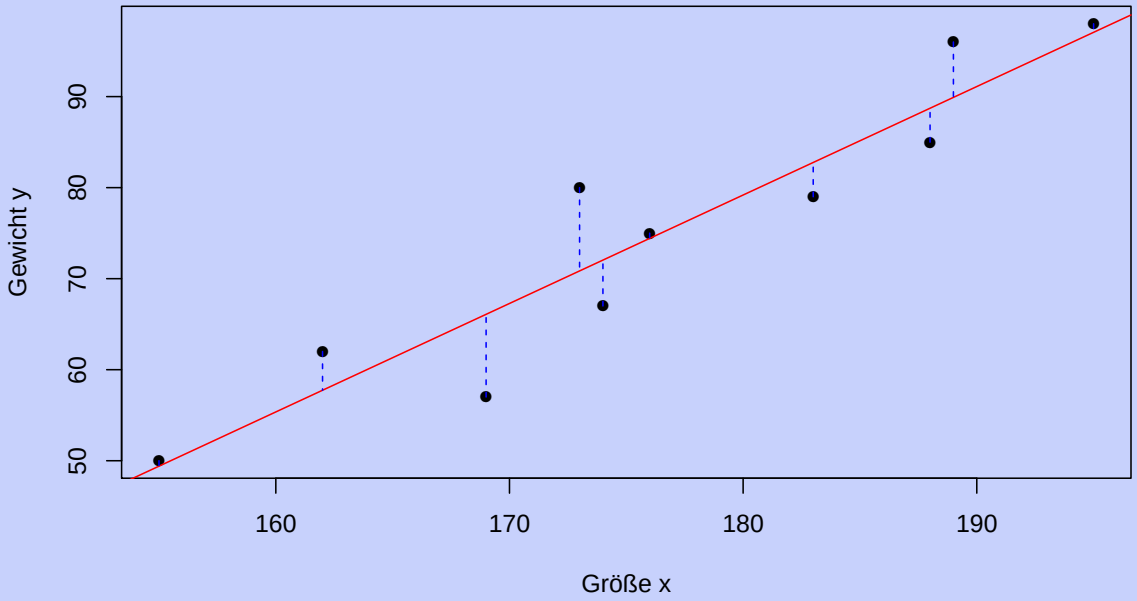
R^2 besagt, dass sich etwa 87,9% der Gewichts-Variabilität durch die (unterschiedlichen) Größen erklären lassen, dazu später mehr.

Größe: $x =$	162	173	169	189	176	188	155	174	183	195
Gewicht: $y =$	62	80	57	96	75	85	50	67	79	98



$$\hat{y} = -135.28 + 1.19 \cdot x$$

Größe: $x =$	162	173	169	189	176	188	155	174	183	195
Gewicht: $y =$	62	80	57	96	75	85	50	67	79	98



$$\hat{y} = -135.28 + 1.19 \cdot x, \; e = y - \hat{y}$$

Umsetzung in R:

```
persdata=data.frame(
  x=c(162,173,169,189,176,188,155,174,183,195),
  y=c(62,80,57,96,75,85,50,67,79,98)
)
#pdf(file="LM_Bsp3.pdf",width=8,height=5)
lm.persdata<-lm(y~x,data=persdata) #die Funktion zur Regression.
    Standardausgabe: Koeffizienten
plot(persdata,xlab="Groesse x",ylab="Gewicht y",main=NULL,pch=16)
abline(lm.persdata,col="red") # setzt die Koeffizienten in die
    Regressionsgerade um

sapply(1:NROW(persdata),function(i){lines(rep(persdata$x[i],2),c(
    persdata$y[i],predict(lm.persdata)[i]),col="blue",lty="dashed")
    }) # zeichnet die Residuen ein
#dev.off()
```

Übung: Der Inhaber einer Kette von 5 freien Tankstellen möchte wissen, ob und wie sich der Tagesgewinn aus den Kraftstoffumsätzen erklären lässt. An einem Tag beobachtet er folgende Werte (in €):

Ums. K. x	6000	2500	8500	6500	9500
Gewinn y	3000	4000	2000	3000	3500

Führe eine einfache lineare Regression durch.

$\hookrightarrow$ Vergleich der empirischen und theoretischen Lösungen

- empirisch: $\sum_{i=1}^{n}(y_i - (ax_i + b))^2 \stackrel{!}{=} \min_{a,b}$

$$\begin{aligned} a &= \frac{s_y}{s_x} r_{xy} \\ b &= \bar{y} - a\bar{x} \end{aligned}$$

- theoretisch (s.o.): $E((Y - (aX + b))^2) \stackrel{!}{=} \min_{a,b}$

$$\begin{aligned} a &= \frac{\sigma_Y}{\sigma_X} \rho_{XY} \\ b &= E(Y) - aE(X) \end{aligned}$$

- Unter Normalverteilungsannahme (s.o.) $E((Y - f(X))^2) \stackrel{!}{=} \min_{f}$

$f(x) = E(Y|X = x) = ax + b$ mit

$$\begin{aligned} a &= \frac{\sigma_Y}{\sigma_X} \rho_{XY} \\ b &= E(Y) - aE(X) \end{aligned}$$

$\hookrightarrow$ In allen Fällen ergibt sich dieselbe Lösung (im empirischen Fall mit empirischen Kennzahlen zu Lage, Streuung, Zusammenhang).

$\hookrightarrow$ Die empirische Lösung ergibt sich aus der theoretischen, wenn man als Modell die Stichprobenverteilung der y_i, x_i annimmt.

Übung: Oben wurde die Funktion $K(\beta_0, \beta_1) = \sum(y_i - \beta_0 - \beta_1 x_i)^2$ minimiert. Dabei war (I) $\frac{\partial K}{\partial \beta_0} = -2n(\bar{y} - \beta_0 - \beta_1 \bar{x})$ (II) $\frac{\partial K}{\partial \beta_1} = -2(\sum x_i y_i - n\beta_0 \bar{x} - \beta_1 \sum x_i^2)$
Prüfe mittels Hesse-Matrix, ob die berechnete Lösung minimal ist.

9.2 Multiple Lineare Regression

↪ Mehrere Regressoren : $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$

↪ Zusammenfassung in Matrizenschreibweise:

$$
\begin{aligned}
\mathrm{Y} &= (Y_1, \ldots, Y_n)^T \\
&= \beta_0 \mathrm{X}_0 + \beta_1 \mathrm{X}_1 + \ldots + \beta_k \mathrm{X}_\mathrm{k} + \varepsilon \\
&= \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
\boldsymbol{\beta} &= (\beta_0, \beta_1, \ldots, \beta_k)^T \\
\boldsymbol{\varepsilon} &= (\varepsilon_1, \ldots, \varepsilon_n)^T
\end{aligned}
$$

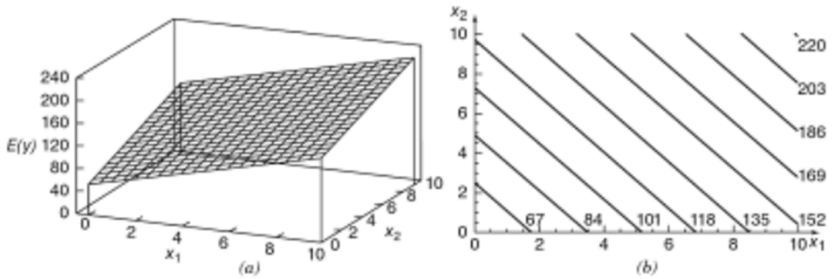
	A	B	C	D	E	F
1	beta	2,000	1,000	0,100		
2	delivery	const	n.prod	distance	delTime	hat_deltime
3	1	1	7	560	16,68	65
4	2	1	3	220	11,5	27
5	3	1	3	340	12,03	39
6	4	1	4	80	14,88	14
7	5	1	6	150	13,75	23
23	21	1	10	140	17,9	26
24	22	1	26	810	52,32	109
25	23	1	9	450	18,75	56
26	24	1	8	635	19,83	73,5
27	25	1	4	150	10,75	21

$$
\mathrm{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix}, \quad \mathrm{X}_0 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \mathrm{X}_\mathrm{j} = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{pmatrix}, 1 \leq j \leq k
$$

↪ Umfasst auch den Spezialfall der einfachen linearen Regression.

↪ Modell beschreibt Hyperebene im Raum der erklärenden Variablen.

$$E(Y|X_1, X_2) = 50 + 20X_1 + 7X_2$$

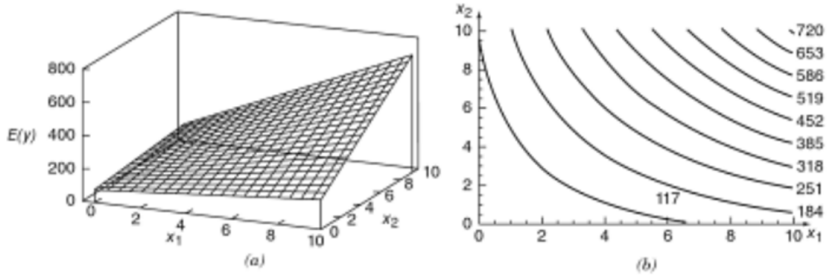


↪ β_i: Änderung in $E(Y)$ bei Erhöhung von X_i um 1 Einheit (ceteris paribus)

↪ Auch polynomielle Zusammenhänge und Interaktionen zwischen Variablen lassen sich abbilden, solange die Funktion linear in den Parametern ist, z.B.

1) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \varepsilon$
2) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \ldots + \varepsilon$

$$E(Y|X_1, X_2) = 50 + 20 \cdot X_1 + 7 \cdot X_2 + 5 \cdot X_1 X_2$$



Parameterschätzung wieder mit KQ-Methode

Mit $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})$, d.h. $\boldsymbol{\varepsilon} = \mathrm{Y} - \boldsymbol{X}\boldsymbol{\beta}$

Minimiere in $\beta = (\beta_0, \ldots, \beta_k)$: $K(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$

$$\begin{aligned}
&= (\mathrm{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}) \\
&= \mathrm{Y}^T\mathrm{Y} - \boldsymbol{\beta}^T\mathrm{X}^T\mathrm{Y} - \mathrm{Y}^T\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathrm{X}^T\mathrm{X}\boldsymbol{\beta} \\
&= \mathrm{Y}^T\mathrm{Y} - \boldsymbol{\beta}^T\mathrm{X}^T\mathrm{Y} - \boldsymbol{\beta}^T\mathrm{X}^T\mathrm{Y} + \boldsymbol{\beta}^T\mathrm{X}^T\mathrm{X}\boldsymbol{\beta} \\
&= \mathrm{Y}^T\mathrm{Y} - 2\boldsymbol{\beta}^T\mathrm{X}^T\mathrm{Y} + \boldsymbol{\beta}^T\mathrm{X}^T\mathrm{X}\boldsymbol{\beta}
\end{aligned}$$

Das führt auf die sog. Normalgleichungen:

$$\frac{\partial K}{\partial \boldsymbol{\beta}} = -2\mathrm{X}^T\mathrm{Y} + 2\mathrm{X}^T\mathrm{X}\boldsymbol{\beta} \stackrel{!}{=} \bar{0} \quad \Leftrightarrow \quad \mathrm{X}^T\mathrm{X}\boldsymbol{\beta} = \mathrm{X}^T\mathrm{Y}$$

$\Rightarrow$ KQ-Schätzer ist $\hat{\boldsymbol{\beta}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{Y}$, falls $(\mathrm{X}^T\mathrm{X})^{-1}$ existiert.

Optimierung mit Excel (Matrixfunktionen)

	A	B	C	D	E	F	G	H	I	J
1	beta	2,341	1,616	0,014			R^2	0,9596		
2	delivery	const	n.prod	distance	delTime	hat_deltime	residuals	SS_Res	SS_T	SS_R
3	1	1	7	560	16,68	21,708	-5,0281	25,282	32,536	0,457
4	2	1	3	220	11,5	10,354	1,1464	1,314	118,461	144,730
5	3	1	3	340	12,03	12,080	-0,0498	0,002	107,205	106,177
6	4	1	4	80	14,88	9,956	4,9244	24,249	56,310	154,464
7	5	1	6	150	13,75	14,194	-0,4444	0,197	74,546	67,070
23	21	1	10	140	17,9	20,514	-2,6142	6,834	20,106	3,496
24	22	1	26	810	52,32	56,007	-3,6865	13,590	896,164	1130,474
25	23	1	9	450	18,75	23,358	-4,6076	21,230	13,206	0,948
26	24	1	8	635	19,83	24,403	-4,5729	20,911	6,523	4,076
27	25	1	4	150	10,75	10,963	-0,2126	0,045	135,350	130,449
28	sum	25	219	10232	559,6	559,6	0,0000	233,732	5784,543	5550,811
29		X^T*X			X^T*Y					
30	normal	25	219	10232	559,6					
31	equations	219	3055	133899	7375,44					
32		10232	133899	6725688	337071,69					

B30:E32 =MMULT(MTRANS(B3:D27);B3:E27)

B1:D1 =MTRANS(MMULT(MINV(B30:D32);E30:E32))

F3:F27 =MMULT(B3:D27;MTRANS(B1:D1))

Lieferzeitenanalyse für Getränke-/Süßigkeitenautomaten

Y: Lieferzeit (Minuten), `delTime`

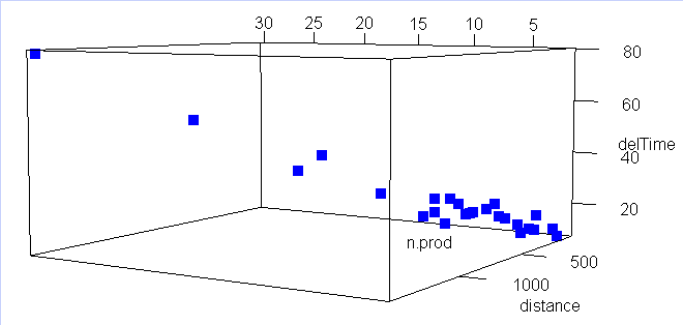
X_1: Anzahl nachzufüllender Produkte, `n.prod`

X_2: Fußweg (ft), `distance`

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

	n.prod	distance	delTime		n.prod	distance	delTime
1	7	560	16.68	14	6	462	19.75
2	3	220	11.50	15	9	448	24.00
3	3	340	12.03	16	10	776	29.00
4	4	80	14.88	17	6	200	15.35
5	6	150	13.75	18	7	132	19.00
6	7	330	18.11	19	3	36	9.50
7	2	110	8.00	20	17	770	35.10
8	7	210	17.83	21	10	140	17.90
9	30	1460	79.24	22	26	810	52.32
10	5	605	21.50	23	9	450	18.75
11	16	688	40.33	24	8	635	19.83
12	10	215	21.00	25	4	150	10.75
13	4	255	13.50				

$$\mathrm{X} = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ \vdots & \vdots & \vdots \\ 1 & 8 & 635 \\ 1 & 4 & 150 \end{bmatrix}, \quad \mathrm{Y} = \begin{bmatrix} 16.68 \\ 11.50 \\ 12.03 \\ \vdots \\ 19.83 \\ 10.75 \end{bmatrix}$$



$$
\begin{aligned}
X^T X &= \begin{bmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{bmatrix} \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix} \\
&= \begin{bmatrix} 25 & 219 & 10.232 \\ 219 & 3.055 & 133.899 \\ 10.232 & 133.899 & 6725.688 \end{bmatrix} \\
X^T Y &= \begin{bmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{bmatrix} \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix} = \begin{bmatrix} 559.60 \\ 7375.44 \\ 337072 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
X^T X &= \begin{bmatrix} 25 & 219 & 10.232 \\ 219 & 3.055 & 133.899 \\ 10.232 & 133.899 & 6725.688 \end{bmatrix} \\
X^T Y &= \begin{bmatrix} 559.60 \\ 7375.44 \\ 337072 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T Y = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2.341 \\ 1.616 \\ 0.014 \end{bmatrix} \\
\hat{y} &= 2.341 + 1.616 \cdot x_1 + 0.014 \cdot x_2
\end{aligned}
$$

Somit beträgt die durchschnittliche Lieferzeit 2.341 Minuten und verlängert sich im Schnitt mit jedem nachzufüllendem Produkt um 1.616 Minuten und mit jedem zusätzlichen Fuß (ft) Wegstrecke um 0.014 Minuten.

R-Code und zugehöriger Output

```
45
46  require(robustbase)
47
48  data(delivery)
49
50  summary(lm.deli <- lm(delTime ~ n.prod+distance, data = delivery))
51  summary(lm.deli <- lm(delTime ~ ., data = delivery))
48:15   (Untitled)
```

Console ~/

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
n.prod      1.615907   0.170735   9.464 3.25e-09 ***
distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

Wie ist der Output zu interpretieren? (siehe Hypothesentests)

Umsatz an einer Tankstelle

Der Inhaber einer Kette von 5 freien Tankstellen möchte wissen, ob und wie sich der Tagesgewinn aus seinen Umsätzen erklären lässt. Hierbei unterteilt er diese in Kraftstoff (K) und in Sonstige (S) Umsätze. Ergebnisse eines Tages: (in €):

Umsatz K x_1	6 000	2 500	8 500	6 500	9 500
Umsatz S x_2	7 000	6 500	3 000	7 000	7 500
Gewinn y	3 000	4 000	2 000	3 000	3 500

Modellieren Sie den Gewinn des Tankstellenbetreibers unter gemeinsamer Berücksichtigung beider Umsatzkategorien!

Eigenschaften der KQ-Schätzer

Es gilt:

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E\left[(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{Y}\right] \\
&= E\left[(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T(\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right] \\
&= E\left[(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{X}\boldsymbol{\beta}+(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\boldsymbol{\varepsilon}\right] \\
&= E\left[\mathrm{I}\boldsymbol{\beta}\right]+(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T E[\boldsymbol{\varepsilon}] \\
&= \boldsymbol{\beta}+0=\boldsymbol{\beta}
\end{aligned}
$$

$\Rightarrow \hat{\boldsymbol{\beta}}$ ist ein <u>erwartungstreuer</u> Schätzer für $\boldsymbol{\beta}$

$\hat{\boldsymbol{\beta}}$ ist <u>bester linearer erwartungstreuer Schätzer</u>, d.h. $\hat{\boldsymbol{\beta}}$ hat die kleinste Varianz in der Klasse aller erwartungstreuen Schätzer, die Linearkombinationen der erklärenden Variablen darstellen (ohne Beweis, [4]).

Eigenschaften der KQ-Schätzer (Forts.)

Weiterhin gilt für die Kovarianzmatrix:

$$
\begin{aligned}
var(\hat{\boldsymbol{\beta}}) &= var\left[(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{Y}\right] \\
&= (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T var(\mathrm{Y})\left[(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\right]^T \\
&= (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T \cdot (\sigma^2\mathrm{I}) \cdot \left[(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\right]^T \\
&= \sigma^2(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1} \\
&= \sigma^2(\mathrm{X}^T\mathrm{X})^{-1} = \sigma^2 C, \qquad \text{mit } C = (\mathrm{X}^T\mathrm{X})^{-1}, \text{ d.h.}
\end{aligned}
$$

$$
\begin{aligned}
\text{d.h. } var(\hat{\beta}_j) &= \sigma^2 C_{jj} \\
cov(\hat{\beta}_i, \hat{\beta}_j) &= \sigma^2 C_{ij}
\end{aligned}
$$

$\rightarrow$ Schätzung von σ?

Hat-Matrix $\mathrm{H} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T \in \mathbb{R}^{n\times n}$

H projiziert Y in vorhergesagte Werte: $\hat{\mathrm{Y}} = \mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{Y} = \mathrm{HY}$,

(Vulgo: „Die Hat-Matrix setzt (der Zielgröße) y den Hut auf.“)

Eigenschaften der Hat-Matrix

- $\mathrm{H}^2 = \mathrm{H}$ und $(\mathrm{I}-\mathrm{H})^2 = \mathrm{I}-\mathrm{H}$ (Übung)
- $tr(\mathrm{H}) = k+1$ (die Spur $tr(H)$ ist die Summe der Diagonalelemente von H)
- $\varepsilon = \mathrm{Y} - \hat{\boldsymbol{Y}} = \mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{Y} - \mathrm{HY} = (\mathrm{I}-\mathrm{H})\mathrm{Y}$

Dazu Eigenschaften der Spur:

- $\hookrightarrow$ $tr(A \pm B) = tr(A) \pm tr(B)$ falls A, B quadratisch sind.
- $\hookrightarrow$ $tr(AB) = tr(BA)$, falls beide Matrixprodukte gebildet werden können und AB, BA quadratisch sind.
- $\hookrightarrow$ $tr(\mathrm{H}) = tr(\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T) = tr(\mathrm{X}^T\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}) = tr(I_{k+1}) = k+1$

Übung: Zeigen Sie:

$$H^2 = H \text{ sowie } (I - H)^2 = I - H.$$

Schätzung von σ^2

$\hat{\sigma}^2 = \dfrac{SS_{res}}{n-p}$ (mit $p = k+1$) ist ein erwartungstreuer Schätzer für σ^2.

$$
\begin{aligned}
SS_{res} &= \varepsilon^T \varepsilon = ((\mathrm{I}-\mathrm{H})\mathrm{Y})^T(\mathrm{I}-\mathrm{H})\mathrm{Y} = \mathrm{Y}^T(\mathrm{I}-\mathrm{H})\mathrm{Y} = \mathrm{Y}^T\mathrm{Y} - \hat{\beta}^T\mathrm{X}^T\mathrm{Y} \\
E(SS_{res}) &= E(\mathrm{Y}^T(\mathrm{I}-\mathrm{H})\mathrm{Y}) \\
&= E((\mathrm{Y}-E(\mathrm{Y}))^T(\mathrm{I}-\mathrm{H})(\mathrm{Y}-E(\mathrm{Y})) \\
&= E(tr((\mathrm{Y}-E(\mathrm{Y}))^T(\mathrm{I}-\mathrm{H})(\mathrm{Y}-E(\mathrm{Y})))) \\
&= E(tr((\mathrm{I}-\mathrm{H})(\mathrm{Y}-E(\mathrm{Y}))(\mathrm{Y}-E(\mathrm{Y}))^T) \\
&= tr((\mathrm{I}-\mathrm{H}) \cdot E((\mathrm{Y}-E(\mathrm{Y}))(\mathrm{Y}-E(\mathrm{Y}))^T)) \\
&= tr((\mathrm{I}-\mathrm{H}) \cdot Cov(Y)) \\
&= tr((\mathrm{I}-\mathrm{H}) \cdot \sigma^2 I_n) \\
&= (n-k-1)\sigma^2
\end{aligned}
$$

Lieferzeitenanalyse

Schätzung der Varianz σ^2 der Residuen.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
n.prod      1.615907   0.170735   9.464 3.25e-09 ***
distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

$$
\begin{aligned}
SS_{res} &= \mathrm{Y}^T\mathrm{Y} - \hat{\beta}^T\mathrm{X}^T\mathrm{Y} \\
\mathrm{Y}^T\mathrm{Y} &= \sum_{i=1}^{25} y_i^2 = 18\,310.63 \\
\hat{\beta}^T\mathrm{X}^T\mathrm{Y} &= [2.341 \quad 1.616 \quad 0.014] \begin{bmatrix} 559.60 \\ 7\,375.44 \\ 337\,072 \end{bmatrix} = 18\,076.9 \\
SS_{res} &= 18\,310.63 - 18\,076.9 = 233.73 \\
\hat{\sigma}^2 &= \frac{SS_{res}}{n-p} = \frac{233.73}{25-3} = 10.62, \quad \hat{\sigma} = 3.259
\end{aligned}
$$

Umsatz an einer Tankstelle

Schätzen Sie die Varianz σ^2 der Residuen im Tankstellen-Beispiel.

Maximum-Likelihood-Schätzer

- die KQ-Schätzer im multiplen linearen Regressionsmodell entsprechen den Maximum-Likelihood-Schätzern bei Gültigkeit der Modellannahmen

$$
\begin{aligned}
\mathrm{Y} &= \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
\boldsymbol{\varepsilon} &\sim \mathcal{N}(0, \sigma^2 \mathrm{I}), \quad \text{d.h.} \\
f(\varepsilon_i) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right)
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right) \\
L(\mathrm{Y}, \mathrm{X}, \boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})^T(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})\right)
\end{aligned}
$$

Log-Likelihood:

$$
\begin{aligned}
\ln L(\mathrm{Y}, \mathrm{X}, \boldsymbol{\beta}, \sigma^2) &= \ln\left(\frac{1}{(2\pi)^{n/2}\sigma^n}\right) - \frac{1}{2\sigma^2}(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})^T(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}) \\
&= 0 - \left(\frac{n}{2}\ln(2\pi) + n\ln(\sigma)\right) - \frac{1}{2\sigma^2}(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})^T(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}) \\
&= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})^T(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})
\end{aligned}
$$

$\Rightarrow$ Für fixes σ wird $\ln L(\mathrm{Y}, \mathrm{X}, \boldsymbol{\beta}, \sigma^2)$ maximal für minimales $(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})^T(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})$

$\Rightarrow$ ML-Schätzer ist $\hat{\boldsymbol{\beta}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{Y}$ (s.o.)

ML-Schätzer für σ^2:

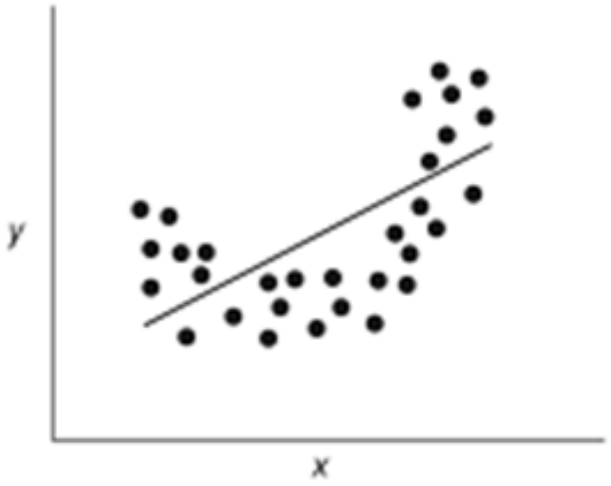
$$\begin{aligned}
\frac{\partial \ln(\mathrm{Y}, \mathrm{X}, \hat{\boldsymbol{\beta}}, \sigma^2)}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})^T(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}}) \stackrel{!}{=} 0 \\
\Leftrightarrow \quad \frac{n}{\sigma} &= \frac{1}{\sigma^3}(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})^T(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}}) \\
\Leftrightarrow \quad n\sigma^2 &= (\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})^T(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})
\end{aligned}$$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})^T(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})}{n} = \frac{SS_{res}}{n}$$

- Der ML-Schätzer für σ^2 ist nicht erwartungstreu.

9.3 Hypothesentests in der linearen Regression

- Wie ist die Modellgüte? Ist das Modell geeignet, um die Zusammenhänge zu beschreiben?
- Von welcher Bedeutung sind die erklärenden Variablen?
- Überprüfung anhand statistischer Hypothesentests
- Annahmen: $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$



Varianzzerlegung

$$SS_T = SS_R + SS_{res} \qquad \text{d.h.}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{d.h.}$$

$$\mathrm{Y}^T\mathrm{Y} - n\bar{\mathrm{Y}}^2 = \hat{\beta}^T\mathrm{X}^T\mathrm{Y} - n\bar{\mathrm{Y}}^2 + \mathrm{Y}^T\mathrm{Y} - \hat{\beta}^T\mathrm{X}^T\mathrm{Y}$$

SS_T : Gesamtvariabilität (Total sum of squares)

SS_R : Variabilität der Regression (sum of sq. due to regression)

SS_{res} : Variabilität der Residuen (residual sum of squares)

Die Varianzzerlegung ist (nur) korrekt, wenn der konstante Effekt β_0 im Modell ist (Eins-Spalte in X)

Beispiel mit Excel (fortges.)

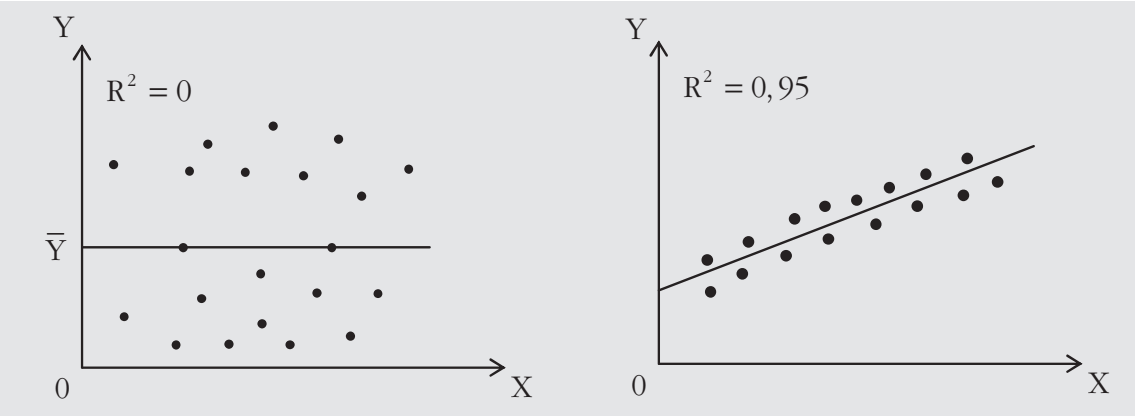
	A	B	C	D	E	F	G	H	I	J
1	beta	2,341	1,616	0,014			R^2	0,9596		
2	delivery	const	n.prod	distance	delTime	hat_deltime	residuals	SS_Res	SS_T	SS_R
3	1	1	7	560	16,68	21,7081	-5,02805188	25,281	32,536	0,457
4	2	1	3	220	11,5	10,3536	1,14636478	1,314	118,461	144,730
5	3	1	3	340	12,03	12,0798	-0,04978497	0,002	107,205	106,177
6	4	1	4	80	14,88	9,95571	4,9242914	24,249	56,310	154,462
7	5	1	6	150	13,75	14,1945	-0,44445881	0,198	74,546	67,069
23	21	1	10	140	17,9	20,5143	-2,61427203	6,834	20,106	3,496
24	22	1	26	810	52,32	56,0066	-3,68657762	13,591	896,164	1130,478
25	23	1	9	450	18,75	23,3576	-4,60757746	21,230	13,206	0,948
26	24	1	8	635	19,83	24,4028	-4,57281023	20,911	6,523	4,076
27	25	1	4	150	10,75	10,9626	-0,21262929	0,045	135,350	130,448
28	sum	25	219	10232	559,6	559,6	-0,00044424	233,732	5784,543	5550,795

g

Bestimmtheitsmaß R^2

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T}, 0 \leq R^2 \leq 1$$

- die Fähigkeit der Regressionsgeraden, die Variabilität in Y zu erklären, ist umso besser, je kleiner die Residualvariabilität im Vergleich zur Gesamtvariabilität ist.



- $R^2 = 0$: Regressionsgerade erklärt nicht besser als der Mittelwert.
- Je größer R^2, desto besser ist der "Fit" der Regressionsgeraden.

Lieferzeitenanalyse

```
58  summary(lm.deli <- lm(delTime ~ n.prod+distance, data = delivery))
```

58:67 (Untitled)

Console C:/DatenM/Repositories/Datenanalyse/R/

```
Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

```
53  summary(lm.deli <- lm(delTime ~ n.prod, data = delivery))
```

53:58 (Untitled)

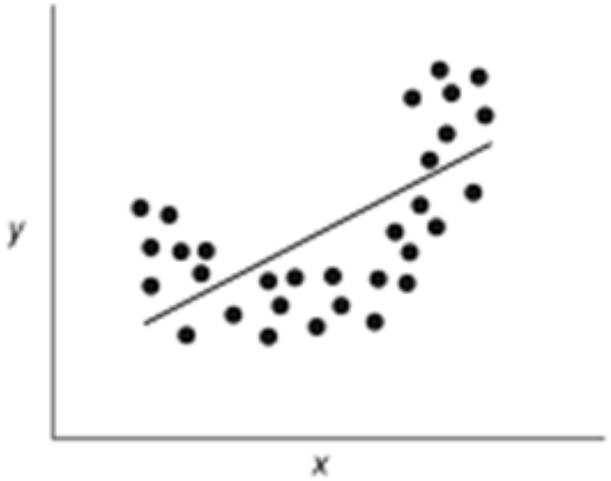
Console C:/DatenM/Repositories/Datenanalyse/R/

```
Residual standard error: 4.181 on 23 degrees of freedom
Multiple R-squared:  0.9305, Adjusted R-squared:  0.9275
F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15
```

- Vorsicht bei der Interpretation von R^2
- R^2 kann sich nicht verschlechtern bei Hinzunahme weiterer erklärender Variablen

Varianzzerlegung und Bestimmtheitsmaß

- Mit weiteren erklärenden Variablen lässt sich R^2 künstlich erhöhen
- Eine Erhöhung von k führt zu einer Verringerung der Freiheitsgrade, d.h. zu ungenaueren Koeffizientenschätzungen
- R^2 kann nicht die Angemessenheit des linearen Modelles beurteilen, d.h. es können nichtlineare Zusammenhänge vorliegen



Adjustiertes Bestimmtheitsmaß $\tilde{R}^2$

- Verwende adjustiertes Bestimmtheitsmaß $\tilde{R}^2$

$$R^2 = 1 - \frac{SS_{res}}{SS_T} \qquad \rightsquigarrow \qquad \tilde{R}^2 = 1 - \frac{SS_{res}/(n-k-1)}{SS_T/(n-1)}$$

- „Bestrafe“ die Hinzunahme weiterer erklärender Variablen.
- Verwendung von erwartungstreuen Schätzern für die wahren Variabilitätswerte.

Lieferzeitenanalyse

```
52
53  summary(lm.deli <- lm(delTime ~ n.prod, data = delivery))
```

53:58 (Untitled)

Console C:/DatenM/Repositories/Datenanalyse/R/

```
Residual standard error: 4.181 on 23 degrees of freedom
Multiple R-squared:  0.9305, Adjusted R-squared:  0.9275
F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15
```

```
57
58  summary(lm.deli <- lm(delTime ~ n.prod+distance, data = delivery))
```

58:67 (Untitled)

Console C:/DatenM/Repositories/Datenanalyse/R/

```
Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

- Hier verbessert das Hinzufügen der zweiten Variablen die Erklärung der Gesamtvariabilität tatsächlich (allerdings nur geringfügig).

Umsatz an einer Tankstelle

Prüfen Sie die Modellgüte im Tankstellen-Beispiel ($\hat{\beta} = (2\,161.9, -0.117, 0.276)^T$)

Umsatz K x_1	6 000	2 500	8 500	6 500	9 500
Umsatz S x_2	7 000	6 500	3 000	7 000	7 500
Gewinn y	3 000	4 000	2 000	3 000	3 500

F-Test auf Modellgüte

Besteht ein linearer Zusammenhang zwischen der abhängigen Variablen und mindestens einer der erklärenden Variablen?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \quad \text{für mind. ein } j$$

Teststatistik/-entscheidung: Es gilt unter H_0

$$\begin{aligned}
SS_T &= SS_R + SS_{res} \quad \text{und } SS_R, SS_{res} \text{ sind st.u.} \\
SS_R/\sigma^2 &\sim \chi^2_k \\
SS_{res}/\sigma^2 &\sim \chi^2_{n-k-1} \\
\Rightarrow F_0 &= \frac{SS_R/k}{SS_{res}/(n-k-1)} = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2} \sim \mathcal{F}_{k,n-k-1}
\end{aligned}$$

H_0 wird abgelehnt, falls $F_0 > \mathcal{F}_{1-\alpha,k,n-k-1}$

	A	B	C	D	E	F	G	H	I	J
1	beta	2,341	1,616	0,014			R^2	0,9596		
2	delivery	const	n.prod	distance	delTime	hat_deltime	residuals	SS_Res	SS_T	SS_R
3	1	1	7	560	16,68	21,7081	-5,02805188	25,281	32,536	0,457
4	2	1	3	220	11,5	10,3536	1,14636478	1,314	118,461	144,730
5	3	1	3	340	12,03	12,0798	-0,04978497	0,002	107,205	106,177
6	4	1	4	80	14,88	9,95571	4,9242914	24,249	56,310	154,462
7	5	1	6	150	13,75	14,1945	-0,44445881	0,198	74,546	67,069
23	21	1	10	140	17,9	20,5143	-2,61427203	6,834	20,106	3,496
24	22	1	26	810	52,32	56,0066	-3,68657762	13,591	896,164	1130,478
25	23	1	9	450	18,75	23,3576	-4,60757746	21,230	13,206	0,948
26	24	1	8	635	19,83	24,4028	-4,57281023	20,911	6,523	4,076
27	25	1	4	150	10,75	10,9626	-0,21262929	0,045	135,350	130,448
28	sum	25	219	10232	559,6	559,6	-0,00044424	233,732	5784,543	5550,795

$$F_0 = \frac{SS_R/k}{SS_{res}(n-k-1)} = \frac{5550.8/2}{233.732/(24-2-1)} \approx 261.2 \approx \frac{22}{2}\frac{0.9596}{1-0.9596} = \frac{n-k-1}{k}\frac{R^2}{1-R^2}$$

```
58  summary(lm.deli <- lm(delTime ~ n.prod+distance, data = delivery))
```

58:67 (Untitled)

Console C:/DatenM/Repositories/Datenanalyse/R/

```
Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

Umsatz an einer Tankstelle

Bestimmen Sie die F-Statistik, sowie den zugehörigen p-Wert zum Tankstellen-Beispiel. Entscheiden Sie anhand Ihrer Ergebnisse, ob ein (linearer) Zusammenhang zwischen (mind.) einem der Umsätze und dem Gewinn nachweisbar ist.

F-Test auf Einfluss von Parametergruppen

Hat ein - festgelegtes - Set von Variablen einen Einfluss auf die abhängige Variable?
Konkret mit $1 \leq i_1 < \cdots < i_m \leq k$:

$$H_0 : \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_m} = 0$$

Teststatistik / Testentscheidung:

- Unter H_0 entfallen m Parameter im Modell, d.h. $\mathrm{Y} = \mathrm{X}_H \beta_H + \varepsilon$, dabei entstehen β_H bzw. X_H durch Streichung der Einträge bzw. Spalten $i_1 + 1, \ldots, i_m + 1$ aus β bzw. X.
  Zugehörige KQ-Schätzung
  $$\hat{\beta}_H = (\mathrm{X}_H^T \mathrm{X}_H)^{-1} \mathrm{X}_H^T \mathrm{Y}, \qquad \hat{\mathrm{Y}}_H = \mathrm{X}_H \hat{\beta}_H$$
- Teststatistik $V = \dfrac{\|\hat{\mathrm{Y}} - \hat{\mathrm{Y}}_H\|^2 / m}{SS_{res}/(n-k-1)} \sim F = \mathcal{F}_{m,n-k-1}$
- H_0 wird abgelehnt, falls $V > \mathcal{F}_{1-\alpha,m,n-k-1}$ (p-value $1 - F(v)$)

Anwendung: Tests für kategorielle Einflussfaktoren (s.u.)

Tests für die Regressionskoeffizienten

- *Welche* Regressionskoeffizienten sind wichtig?

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

Prinzipiell ein Spezialfall des Parametergruppentests (F-Test), aber gleichwertig als t-Test durchführbar.

Teststatistik / Testentscheidung:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (\mathrm{X}^T \mathrm{X})^{-1}_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

- H_0 wird abgelehnt, falls $|t_0| > t_{1-\alpha/2, n-k-1}$
- Test beurteilt den Beitrag von β_j bedingt auf das Vorliegen der anderen Koeffizienten

```r
require(robustbase)

data(delivery)

summary(lm.deli <- lm(delTime ~ n.prod+distance, data = delivery))
summary(lm.deli <- lm(delTime ~ ., data = delivery))
```

48:15 (Untitled)

Console ~/

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
n.prod      1.615907   0.170735   9.464 3.25e-09 ***
distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

Umsatz an einer Tankstelle

Welche der Koeffizienten $\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2\,161.889 & -0.117 & 0.276 \end{pmatrix}^T$ haben einen signifikanten Einfluss ($\alpha = 5\%$) auf das Modell?

Konfidenz- und Prognoseintervalle

Konfidenzintervalle für Regressionskoeffizienten

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}) \text{ und } y_i \sim \mathcal{N}(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2)$$

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 C_{jj}) \text{ und } \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-k-1)$$

- $\hookrightarrow$ $\sqrt{\hat{\sigma}^2 C_{jj}}$ wird von R als „standard error“ angegeben
- $\hookrightarrow$ Nach dem in Abschnitt 5 angegebenen Verfahren (vgl. 5-7) wird damit folgendes KI bestimmt:

$(1-\alpha)$-Konfidenzintervall für den Regressionskoeffizienten β_j

$$\hat{\beta}_j - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 C_{jj}}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
n.prod      1.615907   0.170735   9.464 3.25e-09 ***
distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

$t_{0.975,22} = 2.074$

<u>95%-Konfidenzintervall:</u>

$$1.615907 - 2.074 \cdot 0.170735 \leq \beta_1 \leq 1.615907 + 2.074 \cdot 0.170735$$

$$1.26181 \leq \beta_1 \leq 1.97001$$

Konfidenzintervalle für den erwarteten Wert am Punkt x_0 (im Datensatz vorhanden)

$$
\begin{aligned}
x_0 &:= (1, x_{01}, \ldots, x_{0k}) \\
\hat{y}_0 &= x_0^T \hat{\beta} \\
E[\hat{y}_0] &= E(y|x_0) \\
var(\hat{y}_0) &= \hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0
\end{aligned}
$$

$(1-\alpha)$-Konfidenzintervall für den erwarteten Wert

$$\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0} \leq E(y|x_0) \leq \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
n.prod      1.615907   0.170735   9.464 3.25e-09 ***
distance    0.014385   0.003613   3.981 0.000631 ***
---
```

$t_{0.975,22} = 2.074$

• 95%-Konfidenzintervall für das Geschäft mit 8 nachzufüllenden Produkten und Fußweg von 635 ft (Fall 24)

$$\begin{aligned} x_0 &= (1\ 8\ 635)^T \\ \hat{y}_0 &= x_0^T \hat{\beta} = (1\ 8\ 635) \begin{bmatrix} 2.341231 \\ 1.615907 \\ 0.014385 \end{bmatrix} = 24.40 \text{ Minuten} \end{aligned}$$

```
Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

$$
\begin{aligned}
var(\hat{y}_0) &= \hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0 \\
&= 10.6239 \cdot (1\ 8\ 635) \times (X^T X)^{-1} \times (1\ 8\ 635)^T = 1.2814
\end{aligned}
$$

95%-Konfidenzintervall für die erwartete Lieferzeit:

$$
\begin{aligned}
\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0} \leq \quad & E(y|x_0) \quad \leq \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots} \\
24.40 - 2.074 \cdot \sqrt{1.2814} \leq \quad & E(y|x_0) \quad \leq 24.40 + 2.074 \cdot \sqrt{1.2814} \\
22.06 \leq \quad & E(y|x_0) \quad \leq 26.75
\end{aligned}
$$

95% der so konstruierten Intervalle werden die wahre erwartete Lieferzeit überdecken.

Umsatz an einer Tankstelle

Konstruieren Sie ein 95%-Konfidenzintervall für den erwarteten Gewinn des Tankstellenbetreibers bei einem Kraftstoffumsatz von 7 000€ sowie sonstigen Einnahmen in Höhe von 6 000€.

Prognoseintervall für den Wert y_0 an einem neuen Punkt x_0

- das KI für den erwarteten Wert ist nicht geeignet
- benötigt wird eine Wahrscheinlichkeitsaussage für einen konkreten Wert aus der Verteilung

$$\begin{aligned} \kappa &= y_0 - \hat{y}_0 \sim \mathcal{N}(0, var(\kappa)) \\ var(\kappa) &= \underbrace{var(y_0)}_{=\sigma^2} + \underbrace{var(\hat{y}_0)}_{=\sigma^2 \cdot x_0^T (X^T X)^{-1} x_0} = \sigma^2 (1 + x_0^T (X^T X)^{-1} x_0) \end{aligned}$$

Durch Schätzen von $var(\kappa)$ Übergang zur t-Verteilung:

$$-t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1 + x_0^T (X^T X)^{-1} x_0)} \leq \quad y_0 - \hat{y}_0 \quad \leq t_{1-\alpha/2,n-k-1}\sqrt{\cdots}$$

$$\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1 + x_0^T (X^T X)^{-1} x_0)} \leq \quad y_0 \quad \leq \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots}$$

95%-Prognoseintervall für die Lieferzeit zu einem Geschäft mit 8 nachzufüllenden Produkten und Fußweg von 275 ft.

$$\begin{aligned}
\hat{y}_0 - t_{1-\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2(1+\mathrm{x}_0^T(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{x}_0)} \leq \quad & y_0 \quad \leq \hat{y}_0 + t_{1-\alpha/2,n-k-1}\sqrt{\cdots} \\
19.22 - 2.074 \cdot \sqrt{10.6239(1+0.05346)} \leq \quad & y_0 \quad \leq 19.22 + 2.074 \cdot \sqrt{\cdots} \\
12.28 \leq \quad & y_0 \quad \leq 26.16
\end{aligned}$$

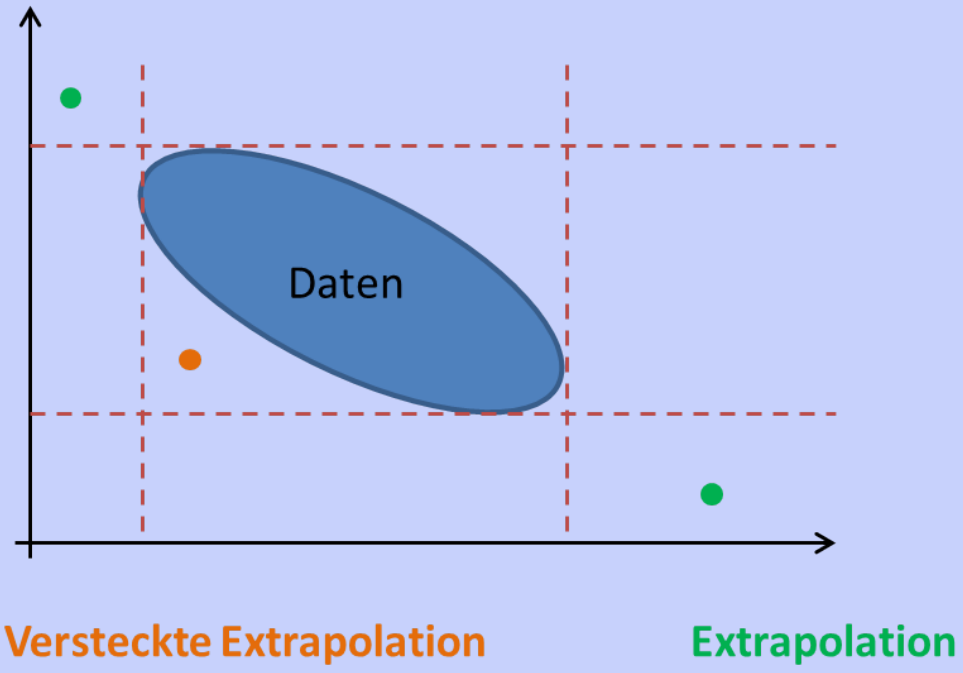
- das Prognoseintervall ist immer breiter als das Konfidenzintervall für den erwarteten Wert
- das PI basiert sowohl auf der Unsicherheit des geschätzten Modells als auch auf der Unsicherheit der wahren zukünftigen Werte

Umsatz an einer Tankstelle

Konstruieren Sie nun ein 95%-Prognoseintervall für den Gewinn des Tankstellenbetreibers gegeben den Umsätzen (7 000€und 6 000€) der vorherigen Übungsaufgabe.

Extrapolation

- Vorsicht bei Extrapolation!



- die Prognosequalität kann außerhalb des Datenbereiches u.U. sehr schlecht sein

Multikollinearität

Vorsicht bei Multikolinearität

- beschreibt lineare bzw. fast lineare Abhängigkeiten zwischen den erklärenden Variablen (d.h. von Spalten der Matrix X)
- exakte lineare Abhängigkeit resultiert in Singularität von $(X^T X)$ $\rightsquigarrow (X^T X)^{-1}$ und Parameterschätzungen können nicht berechnet werden
- Näherungsweise lineare Abhängigkeit resultiert in numerischen Ungenauigkeiten bei der Schätzung
- Solche Modelle haben üblicherweise schlechte Prognosequalität
- Solche Modelle sind sensitiv gegenüber kleinen Änderungen in den erklärenden Variablen
- Einflüsse der Variablen können nicht mehr korrekt unterschieden werden

9.4 Überprüfung der Modellannahmen

Ein lineares Modell kann praktisch immer angepasst werden. Die Frage ist wie gut die Annahmen des Modells – zumindest approximativ – erfüllt sind.

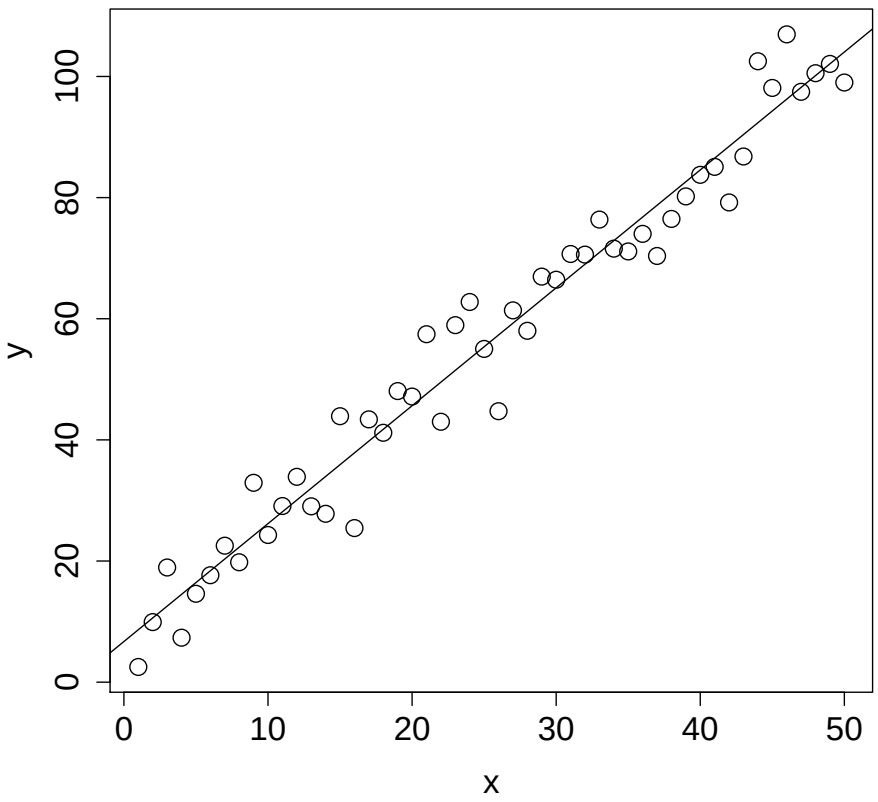
Annahmen:

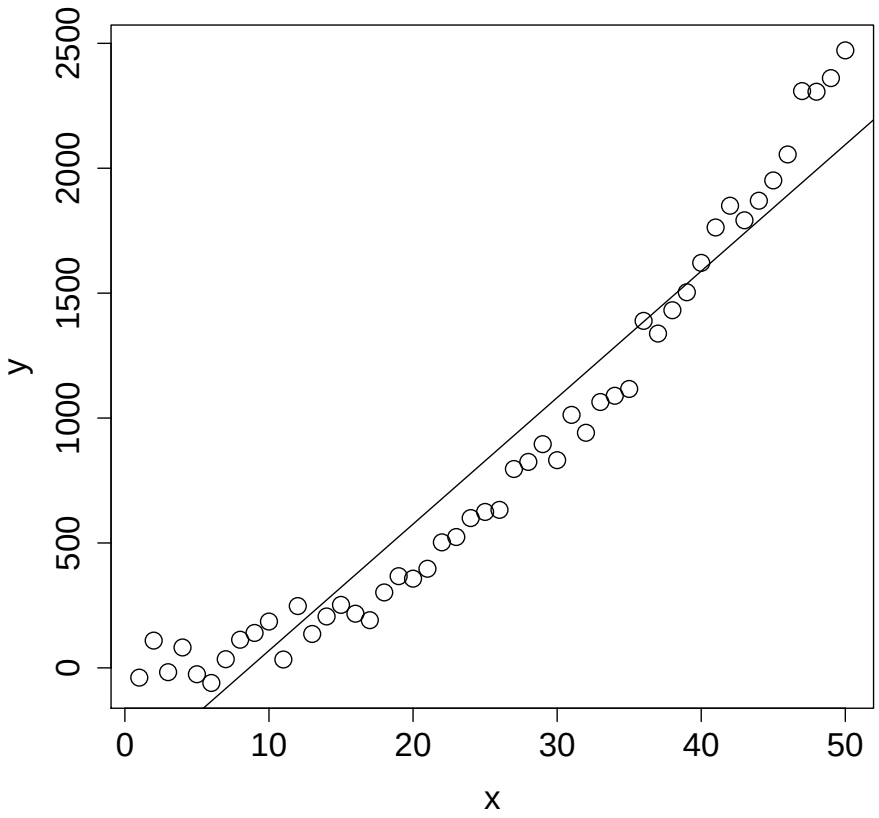
- der Zusammenhang zwischen der abhängigen und den erklärenden Variablen ist linear
- $E(\varepsilon) = 0$
- $var(\varepsilon) = \sigma^2 \mathrm{I}$, d.h.
  $cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j$ und $var(\varepsilon_i) = \sigma^2,\ i = 1, \ldots, n$
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$

Bei Verletzungen der Modellannahmen können instabile Modelle entstehen sowie wahre Zusammenhänge nicht richtig erkannt werden.

Beispiel

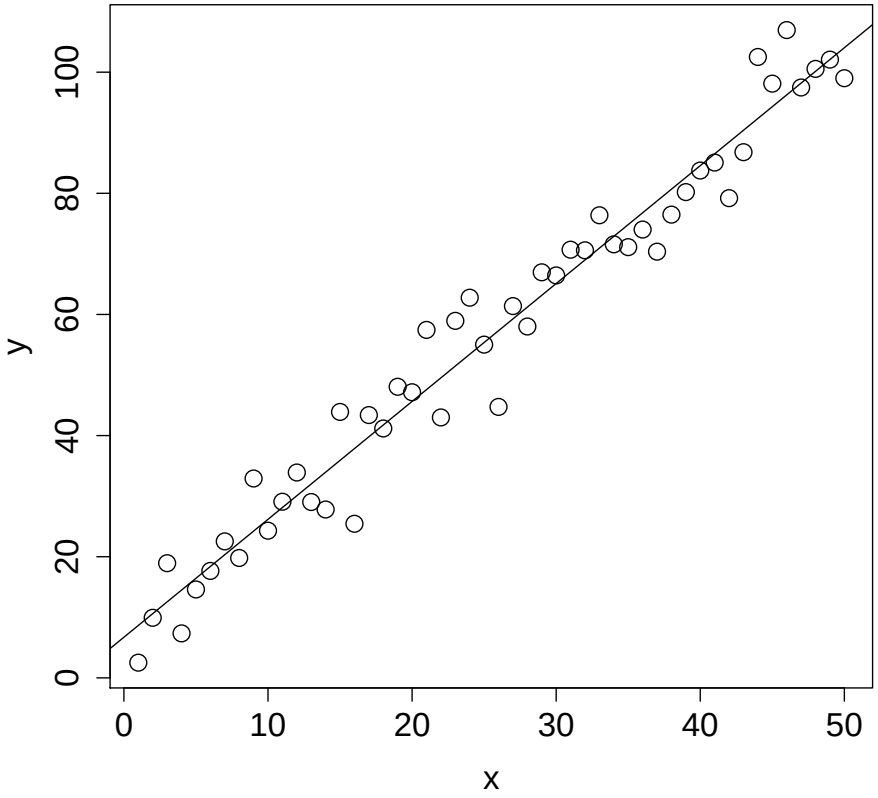
Sind die Modellannahmen in diesen Fällen wohl erfüllt?
Beachte: Wir unterstellen $Y = \beta_0 + \beta_1 X + \varepsilon$.

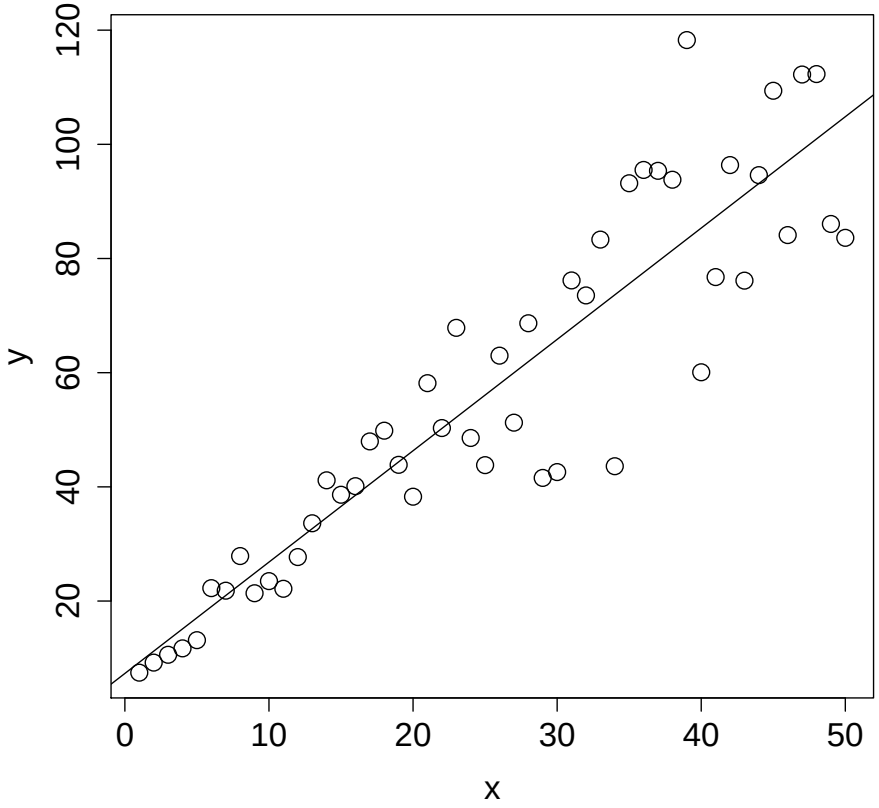




Beispiel

Sind die Modellannahmen in dem folgenden Fall erfüllt?





Residualanalyse

Residuen werden als Realisationen der Modellfehler ε betrachtet

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

D.h. die Differenz aus tasächlichem y-Wert und der Vorhersage. Es gilt:

$$E\left(\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}\right) = E\left(\frac{SS_{res}}{n-k-1}\right) = E\left(MS_{res}\right) = \sigma^2$$

- häufig werden Residuen standardisiert
- Ausreisser bzw. extreme Werte können identifiziert werden
- sogenannte Diagnoseplots geben Hinweise auf die Adäquatheit der Modellannahmen

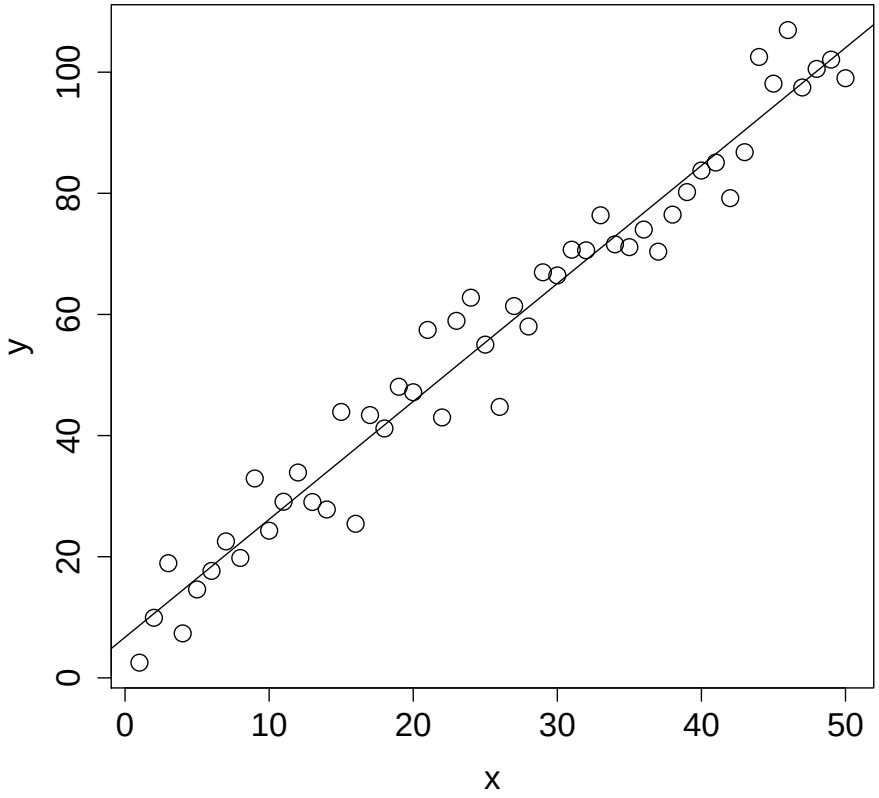
Standardisierte Residuen

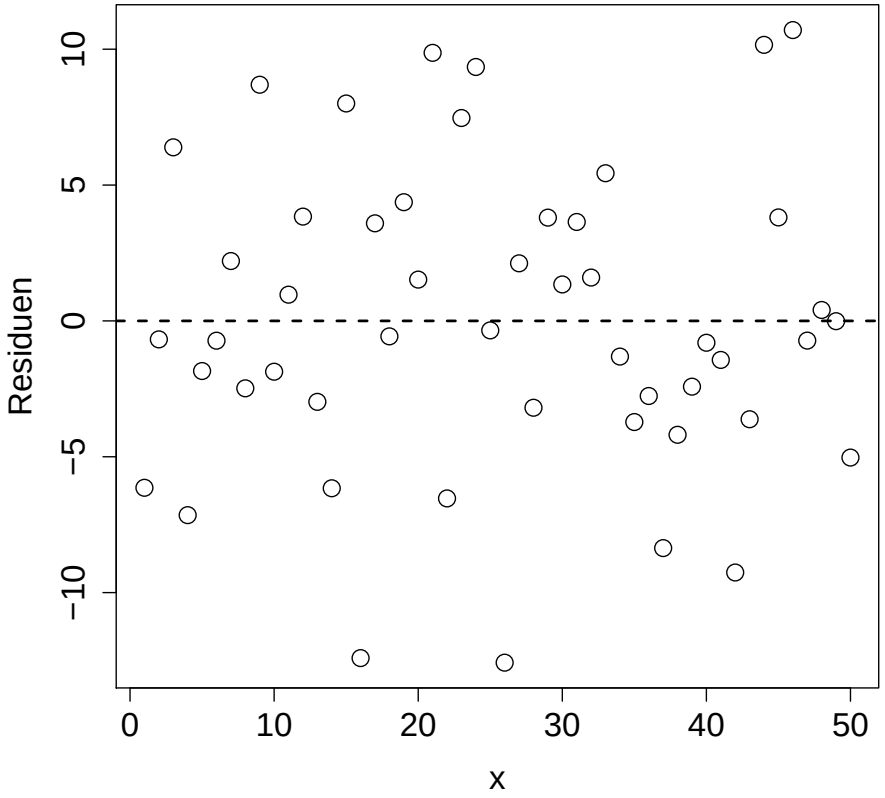
Standardisierung mittels der Residualvarianz (siehe reguläre Standardisierung; dadurch ist die Varianz ungefähr bei 1):

$$d_i = \frac{e_i}{\sqrt{MS_{res}}} \quad i = 1, \ldots, n \text{ mit } E(d_i) = 0, var(d_i) \approx 1$$

Beispiel - Residuen

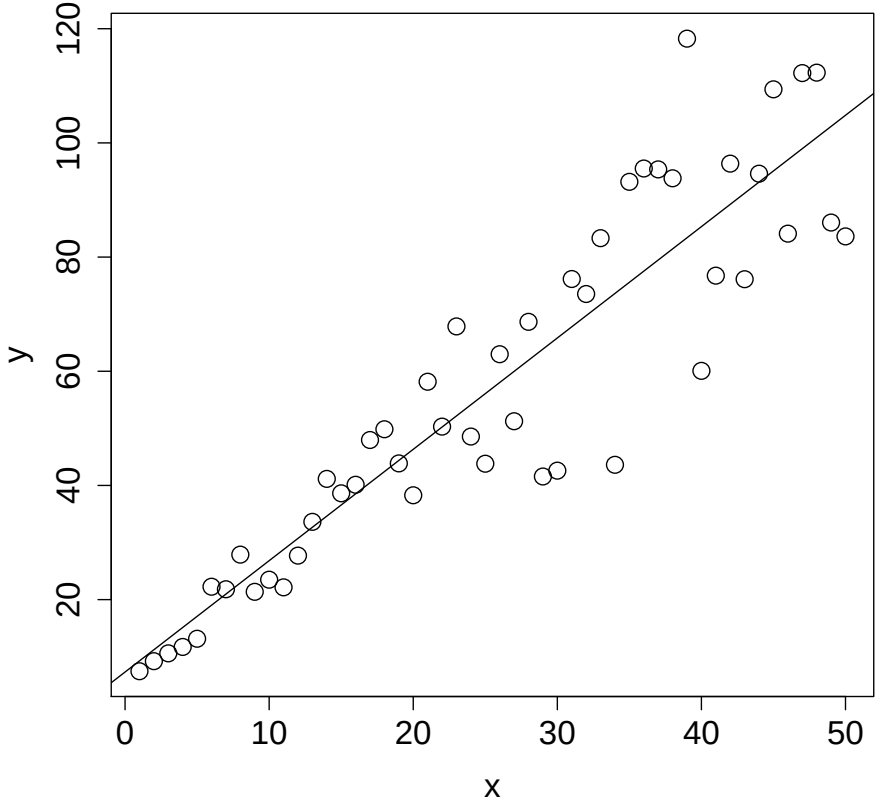
Im folgenden Fall ist alles okay:

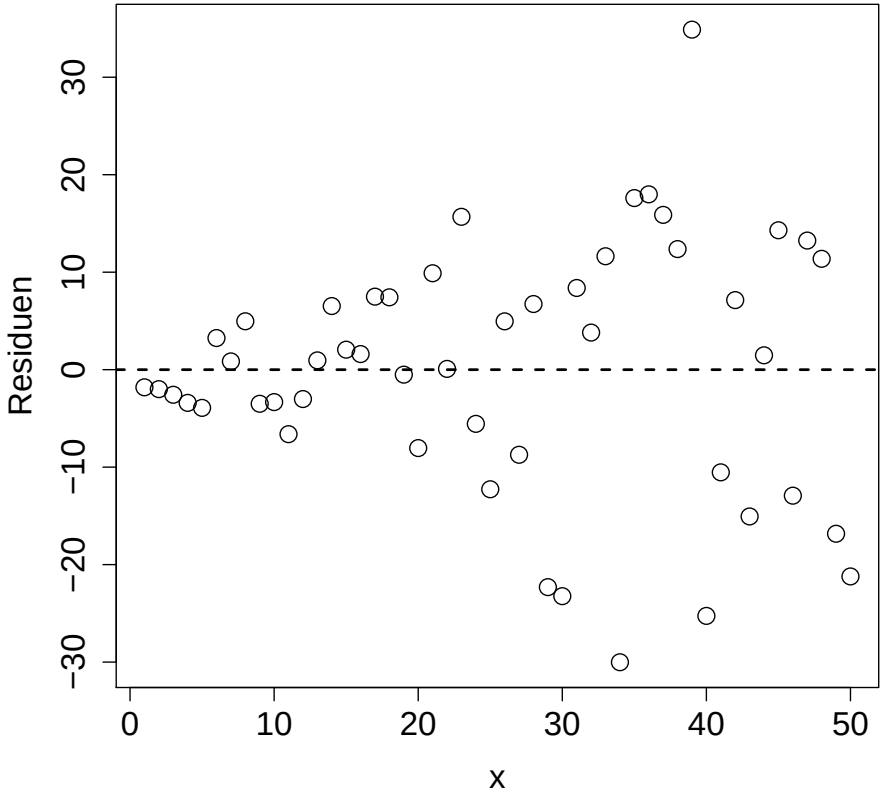




Beispiel - Residuen

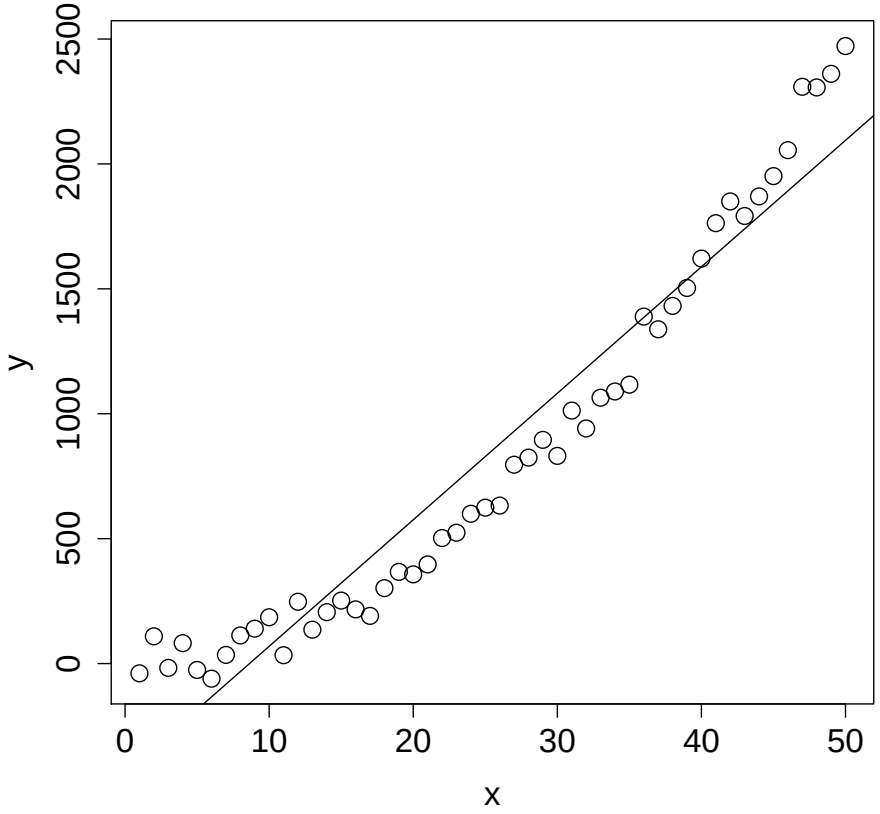
Hier scheint eine Modellannahme nicht erfüllt zu sein. Welche?

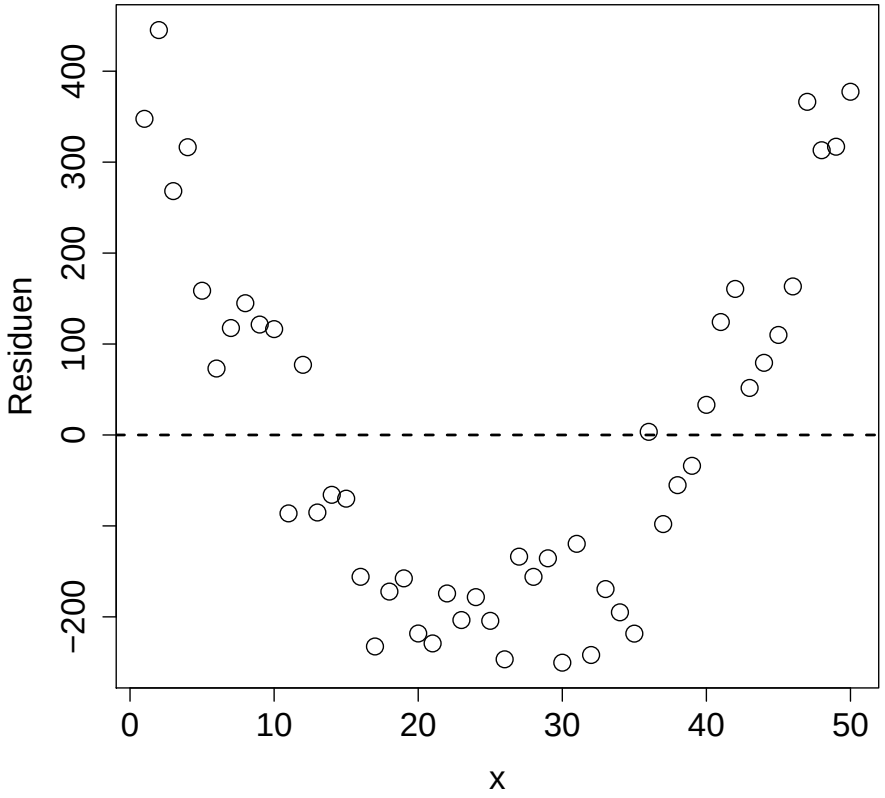




Beispiel - Residuen

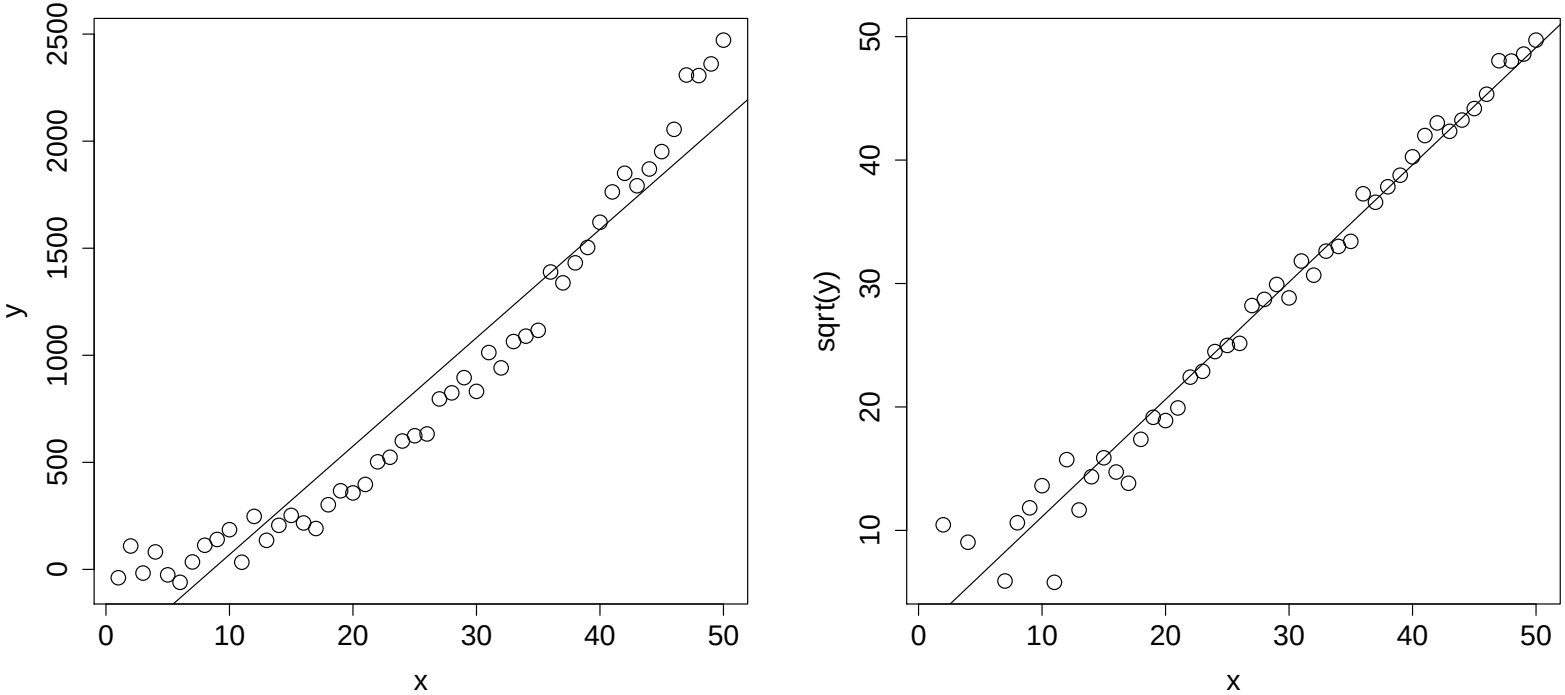
Hier erkennen wir ein eindeutiges Muster was auf nicht-Linearität schließen lässt:





Transformationen als Hilfsmittel

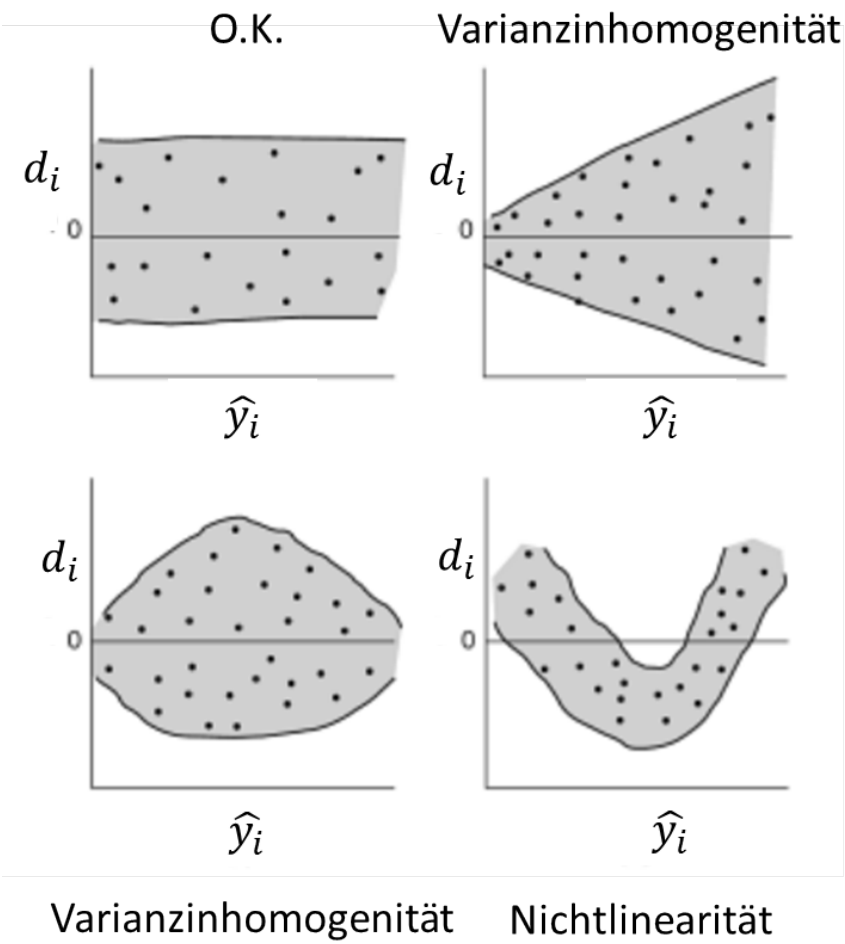
Betrachten wir erneut die offensichtlich nicht-linearen Daten aus dem letzten Beispiel. Hier lässt ein Blick auf die Daten selbst (oder die Residuen) den Schluss zu, dass ein linearer Zusammenhang nicht angebracht. Vielmehr scheint es einen quadratischen Zusammenhang zu geben. Links: originale Daten, rechts: $\tilde{Y} = \sqrt{Y}$



Abschließende Worte zu Residuen

(Standardisierten) Residuen vs. vorhergesagte Werte. Wie hilft uns das?

- verschiedene Abweichungen von den Modellannahmen könnten auffällig werden
- U.U. sind Transformationen oder Hinzunahme weiterer Variablen nötig.
- In der linearen Einfachregression offenbaren Residuen-Plots nicht wirklich mehr als die Plots der Daten selbst. Im multiplen Fall sind sie hingegen ein probates Mittel.



Einordnung

Mit der Analyse von (standardisierten) Residuen können wir einige Annahmen testen:

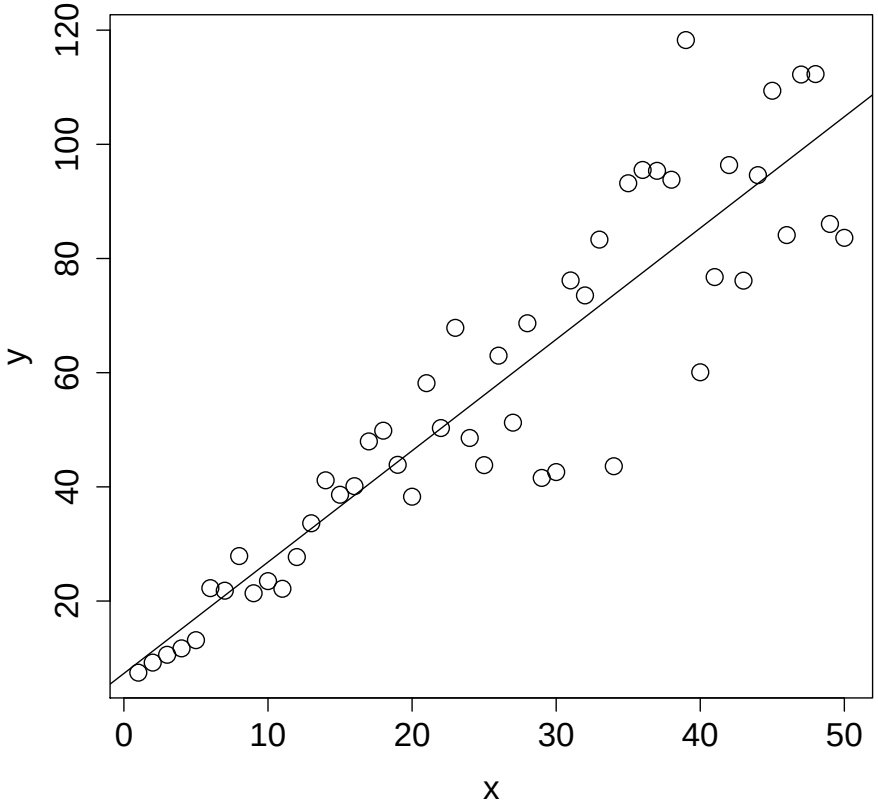
$$E(\varepsilon) = 0 \quad \text{(Im Mittel richtig?)} \quad \checkmark$$
$$var(\varepsilon) = \sigma^2 \mathrm{I} \quad \text{(Unkorreliert und mit gleicher Varianz)} \quad \checkmark$$

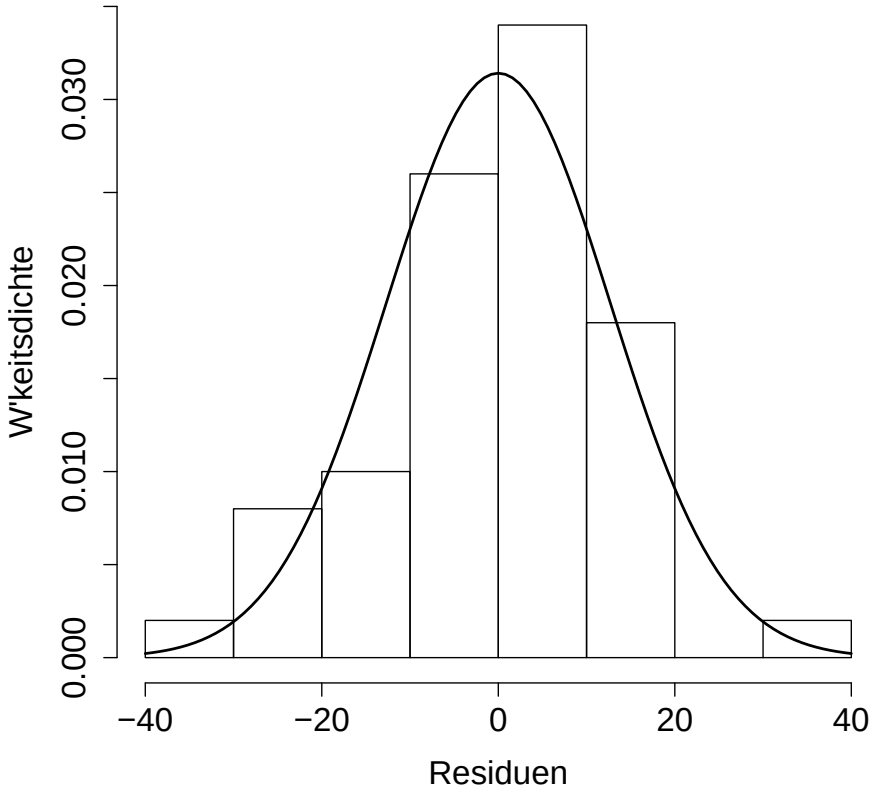
Was fehlt noch? Die Überprüfung auf Normalverteilung:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$$

Erster Ansatz - Histogramme

Wir plotten ein Histogramm der Residuen e_i bzw. d_i, $i = 1, \ldots, n$





Quantile-Quantile-Plots (QQ-Plots)

- Vergleich der realisierten Werte mit Realisationen einer $\mathcal{N}(0,1)$-Verteilung
- Beobachtete Werte d_i werden sortiert: $(d_{(1)} < d_{(2)} < \cdots < d_{(n)})$
- Genau j/n der Werte sind kleiner oder gleich $d_{(j)}$,

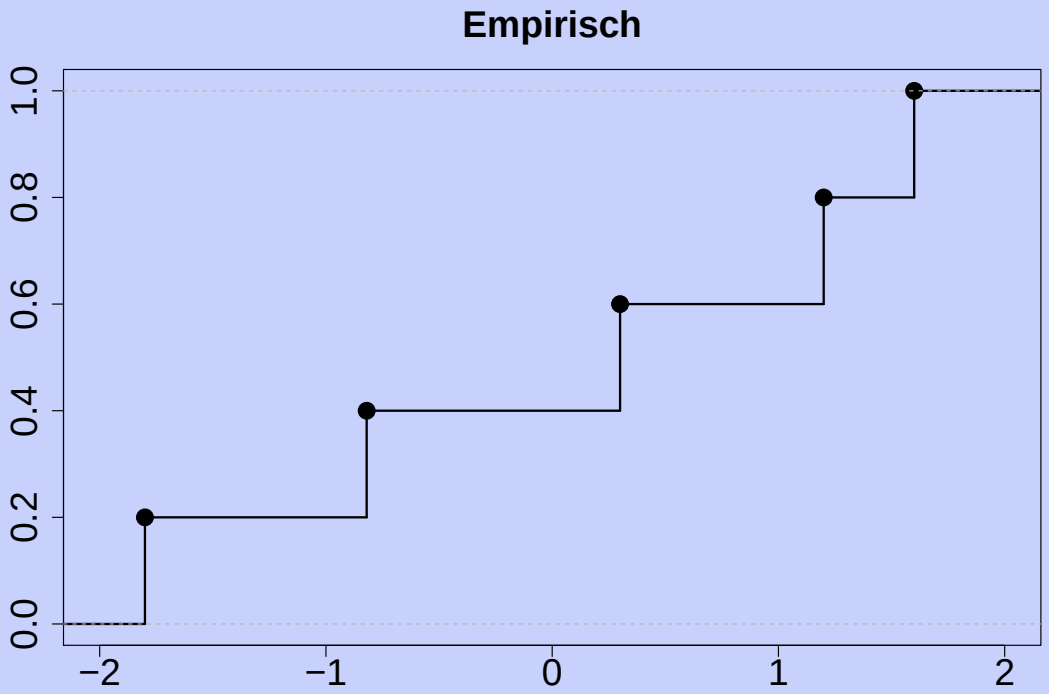
QQ-Plot

- Fasse $d_{(j)}$ als empirisches j/n-Quantil auf und stelle es dem entsprechenden Quantil von $\mathcal{N}(0,1)$ gegenüber
- **Stetigkeitskorrektur**: Anteil j/n wird approximiert durch $j^* = \frac{j-1/2}{n}$.
- Schließlich: n Punktpaare werden gegeneinander aufgetragen:

$$q_j = \left(\Phi^{-1}\left(\frac{j-1/2}{n}\right), d_{(j)}\right)$$

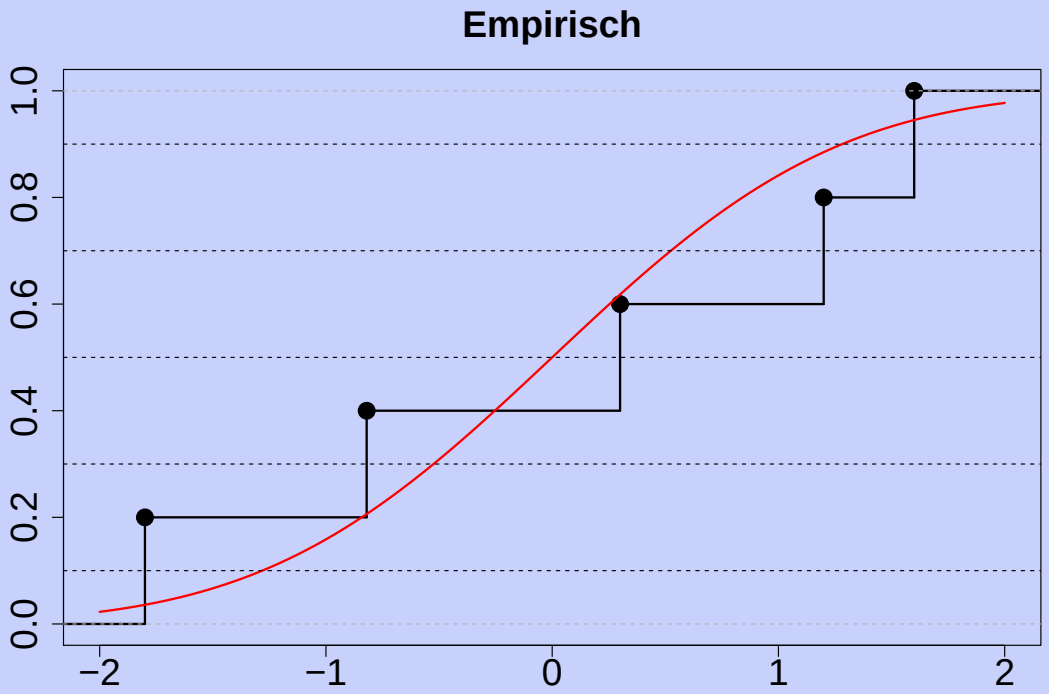
Quantile-Quantile-Plots (QQ-Plots)

d	Ränge	j^*
-1.8	1	0.1
-0.82	2	0.3
0.3	3	0.5
1.2	4	0.7
1.6	5	0.9



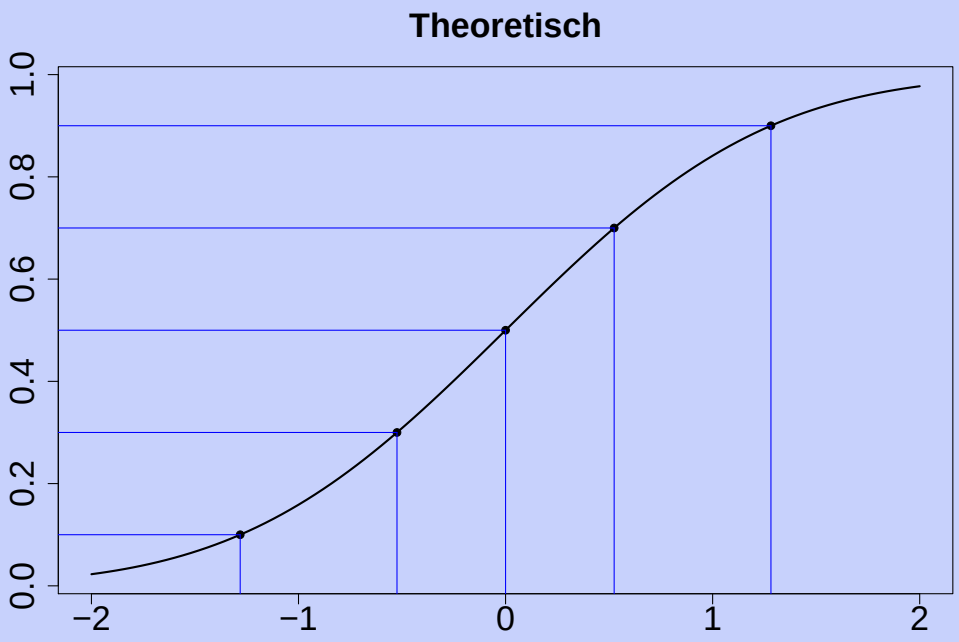
Quantile-Quantile-Plots (QQ-Plots)

d	Ränge	j^*
-1.8	1	0.1
-0.82	2	0.3
0.3	3	0.5
1.2	4	0.7
1.6	5	0.9



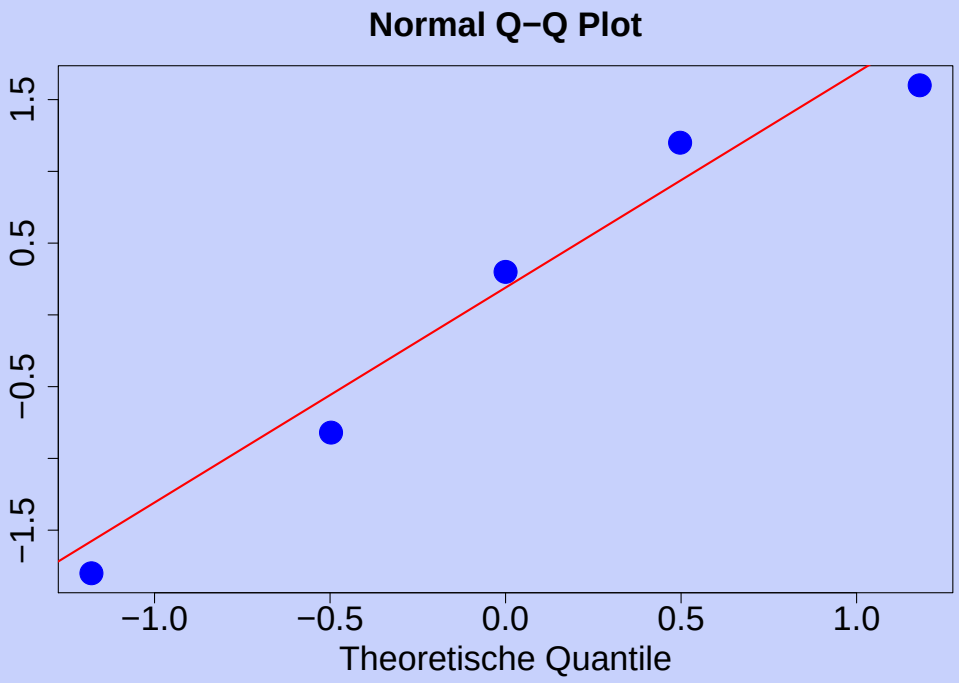
Quantile-Quantile-Plots (QQ-Plots)

d	Ränge	j^*	q (normal)
-1.8	1	0.1	-1.28
-0.82	2	0.3	-0.52
0.3	3	0.5	0.00
1.2	4	0.7	0.52
1.6	5	0.9	1.28



Quantile-Quantile-Plots (QQ-Plots)

d	Ränge	j^*	q (normal)
-1.8	1	0.1	-1.28
-0.82	2	0.3	-0.52
0.3	3	0.5	0.00
1.2	4	0.7	0.52
1.6	5	0.9	1.28

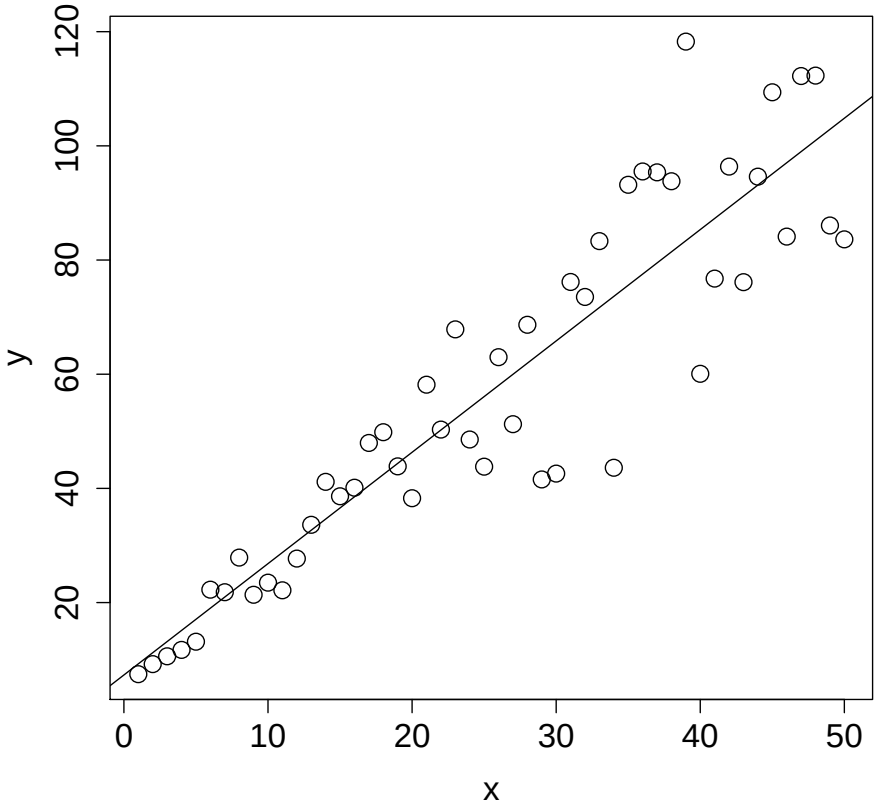


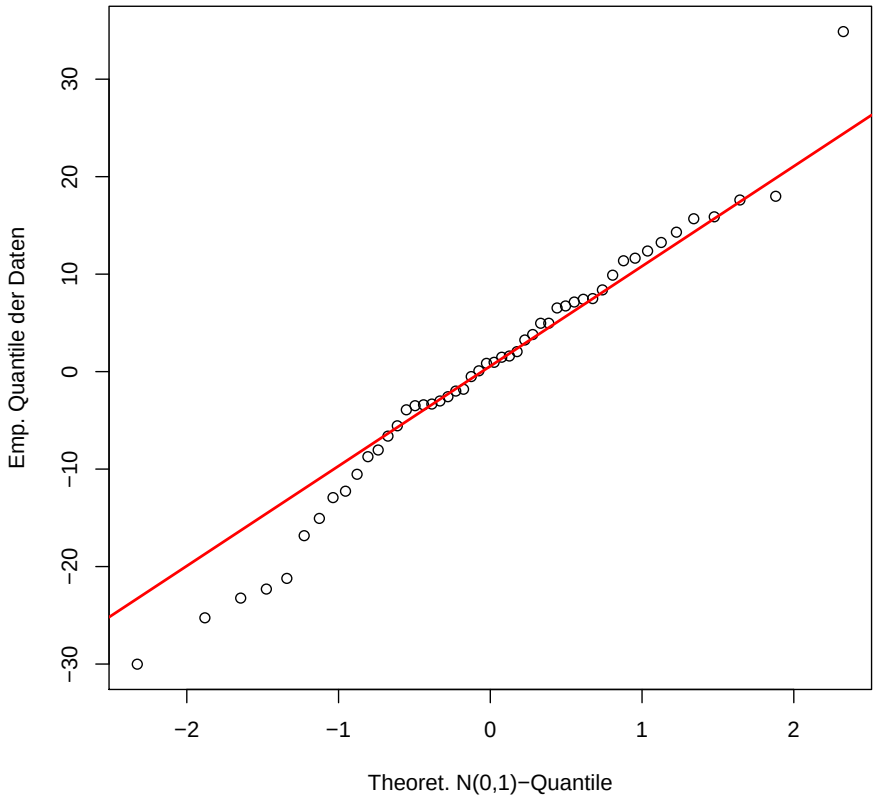
- Plots auf der Originalskala enthalten die folgenden Punktepaare:

$$\left(\Phi^{-1}\left(\frac{j-0.5}{n}\right),\ d_{(j)}\right) \text{ mit der Geraden } y_j = \bar{d} + s_d \cdot d_{(j)}$$

Beispiel: QQ-Plots

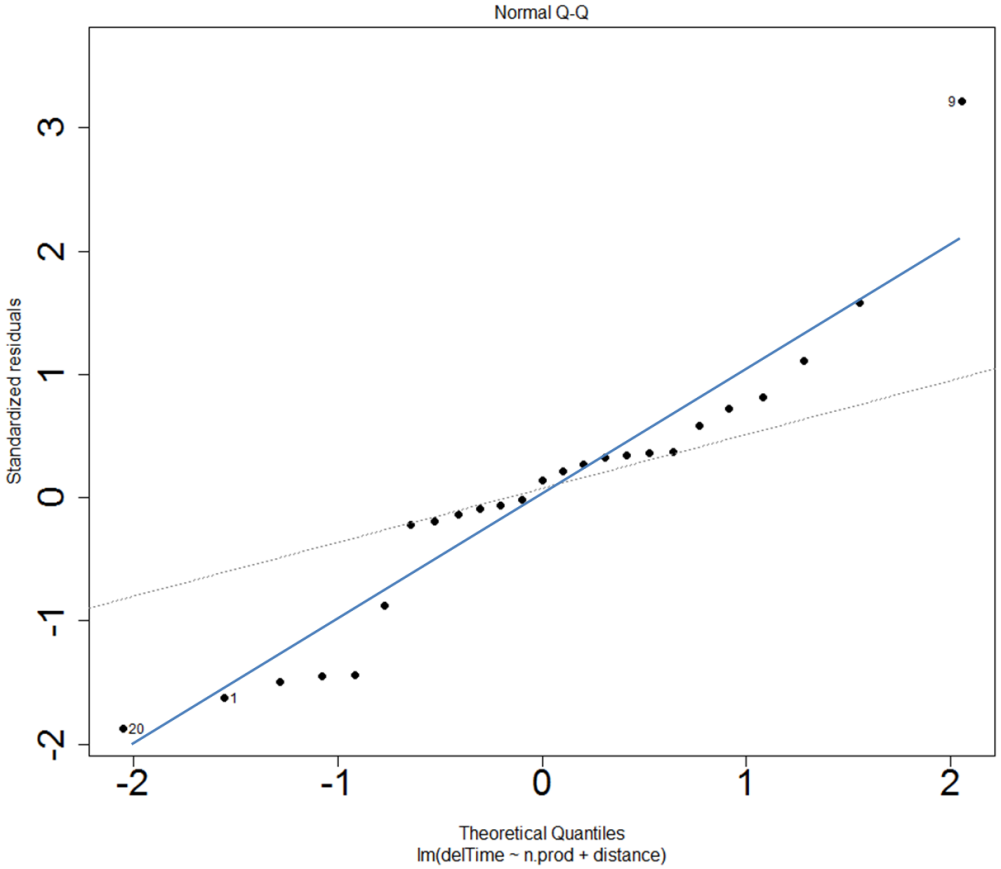
Wir zeichnen einen QQ-Plot der Residuen e_i (bzw. d_i), $i = 1, \ldots, n$





Lieferzeiten, QQ-Plot

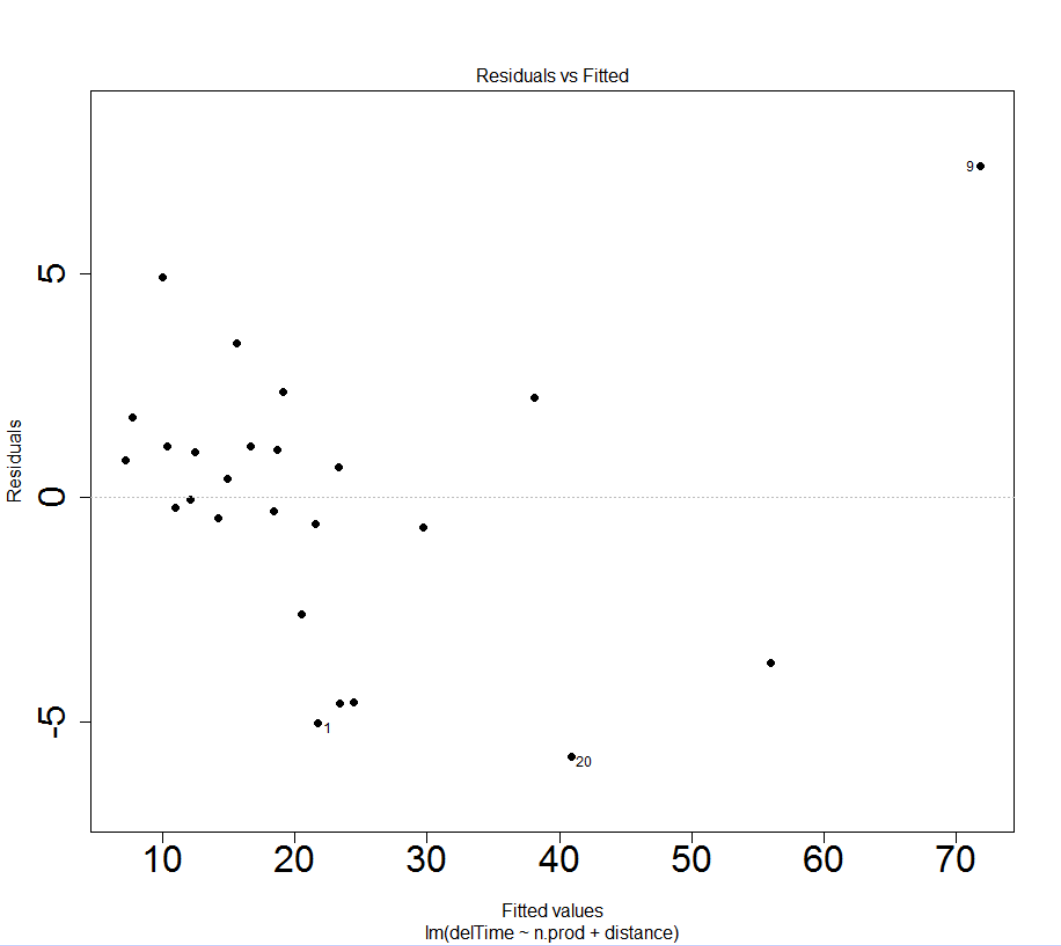
```
lm.deli = lm(delTime ~ n.prod + distance, data = delivery)

plot(lm.deli)
```



Umsatz an einer Tankstelle

Untersuchen Sie, ob für das lineare Modell, welches den Gewinn des Tankstellenbetreibers in Abhängigkeit der Umsätze modelliert, die Normalverteilungsannahme der Residuen erfüllt ist.

Lieferzeiten



Umsatz an einer Tankstelle

Ist ein Zusammenhang zwischen den prognostizierten Gewinnen ($\hat{y}$) und den (standardisierten) Residuen des linearen Modells (e) erkennbar?

9.4.2 Residuenanalyse

9.5 Indikatorvariablen

↪ neben quantitativen erklärenden Variablen (z.B. Temperatur, Gewicht, etc.) können auch qualitative oder kategorielle Variablen auftreten (z.B. Geschlecht, Schicht (morgens, mittags, abends), etc.)

↪ Modellierung mit Hilfe von Indikator- bzw. Dummyvariablen.

Beispiel: Untersuchung des Einflusses der Art von Schneidwerkzeug und der Umdrehungen/min. einer Drehmaschine auf die Lebensdauer des Werkzeugs. [5]

- □ Y: Lebensdauer des Schneidwerkzeuges
- □ X_1: Umdrehungen pro Minute
- □ X_2: $\begin{cases} 0 & \text{Werkzeugtyp } A \\ 1 & \text{Werkzeugtyp } B \end{cases}$

	y	x1	x2		y	x1	x2
1	18.73	610	0	11	30.16	670	1
2	14.52	950	0	12	27.09	770	1
3	17.43	720	0	13	25.40	880	1
4	14.54	840	0	14	26.05	1000	1
5	13.44	980	0	15	33.49	760	1
6	24.39	530	0	16	35.62	590	1
7	13.34	680	0	17	26.07	910	1
8	22.71	540	0	18	36.78	650	1
9	12.68	890	0	19	34.95	810	1
10	19.32	730	0	20	43.67	500	1

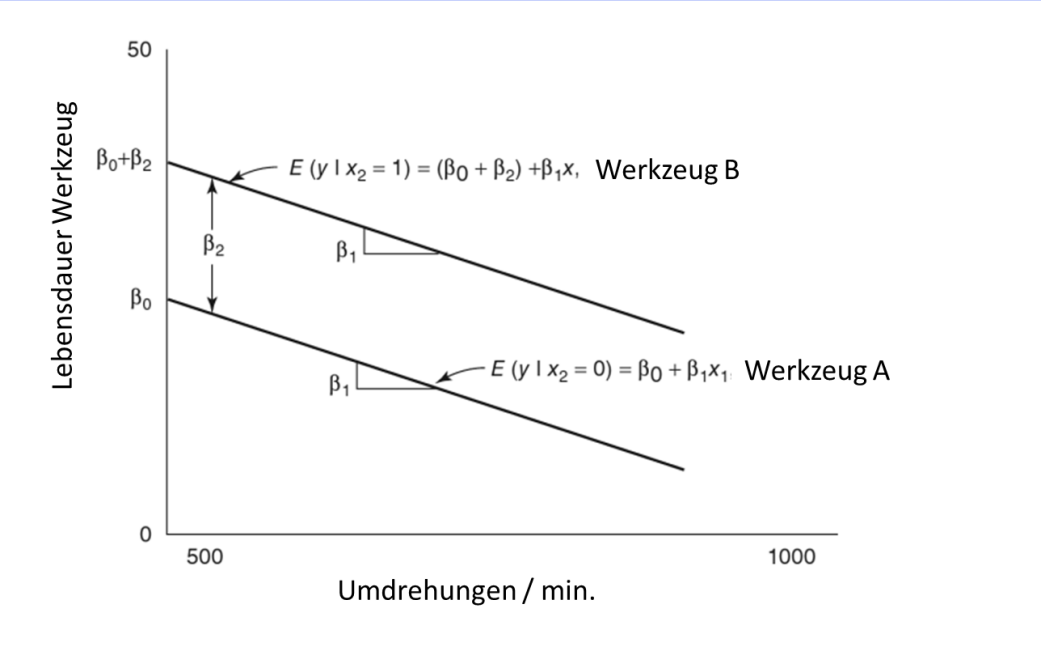
↪ die Reihenfolge der Zuweisung bei X_2 ist unerheblich.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Für die verschiedenen Werkzeugtypen ergibt sich:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \varepsilon = \beta_0 + \beta_1 X_1 + \varepsilon \qquad \text{für Typ A: } X_2 = 0$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \varepsilon = (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon \qquad \text{für Typ B: } X_2 = 1$$



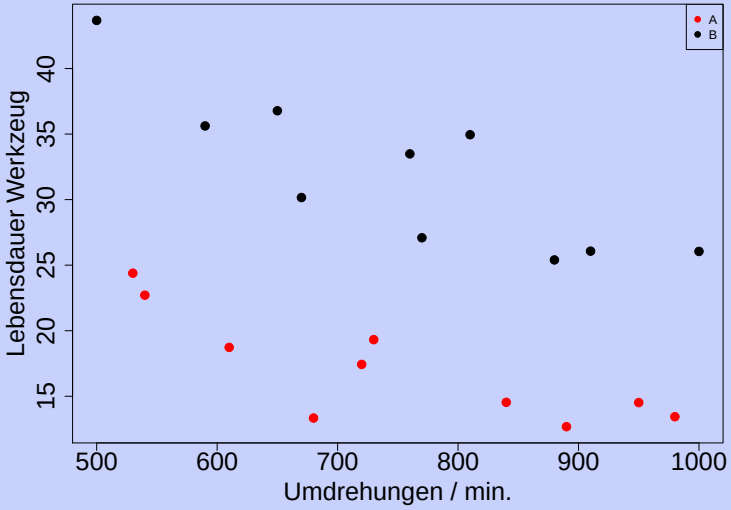
[5]

- ↪ parallele Regressionsgeraden
- ↪ Fehlervarianzen werden identisch angenommen
- ↪ β_2 repräsentiert den erwarteten Unterschied in der Lebensdauer, wenn von Typ A auf Typ B gewechselt wird.
- ↪ Generalisierung auf a verschiedene Typen möglich, es werden $a-1$ Indikatorvariablen benötigt.

X_2	X_3	Typ
0	0	Beobachtung von Werkzeugtyp A
1	0	Beobachtung von Werkzeugtyp B
0	1	Beobachtung von Werkzeugtyp C

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Schätzung des Regressionsmodells:



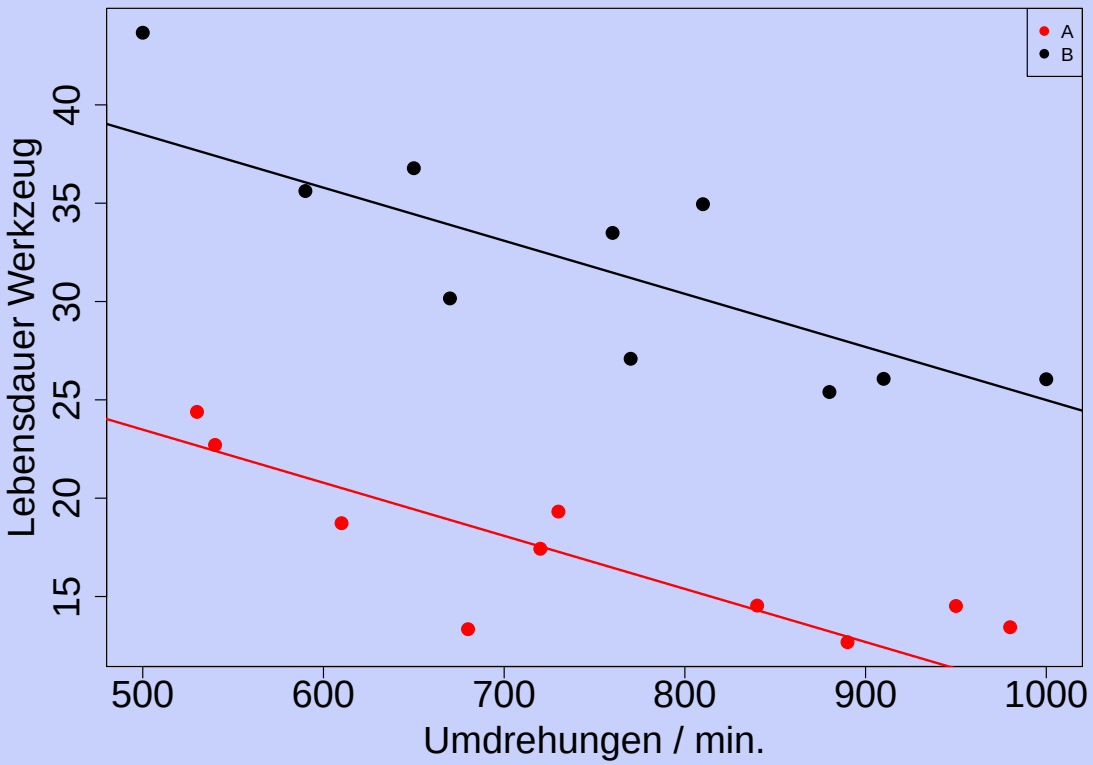

```
17
18  summary(lm(y ~ x1+factor(x2),data=ToolLife))
19
17:1   (Top Level)   R S
Console   Terminal   Jobs
~/

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
x1          -0.02661    0.00452  -5.887 1.79e-05 ***
factor(x2)1 15.00425    1.35967  11.035 3.59e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared:  0.9003,    Adjusted R-squared:  0.8886
F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09
```

$$\hat{Y} = 36.986 - 0.027X_1 + 15.004X_2$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 = 36.986 - 0.027X_1 \qquad \text{für Typ A: } X_2 = 0$$

$$\hat{Y} = (\beta_0 + \beta_2) + \beta_1 X_1$$

$$= (36.986 + 15.004) - 0.027X_1 = 51.99 - 0.027X_1 \qquad \text{für Typ B: } X_2 = 1$$



$$\hat{Y} = 36.986 - 0.027X_1 + 15.004X_2$$

$$\hat{Y} = 36.986 - 0.027X_1 \quad \text{für Typ A}$$

$$\hat{Y} = 51.99 - 0.027X_1 \quad \text{für Typ B}$$

$\hookrightarrow$ β_2 repräsentiert den erwarteten Unterschied in der Lebensdauer, wenn von Typ A auf Typ B gewechselt wird.

$\hookrightarrow$ 95%-Konfidenzintervall für β_2

$$\begin{aligned} \hat{\beta}_2 - t_{0.975,17} \cdot se(\hat{\beta}_2) \leq \quad & \beta_2 \quad \leq \hat{\beta}_2 + t_{0.975,17} \cdot se(\hat{\beta}_2) \\ \Leftrightarrow \quad 15.004 - 2.110 \cdot 1.360 \leq \quad & \beta_2 \quad \leq 15.004 + 2.110 \cdot 1.360 \\ \Leftrightarrow \quad 12.135 \leq \quad & \beta_2 \quad \leq 17.873 \end{aligned}$$

Aspekte der gemeinsamen Modellierung:

+ einfacher zu interpretieren

$\hookrightarrow$ Schätzung einer gemeinsamen Fehlervarianz
  + mehr Freiheitsgrade für die Schätzung
  – nur sinnvoll, wenn Annahme gleicher Fehlervarianz korrekt ist.

+ Änderung in der Modellierung möglich, um unterschiedliche Steigungen zu erhalten.

<u>Annahme:</u> Unterschiede im Achsenabschnitt und in der Steigung der Regressionsgeraden je nach Werkzeugtyp.

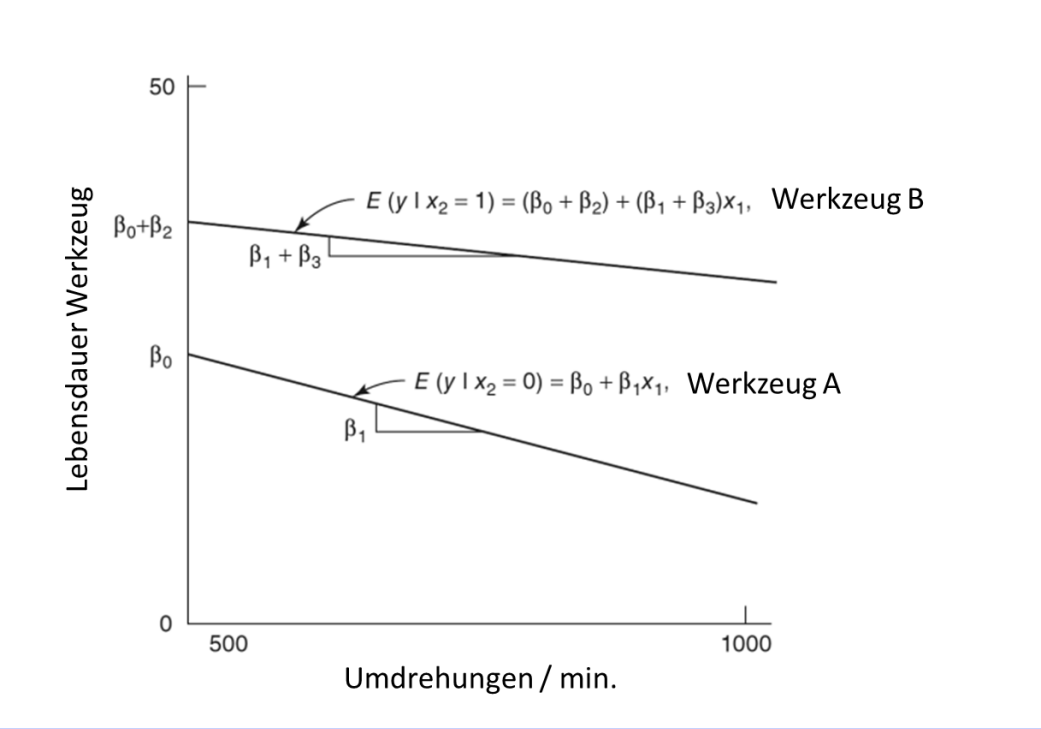
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Für Typ A (d.h. $X_2 = 0$):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \beta_3 X_1 \cdot 0 + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \varepsilon \end{aligned}$$

Für Typ B (d.h. $X_2 = 1$):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \beta_3 X_1 \cdot 1 + \varepsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \varepsilon \end{aligned}$$



$\hookrightarrow$ β_2 repräsentiert den erwarteten Unterschied im Achsenabschnitt, wenn von Typ A auf Typ B gewechselt wird.

$\hookrightarrow$ β_3 repräsentiert den erwarteten Unterschied in der Steigung, wenn von Typ A auf Typ B gewechselt wird.

$\hookrightarrow$ Tests können auf dem gemeinsamen Modell durchgeführt werden

Schätzung des Regressionsmodells:

$$\hat{Y} = 32.775 - 0.021 \cdot X_1 + 23.971 \cdot X_2 - 0.012 \cdot X_1 X_2, \quad \text{d.h.}$$

$$\hat{Y} = 32.775 - 0.021 \cdot X_1 \quad \text{für Typ A}$$

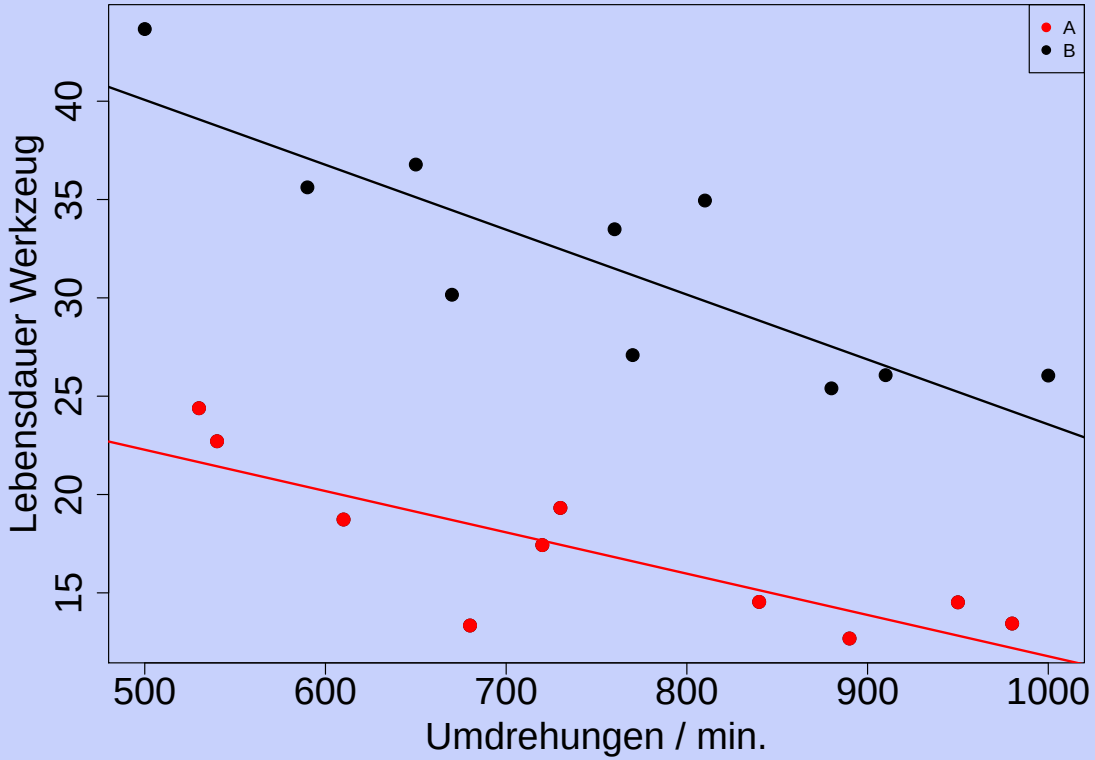
$$\hat{Y} = (32.775 + 23.791) + (-0.021 - 0.012) \cdot X_1$$

$$= 56.566 - 0.033 \cdot X_1 \quad \text{für Typ B}$$

```
summary(lm(y ~ x1*factor(x2),data=ToolLife))

~/
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    32.774760   4.633472   7.073 2.63e-06 ***
x1             -0.020970   0.006074  -3.452  0.00328 **
factor(x2)1    23.970593   6.768973   3.541  0.00272 **
x1:factor(x2)1 -0.011944   0.008842  -1.351  0.19553
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared:  0.9105,    Adjusted R-squared:  0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08
```

$$\hat{Y} = 32.775 - 0.021 \cdot X_1 \quad \text{für Typ A}$$

$$\hat{Y} = 56.566 - 0.033 \cdot X_1 \quad \text{für Typ B}$$



$H_0 : \beta_3 = 0$ vs. $\beta_3 \neq 0$ $\rightarrow$ wird nicht abgelehnt.

Variablen mit mehr als 2 Levels

↪ 4 Levels können über 3 Indikatorvariablen realisiert werden.

X_2	X_3	X_4
0	0	0
1	0	0
0	1	0
0	0	1

↪ Regressionsmodell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

↪ durch Hinzufügen von Interaktionen können verschiedene Steigungen realisiert werden

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 + \varepsilon$$

Beispiel Klimasysteme, vgl. [5]

$$
\begin{aligned}
Y &: \text{Energieverbrauch in den Sommermonaten Juni-September (kw/h)} \\
X_1 &: \text{Größe des Hauses } (ft^2) \\
X_2 &: \text{Klimasystem} = \begin{cases} 1 & \text{kein Klimasystem} \\ 2 & \text{Fensterklimasysteme} \\ 3 & \text{Wärmepumpe} \\ 4 & \text{zentrale Klimaanlage} \end{cases}
\end{aligned}
$$

$\hookrightarrow$ Modellierung von X_2 mit Dummyvariablen:

Klimasystem	X_{22}	X_{23}	X_{24}
kein Klimasystem	0	0	0
Fensterklimasysteme	1	0	0
Wärmepumpe	0	1	0
zentrale Klimaanlage	0	0	1

Eine Variable X_{21} wird nicht benötigt: Mit X_{21} wäre zudem $\mathrm{X}^T\mathrm{X}$ nicht invertierbar (Überparametrisierung). Andere Formen der Kodierung sind möglich, im Falle von Überparametrisierung werden die Lösungen durch zusätzliche (lineare) Nebenbedingungen an die Parameter eindeutig gemacht.

Modellierungsmöglichkeit I

$$Y = \beta_0 + \beta_1 X_1 + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \varepsilon$$

Wie sehen die resultierenden individuellen Regressionsmodelle aus?

$$\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 1 \\
Y &= (\beta_0 + \beta_{22}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 2 \\
Y &= (\beta_0 + \beta_{23}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 3 \\
Y &= (\beta_0 + \beta_{24}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 4
\end{aligned}$$

$\hookrightarrow$ Annahme: linearer Grundzusammenhang, Steigung und Varianz hängen nicht vom Klimasystem ab

$\hookrightarrow$ β_{22}, β_{23} und β_{24} beschreiben den Effekt des jeweiligen Systems

Modellierungsmöglichkeit I

$$Y = \beta_0 + \beta_1 X_1 + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \varepsilon$$

Wie sehen die resultierenden individuellen Regressionsmodelle aus?

$$\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 1 \\
Y &= (\beta_0 + \beta_{22}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 2 \\
Y &= (\beta_0 + \beta_{23}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 3 \\
Y &= (\beta_0 + \beta_{24}) + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 4
\end{aligned}$$

- $\hookrightarrow$ relative Effizienz kann gemessen werden, z.B. misst $\beta_{22} - \beta_{24}$ die relative Effizienz von Fensterklimasystemen im Vergleich mit einer zentralen Klimaanlage
- $\hookrightarrow$ Modellgleichung in `R`:
  `y ~ x1 + factor(x2)` (R generiert dann die Dummy-Variablen)

Modellierungsmöglichkeit II

- ↪ Sind die Annahmen realistisch?
- ↪ wahrscheinlicher ist eine Interaktion zwischen Hausgröße und Typ des Klimasystems
- ↪ d.h. unterschiedliche Steigungen realistischer

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} \\ &\quad + \beta_{122} X_1 X_{22} + \beta_{123} X_1 X_{23} + \beta_{124} X_1 X_{24} + \varepsilon \end{aligned}$$

Modellierungsmöglichkeit II

$\hookrightarrow$ d.h. unterschiedliche Steigungen realistischer

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} \\ &\quad + \beta_{122} X_1 X_{22} + \beta_{123} X_1 X_{23} + \beta_{124} X_1 X_{24} + \varepsilon \end{aligned}$$

Wie sehen die resultierenden individuellen Regressionsmodelle aus?

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \varepsilon && \text{für } X_2 = 1 \\ Y &= (\beta_0 + \beta_{22}) + (\beta_1 + \beta_{122}) X_1 + \varepsilon && \text{für } X_2 = 2 \\ Y &= (\beta_0 + \beta_{23}) + (\beta_1 + \beta_{123}) X_1 + \varepsilon && \text{für } X_2 = 3 \\ Y &= (\beta_0 + \beta_{24}) + (\beta_1 + \beta_{124}) X_1 + \varepsilon && \text{für } X_2 = 4 \end{aligned}$$

Modellgleichung in R:

```
y ~ x1 * factor(x2)
```

Übung: Im Herzviertel wurden Wohnfläche, Zimmerzahl und Kaltmiete von 1-4-Zimmer-Wohnungen erhoben:

Whg	qm	Zi	KM	Whg	qm	Zi	KM	Whg	qm	Zi	KM	Whg	qm	Zi	KM
1	40	1	590	9	50	2	590	17	115	3	1030	25	120	4	960
2	35	1	580	10	45	2	560	18	115	3	1020	26	105	4	870
3	35	1	540	11	50	2	600	19	100	3	860	27	110	4	960
4	55	1	820	12	45	2	540	20	90	3	810	28	120	4	1010
5	40	1	630	13	80	2	930	21	95	3	860	29	105	4	830
6	40	1	620	14	75	2	860	22	75	3	610	30	120	4	950
7	45	1	720	15	105	3	940	23	105	3	890				
8	45	2	540	16	115	3	1040	24	90	3	850				

Die Kaltmiete soll anhand der Wohnfläche und Zimmerzahl erklärt werden. Wie lautet ein mögliches Regressionsmodell? Stellen Sie sinnvolle Hypothesen auf