Stochastik für Ingenieure

- Einführung
- Wahrscheinlichkeitsrechnung
- Simulation zufälliger Ereignisse
- Stochastische Prozesse
- Grundlagen der Statistik
- Testtheorie
- Regressionsanalyse
- Aufgaben
- Anhang

Peter Junglas 01.12.2017

Inhaltsverzeichnis

Übersicht

- Einführung
- Wahrscheinlichkeitsrechnung
    - Definition der Wahrscheinlichkeit
    - Bedingte Wahrscheinlichkeit
    - Diskrete Zufallsverteilungen
        - Zufallsvariablen
        - Spezielle diskrete Verteilungen
        - Eigenschaften diskreter Verteilungen
    - Stetige Zufallsverteilungen
        - Dichte- und Verteilungsfunktion
        - Eigenschaften stetiger Verteilungen
        - Spezielle stetige Verteilungen
    - Mehrdimensionale Zufallsvariablen
        - Zufallsvektoren
        - Unabhängige Zufallsvariablen
        - Grenzwertsätze
- Simulation zufälliger Ereignisse
    - Problemstellung
    - Bestimmung von U(0,1)-verteilten Zufallszahlen
    - Zufallszahlen mit vorgegebener Verteilung
- Stochastische Prozesse
    - Markov-Ketten
    - Bediensysteme
- Grundlagen der Statistik
    - Deskriptive Statistik
        - Grundbegriffe
        - Eindimensionale Daten
        - Zweidimensionale Daten
    - Parameterschätzung
    - Konfidenzintervalle
- Testtheorie
    - Statistische Tests von Hypothesen
    - Parametrische Tests
        - Parametertests bei Normalverteilung mit einer Stichprobe
        - Parametertests bei Normalverteilung mit zwei Stichproben
        - Parametertests bei Bernoulli-Experimenten
    - Nicht-parametrische Tests
    - Tests mit mehr als zwei Stichproben
- Regressionsanalyse
    - Einfache lineare Regression
    - Statistische Analyse der Regressionsparameter
    - Ergänzende Analysen bei linearer Regression
    - Mehrfache lineare Regression
- Aufgaben
    - Aufgabe 1
    - Aufgabe 2
    - Aufgabe 3
    - Aufgabe 4
    - Aufgabe 5
    - Aufgabe 6
    - Aufgabe 7
    - Aufgabe 8
    - Aufgabe 9
    - Aufgabe 10
    - Aufgabe 11
    - Aufgabe 12
    - Aufgabe 13
    - Aufgabe 14

- Aufgabe 15
- Aufgabe 16
- Aufgabe 17
- Aufgabe 18
- Aufgabe 19
- Aufgabe 20
- Aufgabe 21
- Aufgabe 22
- Aufgabe 23
- Aufgabe 24
- Aufgabe 25
- Aufgabe 26
- Aufgabe 27
- Aufgabe 28
- Aufgabe 29
- Aufgabe 30
- Aufgabe 31
- Aufgabe 32
- Aufgabe 33
- Aufgabe 34
- Aufgabe 35
- Aufgabe 36
- Aufgabe 37
- Aufgabe 38
- Aufgabe 39
- Aufgabe 40
- Aufgabe 41
- Aufgabe 42
- Aufgabe 43
- Aufgabe 44
- Aufgabe 45
- Aufgabe 46
- Aufgabe 47
- Anhang
  - Literatur
  - Nachweise
  - Herleitungen
    - Poisson-Grenzwertsatz
    - Erwartungswert bei der Binomialverteilung
    - Randverteilung der bivariaten Normalverteilung
    - Erwartungswert E(X Y) der bivariaten Normalverteilung
    - Herleitung des Box-Muller-Verfahrens
    - Eigenschaften der geometrischen Verteilung
    - Erwartungswert des ML-Schätzers S^2 für die Varianz
    - Aufteilung der Varianzen bei einfaktorieller ANOVA
    - Berechnung der Varianz des ML-Schätzers A bei einfacher linearer Regression
    - Berechnung der Kovarianzmatrix des ML-Schätzers B bei mehrfacher linearer Regression
    - Formel zur Berechnung des Schätzers $s_{2y|x}$ bei mehrfacher linearer Regression
  - Matlab-Beispiele
    - Beispiele für Kap 1.3.2
    - Beispiele für Kap 1.4.3
    - Beispiele für Kap 1.5.1
    - Beispiele für Kap 1.5.2
    - Beispiele für Kap 1.5.3
    - Beispiele für Kap 3.1
    - Beispiele für Kap 4.1.2a
    - Beispiele für Kap 4.1.2b
    - Beispiele für Kap 4.1.2c
    - Beispiele für Kap 4.1.3a
    - Beispiele für Kap 4.1.3b
    - Beispiele für Kap 4.3

- Beispiele für Kap 5.1
- Beispiele für Kap 5.2.1f
- Beispiele für Kap 5.2.2
- Beispiele für Kap 5.2.3
- Beispiele für Kap 5.3
- Beispiele für Kap 5.4
- Beispiele für Kap 6.1ff
- Beispiele für Kap 6.4
- ranksumcdf.m
- erzeugeZugversuchDaten.m

Einführung

- Zufällige Größen:

  Ergebnis einer (beliebigen) Messung

  Anzahl defekter Bauteile in einer Fertigung

  Zahl der Erkrankungen am Tag einer Klausur

  Belastung einer Fahrzeugfederung über die Lebensdauer

  Aktienkurs einer Firma

  Lottozahlen

- Wahrscheinlichkeit - was ist das:

  historisch
  - Grenzwert relativer Häufigkeit (**Frequentismus**)

  - Maß für die Überzeugung des Eintretens (**Bayesianismus**)

  mathematisch
  - Modellbildungen etwa seit 1650 (Pascal, Fermat)
  - große Probleme bei stetigen Verteilungen → Maßtheorie etwa ab 1900 (Borel, Lebesgue)
  - axiomatische Definition 1933 (Kolmogorov)

- Stochastik:

  mathematische Beschreibung und Untersuchung von Zufallsexperimenten

  völlig präzise, mathematisch exakte Theorie

  Teilgebiet Wahrscheinlichkeitstheorie
  - Verteilung bekannt → Plausibilität zukünftiger Ereignisse

  Teilgebiet Statistik
  - beobachtete Ereignisse → Bestimmung der Verteilung

  beides gehört in Anwendungen eng zusammen

- Liefert Antworten auf praktische Fragen:

  Wie häufig tritt während der Garantiezeit durchschnittlich ein Defekt auf?

  Ist eine aufgetretene Abweichung nur Zufall oder Hinweis auf Fertigungsfehler?

  Wie viele Bauteile müssen getestet werden, um zuverlässige Aussagen über die Qualität zu bekommen?

  Habe ich bei positivem Testergebnis wirklich Krebs oder ist das reiner Zufall?

Wahrscheinlichkeitsrechnung

- Definition der Wahrscheinlichkeit
- Bedingte Wahrscheinlichkeit
- Diskrete Zufallsverteilungen
- Stetige Zufallsverteilungen
- Mehrdimensionale Zufallsvariablen

Definition der Wahrscheinlichkeit

- Beispiel "idealer Würfel":

  Ergebnisse eines Wurfs: {1, 2, 3, 4, 5, 6}

  Ereignisse
  - E_1: Es wird eine 6 gewürfelt
  - E_2: Es wird eine Zahl < 3 gewürfelt
  - E_3: Es wird eine ungerade Zahl gewürfelt

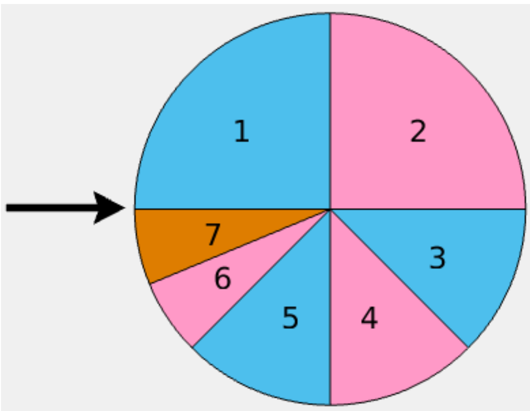
  Elementarereignis: Ereignis, das nur aus einem einzigen Ergebnis besteht

  Annahme: Ergebnisse gleichwahrscheinlich, also
  - $P(\{1\}) = \ldots = P(\{6\}) = 1/6$

  Wahrscheinlichkeit eines Ereignisses durch Abzählen der enthaltenen Elementarereignisse
  - $P(E_1) = P(\{6\}) = 1/6$
  - $P(E_2) = P(\{1,2\}) = 2/6$
  - $P(E_3) = P(\{1,3,5\}) = 3/6$

- Beispiel "Drehen eines Glücksrads":

  kreisförmige Scheibe mit Sektor-Markierungen

  - wird kräftig in Drehung versetzt und kommt durch Reibung zum Stehen
  - feststehender Pfeil zeigt auf die Scheibe und definiert das Ergebnis

  Ergebnisse: jeder beliebige Winkel $\varphi \in [0, 360°)$

  Ereignisse
  - E_1: der Winkel beträgt 90°
  - E_2: der Winkel liegt zwischen 45° und 90°
  - E_3: der Sektor "7" wird erreicht
  - E_4: ein roter Sektor wird erreicht

  Annahme: alle Ergebnisse gleichwahrscheinlich
  - unendlich viele Ergebnisse, sogar überabzahlbar viele!
  - $P(E_1) = P(\{\varphi, \text{beliebig}\}) = 0$

  Wahrscheinlichkeit eines Ereignisses = eingeschlossene Fläche / Kreisfläche
  - $P(E_2) = 1/8$
  - $P(E_3) = 1/16$
  - $P(E_4) = 1/4 + 1/8 + 1/16 = 7/16$

- Definition des Wahrscheinlichkeitsraums $(\Omega, \mathcal{F}, P)$:

  Menge Ω (**Ergebnisraum**)

Menge $\mathcal{F}$ von Teilmengen von Ω (**Ereignisraum**) mit folgenden Eigenschaften (**σ-Algebra**)

$$\Omega \in \mathcal{F}$$

$$A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F} \qquad (A^c := \Omega \setminus A)$$

$$A_i \in \mathcal{F} \Rightarrow \bigcup_i A_i \in \mathcal{F} \qquad (i \in \mathbb{N})$$

Abbildung P: $\mathcal{F} \to [0,1]$ (**Wahrscheinlichkeitsmaß**) mit

$$P(\Omega) = 1$$

$$A_i \in \mathcal{F} \text{ paarweise disjunkt } \Rightarrow P(\bigcup_i A_i) = \sum_i P(A_i) \qquad (i \in \mathbb{N})$$

- Anmerkung zu $\mathcal{F}$:

  bei endlichem oder abzählbarem Ω oft einfach

  $\mathcal{F} = \mathcal{P}(\Omega)$ (**Potenzmenge** von Ω)

  bei überabzahlbarem Ω (z.B. Intervall reeller Zahlen)
  - Potenzmenge "zu groß" → keine sinnvolle Definition von P möglich
  - Ausweg: nur "vernünftige" Teilmengen zulassen

  bei reellen Zahlen z.B. Intervalle und ihre abzählbaren Durchschnitte und Vereinigungen

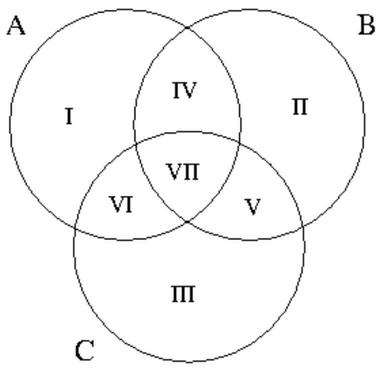
  i. F. wird $\mathcal{F}$ immer so angenommen

- Einfache Folgerungen:

  1. $P(A^C) = 1 - P(A)$
  2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  3. $A \subset B \Rightarrow P(B \setminus A) = P(B) - P(A)$
  4. $P(B \setminus A) = P(B) - P(A \cap B)$
  5. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

- Beweis von 5:

  Ereignisse als Flächen symbolisiert



  Zerlegung in disjunkte Teile liefert sofort

$$\begin{aligned}
P(A) &= P(I) + P(IV) + P(VI) + P(VII) \\
P(B) &= P(II) + P(IV) + P(V) + P(VII) \\
P(C) &= P(III) + P(V) + P(VI) + P(VII) \\
P(A \cap B) &= P(IV) + P(VII) \\
P(A \cap C) &= P(VI) + P(VII) \\
P(B \cap C) &= P(V) + P(VII) \\
P(A \cap B \cap C) &= P(VII)
\end{aligned}$$

Einsetzen in (5) ergibt dann

$P(A \cup B \cup C) = P(I) + P(II) + P(III) + P(IV) + P(V) + P(VI) + P(VII)$

- Beispiel n-faches Würfeln:

  $\Omega = \{(a_1, a_2, .., a_n) \mid a_i \in \{1,2,3,4,5,6\} \}$

  $P(\omega) = 1/6^n$ für alle $\omega \in \Omega$

  Wie groß ist die Wahrscheinlichkeit, beim Mensch-ärger-dich-nicht herauszukommen?
  - konkret: A = bei dreimaligem Würfeln mindestens eine 6 würfeln

  Trick: Berechnung des Gegenereignisses
  - A^c = bei dreimaligem Würfeln keine 6 würfeln
  - $P(A^c) = 5^3/6^3$
  - $P(A) = 1 - P(A^c) = 1 - 5^3/6^3 = 42.13\ \%$

- Laplace-Experiment:

  Ω endlich

  Annahme: alle Elementarereignisse gleichwahrscheinlich

  Wahrscheinlichkeit eines Ereignisses durch Abzählen der enthaltenen Elementarereignisse

  $P(A) = |A| / |\Omega|$

  Kombinatorik nützlich
  - Zahl der Permutationen (Vertauschungen) von n Elementen

    $P(n) = n! = 1 \cdot 2 \cdot ... \cdot n$
  - Zahl der Möglichkeiten, k Elemente von n auszuwählen

    $$\binom{n}{k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{n!}{k!\,(n-k)!}$$

- Beispiel Kniffel:

  berechne P(FullHouse) beim Kniffel

  Anzahl von FullHouse-Ergebnissen, sortiert: XXXYY, 6*5 Möglichkeiten

  Anzahl verschiedener Reihenfolgen
  - generell: 5! = 120
  - Vertauschen der X bzw. Y untereinander liefert keine neuen Reihenfolgen!
  - es bleiben: 5!/(3! 2!) = 120/(6*2) = 10

  Damit
  - |A| = 10*30 = 300
  - $|\Omega| = 6^5 = 7776$
  - ⇒ P(A) = 300/7776 = 3.858 %

- Aufgaben:

  Aufgabe 1

  Aufgabe 2

Bedingte Wahrscheinlichkeit

- Definition der bedingten Wahrscheinlichkeit:

  Seien A, B zwei Ereignisse mit P(A) > 0, dann

  $$P(B|A) := \frac{P(A \cap B)}{P(A)}$$

  Wahrscheinlichkeit, dass B eintritt, falls A eingetreten ist

  anschaulich:
  - A eingetreten ⇒ nur Elementarereignisse aus A sind möglich
  - in B können nur die Elementarereignisse aus A eintreten

  erlaubt Berücksichtigung zusätzlicher Information

- Beispiel Fertigung mit zwei Fehlertypen:

  bei einer Fertigung treten zwei Fehlerarten A und B auf

  langfristige Beobachtungen liefern

P(nur A)	2%
P(nur B)	2.5%
P(A und B)	0.5%

  Wahrscheinlichkeit für Fehler B

  $P(B) = P(B \setminus A) + P(B \cap A) = 2.5\% + 0.5\% = 3\%$

  Wahrscheinlichkeit für Fehler B, wenn Fehler A aufgetreten ist

  $$\begin{aligned} P(A) &= P(A \setminus B) + P(A \cap B) = 2\% + 0.5\% = 2.5\% \\ P(B|A) &= \frac{P(B \cap A)}{P(A)} = \frac{0.5}{2.5} = 20\% \end{aligned}$$

- Multiplikationsregel:

  Umformen liefert (für $P(A_1) \neq 0$)

  $P(A_1 \cap A_2) = P(A_1)\, P(A_2 \mid A_1)$

  zweimal (für $P(A_1 \cap A_2) \neq 0$)

  $P(A_1 \cap A_2 \cap A_3) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2)$

  Iteration (für $P(A_1 \cap \ldots \cap A_{n-1} \neq 0)$

  $P(A_1 \cap \ldots \cap A_n) = P(A_1)\, P(A_2 \mid A_1) \ldots P(A_n \mid A_1 \cap \ldots \cap A_{n-1})$

- Beispiel Urne ohne Zurücklegen:

  Urne enthalte 40 rote und 60 schwarze Kugeln

  3 Kugeln werden nacheinander (ohne Zurücklegen) gezogen

  Ereignis A_i = i. Kugel ist rot

  Wie groß ist die Wahrscheinlichkeit, dass drei rote Kugeln gezogen werden?

  $$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \\ P(A_1) &= 40/100 \\ P(A_2|A_1) &= 39/99 \qquad \text{(eine rote ist ja schon raus)} \\ P(A_3|A_1 \cap A_2) &= 38/98 \qquad \text{(zwei rote sind ja schon raus)} \end{aligned}$$

  also

$P(A_1 \cap A_2 \cap A_3) = 40/100 \cdot 39/99 \cdot 38/98 = 6.110\ \%$

- Formel von der totalen Wahrscheinlichkeit:

  Ereignisse $A_1, \dots, A_n$ bilden eine disjunkte Zerlegung von Ω, d.h.

  A_i paarweise disjunkt, $A_1 \cup \dots \cup A_n = \Omega$

  dann

  $$P(B) = \sum_{k=1}^{n} P(B|A_k)P(A_k)$$

  Beweis

  $$\begin{aligned} P(B) &= P(B \cap \Omega) \\ &= P(B \cap (A_1 \cup \cdots \cup A_n)) \\ &= P((B \cap A_1) \cup \dots (B \cap A_n)) \\ &= P(B \cap A_1) + \cdots + P(B \cap A_n) \\ &= \sum_{k=1}^{n} P(B|A_k)P(A_k) \end{aligned}$$

- Beispiel Fußball-Ergebnis:

  Mannschaft BD spielt im Halbfinale gegen eine der drei Mannschaften HV, WB und BM, Gegner wird ausgelost.

  Trainer schätzt Siegchancen folgendermaßen ein
  - 80% gegen HV, 70% gegen WB, 30% gegen BM

  Wie groß ist die Wahrscheinlichkeit, das Finale zu erreichen?

  Ereignisse

B	BD gewinnt Halbfinale
A_1,A_2,A_3	Gegner ist HV, WB, BM

  Wahrscheinlichkeiten

  $P(A_1) = 1/3 = P(A_2) = P(A_3)$

  $P(B|A_1) = 0.8$, $P(B|A_2) = 0.7$, $P(B|A_3) = 0.3$

  damit

  $$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) \\ &= 0.8 * (1/3) + 0.7 * (1/3) + 0.3 * (1/3) = 0.6 \end{aligned}$$

- Satz von Bayes:

  Ereignisse $A_1, \dots, A_n$ bilden eine disjunkte Zerlegung von Ω, dann gilt

  $$\begin{aligned} P(A_i|B) &= \frac{P(A_i \cap B)}{P(B)} \\ &= \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)} \end{aligned}$$

  erlaubt "rückwärts Schließen", d. h. Berechnen von $P(A_i|B)$ aus $P(B|A_i)$

- Beispiel Krebstest:

  Ein Krebstest erkenne das Vorhandensein von Prostatakrebs mit 95% Wahrscheinlichkeit. Umgekehrt liefere er ein (falsches) positives Ergebnis bei 1% der getesten gesunden Männer. Bei 0.3% aller Männer trete der Krebs auf.

  Wie groß ist die Wahrscheinlichkeit, Prostatakrebs zu haben, wenn der Test positiv ist?

Ereignisse:

- B = Test ist positiv, A_1 = Person ist erkrankt, A_2 = Person ist gesund
- A_1, A_2 bilden disjunkte Zerlegung von Ω

Wahrscheinlichkeiten

- $P(A_1) = 0.3\%$, $P(A_2) = 99.7\%$
- $P(B|A_1) = 95\%$, $P(B|A_2) = 1\%$

Satz von Bayes liefert

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} \\ &= \frac{0.95 * 0.003}{0.95 * 0.003 + 0.01 * 0.997} \\ &= 22.23\% \end{aligned}$$

- Unabhängigkeit:

  A, B **unabhängig** :⇔ $P(A \cap B) = P(A)\, P(B)$

  direkte Folgerung (für $P(B) \neq 0$)

  - A, B unabhängig ⇔ $P(A) = P(A|B)$

  anschaulich: Eintreten von B ändert die Wahrscheinlichkeit für A nicht

  vgl. Beispiel Fertigung mit zwei Fehlertypen

  - $P(B) = 3\% \neq P(B|A) = 20\%$
  - ⇒ Fehler A und B sind nicht unabhängig

- Beispiel Würfelwurf:

  zweimaliger Würfelwurf, betrachen

  - A = 1. Wurf ergibt 6
  - B = 2. Wurf ergibt 6

  $P(A) = 1/6 = P(B)$, $P(A \cap B) = 1/36 = P(A)\, P(B)$

  also: A und B unabhängig

- Aufgaben:

  Aufgabe 3

  Aufgabe 4

  Aufgabe 5

Diskrete Zufallsverteilungen

- Zufallsvariablen
- Spezielle diskrete Verteilungen
- Eigenschaften diskreter Verteilungen

Zufallsvariablen

- Definition von reellen Zufallsvariablen:

  anschaulich: (numerisches) Ergebnis eines Zufallsexperiments

  mathematisch:

  Sei (Ω, $\mathcal{F}$, P) ein Wahrscheinlichkeitsraum. Eine **reelle Zufallsvariable** X ist eine Abbildung

  X: Ω → ℝ

  wobei für jedes a ∈ ℝ

  $$\{\omega \in \Omega | X(\omega) \leq a\} \in \mathcal{F}$$

  Bedingung wichtig für überabzahlbares Ω

  auch andere Zielräume als ℝ möglich
  - betrachten später $\mathbb{R}^n$
  - bis dahin hier immer ℝ

- Beispiele für Zufallsvariablen:

  Augensumme beim Wurf mit zwei Würfeln

  $$\begin{aligned} \Omega &= \{(\omega_1, \omega_2) | \omega_i \in 1 \ldots 6\} \\ X((\omega_1, \omega_2)) &= \omega_1 + \omega_2 \end{aligned}$$

  Anzahl der defekten Bauteile bei einer Stichprobe

  erzielte Sektornummer beim Glücksrad

  Lebensdauer eines Produkts

- Diskrete Zufallsvariable:

  Zufallsvariable, die nur abzählbar viele verschiedene Werte annimmt

  Beispiele oben
  - 1, 2, 3 diskret, 4 nicht
  - 3 diskret, obwohl Ω überabzählbar!

- Verteilungsfunktion:
  - Sei X eine Zufallsvariable, I ⊂ ℝ ein Intervall, dann heißt die Abbildung

    $$I \to P(\{\omega | X(\omega) \in I\})$$

    die **Verteilung von X**. Man schreibt vereinfacht

    $$\begin{aligned} P(X \in A) &:= P(\{\omega | X(\omega) \in A\}) \\ P(X = a) &:= P(\{\omega | X(\omega) = a\}) \end{aligned}$$

  - Für a ∈ ℝ definiert man die **kumulative Verteilungsfunktion** (**cdf**)

    F(a) := P(X ≤ a)

  - Bei einer diskreten Zufallsvariablen mit Wertebereich a_1 a_2, ... definiert man noch die **Wahrscheinlichkeitsfunktion**

    $p(a_i) := P(X = a_i)$

- Beispiel Augensumme zweier Würfel:

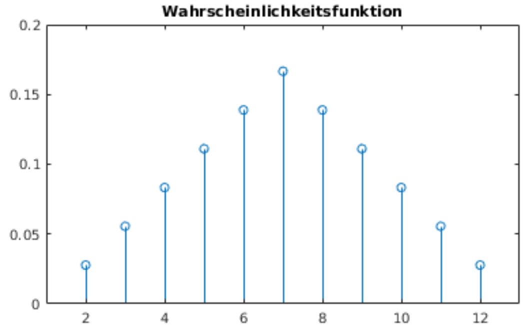
  Gleichverteilung aller Ergebnisse aus Ω

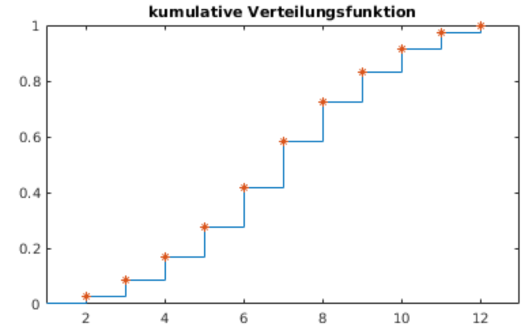
  mögliche Werte 2, 3, ..., 12

  Wahrscheinlichkeitsfunktion als Tabelle (durch Abzählen)

a_i	2	3	4	5	6	7	8	9	10	11	12
$p(a_i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

kumulative Verteilungsfunktion





Spezielle diskrete Verteilungen

- Gleichverteilung:

  mögliche Werte 1, ..., N

  p(k) = 1/N für k = 1, ..., N

  Beispiele: Münzwurf, Würfeln

- Binomialverteilung B(n,p):

  Bernoulli-Experiment
  - Experiment mit zwei Ausgängen ("Erfolg", "Misserfolg")
  - Wahrscheinlichkeit für Erfolg sei p

  X = Anzahl von Erfolgen bei n-maliger Wiederholung

  Wahrscheinlichkeitsfunktion

  $$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \le k \le n$$

  Begründung
  - ein bestimmtes Ereignis mit k Erfolgen hat Wahrscheinlichkeit p^k $(1\text{-}p)^{n\text{-}k}$
  - es gibt $\binom{n}{k}$ viele solcher Ereignisse

  Schreibweise
  - X ~ B(n,p) :⇔ X ist verteilt nach B(n,p)
  - analog für andere Verteilungen

- Anwendung Produktionsfehler:

  Bei einer Stanzmaschine seien 10% der gefertigten Teile außerhalb der benötigten Toleranz und müssen nachgearbeitet werden.

  Stichprobe mit 15 Teilen werde entnommen, X = Zahl der defekten Bauteile

  Wahrscheinlichkeiten gesucht für
  - genau zwei defekte Teile
  - Zahl der defekten Teile zwischen 5% und 15%
  - höchstens 12 defekte Teile

  Lösungen

  $$\begin{aligned}
  P(X=2) &= \binom{15}{2} 0.1^2 0.9^{13} = 26.69\% \\
  P(0.75 \le X \le 2.25) &= P(X=1) + P(X=2) \\
  &= \binom{15}{1} 0.1^1 0.9^{14} + \binom{15}{2} 0.1^2 0.9^{13} = 61.00\% \\
  P(X \le 12) &= 1 - P(X > 12) \\
  &= 1 - (P(X=13) + P(X=14) + P(X=15)) \\
  &= 1 - \left(\binom{15}{13} 0.1^{13} 0.9^2 + \binom{15}{14} 0.1^{14} 0.9^1 + \binom{15}{0} 0.1^{15} 0.9^0 \right) \\
  &= 1 - 8.6410 \cdot 10^{-12}
  \end{aligned}$$

- Geometrische Verteilung G(p):

  X = Zahl der Versuche, bis Bernoulli Erfolg liefert

  Wahrscheinlichkeitsfunktion

  $p(k) = (1 - p)^{k-1}$ p für k = 1, 2, ...

Beispiel Produktionsfehler

- Wie groß ist die Wahrscheinlichkeit, dass erst das 5. Teil defekt ist?

  $p(X = 5) = (1 - p)^4 \cdot p = 6.56\%$

manchmal zählt man nur Zahl der Fehlversuche Y

  $p(Y = k) = (1 - p)^k \, p$ für k = 0, 1, ...

- Verteilung von Y heißt auch $G^0(p)$

• Hypergeometrische Verteilung H(N, R, n):

Grundmodell

- Urne enthalte N Kugeln, R rote, N-R weiße
- man zieht n Kugeln ohne Zurücklegen
- X = Anzahl der gezogenen roten Kugeln

Zahl der möglichen Ziehungen insgesamt

$$|\Omega| = \binom{N}{n}$$

Zahl der Ziehungen mit genau r roten Kugeln

$$|A_r| = \binom{R}{r}\binom{N-R}{n-r}$$

Wahrscheinlichkeitsverteilung von X

$$P(X = r) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

• Beispiel Lottozahlen:

wie üblich werden 6 aus 49 gezogen

Vorstellung: die gezogenen werden rot bemalt

X = Zahl der richtigen, hypergeometrisch verteilt

insbesondere für 4 Richtige

$$P(X = 4) = \frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}} = 0.0969\%$$

• Poisson-Verteilung Po(λ):

Verteilung für häufige Experimente mit kleiner Erfolgs-Wahrscheinlichkeit

X = Zahl der Erfolge in einem gegebenen Zeitintervall

Beispiele

- Zahl von Schadensfällen pro Monat bei einer Versicherung
- Zahl radioaktiver Zerfälle pro Sekunde
- Zahl der Fahrzeuge pro Minute auf einer Straße

Parameter λ = Rate, "mittlere" Zahl der Erfolge pro Zeitintervall

Wahrscheinlichkeitsverteilung von X

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

- Beispiel Versicherung:

  Annahme: Zahl der Schadensfälle ist Poisson-verteilt mit Rate λ = 3/Monat

  berechne Wahrscheinlichkeiten folgender Ereignisse
  - genau 3 Schadensfälle im Monat
  - höchstens 2 Schadensfälle im Monat
  - mehr als 3 Schadensfälle im Monat

  Lösungen

$$\begin{aligned} P(X=3) &= e^{-\lambda}\frac{\lambda^3}{3!} = 22.40\% \\ P(X\le 2) &= P(X=0)+P(X=1)+P(X=2) = 42.32\% \\ P(X>3) &= 1-P(X=3)-P(X\le 2) = 35.28\% \end{aligned}$$

- Poisson-Grenzwertsatz:

  Betrachte binomialverteilte Zufallsvariablen X_n mit Parametern n und p_n, so dass

$$\lim_{n\to\infty} np_n = \lambda > 0$$

  X_n geht dann gegen eine Poissonverteilung, d. h.

$$\lim_{n\to\infty} P(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

  Beweis im Anhang

- Aufgaben:

  Aufgabe 6

  Aufgabe 7

  Aufgabe 8

Eigenschaften diskreter Verteilungen

- Erwartungswert einer diskreten Zufallsvariablen:

  X diskrete Zufallsvariable mit Werten a_i und Wahrscheinlichkeitsfunktion $p(a_i)$

  Erwartungswert E(X) (oder μ_X)

  $$E(X) := \sum_i a_i p(a_i)$$

- Beispiele:

  X Gleichverteilung bei Werten 1 ... N

  $$\begin{aligned} E(X) &= \sum_{i=1}^{N} i \frac{1}{N} \\ &= \frac{1}{N}\frac{1}{2} N(N+1) \\ &= \frac{N+1}{2} \end{aligned}$$

  X binomialverteilt nach B(2, 0.1)

  $$\begin{aligned} E(X) &= 0 \cdot p(0) + 1 \cdot p(1) + 2 \cdot p(2) \\ &= 1 \cdot \binom{2}{1} \cdot 0.1 \cdot 0.9 + 2 \cdot \binom{2}{2} \cdot 0.1^2 \\ &= 0.18 + 2 \cdot 0.01 = 0.2 \end{aligned}$$

  X binomialverteilt nach B(n,p) → E(X) = n p
  - Berechnung im Anhang

- Eigenschaften des Erwartungswerts:

  Transformation einer Zufallsvariable mit u: $\mathbb{R} \to \mathbb{R}$

  $$E(u(X)) = \sum_i u(a_i) p(a_i)$$

  linear

  E(aX) = a E(X)

  E(X + Y) = E(X) + E(Y)

  normiert

  E(1) = 1

- Anwendung: Erwartungswert bei B(n,p):

  Zufallsvariable X_i misst Erfolg bei i-tem Versuch

  $$X_i = \begin{cases} 1 & | \quad \text{Erfolg bei Versuch i} \\ 0 & | \quad \text{kein Erfolg bei Versuch i} \end{cases}$$

  Erwartungswert von X_i

  $$E(X_i) = 0 \cdot (1-p) + 1 \cdot p = p$$

  für X ~ B(n,p) ist

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

  also Erwartungswert von X

$$
\begin{aligned}
E(X) &= E\left(\sum_{i=1}^{n} X_i\right) \\
&= \sum_{i=1}^{n} E(X_i) \\
&= \sum_{i=1}^{n} p \\
&= np
\end{aligned}
$$

- Quantile:

  Für p ∈ [0,1] ist das **p-Quantil** von X ein Wert x_p mit

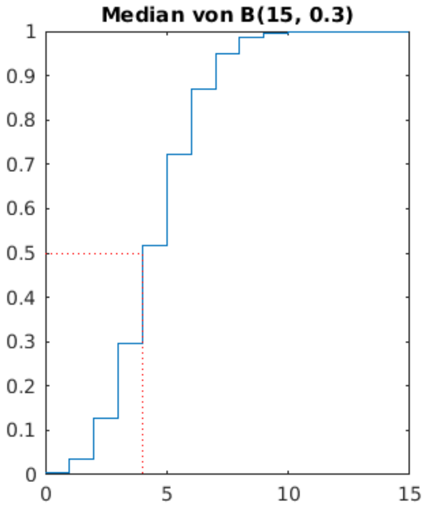
$$
\begin{aligned}
P(X \le x_p) &\ge p \\
P(X \ge x_p) &\ge 1-p
\end{aligned}
$$

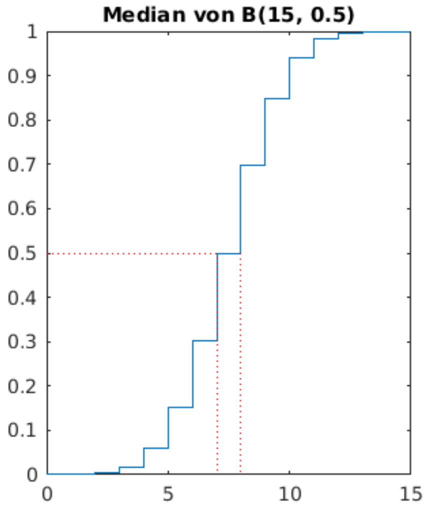
  Das 0.5-Quantil heißt auch **Median**.

  lässt sich aus der kumulativen Verteilungsfunktion F(x) ablesen

  zwei Möglichkeiten
  - F(x) nimmt p nicht an ⇒ p-Quantil eindeutig
  - F(x) nimmt p an ⇒ jeder Wert aus Intervall ist p-Quantil
  - man definiert dann den linken Endpunkt

  Beispiel Binomialverteilung





- Varianz und Standardabweichung:

  Wie stark streuen die Werte (genauer: Realisierungen) einer Zufallsvariablen X um den Mittelwert?

  Varianz

  $\mathrm{Var}(X) := E((X - E(X))^2)$

  Standardabweichung

  $\sigma(X) := \sqrt{\mathrm{Var}(X)}$

- Eigenschaften der Varianz:

  X Zufallsvariable, a, b, ∈ ℝ

$$
\begin{aligned}
\mathrm{Var}(X) &= E(X^2) - (E(X))^2 \\
\mathrm{Var}(aX+b) &= a^2\,\mathrm{Var}(X) \\
E((X-a)^2) &= \mathrm{Var}(X) + (E(X)-a)^2
\end{aligned}
$$

Beweis der ersten Formel

$$\begin{aligned}
\mathrm{Var}(X) &= E((X - E(X))^2) \\
&= E(X^2 - 2XE(X) + E(X)^2) \\
&= E(X^2) - 2E(X)E(X) + E(X)^2 \\
&= E(X^2) - E(X)^2
\end{aligned}$$

andere ähnlich

- Berechnung der Varianz für X gleichverteilt mit Werten {1, ..., N}:

  Berechnen zunächst $E(X^2)$

$$\begin{aligned}
E(X^2) &= \sum_{i=1}^{N} i^2 P(X = i) \\
&= \frac{1}{N} \sum_{i=1}^{N} i^2 \\
&= \frac{1}{N} \cdot \frac{1}{6} N(N+1)(2N+1) \\
&= \frac{1}{6}(N+1)(2N+1)
\end{aligned}$$

  Erwartungswert von X von oben liefert

$$E(X)^2 = \frac{1}{4}(N+1)^2$$

  damit erhält man

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{1}{12}(N+1)(N-1)$$

- Werte für Beispiel-Verteilungen:

	$E(X)$	$\mathrm{Var}(X)$
gleichverteilt	$\frac{N+1}{2}$	$\frac{N^2-1}{12}$
$B(n,p)$	np	$np(1-p)$
$G(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$G^0(p)$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
$H(N,R,n)$	$n\frac{R}{N}$	$n\frac{R}{N}\left(1-\frac{R}{N}\right)\frac{N-n}{N-1}$
$Po(\lambda)$	λ	λ

- Aufgaben:

  Aufgabe 9

Stetige Zufallsverteilungen

- Dichte- und Verteilungsfunktion
- Eigenschaften stetiger Verteilungen
- Spezielle stetige Verteilungen

Dichte- und Verteilungsfunktion

- Stetige Zufallsvariable:

  Dichtefunktion: integrierbare Funktion f: ℝ → ℝ mit

  $$f(t) \geq 0 \quad \forall t \in \mathbb{R}$$

  $$\int_{-\infty}^{\infty} f(t)\, dt = 1$$

  stetige Zufallsvariable X: Ω → ℝ mit

  $$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

  f heißt auch **Wahrscheinlichkeitsdichte von X**

  für kleines ε > 0

  $$P\left(a - \frac{\varepsilon}{2} \leq X \leq a + \frac{\varepsilon}{2}\right) = \int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f(x)\, dx \approx \varepsilon f(a)$$

  anschaulich
  - ε f(a) ist die Wahrscheinlichkeit, dass X im kleinen Intervall der Breite ε um a herum liegt

  insbesondere folgt mit ε → 0 für jedes x ∈ ℝ

  P(X = x) = 0

- (Kumulative) Verteilungsfunktion von X:

  wie oben definiert

  $$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

  falls F differenzierbar, gilt

  F'(x) = f(x)

  mit F hat man sofort

  P(a ≤ X ≤ b) = F(b) - F(a)

- Beispiel Gleichverteilung:

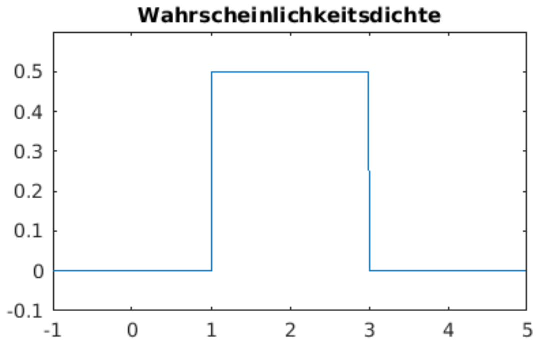
  X sei stetige Zufallsvariable mit der Dichtefunktion

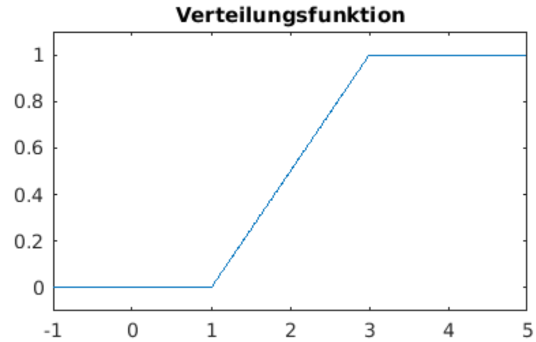
  $$f(x) = \begin{cases} \frac{1}{2} & | \quad 1 < x < 3 \\ 0 & | \quad \text{sonst} \end{cases}$$

  Verteilungsfunktion damit

  $$F(x) = \begin{cases} 0 & | \quad x \leq 1 \\ \frac{1}{2}(x-1) & | \quad 1 < x < 3 \\ 1 & | \quad x \geq 3 \end{cases}$$

  im Bild





damit etwa

$P(X \leq 2) = F(2) = 1/2$

$P(1.8 \leq X \leq 2.2) = F(2.2) - F(1.8) = 0.6 - 0.4 = 0.2$

$P(X \geq 2.5) = 1 - F(2.5) = 1 - 0.75 = 0.25$

- Aufgaben:

  Aufgabe 10

Eigenschaften stetiger Verteilungen

- Transformations-Formel:

  gesucht sei Wahrscheinlichkeitsdichte von Funktion von X, etwa X^2, log(X)

  allgemein Y = u(X) mit u streng monoton

  Satz: Y hat die Dichte

  $$g(y) = f(u^{-1}(y)) \left| \frac{d}{dy} u^{-1}(y) \right|$$

  Beweisidee: berechne Verteilungsfunktion von Y und verwende Substitution

  Beispiel: X habe nur positive Werte, gesucht Dichte von X^2

  $$\begin{aligned} u(x) &= x^2 \\ \Rightarrow \quad u^{-1}(y) &= \sqrt{y} \\ \Rightarrow \quad \frac{d}{dy} u^{-1}(y) &= \frac{1}{2\sqrt{y}} \\ \Rightarrow \quad g(y) &= \frac{f(\sqrt{y})}{2\sqrt{y}} \end{aligned}$$

- Erwartungswert einer stetigen Zufallsvariablen:

  X sei stetige Zufallsvariable mit Wahrscheinlichkeitsdichte f
  - **Erwartungswert** E(X) (oder μ_X)

  $$E(X) := \int_{-\infty}^{\infty} x f(x)\, dx$$

  Eigenschaften wie im diskreten Fall

  E(aX) = a E(X)

  E(X + Y) = E(X) + E(Y)

  E(1) = 1

  außerdem gilt

  $$E(u(X)) = \int_{-\infty}^{\infty} u(x) f(x)\, dx$$

- Varianz und Standardabweichung:

  definiert wie im diskreten Fall

  $\mathrm{Var}(X) := E((X - E(X))^2)$

  $$\sigma(X) := \sqrt{\mathrm{Var}(X)}$$

  Eigenschaften wie oben

  $$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - (E(X))^2 \\ \mathrm{Var}(aX + b) &= a^2\, \mathrm{Var}(X) \\ E((X - a)^2) &= \mathrm{Var}(X) + (E(X) - a)^2 \end{aligned}$$

- Median und Quantile:

  X stetige Zufallsvariable → Verteilungsfunktion macht keine Sprünge

  Für p ∈ [0,1] ist das **p-Quantil** von X ein Wert x_p mit

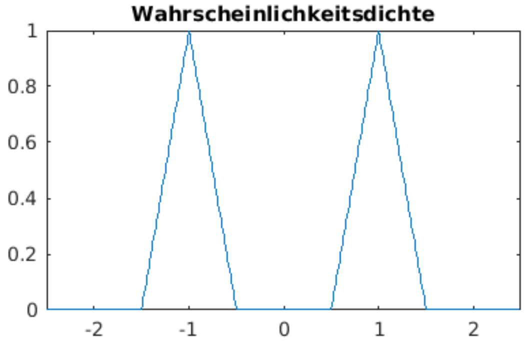
  $F(x_p) = p$

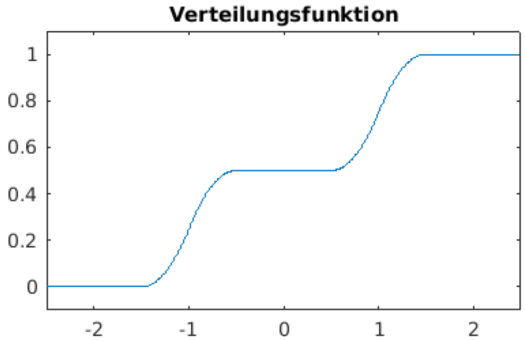
Das 0.5-Quantil heißt wieder **Median**.

F streng monoton → Quantile eindeutig bestimmt

Beispiel für nicht-eindeutigen Median

- Dichtefunktion mit zwei Maxima (**bimodal**) und Nullbereich





- Beispiel Gleichverteilung:

  Beispiel von oben (Gleichverteilung im Intervall [1,3])

  Berechnung des Erwartungswerts

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_1^3 x \cdot \frac{1}{2}\, dx = \frac{1}{4} x^2 \Big|_1^3 = 2$$

  Berechnung der Varianz

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_1^3 x^2 \cdot \frac{1}{2}\, dx = \frac{1}{6} x^3 \Big|_1^3 = \frac{13}{3}$$

  daher

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - (E(X))^2 = \frac{1}{3} \\ \sigma(X) &= \frac{1}{\sqrt{3}} = 0.5774 \end{aligned}$$

  Berechnung von Median, 5%- und 95%-Quantilen

$$\begin{aligned} F(x) &= \frac{1}{2}(x-1) = p \quad \text{für } x \in [1,3] \\ \Rightarrow \qquad x &= 2p + 1 \end{aligned}$$

  also

  $x_{0.5} = 2$

  $x_{0.05} = 1.1$

  $x_{0.95} = 2.9$

- Aufgaben:

  Aufgabe 11

Aufgabe 11

Spezielle stetige Verteilungen

- Gleichverteilung U(a,b) auf Intervall [a,b]:

  Dichtefunktion

  $$f(x) = \begin{cases} \dfrac{1}{b-a} & | \quad x \in [a,b] \\ 0 & | \quad \text{sonst} \end{cases}$$

  Kennwerte

  $$\begin{aligned} E(X) &= \frac{1}{2}(a+b) \\ \mathrm{Var}(X) &= \frac{1}{12}(b-a)^2 \end{aligned}$$

  Anwendung
  - Rundungsfehler
  - Wartezeiten (bei periodischer Abholung)

- Exponentielle Verteilung Ex(λ):

  Parameter $\lambda > 0$

  Dichtefunktion

  $$f(x) = \begin{cases} \lambda e^{-\lambda x} & | \quad x \geq 0 \\ 0 & | \quad x < 0 \end{cases}$$

  Verteilungsfunktion

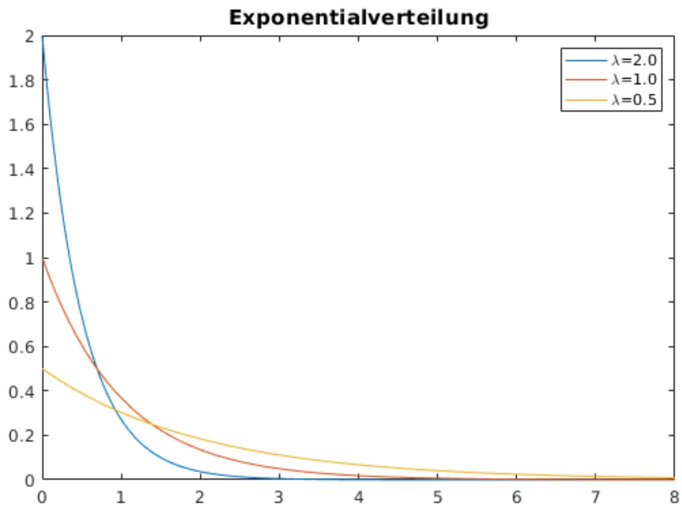
  $$F(x) = 1 - e^{-\lambda x}$$

  Kennwerte

  $$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ \mathrm{Var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

  Anwendung
  - Lebensdauer eines Geräts (ohne Alterung)
  - Wartezeit bis zum nächsten Schadensfall
  - Zeit bis zum nächsten radioaktiven Zerfall

  im Bild



- Besonderheiten der exponentiellen Verteilung:

  Gedächtnislosigkeit: Für X ~ Ex(λ) gilt

  $$P(X \geq x + y \mid X \geq x) = P(X \geq y)$$

  Bedeutung:
  - Hat man schon x Zeiteinheiten gewartet, ohne dass das Ereignis eingetreten ist, verringert das die weitere Wartezeit nicht.

  Zusammenhang mit Poisson-Verteilung:
  - X = Zeitdauer zwischen aufeinander folgenden unabhängigen Ereignissen
  - Y = Anzahl von Ereignissen in einem Zeitintervall
  - X ~ Ex(λ) ⇔ Y ~ Po(λ)

- Beispiel Produkt-Lebensdauer:

  Die Lebensdauer eines Produkts sei exponentiell verteilt mit mittlerer Lebensdauer von 6 Jahren.

  Wegen E(X) = 1/λ ist hier λ = 1/(6 Jahre)

  Wahrscheinlichkeit, dass es länger als 6 Jahre hält

  P(X ≥ 6) = 1 - F(6) = 1/e = 36.79 %

  Wie groß muss die Garantiezeit gewählt werden, wenn man höchstens p = 10% Ausfälle während der Garantiezeit haben möchte?

  $$\begin{aligned} P(X \leq T) &= p \\ &= F(T) = 1 - e^{-\lambda T} \\ \Rightarrow \qquad T &= -\frac{\ln(1-p)}{\lambda} = 0.6322 \text{ Jahre} \end{aligned}$$

- Gammaverteilung γ(r, λ):

  Parameter λ > 0, r > 0

  Dichtefunktion

  $$f(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & | \quad x \geq 0 \\ 0 & | \quad x < 0 \end{cases}$$

  - mit Gammafunktion

  $$\Gamma(r) := \int_0^\infty x^{r-1} e^{-x} \, dx$$

  - speziell für r ∈ ℕ

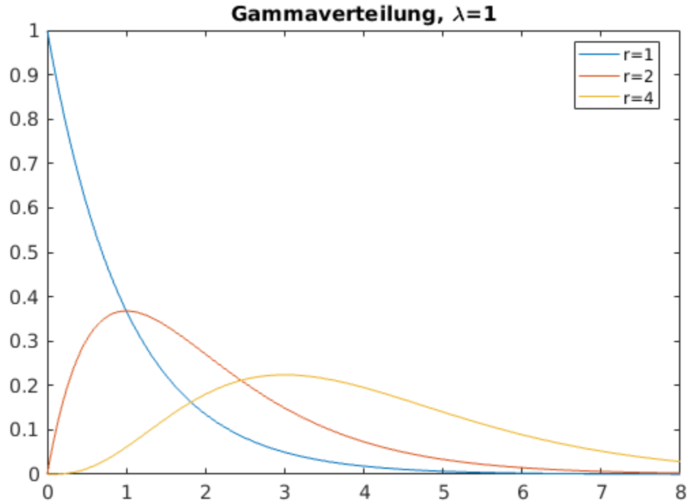
  $$\Gamma(r) = (r-1)!$$

  Kennwerte

  $$\begin{aligned} E(X) &= \frac{r}{\lambda} \\ \text{Var}(X) &= \frac{r}{\lambda^2} \end{aligned}$$

  Anwendung
  - Bedienzeiten oder Reparaturzeiten
  - Zeit zwischen (kleinen) Versicherungsschäden

  im Bild



- Spezialfälle der Gammaverteilung:

  Erlang-Verteilung
  - Erl(n, λ) = γ(n, λ) für n ∈ ℕ
  - Wartezeit auf Kette von n Ereignissen
  - Erl(1,λ) = Ex(λ)

  Chi-Quadrat-Verteilung mit f Freiheitsgraden (f ∈ ℕ)
  - $\chi_f^2 = \gamma(f/2, 1/2)$
  - wichtig bei statistischen Tests

- Normalverteilung N(μ, σ^2):

  Dichtefunktion

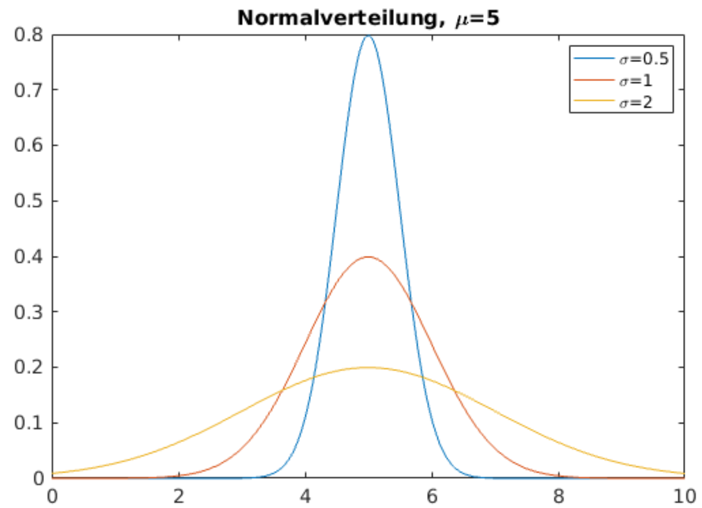
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  Kennwerte
  - E(X) = μ
  - Var(X) = σ^2

  Anwendung
  - wichtigste Verteilung überhaupt
  - Messfehler
  - Produktionsungenauigkeiten
  - Größenverteilung bei Menschen
  - und viele weitere

  in vielen Fällen gute Näherung (s.u.)

  im Bild



- Berechnung der Verteilungsfunktion:

  Problem: Stammfunktion der Dichte lässt sich nicht auf einfache Funktionen zurückführen

  Trick: Standardisierung

  - statt X ~ N(μ, $σ^2$) betrachte Z = (X - μ)/σ ~ N(0,1)

  Verteilungsfunktion für Standard-Normalverteilung

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

  Werte von Φ sind tabelliert, es reichen die für x ≥ 0 wegen

  Φ(-x) = 1 - Φ(x)

  wird oft auf die Fehlerfunktion erf(x) zurückgeführt

  $$\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\, dt$$

  - damit

  $$\Phi(x) = \frac{1}{2}\left(\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) + 1\right)$$

  in Matlab vorhandene Funktionen

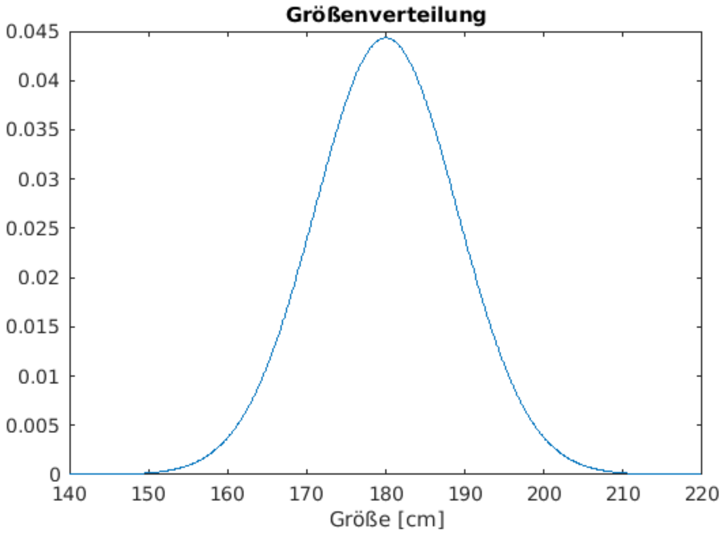
  - `erf(x)`
  - `erfc(x)                  % = 1 - erf(x), sinnvoll für größere x`
  - `erfinv(x)                % Inverse von erf`
  - `erfcinv(x)               % Inverse von erfc`

  in Matlab + Statistik-Toolbox vorhandene Funktionen

  - `normcdf(x)               % Phi(x)`
  - `normcdf(x, mu, sigma)    % F(x) bei X ~ N(mu, sigma)`
  - `norminv(p, mu, sigma)    % Inverse von normcdf`
  - und natürlich mit `makedist` und zugehörigen Funktionen

- Beispiel Größenverteilung:

  Die Körpergröße der erwachsenen Männer in einer Population sei normalverteilt mit einem Mittelwert von 180 cm und einer Standardabweichung von 9 cm.



Wie groß ist der Anteil der Männer, die größer sind als 200 cm?

Gesucht ist A = 1 - P(X ≤ 200) für X ~ N(180, 9^2).

Standardisierung

$$\begin{aligned} P(X \leq 200) &= P\left(\frac{X-180}{9} \leq \frac{20}{9}\right) \\ &= P\left(Z \leq \frac{20}{9}\right) \\ &= \Phi\left(\frac{20}{9}\right) \end{aligned}$$

Berechnung mit Matlab

```
x = 20/9;
A = 1 - 0.5*(erf(x/sqrt(2)) + 1)
```

- alternativ

```
A = 1 - normcdf(200, 180, 9)
```

Ergebnis: 1.31% der Männer sind über 200 cm groß.

- Weitere Eigenschaften der Normalverteilung:

  Quantile der Standard-Normalverteilung

p [%]	50	75	90	95	97.5	99
x_p	0.0000	0.6745	1.2816	1.6449	1.9600	2.3263

  zentrale Schwankungsintervalle

  - Anteil innerhalb mehrerer Standardbreiten, also

k	1	2	3	4
p_k	0.6827	0.9545	0.9973	0.9999

- Aufgaben:

  Aufgabe 12

  Aufgabe 13

  Aufgabe 14

Mehrdimensionale Zufallsvariablen

- Zufallsvektoren
- Unabhängige Zufallsvariablen
- Grenzwertsätze

Zufallsvektoren

- Definition von Zufallsvektoren:

  häufig interessieren mehrere Zufallsvariablen gleichzeitig
  - verschiedene Merkmale (Größe, Gewicht, Alter)
  - Lebensdauern verschiedener Teilkomponenten
  - Ergebnisse wiederholter Messungen

  formal
  - Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Ein **n-dimensionaler Zufallsvektor X** = $(X_1, X_2, \ldots, X_n)$ ist eine Abbildung $\mathbf{X}: \Omega \to \mathbb{R}^n$, wobei für jedes $(a_1, \ldots, a_n) \in \mathbb{R}^n$

    $$\{\omega \in \Omega | X_1(\omega) \le a_1 \cap \ldots X_n(\omega) \le a_n\} \in \mathcal{F}$$

  i. F. häufig nur für n=2 formuliert
  - Bezeichnung **X** = (X, Y)
  - Verallgemeinerung meistens klar

- Verteilungsfunktionen:

  für diskrete Zufallsvariable **X** = (X, Y)
  - **gemeinsame Wahrscheinlichkeitsfunktion** p(x,y) mit

    p(x,y) = P(X = x, Y = y)

  für stetige Zufallsvariable **X** = (X, Y)
  - **gemeinsame Dichtefunktion** f(x,y) mit

    $$P((X,Y) \in R) = \int\int_R f(x,y)\,dx\,dy$$

  - für Rechteck $R \subset \mathbb{R}^2$

  in beiden Fällen
  - **(kumulative) Verteilungsfunktion** F(x,y)

    $F(x,y) = P(X \le x, Y \le y)$

- Rand- oder marginale Verteilungen:

  Verteilung der einen Größe, unabhängig vom Wert der anderen

  X diskret → **marginale Wahrscheinlichkeitsfunktion** p_X

  $$p_X(x) = P(X = x) = \sum_y p(x,y)$$

  - analog p_Y

  X stetig → **marginale Dichtefunktion** f_X

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

  - analog f_Y

  X, Y **unabhängig** ⇔ $P(X \le x, Y \le y) = P(X \le x)\, P(Y \le y)$

  äquivalent dazu:
  - $p(x,y) = p_X(x)\, p_Y(y)$ (diskret)
  - $f(x,y) = f_X(x)\, f_Y(y)$ (stetig)

  Satz: X, Y unabhängig ⇒ E(X Y) = E(X) E(Y)

- Beispiel 3-2-Würfel:

Wurf mit zwei Würfeln A, B

- A: je zweimal die Zahlen 1, 2, 3
- B: je dreimal die Zahlen 1,2

Ergebnisraum

- Ω = {(1,1), (1,2), (2,1), (2,2), (3,1), (3,2)}
- alle ω mit p = 1/6

Zufallsvariable X = Summe der beiden Werte

- mögliche Werte 2,3,4,5

Zufallsvariable Y = Produkt der beiden Werte

- mögliche Werte 1,2,3,4,6

Tabelle der gemeinsamen Wahrscheinlichkeitsfunktion p(x,y)

X/Y	**1**	**2**	**3**	**4**	**6**
2	1/6	0	0	0	0
3	0	2/6	0	0	0
4	0	0	1/6	1/6	0
5	0	0	0	0	1/6

Randverteilungen durch Summe der Zeilen bzw. Spalten

X	2	3	4	5
p_X	1/6	2/6	2/6	1/6

Y	1	2	3	4	6
p_Y	1/6	2/6	1/6	1/6	1/6

offensichtlich nicht unabhängig, z.B.

- $p(3,2) = 2/6 \neq p_X(3)\, p_Y(2) = (2/6) \cdot (2/6)$

- Beispiel bivariate (zweidimensionale) Normalverteilung:

  definiert durch gemeinsame Dichtefunktion

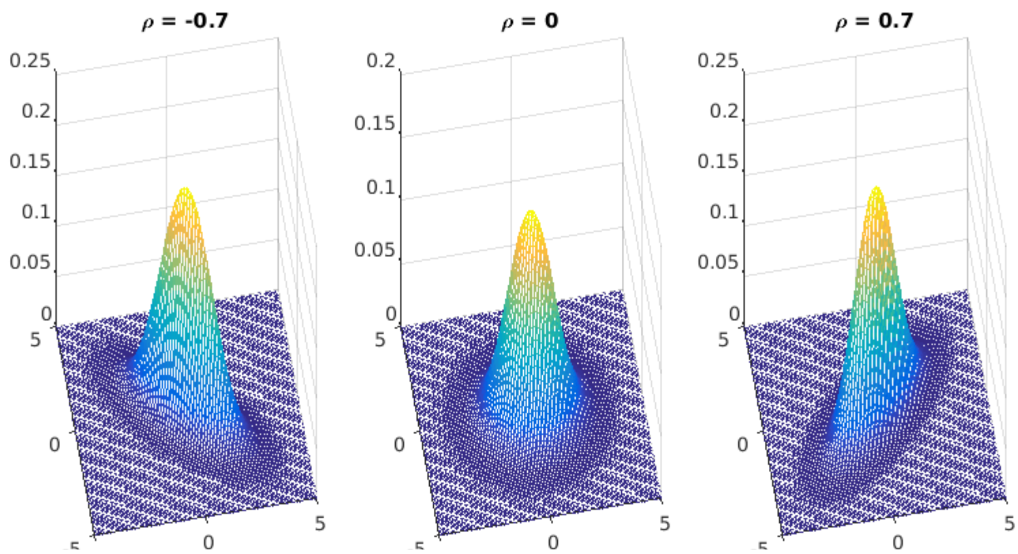
$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)}$$

  Parameter ρ: Korrelationskoeffizient, $|\rho| < 1$

  Berechnung der Randverteilungen ⇒

  - $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ (nachrechnen!)

  im Bild, für $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$



-5 -5 -5

- Mehrdimensionale (multivariate) Normalverteilung:

  n-dimensionale Standard-Normalverteilung mit **z** $\in \mathbb{R}^n$

  $$f_Z(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}}\, e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}}$$

  beliebige Normalverteilung durch lineare Transformation
  - **X = A Z + μ**
  - **μ** $\in \mathbb{R}^n$, **A** nicht-singuläre nxn-Matrix

  damit Dichtefunktion

  $$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{\Sigma})^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

  mit Kovarianzmatrix

  $\mathbf{\Sigma} = \mathbf{A}\,\mathbf{A}^T$

- Kenngrößen von Zufallsvektoren:

  Erwartungswert des Vektors ist einfach der Vektor der Einzelwerte

  E(**X**) := (E(X), E(Y))

  Erwartungswert einer Funktion u(X,Y) bei stetigem **X =** (X,Y)

  $$E(u(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x,y)f(x,y)\,dx\,dy$$

  Kovarianz zweier Zufallsvariablen

  Cov(X,Y) := E((X - E(X)) (Y - E(Y)))

  Eigenschaften der Kovarianz
  1. Cov(X,Y) = Cov(Y,X)
  2. Cov(X,X) = Var(X)
  3. Cov(aX + b, Y) = a Cov(X, Y)
  4. Cov(X,Y) = E(X Y) - E(X) E(Y)
  5. Var(X + Y) = Var(X) + Var(Y) + 2 Cov(X,Y)

  Kovarianz misst den Grad der linearen Abhängigkeit von X und Y

  X, Y **unkorreliert** :⇔ Cov(X,Y) = 0

  X, Y unabhängig ⇒ X, Y unkorreliert
  - Umkehrung gilt nicht!

- Kovarianzmatrix:

  bei zwei Zufallsvariablen

  $$\mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix}$$

  bei n Zufallsvariablen

  $$\mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & \mathrm{Var}(X_n) \end{pmatrix}$$

- Korrelationskoeffizient:

  Cov(X,Y) ändert sich mit Skalierung, daher durch Skalengröße teilen

  Definition des **Korrelationskoeffizienten**

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

  Eigenschaften
  - $|\rho_{X,Y}| \le 1$
  - $\rho_{X,Y} = 1 \Leftrightarrow X = aY + b$ mit $a > 0$
  - $\rho_{X,Y} = -1 \Leftrightarrow X = aY + b$ mit $a < 0$

  Stärke des linearen Zusammenhangs

- Kenngrößen im Beispiel 3-2-Würfel:

  Erwartungswert direkt über die Randverteilungen

$$\begin{aligned} E(X) &= \frac{1}{6}\cdot 2 + \frac{2}{6}\cdot 3 + \frac{2}{6}\cdot 4 + \frac{1}{6}\cdot 5 = 3.5 \\ E(Y) &= \frac{1}{6}\cdot 1 + \frac{2}{6}\cdot 2 + \frac{1}{6}\cdot 3 + \frac{1}{6}\cdot 4 + \frac{1}{6}\cdot 6 = 3.0 \end{aligned}$$

  analog $E(X^2)$ und $E(Y^2)$ ergibt
  - $E(X^2) = 13.1667$
  - $E(Y^2) = 11.6667$

  Berechnung von $E(X\,Y)$ über die Tabelle von $p(x,y)$

$$E(XY) = \frac{1}{6}\cdot 2\cdot 1 + \frac{2}{6}\cdot 3\cdot 2 + \frac{1}{6}\cdot 4\cdot 3 + \frac{1}{6}\cdot 4\cdot 4 + \frac{1}{6}\cdot 5\cdot 6 = 12$$

  damit

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E(XY) - E(X)E(Y) = 1.5 \\ \mathrm{Var}(X) &= E(X^2) - E(X)^2 = 0.9167 \\ \mathrm{Var}(Y) &= E(Y^2) - E(Y)^2 = 2.6667 \\ \rho_{X,Y} &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 0.9594 \end{aligned}$$

  X und Y sind stark positiv korreliert (klar!)

- Kenngrößen der bivariaten Normalverteilung:

  Randverteilungen X,Y (eindimensional) normalverteilt ⇒
  - $E(X) = \mu_1$, $E(Y) = \mu_2$
  - $\mathrm{Var}(X) = \sigma_1^2$, $\mathrm{Var}(Y) = \sigma_2^2$

  Berechnung von $E(X\,Y)$ als Doppelintegral liefert (nachrechnen!)
  - $E(X\,Y) = \sigma_1\,\sigma_2\,\rho + \mu_1\,\mu_2$

  damit sofort
  - $\mathrm{Cov}(X,Y) = \sigma_1\,\sigma_2\,\rho$
  - $\rho_{X,Y} = \rho$

  ρ ist wirklich der Korrelationskoeffizient

  hier gilt die Umkehrung: $\mathrm{Cov}(X,Y) = 0 \Rightarrow X, Y$ unabhängig
  - denn: $\rho = 0 \Rightarrow f(x,y) = f_X(x)\,f_Y(y)$

- Aufgaben:

  Aufgabe 15

Unabhängige Zufallsvariablen

- Unabhängigkeit mehrerer Zufallsvariablen $X_1, .., X_n$:

  Wahrscheinlichkeits- bzw. Dichtefunktion faktorisiert
  - $p(x_1, ..., x_n) = p(x_1) ... p(x_n)$ (diskret)
  - $f(x_1, ..., x_n) = f(x_1) ... f(x_n)$ (stetig)

  vereinfacht Berechnungen erheblich

  X, Y unabhängig ⇒
  - E(X Y) = E(X) E(Y)
  - Var(X + Y) = Var(X) + Var(Y)

  weitere Vereinfachung: alle X_i haben gleiche Verteilung
  - Bezeichnung: i.i.d. = independent identically distributed

  häufig gute Annahme für Messreihen

- Beispiel Rundungsfehler:

  Zahlenwerte im Computer
  - sind auf feste (binäre) Stellenzahl gerundet
  - Rechenergebnisse haben Rundungsfehler
  - Standardgenauigkeit $\delta \approx 2 \cdot 10^{-16}$
  - Abweichung maximal δ/2

  Annahmen (meistens gut erfüllt)
  - Rundungsfehler X gleichverteilt im Intervall [-δ/2, δ/2]
  - Rundungsfehler von Teilrechnungen unabhängig voneinander

  für Einzelfehler
  - E(X) = 0
  - $\mathrm{Var}(X) = \delta^2/12$

  Zufallsvariable S für Summe aus n Größen

$$\begin{aligned} S &= \sum_{i=1}^{n} X_i \\ E(S) &= 0, \quad \mathrm{Var}(S) = n\,\delta^2/12 \\ \sigma(S) &= \delta\sqrt{\frac{n}{12}} \end{aligned}$$

  mittlerer Fehler wächst mit $\sqrt{n}$, nicht mit n
  - Grund: positive und negative Fehler heben sich teilweise auf

- Summe unabhängiger Zufallsgrößen:

  X, Y unabhängige Zufallsvariable mit Dichten f(x), g(y)

  gesucht: Dichtefunktion h(z) von Z = X + Y

  Ergebnis

$$h(z) = \int_{-\infty}^{\infty} dx\, f(x) g(z-x)$$

  Beweis

$$
\begin{aligned}
P(Z \leq z) &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} dy\, f(x) g(y) \\
&\overset{(1)}{=} \int_{-\infty}^{\infty} dx \int_{-\infty}^{z} dt\, f(x) g(t-x) \\
&= \int_{-\infty}^{z} dt \left(\int_{-\infty}^{\infty} dx\, f(x) g(t-x) \right) \\
&= \int_{-\infty}^{z} dt\, h(t)
\end{aligned}
$$

mit der Substitution t = x + y, dt = dy bei (1)

Spezialfälle

$$
\begin{aligned}
X &\sim N(\mu_1, \sigma_1^2), \quad Y \sim N(\mu_2, \sigma_2^2) \quad \Rightarrow \quad X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\
X &\sim \gamma(r_1, \lambda), \quad Y \sim \gamma(r_2, \lambda) \quad \Rightarrow \quad X + Y \sim \gamma(r_1 + r_2, \lambda)
\end{aligned}
$$

- Beispiel Beladung eines Schmelzofens:

  Schmelzofen wird mit Kisten aus Ausschussmaterial beladen
  - Behälter für Schmelze fasst maximal 3000 kg

  unterschiedliche Masse pro Kiste, Verteilung X ~ N(100 kg, 400 kg^2)

  wird mit n = 27 Kisten beladen

  Gesamtladung S_n des Ofens also normalverteilt mit
  - $E(S_n)$ = n E(X) = 2700 kg
  - $Var(S_n)$ = n Var(X) = 10800 kg^2
  - $\sigma(S_n)$ = 103.9 kg

  Wahrscheinlichkeit für Überladung

$$
\begin{aligned}
P(S_n > 3000) &= 1 - P(S_n \leq 3000) \\
&= 1 - P((S_n - 2700)/103.9 \leq 2.8868) \\
&= 1 - \Phi(2.8868) \\
&= 0.1946\%
\end{aligned}
$$

- Maximum und Minimum unabhängiger Zufallsgrößen:

  gegeben: n i.i.d. Zufallsvariablen X_i mit Dichte f(x) und Verteilungsfunktion F(x)

  gesucht: Dichte $h_{max}(z)$ von Z = $Max(X_i)$

  Ergebnis

$$
h_{max}(x) = n f(x) F(x)^{n-1}
$$

  Beweis

$$
\begin{aligned}
H_{max}(x) &= P(Z \leq x) \\
&= P(X_1 \leq x, \ldots, X_n \leq x) \\
&= P(X_1 \leq x) \cdot \ldots \cdot P(X_n \leq x) \\
&= F(x)^n \\
h_{max}(x) &= \frac{d}{dx} H(x) \\
&= n F(x)^{n-1} F'(x) \\
&= n f(x) F(x)^{n-1}
\end{aligned}
$$

  analog für das Minimum

$$
\begin{aligned}
H_{min}(x) &= 1 - (1 - F(x))^n \\
h_{min}(x) &= n f(x)(1 - F(x))^{n-1}
\end{aligned}
$$

$$h_{min}(x) = nf(x)(1-F(x))^{n-1}$$

- Beispiel Gleichverteilung:

  n in [0,1] gleichverteilte Zufallszahlen X_i

  gesucht: Erwartungswert für Maximalwert Z

  Dichte und Verteilungsfunktion der einzelnen X_i

$$f(x) = \begin{cases} 1 & | \quad 0 \le x \le 1 \\ 0 & | \quad \text{sonst} \end{cases}$$

$$F(x) = \begin{cases} 0 & | \quad x < 0 \\ x & | \quad 0 \le x \le 1 \\ 1 & | \quad x > 1 \end{cases}$$

  mit obigen Formeln

$$h(x) = \begin{cases} nx^{n-1} & | \quad 0 \le x \le 1 \\ 0 & | \quad \text{sonst} \end{cases}$$

  Erwartungswert

$$E(Z) = \int_{-\infty}^{\infty} dx\, x h(x) = n \int_0^1 dx\, x^n = \frac{n}{n+1}$$

  z. B. für n = 10
  - E(Z) = 10/11 = 0.9091
  - P(Z ≤ 0.9) = 0.9^{10} = 34.87 %

- Aufgaben:

  Aufgabe 16

  Aufgabe 17

Grenzwertsätze

- Tschebyschoff-Ungleichung:

  X Zufallsvariable mit $E(X) = \mu$, $Var(X) = \sigma^2$, $c > 0$ beliebig ⇒

  $$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

  anschaulich: Abweichung einer Zufallsgröße vom Mittelwert selten größer als einige σ

- Beweis der Tschebyschoff-Ungleichung:

  Betrachte diskrete Zufallsgröße Y gegeben durch

  $$Y = \begin{cases} 0 & | \quad |X-\mu| < c \\ c^2 & | \quad |X-\mu| \geq c \end{cases}$$

  Offensichtlich gilt

  $$\begin{aligned} P(Y=0) &= P(|X-\mu| < c) \\ P(Y=c^2) &= P(|X-\mu| \geq c) \end{aligned}$$

  Daher ist Erwartungswert

  $$E(Y) = c^2 P(|X-\mu| \geq c)$$

  Außerdem ist

  $$Y \leq |X-\mu|^2$$

  Daraus folgt

  $$\begin{aligned} E(Y) \leq E((X-\mu)^2) &= \mathrm{Var}(X) = \sigma^2 \\ \Rightarrow \quad P(|X-\mu| \geq c) &\leq \frac{\sigma^2}{c^2} \end{aligned}$$

- (Schwaches) Gesetz der großen Zahlen:

  X_i Folge von i.i.d. Zufallsvariablen mit $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $\varepsilon > 0$ beliebig ⇒

  $$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) = 0$$

  anschaulich: Mittelwert geht mit Wahrscheinlichkeit 1 gegen Erwartungswert

  Stärkere Version (Mittelwert geht gegen Erwartungswert) ist falsch, da dauerhafte große Abweichungen möglich sind (wenn auch unwahrscheinlich).

  Spezialfall
  - $X_i \sim B(1,p)$, d.h. Eintreffen eines Ereignisses A mit Wahrscheinlichkeit p
  - $E(X_i) = p$
  - Mittelwert der X_i = relative Häufigkeit des Eintretens von A
  - Satz sagt: Relative Häufigkeit geht nach Wahrscheinlichkeit gegen p

  nachträgliche Begründung für Frequentismus (s.o.)

- Beweis des Gesetzes der großen Zahlen:

  Definiere Zufallsgröße $\widetilde{X}_n$

  $$\widetilde{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

Wegen Unabhängigkeit der X_i addieren sich Erwartungswerte und Varianzen. Daher

$$\begin{aligned} E(\widetilde{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \\ \mathrm{Var}(\widetilde{X}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Einsetzen in Tschebyschoff-Ungleichung ⇒

$$P(|\widetilde{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

mit $n \to \infty$ folgt Behauptung

- Zentraler Grenzwertsatz:

  X_i Folge von i.i.d. Zufallsvariablen mit $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

  betrachte Summenverteilung

  $$S_n := \sum_{i=1}^{n} X_i$$

  Fakt: S_n geht gegen Normalverteilung, aber
  - $E(S_n) = n\,\mu$, $Var(S_n) = n\,\sigma^2$
  - gehen im Limes gegen unendlich, daher mathematisch unbequem

  Ausweg: standardisierte Verteilung betrachten

  $$Z_n := \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

  Satz: Verteilung von Z_n geht gegen Standardnormalverteilung

  $$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z)$$

  Beweis aufwändig

  Satz verallgemeinerbar auf nicht-identische Verteilungen (unter bestimmten Voraussetzungen)

  Grund für häufiges Auftreten der Normalverteilung, z.B.
  - Messfehler oder Produktionsungenauigkeiten oft Summe vieler kleiner Einzelfehler
  - Größe von Menschen durch viele verschiedene Parameter bestimmt

- Approximation der Binomialverteilung:

  Einzelverteilung $X_i = B(1,p)$, $i = 1, .. n$
  - Eintreffen eines Ereignisses A mit Wahrscheinlichkeit p

  $X = \Sigma X_i$ = Anzahl der Treffer ~ $B(n, p)$

  Binomialverteilung nähern durch Normalverteilung

  $$P(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)$$

  Stetigkeitskorrektur
  - Treppenfunktion → glatte Dichtefunktion
  - genauer, wenn Dichtefunktion Stufen in der Mitte trifft

  besser also

$$P(X \leq x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(X = x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

Anwendungsbereich

- n p und n (1-p) groß genug
- Faustregel: beide ≥ 5

- Beispiel Produktionsfehler (s.o.):

  Fehlerquote p = 10 %, entnommen werden n = 100 Stück

  Check: n p = 10, n (1-p) = 90 → Näherung ok

  Berechnung von P(X = 10) mit Binomialverteilung

$$P(X = 10) = \binom{100}{10} 0.1^{10}\, 0.9^{90} = 13.19\%$$

  Berechnung von P(X = 10) mit Normalverteilung

$$\begin{aligned} P(X = 10) &\approx \Phi\left(\frac{10 + 0.5 - 10}{\sqrt{9}}\right) - \Phi\left(\frac{10 - 0.5 - 10}{\sqrt{9}}\right) \\ &= \Phi(0.1667) - \Phi(-0.1667) \\ &= 2\Phi(0.1667) - 1 \\ &= \mathrm{erf}(0.1667/\sqrt{2}) \\ &= 13.24\% \end{aligned}$$

- Aufgaben:

  Aufgabe 18

Simulation zufälliger Ereignisse

- Problemstellung
- Bestimmung von U(0,1)-verteilten Zufallszahlen
- Zufallszahlen mit vorgegebener Verteilung

Problemstellung

- Analyse komplexer stochastischer Situationen:

  Beispiel
  - Kaufhaus mit mehreren Kassen
  - Kunden kommen in zufälligen Abständen an (mit Verteilung X)
  - Abfertigungsdauer der einzelnen Kunden schwankt zufällig (mit Verteilung Y)

  Aufgabe
  - berechne mittlere Wartezeit der Kunden und mittlere Queuelänge

- Simulation:

  Alternative, falls explizite Berechnung nicht möglich oder zu aufwändig

  Vorgehensweise
  - "bestimme" Werte für Zufallszahlen mit Verteilungen X und Y
  - führe damit Beispielmodell durch und ermittle gesuchte Werte
  - wiederhole "oft" und berechne Mittelwerte

  Gesetz der großen Zahlen → Werte gehen gegen gesuchte Erwartungswerte

- "Echte" Zufallszahlen:

  reale zufällige Prozesse
  - Lottozahlen
  - Rauschen elektronischer Bauelemente
  - atmosphärisches Rauschen (s. random.org)
  - Quanteneffekte (s. ANU Quantum Optics)

  Probleme
  - für viele Zwecke zu langsam
  - abhörbar (zumindest die Netzdienste)

- Pseudo-Zufallszahlen:

  per Algorithmus erzeugte Zahlenfolge, die zufällig "aussieht"

  ausgefeilte, teilweise auch kryptographisch verwendete Verfahren (RNGs = Random Number Generators)

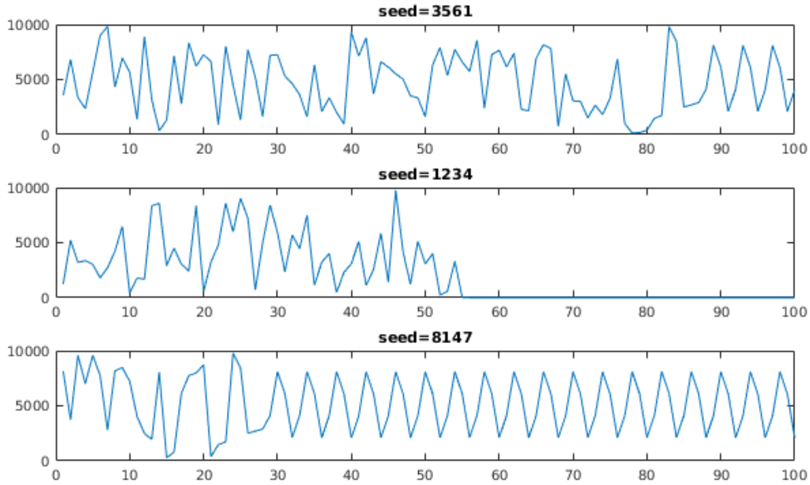
  bestehen - je nach Algorithmus - diverse statistische Tests, aber nie alle

  Vorteil: Reproduzierbarkeit

  einfaches (schlechtes!) Beispiel: Mittelquadratmethode von von Neumann
  - wähle Startwert (**seed**) mit 4 Dezimalen
  - quadriere
  - nächster Wert sind die 4 "mittleren" Dezimalstellen

  Beispielfolge
  - 1234, 5227, 3215, 3362, 3030, 1809, 2724, 4201, 6484, 422, 1780

  Methode i.a. ungeeignet



Anforderungen an geeigneten Algorithmus

- **nicht** möglichst kompliziert und unvorhersehbar **sondern** möglichst einfach
- besteht viele statistische Tests
- optimal: mathematische Analyse der Eigenschaften

Bestimmung von U(0,1)-verteilten Zufallszahlen

- Lineare Kongruenzgeneratoren (**LCG**):

  gegeben seien natürliche Zahlen m, a, c und Startwert X_0 (a, c, X_0 jeweils < m)

  berechne Folge ganzer Zahlen nach

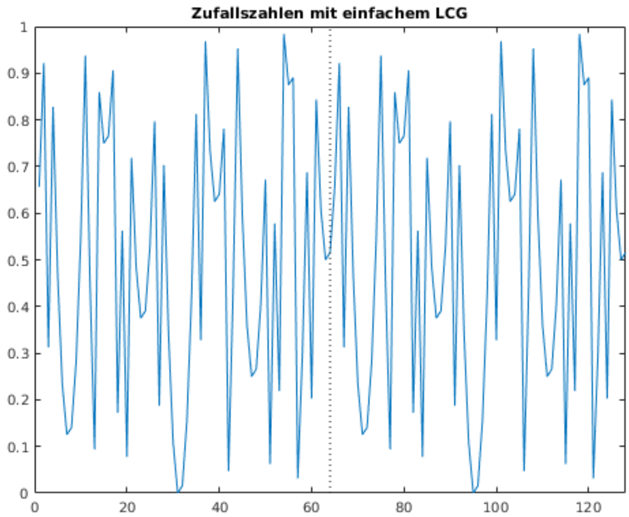
  $X_{n+1} = (a X_n + c) \mod m$

  zur Erinnerung: a mod m = Rest von a bei Teilen durch m

  Zufallszahl zwischen 0 und 1 dann

  $U_n = X_n / m$

  Beispiel mit kleinen Zahlen (m = 64, a = 9, c = 1, X_0 = 42)



- Eigenschaften von LCGs:

  Werte immer rational mit Nenner m

  Folge wiederholt sich mit gewisser Periode ≤ m
  - maximale Periode m möglich (für gewisse Werte von a, c, m)

  also: m möglichst groß und Zweierpotenz (→ mod-Operation sehr schnell)

  einige Beispielwerte

m	**a**	**c**	**Quelle**
2^{31}	1103515245	12345	Ansi-C
2^{24}	1140671485	12820163	MS Visual Basic (bis Version 6)
2^{48}	25214903917	11	Java (einfache Version)
2^{31}	65539	0	RANDU (bis in die 70er weit verbreitet, aber **ganz** schlecht!)

- Tests der Zufallszahlen:

  viele Tests denkbar und wichtig

  sehr ausführliche und vielseitige Testsammlung [11]

  genaue statistische Analyse braucht Testtheorie (s.u.)

  Fehlerniveau α bedeutet
  - Wenn die zu testende Vermutung (z. B. "Zufallszahlen gleichverteilt") abgelehnt wird, ist mit Wahrscheinlichkeit α die Vermutung doch wahr.

  i. f. zwei konkrete Beispiele

- Test auf Gleichverteilung:

Vorgehensweise

- erzeuge n Zufallszahlen U_i (i. F. $n=10^4$)
- zerlege [0,1] in k=10 gleich große Teilintervalle
- f_j (j = 1 .. k) ist Anzahl der Werte im j-ten Teilintervall
- berechne

$$\chi^2 := \frac{k}{n}\sum_{j=1}^{k}\left(f_j - \frac{n}{k}\right)^2$$

Idee: χ^2 sollte klein sein

konkret für α = 0.01: Annahme ablehnen, wenn

$$\chi^2 > \chi^2_{k-1}(1-\alpha) = \chi^2_9(0.99) = 21.67$$

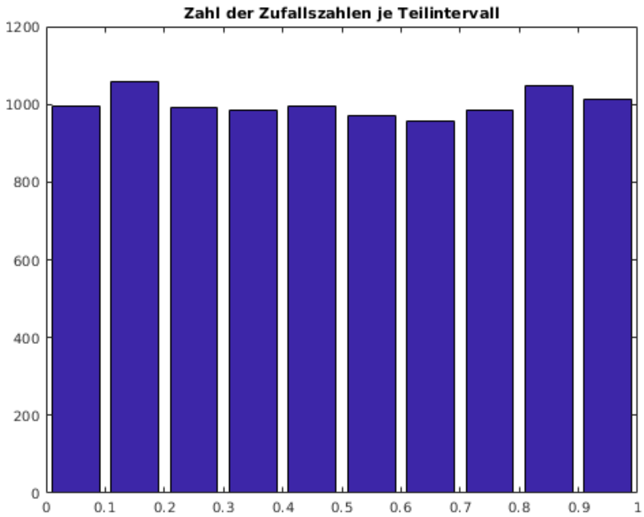
ist man mit 5% Fehlerwahrscheinlichkeit zufrieden, reicht

$$\chi^2 > \chi^2_{k-1}(1-\alpha) = \chi^2_9(0.95) = 16.92$$

zur Erinnerung: χ^2_f = Chi-Quadrat-Verteilung mit f Freiheitsgraden

konkret für Visual-Basic-RNG mit $X_0 = 42$

- $\chi^2 = 8.92$
- Verteilung auf die Teilintervalle



- Test auf Gleichverteilung im Raum:

Idee

- fasse je drei aufeinander folgende U_i zu einem 3d-Vektor V_i zusammen
- prüfe die Gleichverteilung der V_i im Einheitswürfel

konkretes Beispiel

- erzeuge $n=10^4$ Zufallsvektoren mit dem RANDU-Generator
- zerlege das Einheitsinterval in k = 10 gleich große Teile
- → der Einheitswürfel wird in k^3 = 1000 Teilwürfel zerlegt
- bestimme die Zahl $f_{i,j,l}$ der Vektoren im Teilwürfel (i,j,l)
- berechne

$$\chi^2 := \frac{k^d}{n}\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{k}\left(f_{i,j,l} - \frac{n}{k^d}\right)^2$$

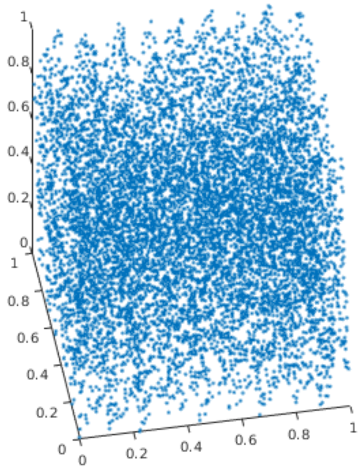
Ergebnis

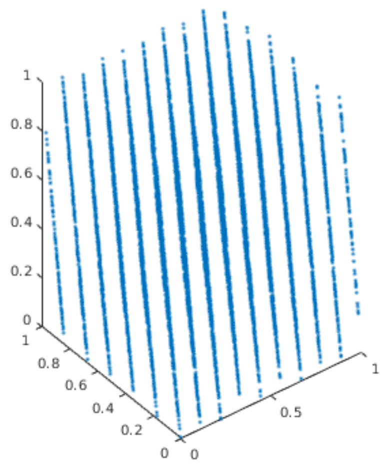
- bei seed = 42 erhält man $\chi^2 = 1175.2$
- Tabellenwerte:

$1-\alpha$	χ^2_{999}
0.99	1105.9
0.9999	1173.9

- Die Annahme der Gleichverteilung wird abgelehnt mit einer Fehlerwahrscheinlichkeit < 0.01%

Ursache
- Vektoren V_i sind auf nur 15 Ebenen im Raum verteilt!

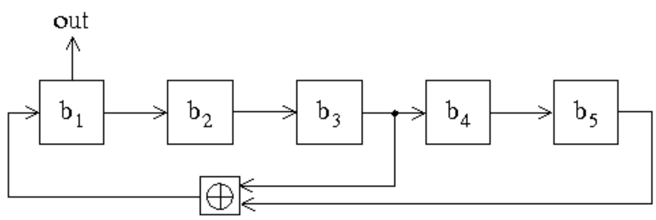




- LFSR-Generator (**Linear Feedback Shift Register**):

  erzeuge Folge von Bits folgendermaßen
  - q Bits werden in Schieberegister gespeichert
  - wähle feste Position r, berechne $b = b_r \oplus b_q$ ($\oplus$ = Summe mod 2 = XOR)
  - schiebe die Bits nach rechts und speichere b im ersten Register
  - gib jeweils Wert in b_1 aus

  Beispiel q = 5, r = 3



  - Startwert 11111 liefert Werte der Bits

11111
01111
00111
00011
10001
11000
01100
10110

1. Bit liefert Ausgangsfolge: 1 0 0 0 1 1 0 1

daraus Zahl in [0,1]
- fasse jeweils l Bits zu einer Binärzahl W_i zusammen

- Zufallszahl $U_i = W_i/2^l$

maximale Periode 2^q-1 (alle Zahlen außer 0) für geeignete q, r und l

- Erweiterungen:

  GFSR (**Generalized LFSR**)

  - arbeitet auf ganzen l-Bit-Worten

    $X_i = X_{i-r} \oplus X_{i-q}$

  - $\oplus$ bitweise berechnen

  Beispiel mit q = 5, r = 3, l = 4 und den ersten 5 Werten als Startwerten

1	1011
2	1010
3	1010
4	1100
5	1110
-	
6	0001
7	0110
8	0100
9	1101
10	1000

  Problem: Güte hängt stark von den Anfangswerten ab

  TGFSR (**Twisted GFSR**)

  - A sei binäre lxl-Matrix, X_i werde als l-Vektor aufgefasst, damit

    $X_i = X_{i-r} \oplus (A*X_{i-q})$

  - A*X wird ebenfalls mod 2 gerechnet

  Mersenne-Twister

  - spezieller TGFSR-Generator mit ein paar zusätzlichen Tricks
  - leicht zu initialisieren
  - Periode 2^{19937} - 1
  - besteht "Würfeltest" in bis zu 623 Dimensionen

  weit verbreitet

  - Programmiersprachen (C++, PHP, Python, ...)
  - numerische Programme (Matlab, Maple, Scilab, Excel, ...)
  - Statistik-Software (SPSS, SAS, R, ...)

- Weitere Verfahren:

  RNGs für mehrere Streams

  - erzeugen viele unabhängige Zufallsfolgen

  kryptographisch sichere RNGs

  - Parameter des zugrundeliegenden RNG lassen sich aus den Zufallszahlen nicht erschließen (in endlicher Zeit)

  Vorsicht: fehlerhafte RNGs haben schon manche Forschungsergebnisse ruiniert!

  - RNG in Excel vor 2003 fehlerhaft, sein Ersatz bis 2007 ebenfalls fehlerhaft [12]
  - seit 2010 Mersenne-Twister, aber in VBA immer noch defekter alter RNG [13]

- RNGs in Matlab:

  Grundfunktion `rand(N,M)`

- erzeugt NxM-Matrix aus U(0,1)-Zufallszahlen
- verwendet standardmäßig Mersenne-Twister

Startwert festlegen

- `rng(seed)` mit `seed` ≥ 0 ganzzahlig
- `rnd('shuffle')` verwendet Uhrzeit zur Initialisierung

RNG-Typ ändern mit `rnd(type)`

`'twister'`	Mersenne Twister
`'simdTwister'`	SIMD-oriented Fast Mersenne Twister
`'combRecursive'`	Combined Multiple Recursive (viele parallele Streams)
`'multFibonacci'`	Multiplicative Lagged Fibonacci (viele parallele Streams)

- Aufgaben:

  Aufgabe 19

Zufallszahlen mit vorgegebener Verteilung

- Aufgabenstellung:

  berechne Zufallszahlenfolge Y_i mit gegebener (diskreter oder kontinuierlicher) Verteilungsfunktion

  genauer: berechnete Folge verhält sich bei (vielen) statistischen Tests wie entsprechend verteilte Zufallsgröße

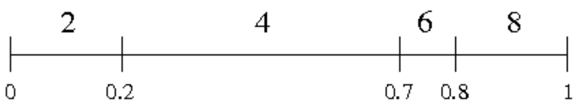
  Startpunkt immer berechnete Zufallszahlen $X_i \sim U(0,1)$

- Grundidee bei diskreter Verteilung:

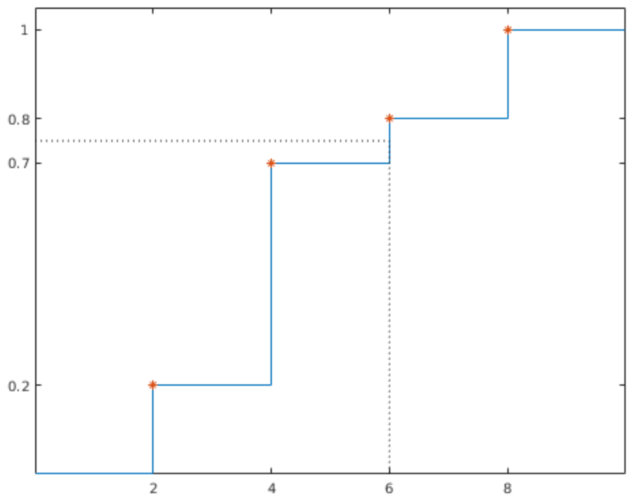
  gegeben sei Wahrscheinlichkeitsverteilung $p_k = P(Y = a_k)$

  zerlege Intervall [0,1] in Strecken I_k der Länge p_k
  - bestimme Wert von $X_i \sim U(0,1)$
  - $X_i \in I_k \;\Rightarrow\; Y_i = a_k$

  Beispiel
  - $a_i = (2, 4, 6, 8)$, $p_i = (0.2, 0.5, 0.1, 0.2)$



  Streckenaufteilung geht direkt mit der kumulativen Verteilungsfunktion $F(a) = P(X \le a)$



  - damit: $Y_i = F^{-1}(X_i)$

- Implementation in Matlab:

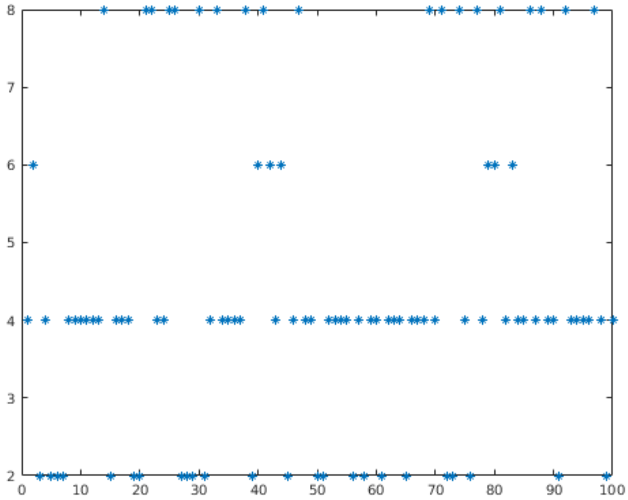
  Definition der Verteilung

```
a = [2, 4, 6, 8];
p = [0.2, 0.5, 0.1, 0.2];
F = cumsum(p)      % -> [0.2, 0.7, 0.8, 1.0]
```

  berechnen von Y

```
X = rand
idx = find(F>X, 1)
Y = a(idx)
```

  `find` klappt nicht ganz so einfach für X Vektor
  - quick and dirty: Schleife
  - Verteilung damit bei 100 Werten



erhaltene Anzahlen

2	4	6	8
24	50	7	19

- ist das ok? vgl. Testtheorie

- Verbesserungen:

  Grundverfahren aufwändig

  Trick für Gleichverteilung in n:m

  Y_i = floor(n + (m - n + 1) X_i)

  Trick für Bernoulli

  $X_i \leq p \rightarrow Y_i = 1$, sonst 0

  - entspricht Grundverfahren

  Trick für Binomial-Verteilung B(n,p)

  - mache n Bernoulli-Würfe Z_k, k = 1..n

  $Y_i = Z_1 + \ldots + Z_n$

- Grundidee bei stetiger Verteilung:

  Ausgangspunkt Verteilungsfunktion F(y) = P(Y ≤ y)

  inverse Transformation (wie bei diskreter Verteilung)

  $Y_i = F^{-1}(X_i)$

  prima, wenn F^{-1} explizit berechenbar

  Gleichverteilung U(a,b)

$$\begin{aligned} F(x) &= \frac{x-a}{b-a} \qquad \text{für } x \in [a,b] \\ F^{-1}(x) &= (b-a)x + a \end{aligned}$$

  Exponentielle Verteilung Ex(λ)

$$\begin{aligned} F(x) &= 1 - e^{-\lambda x} \\ F^{-1}(x) &= -\frac{1}{\lambda}\ln(1-x) \end{aligned}$$

  - kleine Vereinfachung, da auch 1 - X_i ~ U(0,1)

$$Y_i = -\frac{1}{\lambda}\ln(X_i)$$

- Normalverteilung:

  Berechnung von F^{-1} nur numerisch, zu aufwändig

  es genügt Y ~ N(0,1), denn damit

  $Z = \mu + \sigma Y \sim N(\mu, \sigma^2)$

  Box-Muller-Algorithmus

  - bestimme X_1, X_2 ~ U(0,1), damit

  $$\begin{aligned} Y_1 &= \sqrt{-2\ln X_1}\ \cos(2\pi X_2) \\ Y_2 &= \sqrt{-2\ln X_1}\ \sin(2\pi X_2) \end{aligned}$$

  Y_1, Y_2 ~ N(0,1), unabhängig

  Beweisidee:

  - betrachte 2d-Normalverteilung
  - gehe zu Polarkoordinaten über

  Details im Anhang

- Alternative Zikkurat-Verfahren:

  heute verbreitet, weil schneller

  geht grundsätzlich für beliebige Verteilungen

  Grundidee

  - überdecke Fläche unter dem Graphen der Dichtefunktion mit gleichverteilten Zufallspunkten
  - wähle einen Punkt und gebe die x-Koordinate aus

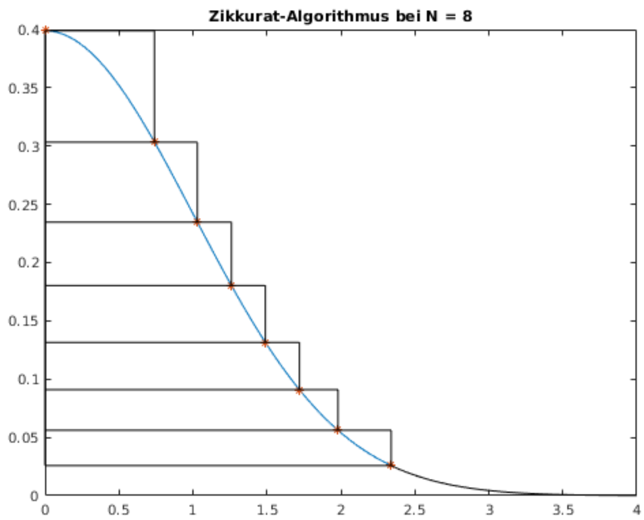
  praktische Umsetzung

  - würfle Punkt auf der Ebene in geeignetem Bereich
  - teste, ob er unterhalb des Graphen der Dichtefunktion liegt
  - ja: gib x-Wert zurück, nein: verwirf den Punkt

  Problem: x-Werte gehen grundsätzlich bis unendlich

  Verbesserung (weniger verworfene Punkte)

  - überdecke Graphen mit N-1 Rechtecken und einem Endstück jeweils gleicher Fläche

  - würfle Nummer 1 .. N und einen Punkt im Rechteck
  - besondere Behandlung für das unterste Stück

- Aufgaben:

  Aufgabe 20

  Aufgabe 21

Stochastische Prozesse

- Markov-Ketten
- Bediensysteme

Markov-Ketten

- Stochastischer Prozess:

  anschaulich: zeitabhängige Zufallsgröße

  etwas genauer: Familie von Zufallsvariablen X(t), t "Zeit"

  - $t \in \mathbb{R}^+$ stetiger stochastischer Prozess
  - $t \in \mathbb{N}$ diskreter stochastischer Prozess

  Wertebereich von X stetig (oft $\mathbb{R}$) oder diskret (oft {1,2,...,n} oder $\mathbb{N}$)

  betrachten hier nur wertediskrete Prozesse, meistens auch zeitdiskret

- Beispiele:

  Roulette-Spieler
  - beginnt mit 20 €
  - setzt beim Roulette immer 10 € auf Rot
  - hört auf, wenn er kein Geld mehr hat oder 50 €
  - X(i) = Besitz nach der i-ten Runde (diskret)

  Kreisprozess
  - System befindet sich in einem von N Zuständen
  - In jedem Schritt wechselt es zum nächst größeren, nach Schritt N zurück auf Zustand 1 (diskret)

  Börsennotierung einer Firma zur Zeit t (stetig oder diskret)

  Länge der Warteschlange an einer Kasse (stetig)

- Markov-Prozess:

  anschaulich: nächster Wert ist nur abhängig vom letzten, nicht von der ganzen Vergangenheit

  diskreter stochastischer Prozess X_n heißt **Markov-Kette** :⇔

  $P(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1} \mid X_n = x_n)$

  homogene Markov-Kette (HMK)
  - Übergangs-Wahrscheinlichkeit $P(X_{n+1} = x_{n+1} \mid X_n = x_n)$ hängt nicht vom Zeitpunkt n ab

  stetiger stochastischer Prozess X(t) heißt **Markov-Prozess** :⇔
  - für t, s > 0, 0 ≤ u < s gilt

    $P(X(t+s)=j \mid X(s) = i, X(u) = k) = P(X(t+s)=j \mid X(s) = i)$
  - homogen :⇔ $P(X(t+s)=j \mid X(s) = i)$ hängt nicht von s ab

  Beispiele
  - Roulette: homogene Markov-Kette
  - Börsenkurs: vermutlich nicht Markov
  - Warteschlange: hängt ab von der Art der Ankunfts- und Abgangsprozesse

- Übergangs-Matrix P einer HMK:

  Definition

  $$\mathbf{P} = (P_{ij}) \quad \text{mit} \quad P_{ij} = P(X_{n+1} = j | X_n = i)$$

  Wahrscheinlichkeit = 1, irgendwohin zu kommen

  $$\sum_{j=0}^{\infty} P_{ij} = 1$$

  Zustandsvektor v ist Zeilenvektor (!) mit

  $v_i = P(X_n = i)$

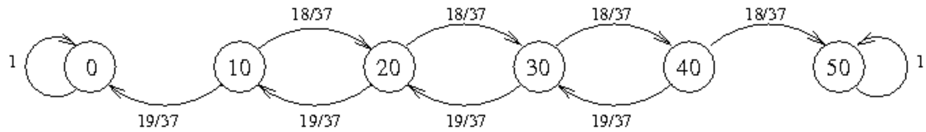
damit Zustandvektor u nach einem Schritt

$$\begin{aligned} u_j &= P(X_{n+1} = j) \\ &= \sum_{i=0}^{\infty} P(X_{n+1} = j, X_n = i) \\ &= \sum_{i=0}^{\infty} P(X_{n+1} = j | X_n = i) \cdot P(X_n = i) \\ &= \sum_{i=0}^{\infty} P_{ij} v_i \\ &= (v \cdot P)_j \end{aligned}$$

- Übergangs-Graph:

  graphische Darstellung einer homogenen Markov-Kette
  - Werte von X_n (Zustände) als Knoten
  - Übergänge zu anderen Zuständen als Pfeile mit Angabe von P_{ij}

  Beispiel Roulette



- Beispiel Autovermietung:

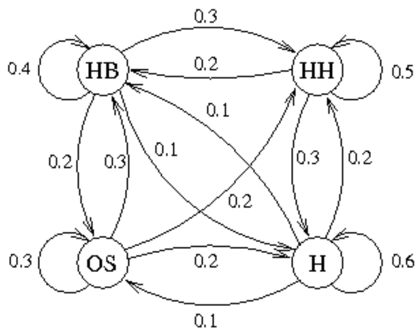
  Niederlassungen in HH, H, HB, OS

  Autos können an beliebigen Niederlassungen zurückgegeben werden

  Statistik des Kundenverhaltens liefert Übergangs-Matrix für Ort eines Autos

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.3 & 0.2 & 0.0 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0.3 & 0.1 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.3 & 0.3 \end{pmatrix}$$

  Übergangs-Graph



- Wahrscheinlichkeit späterer Zustände:

  Frage: Wie groß ist die Wahrscheinlichkeit $P(X_5 = 50)$, beim Roulette-Beispiel nach 5 Schritten gewonnen zu haben?

  allgemein gesucht

  $P_{ij}^n = P(X_{n+k} = j \mid X_n = i)$ $(n \geq 0)$

  Chapman–Kolmogorov-Gleichung

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m \qquad (n, m \geq 0)$$

- Beweis

$$
\begin{aligned}
P_{ij}^{n+m} &= P(X_{n+m} = j | X_0 = i) \\
&= \sum_{k=0}^{\infty} P(X_{n+m} = j, X_n = k | X_0 = i) \\
&= \sum_{k=0}^{\infty} P(X_{n+m} = j | X_n = k, X_0 = i) P(X_n = k | X_0 = i) \\
&= \sum_{k=0}^{\infty} P_{kj}^{m} P_{ik}^{n}
\end{aligned}
$$

geschrieben als Matrizen

$$\mathbf{P}^{(n+m)} = \mathbf{P}^{(n)} \cdot \mathbf{P}^{(m)}$$

- insbesondere

$$
\begin{aligned}
\mathbf{P}^{(1)} &= \mathbf{P} \\
\mathbf{P}^{(2)} &= \mathbf{P}^{(1+1)} = \mathbf{P} \cdot \mathbf{P} = \mathbf{P}^2 \\
&\vdots \\
\mathbf{P}^{(n+1)} &= \mathbf{P}^{(n)} \cdot \mathbf{P}^{(1)} = \mathbf{P}^n \cdot \mathbf{P} = \mathbf{P}^{n+1}
\end{aligned}
$$

im Beispiel

- Zustände (0,1,2,3,4,5) ≙ (0,10,20,30,40,50) €

$$
\mathbf{P} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
19/37 & 0 & 18/37 & 0 & 0 & 0 \\
0 & 19/37 & 0 & 18/37 & 0 & 0 \\
0 & 0 & 19/37 & 0 & 18/37 & 0 \\
0 & 0 & 0 & 19/37 & 0 & 18/37 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

- mit Matlab leicht P^5 berechnen und ablesen

  $P_{2,5}{}^5 = 0.2014$

- Eigenschaften von Zuständen:

  Zustand j heißt **von i aus erreichbar**, wenn man mit Wahrscheinlichkeit > 0 in n ≥ 0 Schritten von i nach j kommt, also

  $P_{i,j}{}^n > 0$ für ein n ≥ 0

  Zustände i, j heißen **verbunden**, wenn i von j aus erreichbar ist und j von i aus

  Übergangs-Graph zerfällt in Erreichbarkeitskomponenten

  - Beispiel Roulette: {0}, {10, 20, 30, 40}, {50}
  - HMK **irreduzibel** :⇔ es gibt nur eine Komponente

  Zustand i heißt **rekurrent**, wenn mit Wahrscheinlichkeit = 1 nach Start bei i in endlich vielen Schritten i wieder erreicht wird, sonst **transient**

  Satz

  - Mit Wahrscheinlichkeit = 1 wird ein rekurrenter Zustand unendlich oft angenommen, ein transienter nur endlich oft.

  Satz

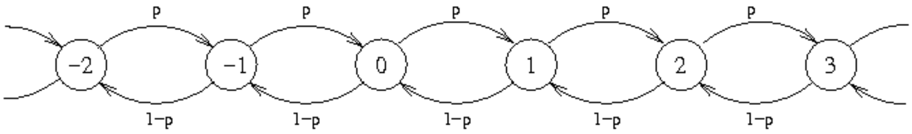
  - Zwei verbundene Zustände sind entweder beide transient oder beide rekurrent.

- Beispiel Zufallspfad:

  eindimensional

  - Zustandsraum seien die ganzen Zahlen $\mathbb{Z}$
  - mit Wahrscheinlichkeit p wechselt man von Zustand i in i+1, mit 1-p von i nach i-1, also

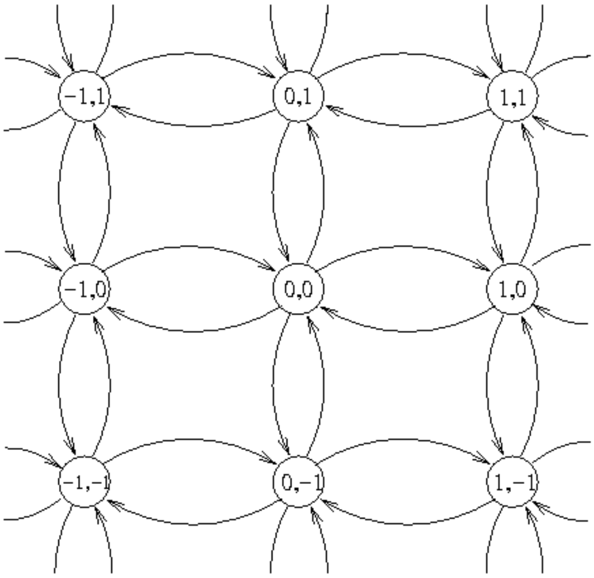
$P_{i,i+1} = p$, $P_{i,i-1} = 1-p$



- besonders interessant ist der symmetrische Fall p=1/2

zweidimensional

- Zustandraum ist $\mathbb{Z}^2$, also ein quadratisches Gitter
- Übergang von einem Punkt (i,j) zu einem der vier Nachbarpunkte
- betrachten nur den symmetrischen Fall mit Wahrscheinlichkeit 1/4 in jeder Richtung



dreidimensional

- analog für $\mathbb{Z}^3$ mit Wahrscheinlichkeit 1/6 zu jedem der sechs Nachbarpunkte

Satz

- Beim unsymmetrischen 1d-Zufallspfad sind alle Zustände transient, beim symmetrischen alle rekurrent.
- Beim symmetrischen 2d-Zufallspfad sind alle Zustände rekurrent.
- Beim symmetrischen 3d-Zufallspfad sind alle Zustände transient.

Merkregel

- Ein betrunkener Mensch findet immer irgendwann nach Hause, ein betrunkener Vogel nicht unbedingt!

- Gleichgewichtsverteilung (GGV):

HMK ist im Gleichgewicht :⇔ $P(X_{n+1} = i) = P(X_n = i)$ (für alle i)

Beispiel Autovermietung

- Verteilung der Autos auf die Niederlassungen, so dass sich diese im Mittel nicht ändert

Gleichgewichtsverteilung π gegeben durch

$$\begin{aligned} \boldsymbol{\pi} &= \boldsymbol{\pi} \cdot \mathbf{P} \\ \sum_{i=0}^{\infty} \pi_i &= 1 \end{aligned}$$

praktische Berechnung

- homogenes Gleichungssystem π (P - 1) = 0 lösen
- in der Regel ein Wert frei, auf 1 setzen
- anschließend Ergebnisvektor auf Gesamt-Wahrscheinlichkeit 1 normieren

Matlab-Trick

- Löse (P' - 1) π' = 0 mit `null(P' - eye(n))`
- n = Zahl der Zustände

- GGV in den Beispielen:

  Autovermietung mit Matlab

  π = [0.3172, 0.3495, 0.2204, 0.1129]

  Roulette (mit Matlab oder leicht per Hand)

  π = [p, 0, 0, 0, 0, 1-p], p ∈ [0,1]

- Existenz einer GGV:

  Satz

  - Sind alle Zustände transient, gibt es keine GGV.

  Voraussetzung nur für unendliche Zustandsräume erfüllbar

  Beispiel: unsysmmetrischer Zufallspfad

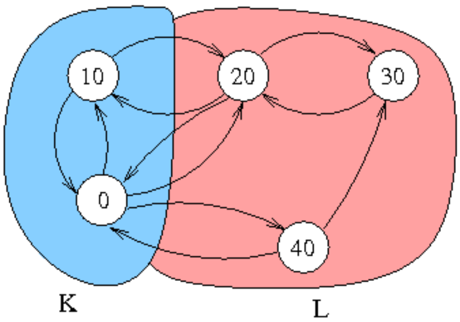
  Satz

  - Irreduzible Markow-Ketten mit endlichem Zustandsraum haben eine eindeutige GGV.

  Roulette-Beispiel nicht irreduzibel, hat zwei unabhängige GGV.

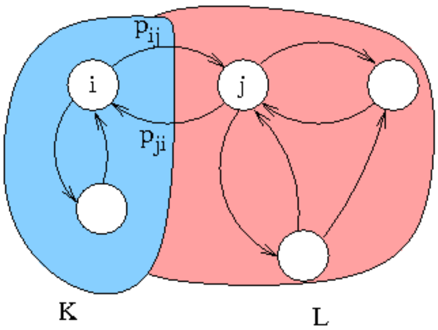
- Schnittprinzip:

  Übergangs-Graph sei in zwei Teile K, L zerlegt



  - Für GGV gilt: Übergangs-Wahrscheinlichkeit von K nach L = Übergangs-Wahrscheinlichkeit von L nach K

  speziell, wenn Zerlegung an einer (Doppel-)Kante möglich

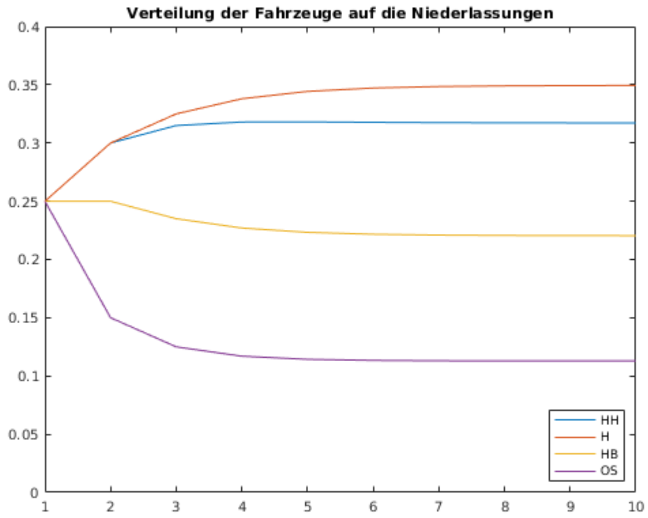


  $\pi_i P_{ij} = \pi_j P_{ji}$

- Langzeitverhalten:

  Beispiel Autovermietung, Start mit Gleichverteilung

  - wiederholte Anwendung von P liefert



viele, aber nicht alle HMK konvergieren gegen die GGV

- Problem sind Schleifen wie im Kreisprozess-Beispiel

hinreichend: irreduzibel und keine Schleifen

- Aufgaben:

  Aufgabe 22

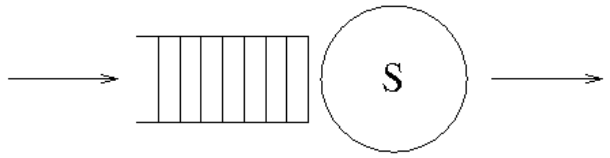
  Aufgabe 23

Bediensysteme

- Bediensystem oder Wartesystem (Queuing system):

  abstraktes stochastisches Modell aus
  - Ankunftsprozess
  - Warteschlangen-Politik
  - Bedienprozess

  graphisch

  Beispiele
  - Kundenabfertigung an Schaltern
  - industrielle Fertigungsprozesse
  - Rechner-Netzwerke
  - Verkehrsplanung

- Kendall-Notation:

  übliches Klassifizierungs-Schema von Wartesystemen

  enthält sechs Parameter A|B|s|c|n|Q

  Bedeutung

Parameter	Bedeutung	wichtige Werte
A	Art des Ankunftsprozesses	M (Markov), G (allgemein)
B	Art des Bedienprozesses	M (Markov), G (allgemein)
s	Zahl der Bediener	1, 2, .., ∞
c	Größe des Warteraums	1, 2, .., ∞
n	max. Anzahl von Kunden	1, 2, .., ∞
Q	Queue-Politik	FIFO, LIFO, Prio, Random

  betrachten i.F. nur $c = \infty$, $n = \infty$, Q = FIFO
  - werden in Notation weggelassen

  wichtigstes Beispiel: M|M|1

- Interessierende Kennzahlen:

  mittlere Zahl von Kunden im System L

  mittlere Zahl von Kunden in der Queue L_Q

  mittlere Zeit eines Kunden im System W

  mittlere Wartezeit eines Kunden in der Queue W_Q

- Formel von Little:

  Ankunftsrate der Kunden sei λ. Dann ist

  $$L = \lambda W$$

  Voraussetzung: Mittelwerte existieren

  anschauliche Begründung (Beweis: [2])
  - Annahme: Jeder Kunde zahlt 1 € pro Zeiteinheit im System
  - Nach großer Zeit T gilt

    Gesamtverdienst des Systems V = L T

Gesamtzahl der Kunden $N = \lambda T$

mittlerer Betrag eines Kunden $B = W$

- also $V = N B \Rightarrow L = \lambda W$

analog auch

$L_Q = \lambda W_Q$

- Stetige oder diskrete Zeit ?

  Warteschlangentheorie meistens mit stetiger Zeit
  - Formeln einfach, Beweise schwer

  hier mit diskreter Zeit
  - Formeln komplizierter, Beweise einfacher

  Trick zur Vereinfachung
  - sehr kleine Zeitschritte h
  - Formeln nur in erster Ordnung h
  - Ergebnisse dann wie bei stetiger Zeit

- Ankunftsprozess:

  feste Rate λ = mittlere Zahl der Kunden pro Zeiteinheit

  Annahme: pro Takt h nur max. eine Ankunft
  - h klein
  - keine Gruppenankünfte
  - → pro Takt ein Bernoulliprozess mit $p_a = \lambda h$

  Eigenschaften des Ankunftsprozesses
  - Zahl der Ankünfte in n Takten ~ $B(n,p_a)$
  - Zahl der Ankünfte in disjunkten Zeitintervallen unabhängig
  - Zahl der Takte zwischen zwei Ankünften ~ $G(p_a)$

- Bedienprozess:

  analog zum Ankunftsprozess

  feste Rate μ = mittlere Zahl der Bedienungen pro Zeiteinheit (falls Kunden da)
  - $1/\mu$ = mittlere Bedienzeit
  - pro Takt ein Bernoulliprozess mit $p_b = \mu h$

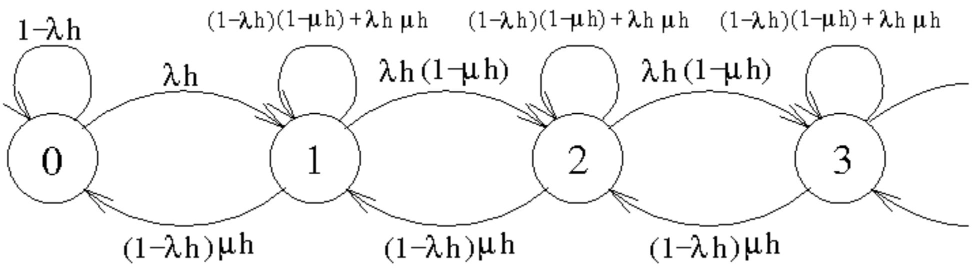
  Eigenschaften des Bedienprozesses
  - Restbedienzeit ab beliebigem Zeitpunkt ebenfalls $G(p_b)$
  - Modell für Bedienprozess nicht besonders realistisch, aber Markov

- Gesamtmodell des M/M/1-Prozesses:

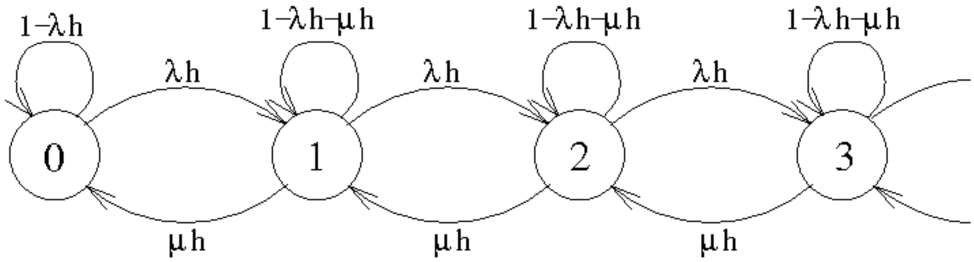
  homogene Markov-Kette

  Zustand = Zahl der Kunden (wartend + bedient)

  beschrieben durch Übergangs-Graph



  Vereinfachung für sehr kleine h



anschaulich: Wahrscheinlichkeit für gleichzeitige Ankunft und Bedienung sehr klein

- wird vernachlässigt
- → Ankunfts- und Bedienprozess nicht mehr unabhängig
- aber kein Problem bei sehr kleinem h

Basis der folgenden Untersuchung

- Bestimmung der Gleichgewichtsverteilung π_i:

Schnittprinzip →

$$\begin{aligned} \pi_i \, P_{i,i+1} &= \pi_{i+1} \, P_{i+1,i} \qquad i = 0, 1, 2, \ldots \\ \pi_i \, \lambda &= \pi_{i+1} \, \mu \\ \pi_{i+1} &= \left(\frac{\lambda}{\mu}\right) \pi_i \end{aligned}$$

Einführung der Auslastung ρ

$$\rho := \frac{\lambda}{\mu}$$

- Name wird bald klar (s.u.)

durch Rekursion erhält man sofort

$$\pi_i = \rho^i \, \pi_0 \qquad i = 0, 1, 2, \ldots$$

Normierungsbedingung

$$1 = \sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} \rho^i$$

Fall $\rho \geq 1$

- Bedingung unerfüllbar
- es existiert keine GGV
- zu viele Kunden, Queue wächst immer weiter

Fall $\rho < 1$ über geometrische Reihe

$$\begin{aligned} 1 &= \pi_0 \cdot \frac{1}{1-\rho} \\ \Rightarrow \qquad \pi_0 &= 1 - \rho \\ \Rightarrow \qquad \pi_i &= \rho^i (1-\rho) \qquad i = 0, 1, 2, \ldots \end{aligned}$$

Zustand im GGV also ~ $G^0(1-\rho)$

- Auslastung A des Servers:

Anteil der Zeit, in der der Server beschäftigt ist

bei GGV also

$$A = 1 - P(\text{System ist leer}) = 1 - \pi_0 = \rho$$

bei $\rho \geq 1$ geht $A \to 1$

- Kennzahlen im Gleichgewicht:

  mittlere Kundenzahl L
  - L = Erwartungswert von X im Gleichgewicht
  - $\pi_i \sim G^0(1-\rho)$, also

    $$L = \frac{\rho}{1-\rho}$$

  - Streuung von L

    $$\sigma_L = \frac{\sqrt{\rho}}{1-\rho}$$

  mittlere Verweilzeit W eines Kunden
  - direkt mit Little

    $$W = \frac{L}{\lambda} = \frac{1}{\lambda}\frac{\rho}{1-\rho} = \frac{1}{\mu-\lambda}$$

  mittlere Wartezeit W_Q eines Kunden

  $$W_Q = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda} = \rho \cdot W$$

  mittlere Queuelänge L_Q über Little

  $$L_Q = \lambda W_Q = \frac{\rho^2}{1-\rho}$$

- Bediensysteme bei stetiger Zeit:

  Ankunftsprozess (bzw. Bedienprozess) sei wie immer Bernoulliprozess mit

  $p_a = \lambda h$, h klein

  Zufallsvariable DT sei Zeit zwischen zwei Ankünften, dann ist

  DT/h ~ G(λh) (Zahl der Takte zwischen zwei Ankünften)

  - für h → 0 gilt dann

    DT ~ Ex(λ)

  - Beweis im Anhang

  Zufallsvariable N sei Zahl der Ankünfte im Zeitintervall, dann

  N ~ B(1/h, λh)

  - für h → 0 gilt dann (s. Poisson-Grenzwertsatz)

    N ~ Po(λ)

  umgekehrt kann man zeigen:
  - Für Ankunftsprozesse mit Markov-Eigenschaft sind die Zeitintervalle zwischen zwei Ankünften exponentiell verteilt

  damit: Ergebnisse dieses Abschnitts gelten auch für stetiges M|M|1-System

- Aufgaben:

  Aufgabe 24

  Aufgabe 25

Grundlagen der Statistik

- Deskriptive Statistik
- Parameterschätzung
- Konfidenzintervalle

Deskriptive Statistik

- Grundbegriffe
- Eindimensionale Daten
- Zweidimensionale Daten

Grundbegriffe

- Statistik:

  "die Kunst, aus Daten zu lernen" [Ross, Kap. 1.1]

  untergliedert in deskriptive, explorative und induktive Statistik

  deskriptive Statistik
  - Beschreiben, Aufbereiten und Zusammenfassen von Daten
  - Verdichten zu Tabellen, graphischen Darstellungen und Kennzahlen

  explorative Statistik ("Data-mining")
  - Suche nach Strukturen und Auffälligkeiten in den Daten
  - formuliert Fragestellungen und Hypothesen
  - bei großen Datenmengen sehr rechner-intensiv

  induktive (auch: schließende, mathematische) Statistik
  - leitet aus Daten Eigenschaften einer umfassenderen Grundgesamtheit ab
  - basiert auf Wahrscheinlichkeitstheorie

- Beispiel A - Lebensdauern:

  Messwerte: Lebensdauern von je 100 internen Festplatten zweier gegebenen Typen

  Datei lebensdauer.dat

  Fragestellungen z.B.
  - mittlere Lebenszeit einer Festplatte
  - Nach welcher Zeit sind 90% noch funktionstüchtig?
  - Welcher Typ ist "besser"?

- Beispiel B - Klausur-Ergebnisse:

  Ergebnisse von Thermodynamik-Klausuren über sechs Jahre
  - jeweils drei Aufgaben, immer gleiche Zuordnung zu Teilthemen

  erfasste Daten
  - pro Klausur: Punktzahlen der Aufgaben 1, 2, 3, Gesamtprozentzahl
  - pro Jahr: Maximalpunktzahlen der Aufgaben

  Datei klausuren.xlsx

  Fragestellungen z.B.
  - Abschneiden der Jahrgänge im Vergleich
  - Vergleich der Aufgaben pro Jahr und insgesamt
  - Sind die einzelnen Klausuren gleich schwer?
  - Wie stark beeinflusst die Gewichtung der Aufgaben das Klausurergebnis?

- Beispiel C - Wahlergebnisse und Arbeitslosenzahlen:

  Wahlergebnisse aller Parteien für die Wahlen zum Deutschen Bundestag seit 1949
  - Ergebnisse nach Wahlkreisen vom Bundeswahlleiter

  Arbeitslosenquote in Deutschland seit 1950 (vor 1991: nur West)
  - saisonal bereinigte Daten vom Statistischen Bundesamt

  Dateien bundestagswahlen.xlsx, arbeitslose.xlsx

  wichtig: genaue Definition von "arbeitslos" und "erwerbstätig"

  Fragestellungen z.B.
  - Abschneiden der Parteien über die Jahre
  - Entwicklung der Arbeitslosenzahlen
  - Korrelationen der Arbeitslosenzahlen und der Wahlergebnisse

- Definitionen:

  Statistische Einheit: Objekt, an dem interessierende Daten erfaßt werden, z.B.
    - Personen (Kunden, Wähler, Studierende)
    - Dinge (Maschinen, Produkte)
    - Ereignisse (Maschinenausfälle, Geburten)

  Grundgesamtheit: Menge aller relevanten statistischen Einheiten, z.B.
    - alle Wahlberechtigten zur Bundestagswahl 2017
    - alle im Jahr 2016 verkauften Diesel-PKWs der Marke Mercedes-Benz
    - alle Geburten in den Landkreisen Vechta und Diepholz im Jahr 2016

  Stichprobe: tatsächlich untersuchte Teilmenge der Grundgesamtheit
    - zufällige Auswahl (alle statistischen Einheiten gleich wahrscheinlich)
    - repräsentative Auswahl (benutzt Quoten von Teilgruppen, basiert auf Vorwissen)

  Merkmal: interessierende Größe
    - qualitativ (nominal, ordinal)
    - quantitativ (diskret, stetig)

  Merkmalsausprägung: Wert des Merkmals bei einer statistischen Einheit
    - Gewinnen sinnvoller Daten nicht trivial, benötigt Versuchsplanung
    - Urliste = ursprüngliche Aufzeichnung der Beobachtungs- oder Messwerte

- Anwendung auf die Beispiele:

  Statistische Einheit
    - A: Festplatte eines Typs
    - B: einzelne Klausur
    - C: ursprünglich: Wahlberechtigte, Arbeitnehmer, hier schon pro Wahlkreis bzw. Jahrgang zusammengefasst

  Grundgesamtheit
    - A: alle verkauften Festplatten eines bestimmten Typs
    - B: alle Thermodynamik-Klausuren der betrachteten Jahre
    - C: alle im jeweiligen Jahr Wahlberechtigten/alle Erwerbspersonen

  Stichprobe
    - A: zufällig (wie bestimmt?)
    - B: komplette Grundgesamtheit
    - C: komplett, repräsentativ (Mikrozensus)

  Merkmal
    - A: Zeit von Inbetriebnahme bis zum Versagen (quantitativ, stetig)
    - B: erreichte Punktzahlen bei den drei Aufgaben (quantitativ, diskret)
    - C: gewählte Partei (qualitativ, nominal), Beschäftigungsverhältnis (qualitativ, nominal)

Eindimensionale Daten

- Häufigkeitstabelle:

  Bezeichnungen
  - n statistische Einheiten, durchnummeriert von 1 bis n
  - m verschiedene Merkmalsausprägungen a_j, j = 1 .. m (m ≤ n)
  - x_i = Ausprägung von X bei Einheit i
  - a_j = j-te Ausprägung des Merkmals

  absolute Häufigkeit von Merkmalsausprägung j
  - h_j = Anzahl der x_i mit $x_i = a_j$

  relative Häufigkeit von Merkmalsausprägung j
  - $f_j = h_j / n$

  graphische Darstellung z. B. als Stab-, Balken- oder Tortendiagramm

- Klassenbildung:

  bei großem m mehrere Ausprägungen zu einer Klasse zusammenfassen

  bei qualitativen Merkmalen häufig Klasse für die seltenen Ausprägungen

  bei stetigen Merkmalen üblicherweise durch Vorgabe von Intervallen
  - h_j = Anzahl der x_i mit Ausprägung in $[c_{j-1}, c_j)$, j = 1, .., m
  - auch bei diskreten Merkmalen mit großem m

  Darstellung als Balkendiagramm (Histogramm)
  - Balkenfläche ~ h_j (nicht Balkenhöhe)

  Tipps
  - möglichst keine unbeschränkten Klassen an den Enden
  - Zahl der Klassen ≤ n

- Beispiel C, Bundestagswahl 2013:

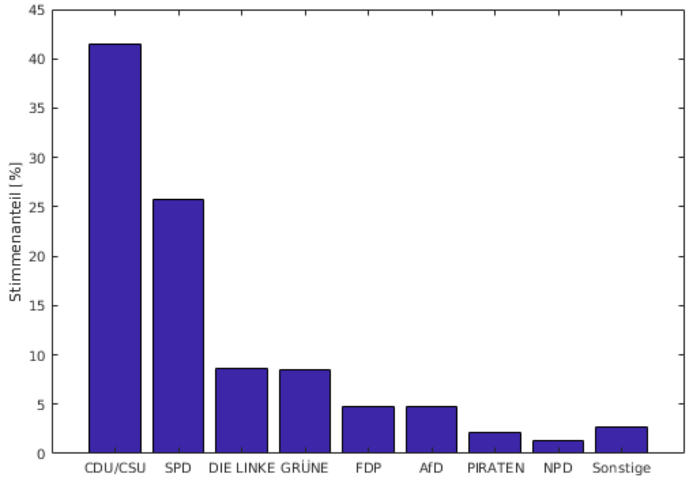
  Daten über Wahlbezirke zusammengefasst (wahl2013.xlsx)
  - Tabelle enthält direkt die h_j
  - gültige Stimmen für 30 Parteien → m = 30

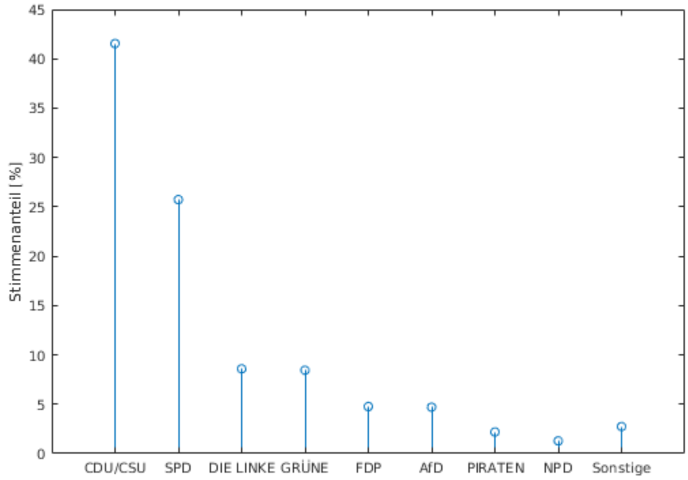
  Klassenbildung
  - Klasse "CDU/CSU"
  - Klasse "Sonstige" für f_i < 1%
  - alle übrigen eine Klasse

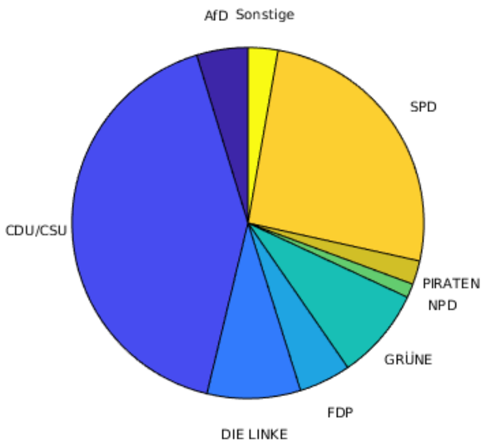
  relative Häufigkeiten der Parteien in %

CDU/CSU	SPD	DIE LINKE	GRÜNE	FDP	AfD	PIRATEN	NPD	Sonstige
41.54	25.73	8.59	8.45	4.76	4.70	2.19	1.28	2.74

  graphische Darstellungen







- Kennzahlen zur Beschreibung der Lage:

  bei quantitativen Merkmalen Kennzahlen analog zu denen von Wahrscheinlichkeitsverteilungen

  Mittelwert

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

  p-Quantil x_p ($p \in [0,1]$) teilt die nach Größe sortierten Daten so, dass Anteil p der Werte links liegt, 1-p rechts; genauer

  - Anzahl($x_i \leq x_p$) $\geq$ n p
  - Anzahl($x_i \geq x_p$) $\geq$ n (1-p)

  x_p ist eindeutig $\Leftrightarrow$ n p nicht ganzzahlig

  - anschaulich sofort klar für p = 0.5
  - bei mehrdeutigem x_p wird meist der Mittelwert der beiden umliegenden Werte benutzt

wichtige Spezialfälle

- **1. Quartil** $x_{1/4} = Q_1$
- **Median** $x_{1/2} = x_{med}$
- **3. Quartil** $x_{3/4} = Q_3$

Modus oder **Modalwert**

- Merkmalsausprägung mit der größten Häufigkeit
- auch für qualitative Merkmale definiert

• Kennzahlen zur Beschreibung der Streuung:

empirische Varianz

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

- ähnlich zur Varianz, aber mit Nenner (n-1) statt n
- gute mathematische Gründe für diese Definition (s.u.)
- anschaulich: ein Wert wird zur Berechnung des Mittelwerts "verbraucht"
- entsprechend **empirische Standardabweichung** s

Interquartilsabstand (IQR)

$d_Q := Q_3 - Q_1$

Spannbreite

$R = x_{max} - x_{min}$

- mit kleinster/größter Ausprägung x_{min}/x_{max}

Faustregel für Ausreißer-Kandidaten

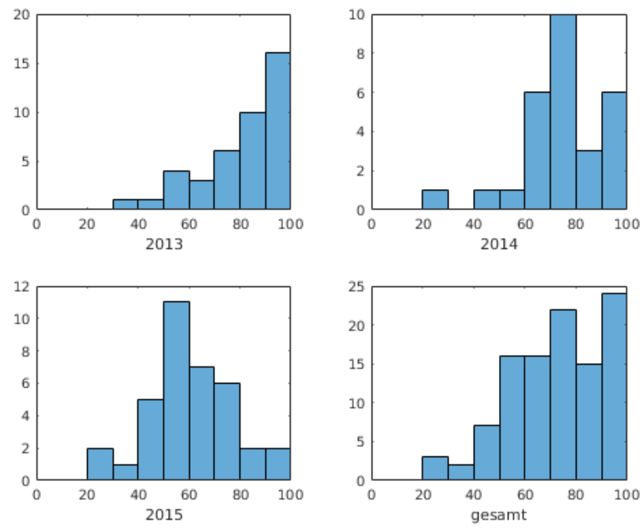
- unterer Zaun $z_u = Q_1 - 1.5\ d_Q$
- oberer Zaun $z_o = Q_3 + 1.5\ d_Q$
- Ausreißer: kleiner als z_u oder größer als z_o

• Beispiel B, Klausuren der Jahrgänge 2013-2015:

betrachtetes Merkmal: Gesamtprozent, Jahr

Auswertungen pro Jahr und gesamt

Histogramme mit 10er-Gruppen



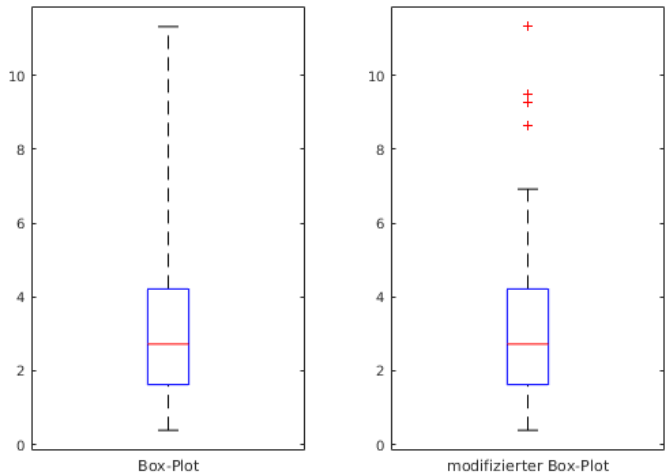
Ergebnisse

Kennzahl	2013	2014	2015	gesamt
mean	80.56	74.36	59.89	71.82
Q_1	74.00	65.00	50.50	57.00
median	83.00	72.00	57.00	72.00
Q_3	93.00	88.00	70.00	89.00
IQR	19.00	23.00	19.50	32.00
s	16.70	17.65	17.15	19.19

- Grafische Darstellungen:

  Box-Plot visualisiert x_{min}, Q_1, x_{med}, Q_3, x_{max}
  - Box von Q_1 bis Q_3
  - x_{med} als rote Querlinie oder Punkt in der Box
  - Linien (**Whisker**) bis x_{min}, x_{max}

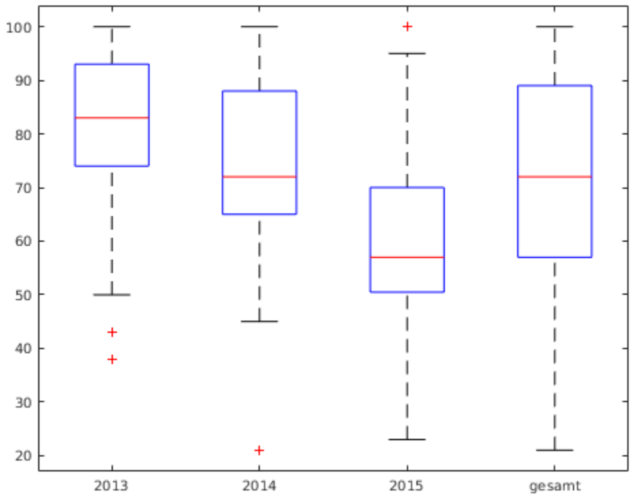
  modifizierter Box-Plot
  - Whisker nur bis zum kleinsten/größten Wert innerhalb $[z_u, z_o]$
  - Ausreißer als Sterne o.ä. markiert



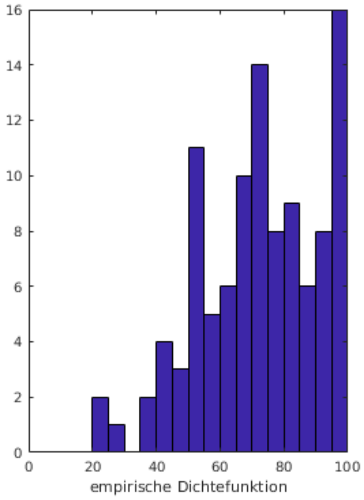
  empirische Verteilungsfunktion
  - $F(x)$ = Anteil der Beobachtungen $\leq x$
  - ganz analog zur kumulativen Verteilungsfunktion
  - daraus Quantile leicht ablesbar wegen $F(x_p) = p$

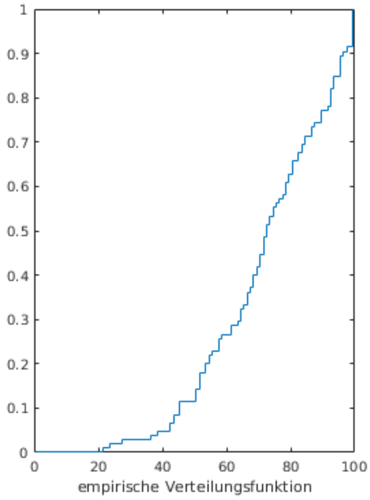
  empirische Dichtefunktion
  - Histogramm mit gleich breiten (aber kleinen) Intervallen
  - sinnvoll bei stetigem Merkmal und vielen Daten

  Beispiel Klausuren
  - Box-Plot der Gesamtprozent-Ergebnisse



- empirische Dichte- und Verteilungsfunktion





- Aufgaben:

  Aufgabe 26

  Aufgabe 27

Zweidimensionale Daten

- Untersuchung zweidimensionaler Daten:

  an statistischen Einheiten werden jeweils zwei Merkmale X, Y gemessen
  - → Daten sind Zahlenpaare (x_i, y_i)

  alle 1d-Methoden anwendbar auf Daten x_i bzw. y_i

  suchen Zusammenhänge zwischen x_i, y_i
  - meistens Tendenzen (großer Wert für X kommt häufig mit großem Wert für Y), nicht funktionale Zusammenhänge
  - falls doch funktionaler Zusammenhang besteht → Regressionsrechnung

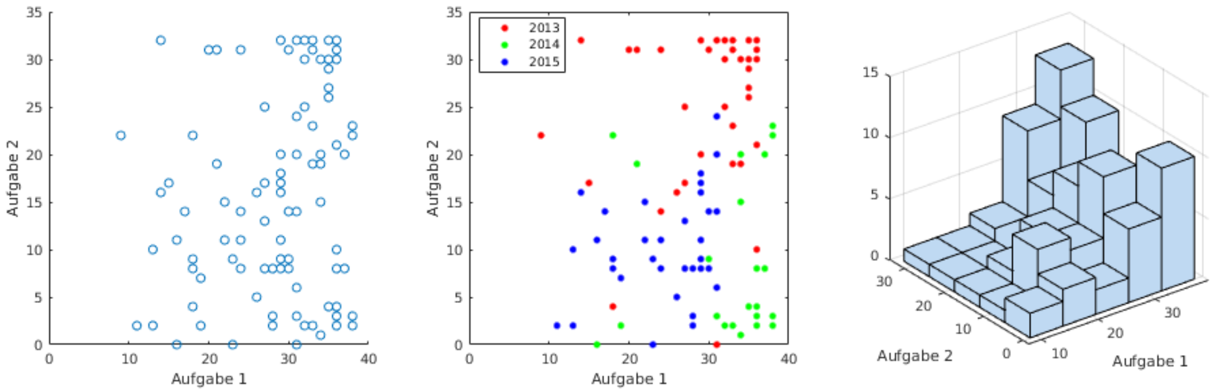
- Darstellung zweidimensionaler Daten:

  Streudiagramm (Punktwolke, scatter plot)
  - jeder Datensatz wird als Punkt oder Symbol in x-y-Diagramm aufgetragen
  - gibt Überblick über Grundverhalten
  - zeigt spezielle Verhältnisse an, z. B. Bildung mehrerer Cluster

  Streudiagramm mit Gruppenzuordnung
  - interessant, wenn Daten in verschiedene Gruppen zerfallen, z.B. Messreihen, Jahrgänge
  - Gruppenzugehörigkeit eines Punkts durch Farbe oder Symboltyp markiert

  2d-Histogramm
  - Klassenbildung in X und Y, insbesondere durch Intervalleinteilung
  - Anzahl der Datensätze in Klasse $KX_i \cap KY_j$ als Balken an Position (i,j)

  Beispiel: Klausurergebnisse in Aufgabe 1 bzw. Aufgabe 2



- Empirische Kovarianz:

  Definition der empirischen Kovarianz s_{xy}

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

  analog zur Kovarianz bei Wahrscheinlichkeitsverteilungen

  s_{xy} skalenabhängig, daher Korrelationskoeffizient r_{xy}

$$r_{xy} := \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

  Beispiel: Klausurergebnisse in Aufgabe 1 bzw. Aufgabe 2

	2013	**2014**	**2015**	**gesamt**
r_{xy}	0.2256	0.1246	0.2600	0.1952

  auffällig: niedriges r_{xy} im Jahr 2014
  - Ursache: viele mit gutem Ergebnis in Aufgabe 1 und schlechtem in Aufgabe 2

- Ursache: viele mit gutem Ergebnis in Aufgabe 1 und schlechtem in Aufgabe 2

- Korrelation ≠ Kausalität:

  Korrelation kann vieles bedeuten
  - manchmal wirklich einen kausalen Zusammenhang
  - häufig einen kausalen Zusammenhang beider Größen zu einer dritten, nicht beobachteten Größe
  - eine voreingenommene Auswahl von Daten
  - oder einfach Zufall

  Beispiel Störche
  - statistische Einheit: deutsche Landkreise und Städte
  - Merkmale: Geburtenrate, Zahl der Störche pro Fläche
  - Korrelation ist hoch: Wo es viele Störche gibt, werden viele Kinder geboren
  - Kausalität ist wohl eher unwahrscheinlich
  - Hintergrundgröße: durchschnittliche Gemeindegröße
  - klar: im ländlichen Raum gibt es mehr Geburten und mehr Störche

  Beispiel Fussball [Maas, Kap. 17]
  - Merkmale: Siegtage von Bayern München bzw. FC Augsburg in der Fußball-Bundeliga-Saison 2012/13 (ausgenommen Spiele gegeneinander)
  - Beobachtung: an jedem Tag, an dem Augsburg gewonnen hat, hat auch München gewonnen
  - Kann das Zufall sein? Ja!
  - Autor hat unter den Daten nach passendem Verein gesucht

  Kausalität lässt sich allein aus statistischer Analyse nicht feststellen!

- Q-Q-Plot:

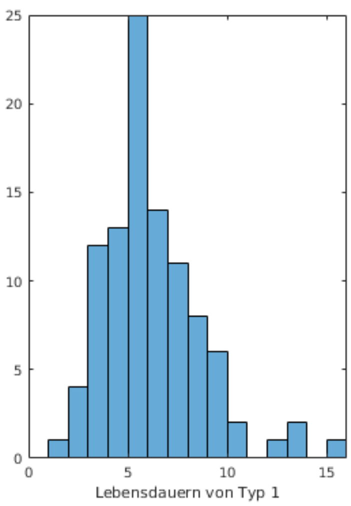
  graphische Methode zum Vergleich zweier Verteilungen
  - haben zwei Datensätze gleiche Verteilung?
  - entsprechen Daten eines Datensatzes einer theoretischen Verteilung?

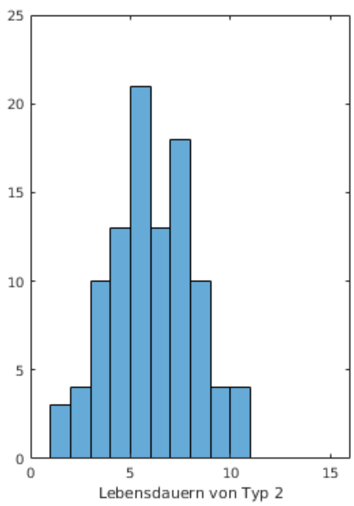
  genauere (quantitative) Aussagen in der Testtheorie

  Vorgehen
  - berechne Quantile x_q, y_q für viele Werte von q
  - plotte Punkte (x_q, y_q)

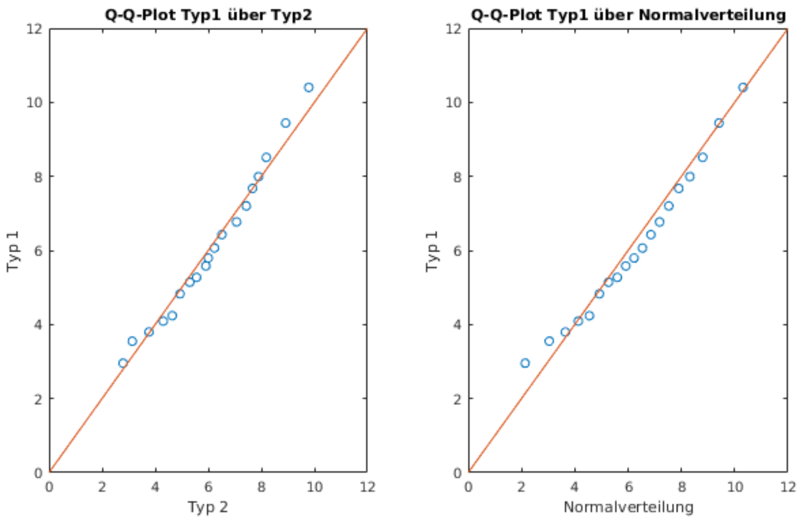
  Auswertung
  - Punkte liegen (ungefähr) auf der Winkelhalbierenden → Verteilungen stimmen (vermutlich) überein
  - Punkte liegen (ungefähr) auf einer Geraden → Verteilungen unterscheiden sich durch Verschiebung und Skalierung

  Beispiel A, Lebensdauern
  - Histogramme der Daten





- Vergleich der beiden Typen, jeweils über i/20-Quantile
- Vergleich des ersten Typs mit Normalverteilung (dazu Mittelwert und Varianz aus Daten berechnen)



Interpretation

- Abweichungen von der Winkelhalbierenden klein → Verteilungen könnten gleich sein
- typische Links-Krümmung, vor allem im rechten Plot
- → Typ 1 ist etwas unsymmetrischer, hat größere Anteile bei hohen Lebensdauern
- → Typ 1 vermutlich nicht normalverteilt

- Aufgaben:

  Aufgabe 28

  Aufgabe 29

Parameterschätzung

- Induktive Statistik:

  nutzt Daten, um auf eine größere Grundgesamtheit zu schließen

  fasst Daten als Realisierungen von Zufallsvariablen auf

  verwendet Wahrscheinlichkeitsrechnung, um auf deren Verteilungen zu schließen

  wichtige Fragestellungen
  - Parameterwerte erschließen ("schätzen")
  - Genauigkeit einer Schätzung quantifizieren (→ Konfidenz-Intervalle)
  - Ja-/Nein-Fragen über Daten beantworten (→ Testtheorie)

- Statistisches Modell:

  Ausgangslage
  - Grundgesamtheit mit N Elementen (N sehr groß oder unendlich)
  - Stichprobe mit n Elementen, zufällig und unabhängig voneinander gewählt
  - beobachtetes Merkmal X

  Annahmen
  - Ergebnisse X_i (i=1,..,n) der i-ten Stichprobe sind Zufallsvariablen
  - X_i sind unabhängig und gleichverteilt (i.i.d.)

  Verteilung der X_i
  - unbekannt oder aufgrund von Modellannahmen gegeben mit unbekannten Parametern
  - Messwerte mit statistischen Messfehlern häufig ~ $N(\mu, \sigma^2)$
  - Zahl (seltener) Ereignisse pro Zeitintervall oft ~ $Po(\lambda)$

  gesuchte Information (Parameter) θ
  - hat festen Wert, der aber unbekannt ist und gesucht wird
  - ist ein Vektor, falls mehrere Größen interessieren

- Beispiele:
  1. Grundgesamtheit: alle von einer Maschine bearbeiteten Werkstücke
     - Merkmal: 1, falls Werkstück defekt, sonst 0
     - X_i Bernoulli-Experiment mit unbekannter Wahrscheinlichkeit p
     - gesuchter Parameter $\theta = p$
  2. Grundgesamtheit: alle an einem Objekt durchführbaren Messungen einer Messgröße
     - Merkmal: Ergebnis der Messung
     - $X_i \sim N(\mu, \sigma^2)$
     - gesuchte Parameter $\theta = (\mu, \sigma^2)$
  3. Grundgesamtheit: alle produzierten Festplatten eines Typs
     - Merkmal: Lebensdauer
     - Verteilung X der X_i unbekannt
     - gesuchte Parameter $\theta = (E(X), Var(X))$

- Mathematische Version: **statistisches Modell** $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$:

  Stichprobenraum $\mathcal{X}$: Menge aller möglichen Stichproben (meistens $\mathbb{R}^n$)

  Ereignismenge $\mathcal{A}$ von Teilmengen von $\mathcal{X}$: σ-Algebra der messbare Teilmengen

  Parameterraum Θ: Menge aller möglichen Werte des Parameters θ

  P_θ (für jedes $\theta \in \Theta$): Wahrscheinlichkeitsmaß auf $(\mathcal{X}, \mathcal{A})$

- **Schätzer** für θ (auch **Schätzfunktion**, **Schätzstatistik**):

Idee

- berechne Größe aus den Ausprägungen x_i der Stichprobe
- verwende dies als Schätzung für θ

Schätzer ist Abbildung $T = t(X_1, ..., X_n)$

- ist selbst Zufallsgröße
- Einsetzen von Ausprägungen x_i liefert **Schätzung**

in Beispiel 1 etwa

$$t(x_1, \ldots, x_n) = \frac{\#\{x_i | x_i = 1\}}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Eigenschaften von Schätzern:

  Schätzer T heißt **erwartungstreu** (unbiased) :⇔

  E(T) = θ

  - anschaulich: im Mittel liefert der Schätzer den gesuchten Parameter

  mittlerer quadratischer Fehler MSE von T

  $MSE(T) = E([T - \theta]^2)$

  - für erwartungstreuen Schätzer folgt

  MSE(T) = Var(T)

  T heißt **konsistent** :⇔

$$\lim_{n \to \infty} \mathrm{Var}(T) = 0$$

  - anschaulich: mit wachsendem n wird die Schätzung immer genauer

- Mittelwert-Schätzer $\overline{X}$:

  gesucht ist Schätzfunktion für Erwartungswert μ der Grundgesamtheit

  naheliegender Schätzer

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

  Erwartungswert von $\overline{X}$

$$E(\overline{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

  - → $\overline{X}$ ist erwartungstreu

  Varianz von $\overline{X}$

  - da X_i i.i.d, erhält man

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

  - → $\overline{X}$ ist konsistent

- Maximum-Likelihood-Schätzer:

  Problem: Wie findet man guten Schätzer?

  Voraussetzung: X_i hat Wahrscheinlichkeits- oder Dichtefunktion $p_\theta(x)$

  Idee: wähle θ mit größter Wahrscheinlichkeit, die Stichprobe zu erhalten

  definiere Likelihood-Funktion L für feste Ausprägung $x = (x_1, ..., x_n)$
  - $L(\theta) := p_\theta(x)$
  - Schätzer ist das θ, das L maximiert

  Eigenschaften eines Maximum-Likelihood-Schätzers
  - annähernd normalverteilt für großes n
  - optimal für großes n (in gewissem Sinn)
  - nicht notwendig erwartungstreu, erst für große n

  praktische Berechnung
  - maximiere $l := \log(L)$ statt L durch Nullsetzen der Ableitung

- ML-Schätzer für θ = (μ, σ) bei Normalverteilung:

  bei gegebenen Messwerten x_i ist die Dichtefunktion

$$\begin{aligned} p_\theta(x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \end{aligned}$$

  - ihr Logarithmus also

$$l(\mu,\sigma) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

  - mit der Ableitung

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

  Nullsetzen und Auflösen liefert die ML-Schätzer

$$\begin{aligned} \mu_{ML} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \\ \sigma^2_{ML} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \end{aligned}$$

  - zur Kontrolle: 2. Ableitungen sind negativ → Maximum

  Schätzer für μ ist das bekannte $\overline{X}$

  Schätzer für σ heißt auch $\tilde{S^2}$

- Erwartungstreuer Schätzer für die Varianz:

  X_i sei beliebig verteilt mit Erwartungswert μ und Varianz σ^2

  Berechnung des Erwartungswerts von $\tilde{S^2}$ liefert

$$E(\tilde{S^2}) = \frac{n-1}{n}\sigma^2$$

  - also nicht erwartungstreu (erst für große n)

- Rechnung einfach, aber länglich, s. Anhang

erwartungstreu ist dagegen

$$S^2 = \frac{n}{n-1}\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- S^2 ist wichtigster Schätzer für die Varianz
- erklärt Definition der empirischen Varianz

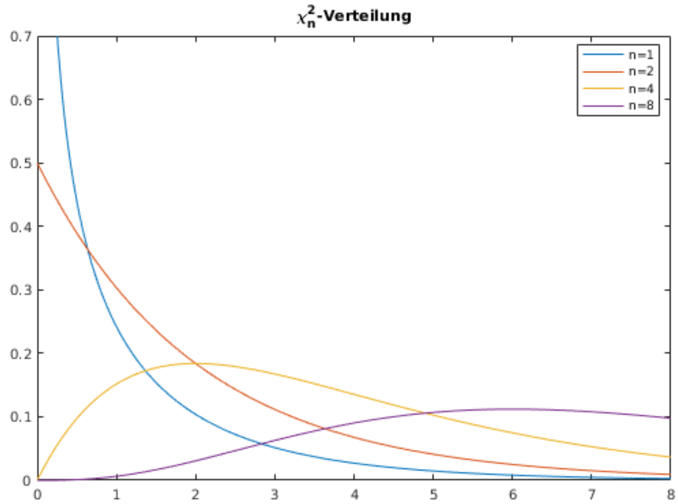
- Chi-Quadrat-Verteilung mit n Freiheitsgraden χ_n^2:

  zur Erinnerung (s.o.): Spezialfall der Gammaverteilung, $\chi_n^2 = \gamma(n/2, 1/2)$

  Dichtefunktion

$$f_n(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \mid \quad x > 0 \\ 0 & \mid \quad x \le 0 \end{cases}$$

  - graphisch



  Eigenschaften

  $E(\chi_n^2) = n$

  $Var(\chi_n^2) = 2\ n$

  χ_n^2 ist die Verteilung der Quadratsumme von n unabhängigen Standard-Normalverteilungen

$$\chi_n^2 \sim \sum_{i=1}^{n} Z_i^2$$

  - mit $Z_i \sim N(0,1)$ i.i.d

- $\overline{X}$ und S^2 für Normalverteilungen:

  sei $X_i \sim N(\mu, \sigma^2)$

  Verteilung von $\overline{X}$
  - Summe von normalverteilten unabhängigen Zustandsgrößen ist normalverteilt (s.o.)
  - X_i i.i.d, daher

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{1}{n}\sigma^2)$$

Verteilung von S^2

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

- insbesondere folgt

  $E(S^2) = \sigma^2$ (gilt schon für allgemeine Verteilungen, s.o.)

  $$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

- → S^2 ist erwartungstreu und konsistent

$\overline{X}$ und S^2 sind unabhängig

Beweise über Berechnungen mit Dichtefunktionen [Kabluchko]

- Aufgaben:

  Aufgabe 30

Konfidenzintervalle

- Wie gut ist ein Schätzwert?

  Anhaltspunkt: Varianz des Schätzers

  anderer Ansatz: Intervall, in dem der wahre Wert θ mit hoher Wahrscheinlichkeit liegt

  genauer
  - gib Irrtums-Wahrscheinlichkeit α vor (oft 5%)
  - suche Intervall $[g_u, g_o]$, in dem θ mit Wahrscheinlichkeit (1-α) liegt (**(1-α)-Konfidenzintervall**)

  gesucht sind Schätzer G_u, G_o für die Intervallgrenzen

  meistens wird Intervall symmetrisch um den Schätzwert von θ konstruiert
  - manchmal stattdessen einseitiges Intervall, etwa $(0, g_o)$ für Toleranz

  in der Regel zusätzliche Annahmen über Verteilung der X_i nötig

- Konfidenzintervall für μ einer Normalverteilung bei bekannter Varianz:

  Annahme: $X_i \sim N(\mu, \sigma^2)$, σ bekannt, μ gesucht

  verwende Schätzer $\overline{X}$ für μ, suche z mit

  $$P(\overline{X} - z \le \mu \le \overline{X} + z) \ge 1 - \alpha$$

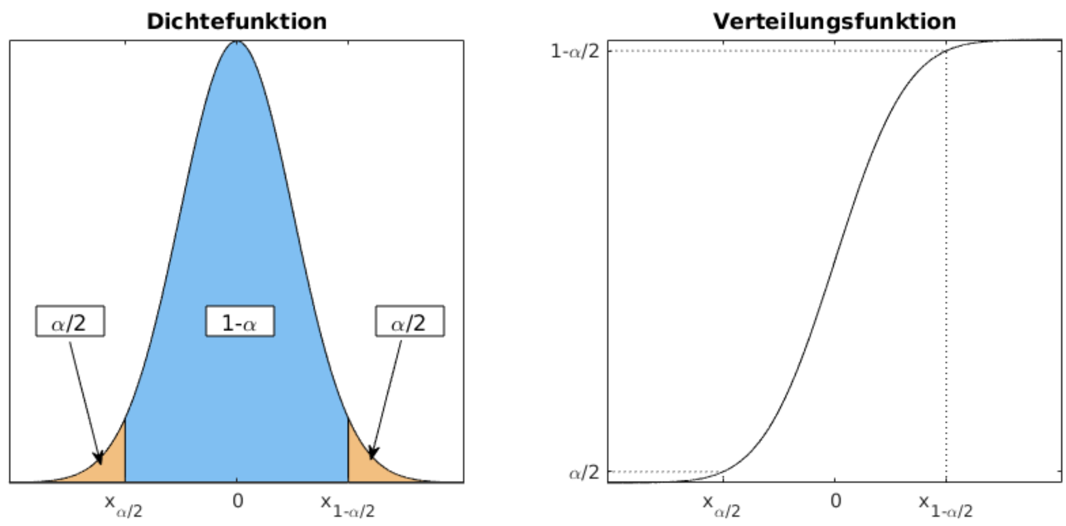
  Verteilung von $\overline{X}$ bekannt

  $$\begin{aligned} \overline{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ \Rightarrow \quad Z := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} &\sim N(0,1) \end{aligned}$$

  einfache Umformung liefert Bedingung für Standard-Normalverteilung

  $$\begin{aligned} P(\overline{X} - z \le \mu \le \overline{X} + z) &= P(\mu - z \le \overline{X} \le \mu + z) \\ &= P\left(-\frac{z}{\sigma/\sqrt{n}} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le \frac{z}{\sigma/\sqrt{n}}\right) \\ &= P\left(-\frac{\sqrt{n}}{\sigma} z \le Z \le \frac{\sqrt{n}}{\sigma} z\right) \ge 1 - \alpha \end{aligned}$$

  - im Bild



  gesucht sind also Werte $x_{\alpha/2}$, $x_{1-\alpha/2}$ mit

$$
\begin{aligned}
\Phi(x_{\alpha/2}) &= \alpha/2 \\
\Phi(x_{1-\alpha/2}) &= 1-\frac{\alpha}{2}
\end{aligned}
$$

- d.h., die α/2- und 1-α/2-Quantile der Standard-Normalverteilung

p-Quantil der Standard-Normalverteilung häufig mit z_p bezeichnet

- Antisymmetrie von Φ → $z_{\alpha/2} = -z_{1-\alpha/2}$
- Werte tabelliert bzw. mit `norminv(1-alpha/2)` in Matlab

α [%]	10	5	2	1	0.5	0.1
$z_{1-\alpha/2}$	1.6449	1.9600	2.3263	2.5758	2.8070	3.2905

damit Schätzer für die Grenzen des Konfidenzintervalls aus

$$
\begin{aligned}
\frac{\sqrt{n}}{\sigma} z &= z_{1-\frac{\alpha}{2}} \\
\Rightarrow \qquad z &= \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \\
\Rightarrow \qquad G_u &= \overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \quad G_o = \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}
\end{aligned}
$$

- Beispiel Füllmengen:

  in einer Abfüllanlage wurden folgende Füllmengen gemessen

  0.7361 0.7750 0.6522 0.7459 0.7296 0.6808 0.7070 0.7303 0.8274 0.8031
  0.6795 0.8110 0.7418 0.7181 0.7414 0.7139 0.7163 0.7647 0.7623 0.7625

  Standardabweichung sei aufgrund langjähriger Erfahrung bekannt: σ = 0.03

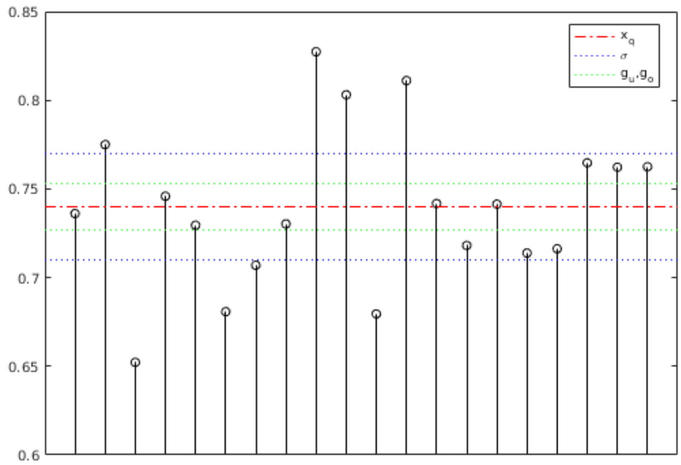
  als Mittelwert erhält man

  $\overline{X}$ = 0.7399

  für α = 5% ist $z_{0.975}$ = 1.9600, also erhält man als Konfidenzintervall

  [0.7268, 0.7531]

  graphisch



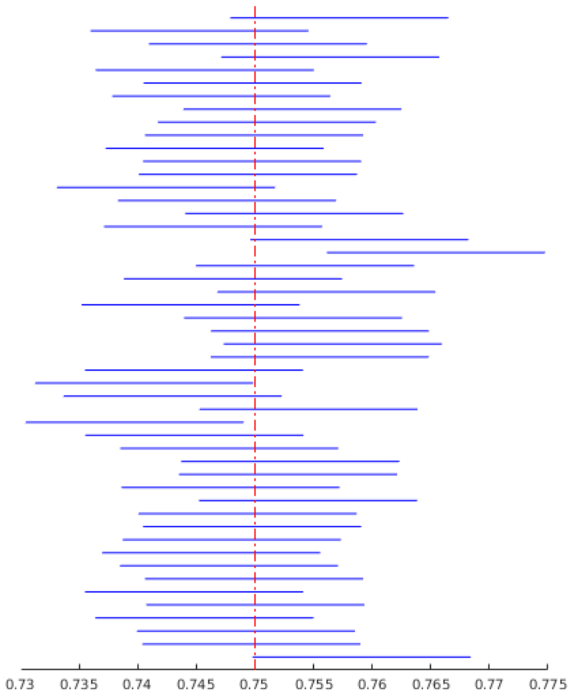
- Bedeutung des Konfidenzintervalls:

  gegeben sei X ~ N(0.75, 0.03^2)

  man nehme 50 Stichproben mit jeweils 40 Werten und berechne dazu die Konfidenzintervalle zu 1-α = 95%

  man erwartet durchschnittlich "2.5" Intervalle, die den wahren Mittelwert μ = 0.75 nicht enthalten

  mit Matlab simuliert



hier drei nicht passende Konfidenzintervalle

- Konfidenzintervall für μ einer Normalverteilung bei unbekannter Varianz:

  Idee: benutze S^2 als Schätzer für σ^2

  $$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \quad \to \quad \frac{\overline{X}-\mu}{\sqrt{S^2}/\sqrt{n}} =: T$$

  Vermutung: Intervall muss breiter werden, da weniger bekannt ist

  T ist nicht normalverteilt, sondern hat eine t_{n-1}-Verteilung

  Quantile $t_{n,p}$ tabelliert bzw. mit `tinv(p,n)`

  - sind antisymmetrisch $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$

  damit Schätzer für die Grenzen des Konfidenzintervalls

  $$G_u = \overline{X} - \sqrt{\frac{S^2}{n}}\, t_{n-1,1-\frac{\alpha}{2}}, \quad G_o = \overline{X} + \sqrt{\frac{S^2}{n}}\, t_{n-1,1-\frac{\alpha}{2}}$$

- Eigenschaften der **t-Verteilung** mit n Freiheitsgraden t_n:

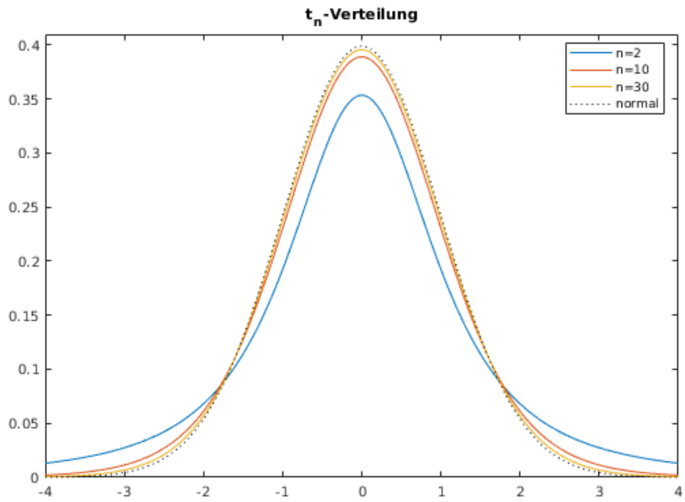
  von William Gosset eingeführt in Veröffentlichungen als "Student"

  - heißt daher auch Student- oder **Student-t-Verteilung**

  Dichtefunktion

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

  - graphisch



Eigenschaften

$$E(t_n) = 0 \qquad \text{für} \quad n > 1$$
$$\mathrm{Var}(t_n) = \frac{n}{n-2} \qquad \text{für} \quad n > 2$$

Satz

- Seien X, Y unabhängige Zufallsgrößen mit

  $X \sim N(0,1)$

  $Y \sim \chi_n^2$

- dann gilt

  $$\frac{X}{\sqrt{Y/n}} \sim t_n$$

erklärt Verteilung von T

t_n geht mit steigendem n gegen N(0,1)

- Näherung wird oft ab n = 30 verwendet

- Beispiel Füllmengen mit geschätzter Varianz:

  Als Schätzwert für die Standardabweichung erhält man s = 0.0444

  Für α = 5% und n = 20 ist $t_{19,0.975} = 2.0930$, also erhält man als Konfidenzintervall

  [0.7192, 0.7607]

  - etwas größer als bei bekanntem σ

- Konfidenzintervall für σ einer Normalverteilung:

  Schätzwert selbst natürlich mit S^2

  es war ja

  $$Z := \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

  - mit den Quantilen $\chi^2_{n,p}$ erhält man daher für den symmetrischen Fall

  $$P(Z \le \chi^2_{n-1,\frac{\alpha}{2}}) = \frac{\alpha}{2}$$
  $$P(Z \ge \chi^2_{n-1,1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

  - man braucht zwei Quantile, da χ^2-Verteilung unsymmetrisch

  daraus

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{n-1}{\sigma^2} S^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow \quad P\left(\frac{n-1}{\chi^2_{n-1,1-\frac{\alpha}{2}}} S^2 \leq \sigma^2 \leq \frac{n-1}{\chi^2_{n-1,\frac{\alpha}{2}}} S^2\right) = 1-\alpha$$

damit Schätzer für die Grenzen des Konfidenzintervalls von σ^2

$$G_u = \frac{n-1}{\chi^2_{n-1,1-\frac{\alpha}{2}}} S^2, \quad G_o = \frac{n-1}{\chi^2_{n-1,\frac{\alpha}{2}}} S^2$$

Schätzer für die Grenzen des Konfidenzintervalls von σ natürlich die Quadratwurzeln

- Beispiel Füllmengen, Konfidenzintervall für Standardabweichung:

  Schätzwert für die Standardabweichung war s = 0.0444

  Quantile aus Tabellen oder mit `chi2inv(p, n)`

  $$\begin{aligned} \chi^2_{19,0.975} &= 32.8523 \\ \chi^2_{19,0.025} &= 8.9065 \end{aligned}$$

  damit Konfidenzintervall für σ^2

  [0.001140, 0.004204]

  - und für σ

    [0.0338, 0.0648]

  nur nebenbei
  - Werte wurden erzeugt mit `normrnd(0.72, 0.03)` und `rng('default')`
  - Intervalle passen nicht - Pech gehabt!
  - mehrere Wiederholungen mit neuen Zufallszahlen klappen (5% Fehlerquote!)

- Konfidenzintervall für den Mittelwert bei beliebiger Verteilung:

  zentraler Grenzwertsatz → $\overline{X}$ nähert sich einer Normalverteilung an

  für n ≥ 30 wählt man daher das Konfidenzintervall wie oben beschrieben

  Varianz unbekannt → t-Verteilung nehmen
  - aber bei n ≥ 30: $t_n \approx N(0,1)$

- Konfidenzintervall für eine unbekannte Wahrscheinlichkeit p:

  X_i Bernoulli-Experiment mit gesuchter Wahrscheinlichkeit p

  Y sei Zahl der Treffer, dann Y ~ B(n,p)

  Schätzer für p natürlich $\overline{X}$ = Y/n

  Konfidenzintervall durch Standardisierung und Grenzwertsatz
  - berechne standardisierte Verteilung

    $$\tilde{Y} = \frac{Y - E(Y)}{\sigma(Y)} = \frac{Y - np}{\sqrt{np(1-p)}}$$

  - und gehe von Normalverteilung aus

    $$P(-z \leq \tilde{Y} \leq z) \approx 1 - \alpha$$

  - mit

    $$z := z_{1-\frac{\alpha}{2}}$$

  löse wie üblich nach dem gesuchten Parameter (hier p) auf
  - dazu zunächst Varianz-Term als bekannt annehmen →

$$P\left(\frac{Y}{n} - \frac{z}{\sqrt{n}}\sqrt{p\,(1-p)} \leq\ p\ \leq \frac{Y}{n} + \frac{z}{\sqrt{n}}\sqrt{p\,(1-p)}\right) \approx 1-\alpha$$

- Varianz-Term los werden:

  brutal: p durch Y/n abschätzen (ok im Grenzwert n → ∞)
  - wird häufig als Standardmethode vorgeschlagen
  - liefert oft *viel* zu kleine Intervalle, auch noch für n > 1000
  - genaue Analyse in [3] → nicht verwenden!

  besser: in quadratische Gleichung umformen →

$$\begin{aligned} &P\left(\overline{X} - \frac{z}{\sqrt{n}}\sqrt{p\,(1-p)} \leq\ p\ \leq \overline{X} + \frac{z}{\sqrt{n}}\sqrt{p\,(1-p)}\right) \\ &\quad = P\left(|\overline{X} - p| \leq \frac{z}{\sqrt{n}}\sqrt{p\,(1-p)}\right) \\ &\quad = P\left((\overline{X} - p)^2 \leq \frac{z^2}{n} p\,(1-p)\right) \\ &\quad = P\left((1 + \frac{z^2}{n})p^2 - (2\overline{X} + \frac{z^2}{n})p + \overline{X}^2 \leq 0\right) \end{aligned}$$

  Auflösen der quadratischen Gleichung liefert Schätzer für die Grenzen des Konfidenzintervalls (**Wilson-Intervall**)

$$\begin{aligned} G_u &= \frac{1}{1+\frac{z^2}{n}}\left(\overline{X} + \frac{z^2}{2n} - \frac{z}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X}) + \frac{z^2}{4n}}\right) \\ G_o &= \frac{1}{1+\frac{z^2}{n}}\left(\overline{X} + \frac{z^2}{2n} + \frac{z}{\sqrt{n}}\sqrt{\overline{X}(1-\overline{X}) + \frac{z^2}{4n}}\right) \end{aligned}$$

  - liegt nicht symmetrisch um den Schätzwert!

- Beispiel Würfel:

  30-maliges Würfeln ergab im Experiment

  6 4 6 5 5 2 6 1 2 3 5 2 6 6 1 4 1 3 1 6 3 6 5 6 2 1 4 2 5 3

  das sind 8 Sechsen, also Schätzwert für die Wahrscheinlichkeit einer Sechs

  $\overline{x}$ = y/n = 0.2667

  Quantil bei α = 5 %

  z = 1.9600

  mit der Wilson-Methode lautet das Konfidenzintervall

  [0.1418, 0.4445]

  - der tatsächliche Wert p = 0.1667 liegt also im Intervall

- Aufgaben:

  Aufgabe 31

  Aufgabe 32

  Aufgabe 33

Testtheorie

- Statistische Tests von Hypothesen
- Parametrische Tests
- Nicht-parametrische Tests
- Tests mit mehr als zwei Stichproben

Statistische Tests von Hypothesen

- Statistischer Test:

  beginnt mit einer Hypothese (Nullhypothese H_0) = Aussage, die anhand von Daten geprüft werden soll

  gegenteilige Aussage heißt Alternative H_1

  Test liefert Kriterium anhand der Stichprobenwerte, ob Hypothese verworfen wird oder nicht

  Testergebnis
  - "Hypothese ist mit Daten verträglich (oder nicht)"
  - nicht "Hypothese ist wahr (oder falsch)"!

  auch äußerst unwahrscheinliches Ergebnis ist möglich → Testergebnis kann falsch sein

- Typische Hypothesen:

  Der Anteil der Ausschussteile einer Produktion liegt unterhalb von 1 %.

  Die produzierten Schrauben sind im Schnitt 5 mm dick.

  Die Streuung der Widerstandswerte beträgt 10 %.

  Der Würfel ist echt, d.h. alle Augenzahlen haben gleiche Wahrscheinlichkeit.

  Zwei mit verschiedenen Methoden gewonnene Messreihen haben die gleiche Verteilung.

  Das Medikament Lernofil ist wirksam gegen die Krankheit Nixocapitose.

- Testergebnisse:

  H_0 ist wahr und Test akzeptiert H_0 (ok)

  H_0 ist falsch und Test verwirft H_0 (ok)

  H_0 ist wahr und Test verwirft H_0
  - **Fehler 1. Art** ("paranoid")

  H_0 ist falsch und Test akzeptiert H_0
  - **Fehler 2. Art** ("leichtgläubig")

  Signifikanzniveau α kontrolliert Fehler 1. Art
  - in medizinischen, sozial- oder geisteswissenschaftlichen Untersuchungen heißt die Ablehnung bei $\alpha = 0.05$ "signifikant"
  - umgekehrt heißt das: Statistische "Ergebnisse" etwa jeder 20. Veröffentlichung sind falsch!
  - zum Vergleich: Eine "Entdeckung" in der Elementarteilchenphysik verlangt 5σ, d.h. α = 5.7330e-07
  - kleines α erfordert natürlich sehr große Datenmengen

  Kontrolle des Fehlers 2. Art oft schwierig bis unmöglich
  - beim Münzwurf: Wahrscheinlichkeiten im Fall $p \neq 0.5$ nicht zu berechnen

  Erkenntnistheorie (Popper): allgemein gültige Hypothese kann durch empirisch gewonnenes Wissen nicht bewiesen ("verifiziert") werden, aber widerlegt ("falsifiziert")

- Beispiel Münzwurf:

  Hypothesen
  - H_0: Münze ist fair, d.h. Wahrscheinlichkeit für Kopf ist $p = 0.5$
  - H_1: Münze ist nicht fair, $p \neq 0.5$

  Münze werde $n = 200$ mal geworfen
  - interessierende Zufallsgröße S = "Anzahl Kopf"
  - Erwartungswert von S unter der Annahme H_0: $S = 100$

  Experiment ergibt: $s = 90$ mal wurde Kopf erzielt.

- Ist das unter der Annahme H_0 ok oder ist die Abweichung zu groß?

Teststatistik T = Stichprobenfunktion, deren Wert über Ablehnung von H_0 entscheidet

- benutze hier: $T = |S - 100|$
- T "zu groß" → H_0 wird abgelehnt
- Wert hier: T = 10

Konkretisierung von "zu groß"

- wähle **Signifikanzniveau** α = Wahrscheinlichkeit, dass H_0 abgelehnt wird, obwohl es wahr ist
- typische Werte: α = 0.05, α = 0.01, wählen hier: 0.05

gesucht ist kritischer Wert c mit

$$P(T > c \mid H_0) \le \alpha$$

- c sehr groß → Fehler 1. Art sehr klein, aber Fehler 2. Art groß
- wähle kleinst mögliches c, also

$$P(T > c \mid H_0) = \alpha$$

falls H_0 gilt, ist S ~ B(n,p) mit bekannter Verteilungsfunktion $FB_{n,p}(k)$ und p = 1/2

$$\mathrm{FB}_{n,p}(k) = \sum_{l=0}^{k} \binom{n}{l} p^l (1-p)^{n-l}$$

- also

$$\begin{aligned} P(T > c \mid H_0) &= P(|S - 100| > c \mid H_0) \\ &= 2\,P(S < 100 - c \mid H_0) \\ &= 2\,P(S \le 99 - c \mid H_0) \\ &= 2\,\mathrm{FB}_{n,p}(99 - c) \end{aligned}$$

mit Matlab rechnet man leicht aus

- $P(T > 13 \mid H_0) = 0.0560$
- $P(T > 14 \mid H_0) = 0.0400$
- optimal also c = 14

bei T = 10 wird H_0 also akzeptiert

Achtung: H_0 ist nicht "bewiesen", sondern es gibt keinen Grund, H_0 abzulehnen.

- Mathematische Präzisierung:

  gegeben sei ein statistisches Modell $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$

  Parameterraum Θ sei in zwei disjunkte Teilmengen Θ_0 und Θ_1 zerlegt

  - Nullhypothese H_0: $\theta \in \Theta_0$
  - Alternativhypothese H_1: $\theta \in \Theta_1$

  ein Test ist eine messbare Funktion $\varphi : \mathcal{X} \to \{0, 1\}$

  - $\varphi(X) = 0$ → H_0 wird akzeptiert
  - $\varphi(X) = 1$ → H_0 wird verworfen
  - **Ablehnungsbereich** $K = \{x \in \mathcal{X} \mid \varphi(x) = 1\}$

  Gütefunktion eines Tests $G : \Theta \to [0, 1]$ mit $G(\theta) = P_\theta(\varphi(X) = 1)$

  - $\theta \in \Theta_0 \to G(\theta)$ = Wahrscheinlichkeit für Fehler 1. Art
  - $\theta \in \Theta_1 \to 1 - G(\theta)$ = Wahrscheinlichkeit für Fehler 2. Art

  Test φ hat **Signifikanzniveau** $\alpha \in (0,1) :\Leftrightarrow G(\theta) \le \alpha$ für alle $\theta \in \Theta_0$

- **Teststatistik** T:

  Stichprobenfunktion $T: \mathcal{X} \to \mathbb{R}$ bzw. über P_θ entsprechende Zufallsvariable

Verteilungsfunktion von T unter der Annnahme H_0 muss berechenbar sein

- Problem für Mathematiker, in der Anwendung werden nur solche T verwendet

typischerweise so gewählt, dass große Werte von T gegen H_0 sprechen, d.h.

- $K = \{x \in \mathcal{X} \mid T(x) > c\}$
- mit einem kritischen Wert c, der von α abhängt

p-Wert einer Beobachtung $x \in \mathcal{X}$

- Wahrscheinlichkeit, dass der Wert T(x) oder ein noch extremerer Wert auftritt, falls H_0 gilt:
- p-Wert(x) = $P(T(X) \geq T(x) \mid H_0)$

beim Münzwurf-Beispiel

- $P(T \geq 10 \mid H_0) = 0.1374$
- Bedeutung: Falls H_0 wahr ist, tritt ein T-Wert von 10 oder mehr in 13.74 % der Fälle ein

- Vorgehensweise bei Tests:
  1. Nullhypothese H_0 und Alternative H_1 formulieren
  2. geeignete Teststatistik T auswählen
  3. Signifikanzniveau α festlegen und kritischen Wert c bestimmen
  4. Stichprobenwerte x in die Statistik T einsetzen
  5. Entscheiden: $x \in K$ (meistens also $T(x) > c$) → H_0 verwerfen, sonst H_0 akzeptieren

- Bemerkungen:

  Soll Fehler 2. Art kontrolliert werden: H_0 und H_1 vertauschen

  sehr viele Testverfahren mit unterschiedlichen Statistiken vorhanden
  - einige wichtige i. F. vorgestellt
  - immer Voraussetzungen prüfen!

  statt c alternativ p-Wert berechnen
  - p-Wert zu klein → verwerfen
  - schlechte Idee: im Nachhinein Signifikanzniveau auf > p-Wert heraufsetzen, um Alternative zu retten

- Probleme bei Tests:

  haben meistens Voraussetzungen, z.B.
  - Messgröße ist normalverteilt
  - σ der Verteilung ist bekannt
  - zwei Stichproben haben gleiche Verteilung
  - zwei Stichproben sind unabhängig

  nicht erfüllt → Test unzuverlässig

  im Zweifelsfall Voraussetzung vorher durch entsprechenden Test prüfen

  weitere typische Probleme
  - Ausreißer = Werte weit weg vom erwarteten Bereich
  - Multimodalität = Verteilung hat mehrere Maxima

- Parametrische Tests:

  Aussagen über Größe eines oder mehrerer Parameter einer (angenommenen) Verteilung, z. B.
  - Wahrscheinlichkeit eines Ereignisses
  - Erwartungswert einer Zufallsvariablen
  - Varianz einer Zufallsvariablen

  zweiseitig = konkreter Wert wird angegeben

  einseitig = entweder obere oder untere Grenze wird angegeben

Präzisierung: Parametrische Tests haben endlichdimensionale Parametermenge $\Theta \subset \mathbb{R}^n$

- Nicht-parametrische Tests:

  allgemeine Aussagen über die Verteilung einer oder mehrerer Zufallsgrößen, z. B.
  - Verteilung einer Zufallsgröße
  - gleiche Verteilung zweier Zufallsgrößen
  - Unabhängigkeit mehrerer Zufallsgrößen

  Präzisierung: Nicht-parametrische Tests haben unendlichdimensionale Parametermenge Θ

  Vorteil: anwendbar bei Daten ohne Kenntnis der Verteilung

  Nachteil: bei bekannter Verteilung ist ein parametrischer Test aussagekräftiger

  häufiges Vorgehen
  - zunächst mit nicht-parametrischem Test auf Verteilung prüfen
  - falls ok: mit parametrischem Test Werte der Parameter testen

Parametrische Tests

- Parametertests bei Normalverteilung mit einer Stichprobe
- Parametertests bei Normalverteilung mit zwei Stichproben
- Parametertests bei Bernoulli-Experimenten

Parametertests bei Normalverteilung mit einer Stichprobe

- Voraussetzung:

  in diesem Abschnitt wird angenommen, alle Stichprobenwerte X_i seien unabhängig normalverteilt mit gleichem μ und σ^2

- Test auf μ bei bekanntem σ^2 (Gauß-z-Test):

  H_0: $\mu \le \mu_0$, H_1: $\mu > \mu_0$

  wähle Teststatistik

  $$T = \frac{\sum_{i=1}^n (X_i - \mu_0)}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0)$$

  qualitativ
  - T groß → viele $X_i > \mu_0$, also bei großem Wert von T verwerfen
  - gesucht: kritischer Wert c für T bei gegebenem α

  H_0 gilt → $T \sim N(0,1)$, daher

  $$\begin{array}{lrcl} & P(T > c \mid H_0) & = & 1 - \Phi(c) \overset{!}{=} \alpha \\ \Leftrightarrow & \Phi(c) & = & 1 - \alpha \\ \Leftrightarrow & c & = & z_{1-\alpha} \end{array}$$

  H_0 wird also verworfen ↔ $T > z_{1-\alpha}$

  analog für andere Seite mit gleichem T
  - H_0: $\mu \ge \mu_0$, H_1: $\mu < \mu_0$
  - H_0 wird verworfen ↔ $T < z_\alpha$

- Zweiseitiger Test:

  H_0: $\mu = \mu_0$, H_1: $\mu \ne \mu_0$

  Abweichungen in beide Richtungen berücksichtigen, also verwende Teststatistik

  $$T = \frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0|$$

  Mit

  $$Z := \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) \sim N(0,1)$$

  - erhält man

  $$\begin{array}{lrcl} & P(T > c \mid H_0) & = & P(|Z| > c \mid H_0) \\ & & = & P(Z > c \mid H_0) + P(Z < -c \mid H_0) \\ & & = & 1 - \Phi(c) + \Phi(-c) \\ & & = & 2(1 - \Phi(c)) \overset{!}{=} \alpha \\ \Leftrightarrow & c & = & z_{1-\frac{\alpha}{2}} \end{array}$$

  H_0 wird verworfen ↔ $T > z_{1-\alpha/2}$

  Rechnungen und Ergebnisse wie beim Konfidenzintervall
  - das gilt oft bei parametrischen Tests

- Beispiel Schraubendurchmesser:

  Voraussetzungen
  - Durchmesser der produzierten Schrauben sind normalverteilt
  - Standardabweichung beträgt σ = 0.05 mm

  Nullhypothese: Die produzierten Schrauben sind im Schnitt 5 mm dick.

- Alternative: mittlerer Durchmesser der produzierten Schrauben ≠ 5 mm

Teststatistik T von oben, Signifikanzniveau $\alpha = 0.05$

kritischer Wert $c = z_{1-\alpha/2} = 1.9600$

Stichprobenwerte x_i (in mm)

- 4.9231 4.9934 4.9988 4.9669 4.9002 4.9238 4.8851 4.9959 4.9588 4.9878

Mittelwert 4.9534, Teststatistik $T(x) = 2.9485 > c$

- $\rightarrow H_0$ wird verworfen

T(x) ist deutlich zu groß, sein p-Wert ergibt sich zu

$$\mathrm{pWert} = P(T \geq T(x) \mid H_0) = 2\,(1 - \Phi(T(x))) = 0.003193$$

- man würde die Nullhypothese also auch bei einem Signifikanzniveau von 0.5 % noch verwerfen!

- Test auf μ bei unbekanntem σ (Student-t-Test):

ganz analog zum Gauß-Test

Nullhypothese H_{0a}: $\mu \leq \mu_0$

- bzw. H_{0b}: $\mu \geq \mu_0$ bzw. H_{0c}: $\mu = \mu_0$

schätze Standardabweichung wie üblich durch

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

im einseitigen Fall wähle Teststatistik

$$T = \frac{\sqrt{n}}{S}(\overline{X} - \mu_0)$$

- im zweiseitigen Fall wähle |T|

H_0 gilt $\rightarrow T \sim t_{n-1}$ (für a, b, bei c ohne Betrag)

kritische Werte für die drei Fälle

- $c_a = t_{n-1,1-\alpha}$
- $c_b = t_{n-1,\alpha}$
- $c_c = t_{n-1,1-\alpha/2}$

H_0 wird verworfen falls $T > c_a$ bzw. $T < c_b$ bzw. $T > c_c$

- Beispiel Schraubendurchmesser:

noch einmal untersuchen unter Annahme, dass σ unbekannt

kritischer Wert $c = t_{n-1,1-\alpha/2} = 2.2622$

- größer als vorher, da weniger Information

Mittelwert 4.9534, Standardabweichung $S = 0.0424$

Teststatistik $T(x) = 3.4801 > c \rightarrow H_0$ wird verworfen

T(x) ist größer als vorher, da $S < \sigma$

- Test auf σ bei unbekanntem μ (χ^2-Streuungstest):

Nullhypothese H_{0a}: $\sigma \leq \sigma_0$

- bzw. H_{0b}: $\sigma \geq \sigma_0$ bzw. H_{0c}: $\sigma = \sigma_0$

schätze σ durch S und wähle Teststatistik

$$T = \frac{n-1}{\sigma_0^2}S^2$$

H_0 gilt → $T \sim \chi^2_{n-1}$

kritische Werte für die drei Fälle

- $c_a = \chi^2_{n-1,1-\alpha}$
- $c_b = \chi^2_{n-1,\alpha}$
- $c_{c1} = \chi^2_{n-1,\alpha/2}$, $c_{c2} = \chi^2_{n-1,1-\alpha/2}$
- bei zweiseitigem Test zwei Werte, da χ^2-Verteilung unsymmetrisch

H_0 wird verworfen falls

- $T > c_a$
- $T < c_b$
- $T < c_{c1}$ oder $T > c_{c2}$

* Test auf σ bei bekanntem $\mu = \mu_0$ (χ^2-Streuungstest):

  wird selten gebraucht

  wie bei unbekanntem μ außer

  - ersetze in Formel für S berechneten Mittelwert $\overline{X}$ durch μ_0
  - H_0 gilt → $T \sim \chi^2_n$ (ein Freiheitsgrad mehr)

* Streuung beim Schraubenbeispiel:

  Streuung der Schraubendurchmesser

  - zunächst angenommen: $\sigma_0 = 0.05$ mm
  - aus Stichprobe bestimmter Schätzer S = 0.0423 mm

  Nullhypothese: Standardabweichung der Schraubendurchmesser beträgt σ = 0.05 mm

  Überprüfen mit χ^2-Streuungstest und obiger Stichprobe

  - kritische Werte: $c_1 = 2.7004$, $c_2 = 19.0228$
  - Teststatistik ergibt T(x) = 6.4606 → H_0 wird akzeptiert

* Aufgaben:

  Aufgabe 34

  Aufgabe 35

Parametertests bei Normalverteilung mit zwei Stichproben

- Zweistichprobentests für die Parameter der Normalverteilung:

  betrachten zwei Messreihen X_i (i = 1...n), Y_j (j = 1...m)

  Voraussetzungen
  - X_i, Y_j alle unabhängig
  - $X_i \sim N(\mu_x, \sigma^2_x)$
  - $Y_j \sim N(\mu_y, \sigma^2_y)$

  Anwendung
  - Merkmal unter zwei unterschiedlichen Bedingungen
  - Merkmal in zwei verschiedenen Teilgesamtheiten

  Beispiele
  - Wie wirkt sich eine Änderung in der Fertigung auf die Qualität des Produkts aus?
  - Wie unterscheiden sich Einkommen von männlichen und weiblichen Arbeitnehmern?
  - Wie wirkt sich ein Mathe-Vorkurs auf die Klausurergebnisse aus?

  Auswahl der Testpersonen in der Praxis kritisch
  - Haben die Versuchsgruppen (außer dem untersuchten Unterschied) noch weitere Unterschiede?
  - Haben beim Auswahlverfahren alle in der betrachteten Grundgesamtheit gleiche Chance?

- Test auf $\mu_x = \mu_y$ bei bekannten σ_x, σ_y (Zweistichproben-z-Test):

  wähle Teststatistik

  $$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

  H_0 gilt → $T \sim N(0,1)$

  kritischer Wert $c = z_{1-\alpha/2}$

  $|T| > c \leftrightarrow H_0$ wird verworfen

- Test auf $\mu_x = \mu_y$ bei unbekannten, aber gleichen σ_x, σ_y (Zweistichproben-t-Test):

  schätze S durch

  $$S^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2\right)$$

  - dies S ist bester Schätzer für $\sigma_x = \sigma_y$

  wähle Teststatistik

  $$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

  H_0 gilt → $T \sim t_{n+m-2}$

  kritischer Wert $c = t_{n+m-2,1-\alpha/2}$

  $|T| > c \leftrightarrow H_0$ wird verworfen

  Test auf $\mu_x = \mu_y$ bei unbekannten, nicht notwendig gleichen σ_x und σ_y
  - erstaunlich schwierig (Behrens-Fisher-Problem)
  - Übersicht und ein Lösungsvorschlag in [4]

- Schraubenlänge nach Wartung:

Ist nach Wartung einer Maschine die mittlere Schraubenlänge noch gleich?

nehmen an, dass $\sigma = 0.05$ mm gleich geblieben ist

- führen Zweistichproben-z-Test durch

Werte vorher

- 4.9204 5.0422 5.0408 4.9438 4.8945 4.9160 4.9507 4.9470 4.9169 4.9653

Werte nachher

- 4.9795 4.9359 4.9858 4.9968 5.0500 4.9599 5.0046 5.0137 4.8933 5.0202 4.9521 4.9837

Nullhypothese: mittlere Schraubenlänge vorher/nachher ist gleich

Signifikanzniveau $\alpha = 0.05$

kritischer Wert $c = z_{1-\alpha/2} = 1.9600$

Ergebnisse

- $\overline{X} = 4.9538$ mm
- $\overline{Y} = 4.9813$ mm
- Teststatistik $|T| = 1.2860$
- $|T| \le c \rightarrow H_0$ wird akzeptiert

- Test auf $\sigma_x = \sigma_y$ bei unbekannten μ_x und μ_y (F-Test):

schätze S_x, S_y wie immer durch

$$\begin{aligned} S_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ S_y^2 &= \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2 \end{aligned}$$

wähle Teststatistik

$$T = \frac{S_x^2}{S_y^2}$$

H_0 gilt $\rightarrow T \sim F_{n-1,\, m-1}$

- F-Verteilung mit n-1 und m-1 Freiheitsgraden (s.u.)
- Quantile bezeichnet mit $F_{n-1,m-1;\alpha}$

F-Verteilung unsymmetrisch, daher zwei kritische Werte

- $c_1 = F_{n-1,m-1;\alpha/2}$
- $c_2 = F_{n-1,m-1;1-\alpha/2}$

$T < c_1$ oder $T > c_2 \leftrightarrow H_0$ wird verworfen

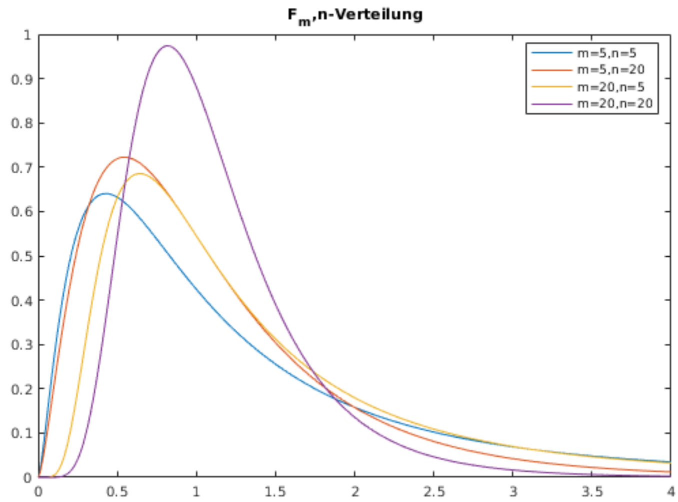
- $F_{m,n}$-Verteilung:

stetige Verteilungsfunktion mit Parametern $m > 0$, $n > 0$

Dichtefunktion

$$f(x|m,n) = \begin{cases} m^{\frac{m}{2}} n^{\frac{n}{2}} \dfrac{\Gamma\left(\frac{m}{2}+\frac{n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \dfrac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}} & | \quad x \ge 0 \\ 0 & | \quad \text{sonst} \end{cases}$$

- graphisch



Eigenschaften

$$E(F_{m,n}) = \frac{n}{n-2} \qquad (n > 2)$$

$$\mathrm{Var}(F_{m,n}) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad (n > 4)$$

Satz

- Seien X, Y unabhängige Zufallsgrößen mit

  $X \sim \chi^2_m$

  $Y \sim \chi^2_n$

- dann gilt

  $$\frac{X/m}{Y/n} \sim F_{m,n}$$

• Streuung der Schraubenlänge nach Verbesserung:

neue Maschine soll Qualität verbessern, insbesondere Streuung der Schraubenlängen verringern

Nullhypothese: Streuung der Schraubenlängen ist gleich geblieben

führen F-Test durch mit $\alpha = 0.01$

- kleiner Wert, weil man keine teure Maschine kaufen will, die nichts bringt

Werte vorher

- 5.0739 5.0270 4.9914 4.9420 4.9071 4.9699 4.9388 4.8843 4.9489 4.9060

Werte nachher

- 5.0333 4.9861 5.0199 4.9841 5.0012 5.0047 4.9904 5.0644 5.0240 5.0359

kritische Werte

- $c_1 = F_{n-1,m-1;\alpha/2} = 0.1529$
- $c_2 = F_{n-1,m-1;1-\alpha/2} = 6.5411$

Ergebnisse

- $S_x = 0.0586$ mm
- $S_y = 0.0258$ mm
- Teststatistik T = 5.1416
- $T > c_1$ und $T < c_2 \rightarrow H_0$ wird akzeptiert

Analyse

- Daten rechtfertigen neue Maschine nicht
- erscheint angesichts der Werte σ_x, σ_y unplausibel
- möglicherweise Fehler 2.Art?

        - bei Test der gleichen Daten mit $\alpha = 0.05$ würde H_0 verworfen

    Ausweg: mehr Daten, $n = m = 100$, s. schrauben100.dat

    - kritische Werte $c_1 = 0.5933$, $c_2 = 1.6854$
    - $S_x = 0.0501$ mm, $S_y = 0.0304$ mm
    - Teststatistik $T = 2.7083$
    - $T > c_2 \rightarrow H_0$ wird verworfen
    - p-Wert = 6.2512e-07 → Maschine kann guten Gewissens gekauft werden

- Aufgaben:

    Aufgabe 36

Parametertests bei Bernoulli-Experimenten

- Exakter Binomialtest:

  Ausgangspunkt
  - n unabhängige Bernoulli-Experimente ("Erfolg", "Misserfolg")
  - unbekannte Erfolgswahrscheinlichkeit p

  Nullhypothesen
  - H_{0a}: $p \le p_0$ bzw. H_{0b}: $p \ge p_0$ bzw. H_{0c}: $p = p_0$

  Testgröße T = Zahl der Erfolge, wobei $T \sim B(n, p_0)$

  T hat kumulative Verteilungsfunktion FB_{n,p_0}, s.o.

  Problem
  - FB nimmt nur einige diskrete Wert an, deswegen hat

    $P(T > c | H_0) = \alpha \Leftrightarrow FB_{n,p_0}(c) = 1 - \alpha$

    i.a. keine Lösung

  stattdessen sucht man

  $$c_a = \min\{c \mid \mathrm{FB}_{n,p_0}(c) \ge 1 - \alpha\}$$

  analog für die anderen Fälle

  $$\begin{aligned} c_b &= \max\{c \mid \mathrm{FB}_{n,p_0}(c) \le \alpha\} \\ c_{c1} &= \max\left\{c \mid \mathrm{FB}_{n,p_0}(c) \le \frac{\alpha}{2}\right\} \\ c_{c2} &= \min\left\{c \mid \mathrm{FB}_{n,p_0}(c) \ge 1 - \frac{\alpha}{2}\right\} \end{aligned}$$

  H_0 wird verworfen falls
  - $T \ge c_a$
  - $T \le c_b$
  - $T \le c_{c1}$ oder $T \ge c_{c2}$

  Berechnung der c-Werte
  - für große n etwas mühsam
  - in Matlab/STB kein Problem
  - $FB_{n,p}(k)$ = `binocdf(k, n, p)`
  - `binoinv` ist Inverses von `binocdf`

  statt c zu berechnen, hier leichter p-Wert, etwa im Fall a

  pWert = $P(T \ge T(x) \mid H_0) = 1 - FB_{n,p_0}(T(x))$

- Abschätzen der Ausschussquote:

  bei Schraubenproduktion entsteht Ausschuss (z.B. Grate im Gewinde)
  - Anteil soll mit neuer Maschine ≤ 0.5 % sein

  Nullhypothese: $p_{Ausschuss} \le 0.005$
  - Alternative: $p_{Ausschuss} > 0.005$

  gewählt werde $\alpha = 0.05$

  Stichprobe von n = 2000 liefert T = 13 Ausschussteile

  $c_a = 15 \rightarrow$ Hypothese wird akzeptiert

  pWert = $0.1350 \ge \alpha$, alles ok

- Genäherter Binomialtest:

für große n Rechnung erleichtern durch Standardisierung und zentralen Grenzwertsatz

$$\begin{aligned} P(T > c_a \mid H_0) &= P\Big(\frac{T - np_0}{\sqrt{np_0(1-p_0)}} > \frac{c_a - np_0}{\sqrt{np_0(1-p_0)}} \mid H_0\Big) \\ &\approx P\Big(Z > \frac{c_a - np_0}{\sqrt{np_0(1-p_0)}}\Big) \overset{!}{=} \alpha \\ \Leftrightarrow \quad \frac{c_a - np_0}{\sqrt{np_0(1-p_0)}} &= z_{1-\alpha} \\ \Leftrightarrow \quad c_a &= np_0 + z_{1-\alpha}\sqrt{np_0(1-p_0)} \end{aligned}$$

- mit Z ~ N(0,1)
- verwerfen also für $T > c_a$

analog für die Fälle b, c

Faustregel: ab $n\, p_0\, (1 - p_0) > 9$

noch genauer mit Stetigkeitskorrektur

im Schrauben-Beispiel

- $n\, p_0\, (1 - p_0) = 9.95$ → passt so gerade
- ca = 15.1885
- liefert hier gleiches Ergebnis

- Exakter Test nach Fisher:

Vergleich der Wahrscheinlichkeiten p_x, p_y zweier Bernoulli-Experimente X_i (i = 1...n), Y_j (j = 1...m)

H_0: $p_x = p_y$, H_1: $p_x \neq p_y$

gemessen werde Zahl der Erfolge T_x, T_y

Grundidee

- halte $K = T_x + T_y$ fest (Gesamtzahl an Erfolgen)
- betrachte T_x unter der Annahme von festem K = k (und der Annahme $p_x = p_y$)

man kann zeigen, dass T_x dann hypergeometrisch verteilt ist [Ross2, Zabluchko]

$$P(T_x = x_1 \mid T_x + T_y = k \wedge H_0) = \frac{\binom{n}{x_1}\binom{m}{k - x_1}}{\binom{n+m}{k}} \qquad (x_1 = 0 \ldots k)$$

man wird H_0 ablehnen, falls (mit $x_1 := T_x(X)$)

$$P(T_x \le x_1) \le \frac{\alpha}{2} \vee P(T_x \ge x_1) \le \frac{\alpha}{2}$$

- mit

$$\begin{aligned} p_1 := P(T_x \le x_1) &= \sum_{i=0}^{x_1} P(T_x = i) = \sum_{i=0}^{x_1} \frac{\binom{n}{i}\binom{m}{k-i}}{\binom{n+m}{k}} \\ p_2 := P(T_x \ge x_1) &= 1 - P(T_x \le x_1 - 1) \end{aligned}$$

mit Matlabs `hygecdf` leicht auszurechnen

p-Wert: pW = 2 min $\{p_1, p_2\}$

- Wirksamkeit eines Medikaments:

Studie will Wirksamkeit eines Medikaments prüfen

- n = 100 Personen bekommen Medikament (Behandlungsgruppe), davon werden 40 gesund
- m = 120 Personen bekommen Placebo (Kontrollgruppe), davon werden 42 gesund

Darstellung als Kontingenztafel

	Medikament	Placebo	Summe
Erfolg	40	42	82
Misserfolg	60	78	138
Summe	100	120	220

Erfolgsquoten 40% (Behandlunsgruppe) bzw. 35% (Kontrollgruppe)

- Ist das signifikant (bei $\alpha = 0.05$)?
- Beachte: Unter H_0 ($p_x = p_y$) "gewinnt" in 50% der Fälle die Behandlungsgruppe!

Berechnung mit Matlab ergibt

- $p_1 = 0.8170$
- $p_2 = 0.2663$

Ergebnis

- H_0 wird akzeptiert, d.h. die Wirkung des Medikaments ist nicht signifikant
- Das gilt in beiden Richtungen: Es nützt wohl nichts, aber es schadet wohl auch nichts

- Aufgaben:

  Aufgabe 37

  Aufgabe 38

Nicht-parametrische Tests

- Vorzeichentest:

  Test auf Wert des Median
  - H_0: Median m hat Wert m_0 ($m = m_0$)
  - H_1: $m \neq m_0$

  Idee
  - teste Vorzeichen von $X - m_0$
  - neue Zufallsvariable V_i

    $$V_i = \begin{cases} 1 & | & x_i < m_0 \\ 0 & | & x_i \geq m_0 \end{cases}$$

  - ist Bernoulli-Größe mit p = 1/2

  Vorsicht
  - gilt zunächst nur bei stetiger Zufallsgröße
  - unstetige Zufallsgröße → möglicherweise $P(X_i = m_0 \mid H_0) \neq 0$
  - einfache Abhilfe: Stichprobenwerte mit $X_i = m_0$ werden verworfen

  Teststatistik T = Anzahl der V_i mit Wert 1

  $T \sim B(n, 1/2)$

- Beispiel Saurer Regen:

  sauberes Regenwasser hat pH-Wert 5.2 - 6.0 (wegen gelöstem CO_2)

  wir definieren "Saurer Regen": pH-Wert < 5.2
  - (üblich: 4.2 - 4.8)

  Untersuchung: Ist Regen im Naturschutzpark Wölfenhausen sauer?

  Messwerte x_i des pH-Werts an 16 Tagen
  - 4.9387 4.8816 5.2655 5.2952 1.6869 4.9456 5.1463 5.2094 5.2547 4.7760 5.1551 4.6190 4.9984 5.4597 5.0853 4.7238

  H_0: Median(X_i) ≥ 5.2, H_1: Median(X_i) < 5.2

  Teststatistik T = Anzahl der Messwerte < 5.2
  - diese Richtung, damit großer Wert verwirft (grundsätzlich egal)
  - Gleichheit hier kein Problem (stetig, 5.2 wird nicht angenommen)

  Auswertung: T(x) = 11

  Berechnung des p-Werts

  $$\mathrm{pWert} = P(T \geq 11 \mid H_0) = 1 - \mathrm{FB}_{16,0.5}(10) = 0.1051$$

  bei $\alpha = 0.05$ wird H_0 akzeptiert
  - → Saurer Regen kann nicht nachgewiesen werden

- Wilcoxon-Rangsummen-Test:

  heißt auch Wilcoxon-Mann-Whitney-Test oder Mann-Whitney-U-Test

  Voraussetzung: X_i, Y_j alle unabhängig, ordinal

  Test auf gleiche Verteilung
  - H_0: X_i, Y_j haben gleiche Verteilung
  - H_1: X_i, Y_j haben verschiedene Verteilung

  Spezialfall
  - X_i, Y_j stetig, ihre Verteilungen unterscheiden sich nur um Verschiebung

- H_0: X_i, Y_j haben gleichen Median, H_1: X_i, Y_j haben verschiedenen Median
- etwa zum Test auf Änderung der Kalibrierung bei zwei Messreihen

Vorgehensweise

- ordne alle Elemente X_i, Y_j gemeinsam der Größe nach
- R_i = Rang von X_i = Position in der gemeinsamen Rangfolge, $R_i \in [1, n+m]$
- bei gleichem Rang mehrerer Elemente (**Bindung**): verwende gemittelte Position

Teststatistik

$$T = \sum_{i=1}^{n} R_i$$

- mit Verteilungsfunktion

  $$\mathrm{wr}_{n,m}(k) := P(T \le k \mid H_0)$$

$\mathrm{wr}_{n,m}$ lässt sich (aufwändig) genau berechnen und einfach approximieren, s.u.

H_0 wird verworfen, falls T zu klein oder zu groß

- konkret: falls für k := T(x) gilt

  $$\begin{array}{rl} & P(T \le k \mid H_0) \le \frac{\alpha}{2} \;\vee\; P(T \ge k \mid H_0) \le \frac{\alpha}{2} \\ \Leftrightarrow & \mathrm{wr}_{n,m}(k) \le \frac{\alpha}{2} \;\vee\; \mathrm{wr}_{n,m}(k-1) \ge 1 - \frac{\alpha}{2} \end{array}$$

entsprechender p-Wert daher

$$\mathrm{pWert} = 2 \min\{\mathrm{wr}_{n,m}(k), 1 - \mathrm{wr}_{n,m}(k-1)\}$$

- Berechnen von $\mathrm{wr}_{n,m}$:

  man bestimmt eine Rekursionsformel

  $$\begin{aligned} \mathrm{wr}_{n,m}(k) &= P(T \le k \mid H_0) \\ &= P(T \le k \mid H_0 \wedge \text{größter Wert ist aus X})\, P(\text{größter Wert ist aus X} \mid H_0) \\ &\quad + P(T \le k \mid H_0 \wedge \text{größter Wert ist aus Y})\, P(\text{größter Wert ist aus Y} \mid H_0) \\ &= \mathrm{wr}_{n-1,m}(k-n-m)\, \frac{n}{n+m} + \mathrm{wr}_{n,m-1}(k)\, \frac{m}{n+m} \end{aligned}$$

  - mit den Anfangswerten

    $$\begin{aligned} \mathrm{wr}_{1,0}(k) &= \begin{cases} 0 & | \quad k \le 0 \\ 1 & | \quad k > 0 \end{cases} \\ \mathrm{wr}_{0,1}(k) &= \begin{cases} 0 & | \quad k < 0 \\ 1 & | \quad k \ge 0 \end{cases} \end{aligned}$$

  für kleine Zahlen per Hand, sonst als Programm

  für "große" Werte von n, m über Approximation durch Normalverteilung

  - speziell, falls keine Werte mit gleichem Rang auftreten, gilt

    $$\begin{aligned} E(T \mid H_0) &= \frac{1}{2} n(n+m+1) \\ \mathrm{Var}(T \mid H_0) &= \frac{1}{12} nm(n+m+1) \end{aligned}$$

  - also

    $$\frac{T - E(T)}{\sqrt{\mathrm{Var}(T)}} \sim N(0,1)$$

  was heißt "groß"?

  - n, m > 7 [Ross2]
  - n oder m > 25 [Fahrmeir]

- Beispiel Lebensdauern:

  Prüfen der Lebensdauern von Festplatten zweier Hersteller

Prüfen der Lebensdauern von Festplatten zweier Hersteller

- H_0: Lebensdauern haben bei beiden gleiche Verteilung

Stichproben

- Hersteller X: 6.9175 7.1546 8.0619 6.7216 7.6289 8.2784
- Hersteller Y: 7.0103 6.1237 6.0309 7.4330 6.8247

Ordnen der Werte

Rang	1	2	3	4	5	6	7	8	9	10	11
Wert	6.0309	6.1237	6.7216	6.8247	6.9175	7.0103	7.1546	7.4330	7.6289	8.0619	8.2784
Herst.	Y	Y	X	Y	X	Y	X	Y	X	X	X

Teststatistik berechnen

$T = 3 + 5 + 7 + 9 + 10 + 11 = 45$

Verteilungsfunktion per Rechner

$$\begin{aligned} \mathrm{wr}_{6,5}(45) &= 0.9589 \\ \mathrm{wr}_{6,5}(44) &= 0.9372 \end{aligned}$$

entsprechender p-Wert = 0.1255

H_0 wird bei $\alpha = 0.05$ akzeptiert

- Chi-Quadrat-Anpassungstest:

  betrachten n diskrete i.i.d Zustandsgrößen X_i, die r Ergebnisse haben können

  vermuten Wahrscheinlichkeitsfunktion p_j, $j = 1 .. r$

  H_0: X_i sind gemäß dieser Wahrscheinlichkeitsfunktion verteilt

  Teststatistik

  - N_j = Zahl der Experimente mit Ergebnis j, $j = 1 .. r$

  $$T = \sum_{j=1}^{r} \frac{(N_j - np_j)^2}{np_j}$$

  Satz (Pearson, 1900)

  - Für $n \to \infty$ ist $T \sim \chi^2_{r-1}$
  - Beweis: [Kabluchko]

  also: verwirf H_0 bei Signifikanzniveau $\alpha \leftrightarrow T > \chi^2_{r-1,1-\alpha}$

  Faustregel

  - alle $N_j \geq 1$
  - mindestens 80% davon > 5

- Beispiel Mendelsche Vererbungsexperimente:

  Mendel züchtete Erbsen, die rund oder kantig und gelb oder grün sein konnten

  er vermutete folgende Wahrscheinlichkeiten

  - p_1 = P(rund und gelb) = 9/16
  - p_2 = P(kantig und gelb) = 3/16
  - p_3 = P(rund und grün) = 3/16
  - p_4 = P(kantig und grün) = 1/16

  Versuche mit n = 556 Erbsen ergaben

N_1	N_2	N_3	N_4
315	101	108	32

  Chi-Quadrat-Anpassungstest liefert

- T = 0.4700
- $\chi^2_{3,0.95} = 7.8147$
- H_0 wird akzeptiert

pWert

$$\text{pWert} = P(T \geq 0.47) \approx 1 - P(\chi_3^2 < 0.47) = 0.9254$$

Ergebnis ist erstaunlich gut
- wurde von Statistiker Fisher angezweifelt (Mendel-Fisher-Kontroverse)
- wird bis heute noch diskutiert [6, 7]

- Ergänzungen zum Chi-Quadrat-Anpassungstest:

  falls X_i stetig verteilt oder r sehr groß
  - bilde Klassen, z. B. durch Intervalle
  - beachte Faustregel

  falls p Parameter der Verteilung aus Daten bestimmt werden
  - verwende entsprechend weniger Freiheitsgrade, $T \sim \chi^2_{r-1-p}$

- Chi-Quadrat-Unabhängigkeitstest:

  Zufallsgrößen X, Y können nur r bzw. s verschiedene Werte annehmen

  machen (viele) Experimente und zählen $N_{i,j}$ = Anzahl der Ergebnisse mit X = i und Y = j

  Frage: Sind X, Y unabhängig?

  Grundidee: bei Unabhängigkeit gilt

  $$P(X = i \wedge Y = j) = P(X = i)\, P(Y = j)$$

  berechne Randverteilungen

  $$\begin{aligned} n_i &:= \sum_{j=1}^{s} N_{i,j} \\ m_j &:= \sum_{i=1}^{r} N_{i,j} \\ n &:= \sum_{i=1}^{r} n_i = \sum_{j=1}^{s} m_j \end{aligned}$$

  und schätze die Einzelwahrscheinlichkeiten

  $$\begin{aligned} \hat{p}_i &:= \frac{n_i}{n} \qquad i = 1 \ldots r \\ \hat{q}_j &:= \frac{m_j}{n} \qquad j = 1 \ldots s \end{aligned}$$

  Teststatistik

  $$T = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{i,j} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}$$

  Für $n \to \infty$ ist $T \sim \chi^2_{(r-1)(s-1)}$

  also: verwirf H_0 ("X, Y, sind unabhängig") bei Signifikanzniveau $\alpha \leftrightarrow T > \chi^2_{(r-1)(s-1),1-\alpha}$

- Beispiel Maschinenausfall im Schichtbetrieb:

  In einer Firma werden vier Maschinen in drei Schichten betrieben.

  Das Muster der Ausfälle scheint schicht-abhängig zu sein
  - Anzahl der Ausfälle pro Schicht und Maschine über ein Jahr

Maschine Schicht	A	B	C	D
1	12	14	13	12
2	16	13	15	11
3	11	10	9	19

H_0: Ausfälle pro Schicht und pro Maschine sind unabhängig

man erhält

- mj = 39, 37, 37, 42
- ni = 51, 55, 49
- n = 155
- T = 5.7287
- $\chi^2_{6,0.95}$ = 12.5916
- pWert = 0.4543

H_0 wird akzeptiert, die Schwankungen sind anscheinend Zufall

- Kolmogorow-Smirnow-Test:

  Anpassungstest (Test auf Verteilung), also

  - H_0: X_i haben die (stetige) Verteilung F_0

  vergleicht die empirische Verteilungsfunktion F_n mit der vorgegebenen

  Vorgehensweise

  - bestimme F_n

    $$F_n(x) = \frac{1}{n}|\{i \in [1,n] \mid x_i \leq x\}|$$

  - berechne Teststatistik

    $$T = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

  Verteilung von T

  - unabhängig von F_0!
  - bekannt (Kolmogorov-Verteilung), aber nicht ganz leicht zu berechnen [8]
  - Quantile $KS_{\alpha,n}$ tabelliert [Rooch] (oder kompletter Test mit `kstest`)
  - für n > 35 Näherungsformel

    $$\mathrm{KS}_{\alpha,n} \approx \frac{1}{\sqrt{n}}\sqrt{\frac{\ln\frac{2}{\alpha}}{2}}$$

  verwirf H_0 ↔ T > $KS_{\alpha,n}$

  Variante für zwei Stichproben X_i, Y_j

  - H_0: X, Y haben gleiche Verteilungsfunktion
  - empirische Verteilungsfunktionen $F_n(x)$, $G_m(x)$
  - Teststatistik nun

    $$T = \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)|$$

  - Näherungsformel für Quantile

    $$\mathrm{KS}_\alpha \approx \sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{\ln\frac{2}{\alpha}}{2}}$$

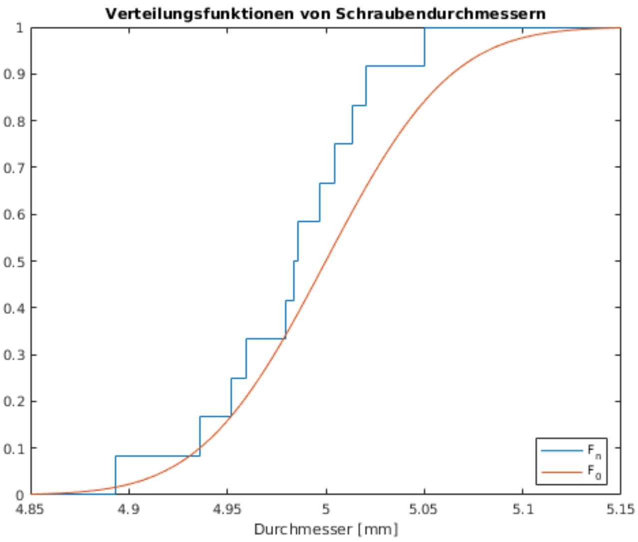
- Beispiel Schraubendurchmesser:

  Stichprobe der Schraubendurchmesser mit neuer Maschine, aufsteigend sortiert

- 4.8933 4.9359 4.9521 4.9599 4.9795 4.9837 4.9858 4.9968 5.0046 5.0137 5.0202 5.0500

H_0: $X \sim N(\mu, \sigma^2)$ mit $\mu = 5$ mm, $\sigma = 0.05$ mm

Plot von empirischer und erwarteter Verteilungsfunktion



praktische Berechnung von T

- F_n stückweise konstant, F monoton steigend
- → Werte an den Sprungstellen (also den Daten) jeweils vorwärts und rückwärts reichen

$$d_{i,o} := |F_n(x_i) - F_0(x_i)|$$
$$d_{i,u} := |F_n(x_{i-1}) - F_0(x_i)|$$

- tabellarisch

i	x_i	$F_n(x_i)$	$F_0(x_i)$	$d_{i,o}$	$d_{i,u}$
1	4.8933	0.0833	0.0164	0.0669	0.0164
2	4.9359	0.1667	0.0999	0.0667	0.0166
3	4.9521	0.2500	0.1690	0.0810	0.0024
4	4.9599	0.3333	0.2113	0.1221	0.0387
5	4.9795	0.4167	0.3409	0.0758	0.0076
6	4.9837	0.5000	0.3722	0.1278	0.0445
7	4.9858	0.5833	0.3882	0.1951	0.1118
8	4.9968	0.6667	0.4745	0.1922	0.1088
9	5.0046	0.7500	0.5367	0.2133	0.1300
10	5.0137	0.8333	0.6080	0.2254	0.1420
11	5.0202	0.9167	0.6569	0.2598	0.1764
12	5.0500	1.0000	0.8413	0.1587	0.0753

- $T = \max \{d_{i,o}, d_{i,u}\} = 0.2598$

für $\alpha = 0.05$ aus Tabelle: $KS_{\alpha,n} = 0.375$ → H_0 wird akzeptiert

- Aufgaben:

  Aufgabe 39

  Aufgabe 40

Tests mit mehr als zwei Stichproben

- Vergleiche mehrerer Stichproben:

  betrachten n Stichproben desselben Merkmals, mit jeweils n_i Werten

  $$X_{ij}, \quad i = 1 \ldots n,\ j = 1 \ldots n_i$$

  Frage: haben alle denselben Mittelwert?

  Idee: mache insgesamt n(n-1)/2 paarweise Tests (z. B. Zweistichproben-t-Tests)

  Problem: Signifikanzniveau des Gesamttests ist riesig
  - etwa bei n = 7 und α = 0.05 hat man 21 Tests
  - man erwartet im Schnitt (etwas mehr als) einen Vergleich, der ablehnt, obwohl H_0 zutrifft

  einfache Abhilfe (**Verfahren von Bonferroni**):
  - bei m Tests mit Gesamt-Signifikanzniveau α führe jeden Test mit α/m aus
  - → Fehler 1. Art des Gesamttests dann ≤ α

  Verbesserung (**Verfahren von Bonferroni-Holm**)
  - ist ein Test bei α/m verworfen worden, wird er gestrichen
  - es bleiben m-1 Tests, also Niveau für diese α/(m-1)
  - usw.

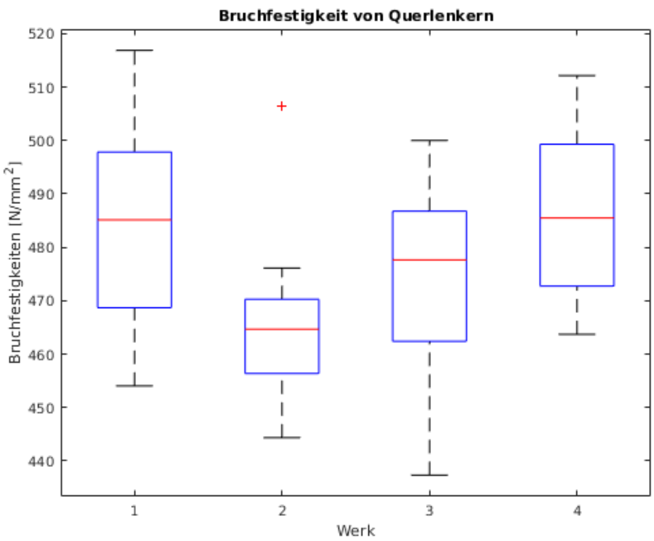
- Beispiel Bruchfestigkeit von Querlenkern:

  eine Firma stellt Querlenker in vier Werken her

  Frage: Ist ihre Biegefestigkeit bei allen vier Werken gleich groß?

  Stichproben (in N/mm^2)
  - Werk 1: 469.24 497.34 499.52 486.75 460.08 469.54 454.05 498.35 483.53 495.10 468.17 516.89
  - Werk 2: 506.34 467.52 447.79 456.38 470.28 468.81 456.78 476.12 461.82 444.36
  - Werk 3: 500.01 463.97 481.84 485.48 437.33 488.08 460.83 473.47
  - Werk 4: 472.85 497.97 485.50 463.71 489.56 472.40 512.17 480.72 503.29

  Boxplot



  Annahme: Werte sind normalverteilt mit gleichem σ (testen!)

  paarweise t-Tests liefern folgende p-Werte

(1,2)	0.0366
(1,3)	0.3018
(1,4)	0.6829
(2,3)	0.3592
(2,4)	0.0149
(3,4)	0.1643

bei einem einzigen Test mit $\alpha = 0.05$ würde $\mu_1 = \mu_2$ (oder $\mu_2 = \mu_4$) verworfen

6 Tests, also ersten Test mit $\alpha/6 = 0.0083$ → keiner wird verworfen

- → Nullhypothese (alle μ_i gleich) wird bei $\alpha = 0.05$ akzeptiert

- Varianzanalyse (**ANOVA** = "ANalysis Of VAriance"):

  n Stichproben X_{ij} wie oben

  gesucht: ein Test, der direkt auf Gleichheit aller μ_i prüft

  Voraussetzung: alle Stichproben $X_{ij} \sim N(\mu_i, \sigma^2)$ für i = 1...n

  H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$

  Idee: vergleiche die Streuungen innerhalb der Stichproben mit Streuung zwischen den Stichproben

  man definiert die Mittelwerte der einzelnen Stichproben

  $$\overline{X}_i := \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

  - und den Mittelwert der Gesamtdaten

  $$\overline{X} := \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} X_{ij} \quad \text{mit} \quad N := \sum_{i=1}^{n} n_i$$

  Für die quadratische Gesamtabweichung gilt

  $$\sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\overline{X}_i-\overline{X})^2$$

  - Beweis im Anhang
  - 1. Term ≙ Abweichung innerhalb der Stichproben
  - 2. Term ≙ Abweichung der Mittelwerte zwischen den Stichproben

  H_0 gilt → Terme sollten gleich groß sein

  - sonst: 2. Term wird größer sein als erster

  man definiert

  $$\begin{aligned} Y_1 &:= \frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2 \sim \chi^2_{N-n} \\ Y_2 &:= \frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\overline{X}_i-\overline{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n} n_i(\overline{X}_i-\overline{X})^2 \sim \chi^2_{n-1} \end{aligned}$$

  - und hat als Teststatistik

  $$T := \frac{Y_2/(n-1)}{Y_1/(N-n)} \sim F_{n-1,N-n}$$

  H_0 wird verworfen ↔ $T > F_{n-1,N-n;1-\alpha}$

  Verfahren heißt einfaktorielle ANOVA, da ein Faktor verschiedene Stichproben erzeugt (≙ verschiedene Werte von i)

  wenn H_0 verworfen wird

- Frage: Welche Reihen sind die unterschiedlichen?
- herausfinden z. B. mit Bonferroni-Holm

- Zweifaktorielle Varianzanalyse:

  zwei Faktoren für unterschiedliche Stichproben, z. B
  - Studienerfolg in Abhängigkeit von Motivation und Familiensituation
  - Fehlerhäufigkeit bei KfZ über Modell und Produktionsstandort

  aus X_{ij} wird X_{ijk}
  - i,j durchlaufen die beiden Faktoren
  - k durchläuft die Stichproben unter festem i und j

  braucht viele Daten!

  funktioniert i. W. wie einfaktoriell, aber mehr Indizes
  - neues Phänomen: Interaktion zwischen den Einflussfaktoren

  auch (beliebig) mehrfaktoriell

- Beispiel Bruchfestigkeit mit Varianzanalyse:

  Berechnen der Mittelwerte
  - $\overline{X}_i$ = (483.21, 465.62, 473.88, 486.46)
  - $\overline{X}$ = 477.5367

  und der Teststatistik
  - Y_1 = 28.5907
  - Y_2 = 6.5773
  - T(x) = 2.6839

  F-Verteilung liefert bei $\alpha = 0.05$
  - $F_{n-1,N-n;1-\alpha}$ = 2.8742
  - pWert = 1 - F(T(x), n-1, N-n) = 0.0616

  H_0 wird also akzeptiert, die Unterschiede sind nicht signifikant

  kompletter Test in Matlab mit `anova1`
  - Trick: x_{ij} als 2d-Matrix, kürzere Datensätze mit NaN füllen

- Kruskal-Wallis-Test:

  Alternative zu ANOVA bei nicht normalverteilten Datensätzen
  - Idee: Erweiterung von Wilcoxon-Rangsummen-Test

  H_0: alle Stichproben haben die gleiche Verteilung F_i

  zunächst sortiere alle x_{ij} der Größe nach
  - jedem Wert x_{ij} wird so sein Rang r_{ij} zugeordnet
  - bei Bindungen mittleren Rang verwenden

  berechne mittleren Rang pro Stichprobe

  $$\overline{r}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}$$

  Idee: Abweichung vom mittleren Rang (N+1)/2 der Gesamtheit sollte nicht zu groß sein

  Teststatistik im Fall ohne Bindungen

  $$\begin{aligned} T_0 &:= \frac{12}{N(N+1)} \sum_{i=1}^{n} n_i \left(\overline{r}_i - \frac{N+1}{2}\right)^2 \\ &= \frac{12}{N(N+1)} \sum_{i=1}^{n} n_i \overline{r}_i^2 - 3(N+1) \end{aligned}$$

bei Bindungen Korrekturfaktor

- es gebe B Bindungsgruppen, in jeder b_k Werte mit gleichem Rang, dann

$$T = \left(1 - \frac{\sum_{k=1}^{B}(b_k^3 - b_k)}{N^3 - N}\right)^{-1} T_0$$

asymptotisch für große Stichproben

$T \sim \chi^2_{n-1}$

- Faustregel: Verteilung ok für $n_i > 6$

H_0 wird verworfen ↔ $T > \chi^2_{n-1,1-\alpha}$

genauere Berechnung der Verteilung von T

- grundsätzlich möglich, reine Kombinatorik (vgl. $wr_{n,m}$)
- im Detail schwierig und sehr rechenintensiv
- Tabellen und bessere asymptotische Formeln in [9]

- Beispiel Bruchfestigkeit mit Kruskal-Wallis-Test:

keine Bindungen (klar: reelle Zahlen)

Ränge der Werte (Stichproben als Spalten)

15	37	35	19
31	12	11	32
34	3	23	26
27	5	25	10
7	17	1	29
16	14	28	18
4	6	8	38
33	21	20	22
24	9		36
30	2		
13			
39			

Rang-Mittelwerte

- $\bar{r}_i$ = (22.75, 12.60, 18.87, 25.56)

Teststatistik

- T(x) = 7.1250

χ^2-Verteilung liefert bei α = 0.05

- $\chi^2_{n-1,1-\alpha}$ = 7.8147
- pWert = 0.0680

H_0 wird akzeptiert

- pWert etwas größer als bei ANOVA
- dort wird mehr Information vorausgesetzt

- Aufgaben:

Aufgabe 41

Aufgabe 42

Regressionsanalyse

- Einfache lineare Regression
- Statistische Analyse der Regressionsparameter
- Ergänzende Analysen bei linearer Regression
- Mehrfache lineare Regression

Einfache lineare Regression

- Problemstellung:

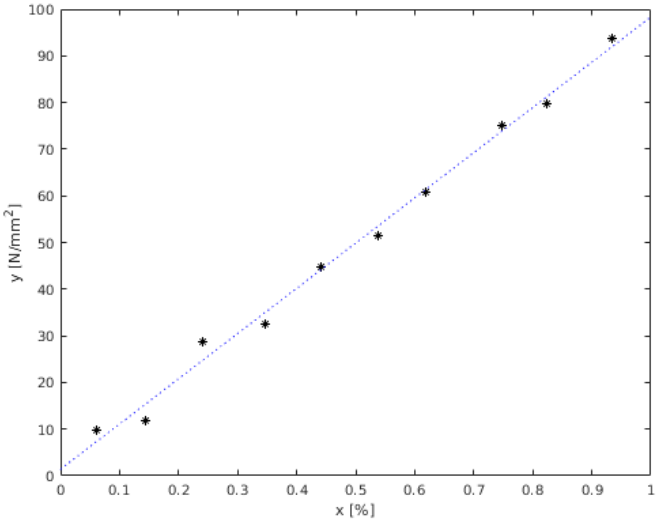
  Messwerte (x_i, y_i) (i = 1, ..., n) einer Größe y in Abhängigkeit von x

  Beispiel: Messwerte eines Zugversuchs, x = Dehnung, y = Spannung

x_i [%]	0.0601	0.1437	0.2399	0.3468	0.4407	0.5386	0.6188	0.7471	0.8235	0.9344
y_i [N/mm^2]	9.68	11.92	28.61	32.42	44.63	51.47	60.91	75.12	79.79	93.80

  Annahme: linearer Zusammenhang zwischen x und y

  y = α + β x



  gesucht: Schätzwerte für α, β

- Ausgleichsrechung:

  für gegebene Messwerte n lineare Gleichungen für 2 Unbekannte
  - in Matrix-Schreibweise

    A z = b
  - mit

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad z = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \qquad b = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

  suche Gerade mit kleinster quadratischer Abweichung

$$S = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

  Lösung gegeben durch

  (A^T A) z = A^T b
  - vgl. Numerische Mathematik

  konkret

$$\begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

  - mit dem Ergebnis

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\alpha = \overline{y} - \beta\,\overline{x}$$

- Alternativer statistischer Ansatz (Regressionsanalyse):

  y_i enthalten unabhängige normalverteilte Störungen $\varepsilon_i \sim N(0, \sigma^2)$

  Messergebnisse daher Zufallsvariablen

  $Y_i = \alpha + \beta\, x_i + \varepsilon_i$

  x_i deterministisch angenommen
  - vereinfacht Theorie deutlich
  - gute Näherung, wenn x_i viel genauer bestimmbar als y_i
  - Theorie erweiterbar auf zufällige x_i

- Schätzer für α, β und σ:

  verwenden für α und β Maximum-Likelihood-Schätzer

  Y_i sind i.i.d. mit $Y_i \sim N(\alpha + \beta\, x_i, \sigma^2)$

  Dichtefunktion also

  $$f_{\alpha\beta\sigma}(y_1, \ldots, y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}$$

  - ihr Logarithmus l = ln f

  $$l(\alpha, \beta, \sigma) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

  l als Funktion von α und β maximal ⇔

  $$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \to \text{minimal}$$

  Ergebnis dasselbe wie bei Ausgleichsrechnung, also

  $$B = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  $$A = \overline{Y} - B\,\overline{x}$$

  ML-Schätzer für σ nicht erwartungstreu (wie üblich)
  - kleine Nennerkorrektur liefert erwartungstreuen Schätzer

  $$S_{y|x}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - A - Bx_i)^2$$

  mit Zahlenwerten von oben
  - a = 1.4088
  - b = 96.9147
  - $s_{y|x}$ = 2.5538

Statistische Analyse der Regressionsparameter

- Verteilungen der Schätzer A und B:

  als Linearkombinationen normalverteilter Größen sind A und B normalverteilt

  Berechnung der Erwartungswerte

  - mit
    $$E(Y_i) = E(\alpha + \beta\, x_i + \varepsilon_i) = \alpha + \beta\, x_i$$
  - und dem bekannten
    $$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$
  - erhält man
    $$\begin{aligned} E(B) &= \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2} \sum_{i=1}^{n} (x_i - \overline{x}) E(Y_i) \\ &= \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2} \sum_{i=1}^{n} (x_i - \overline{x})(\alpha + \beta\, x_i) \\ &= \beta \frac{\sum_{i=1}^n (x_i - \overline{x})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ &= \beta \end{aligned}$$
  - sowie für A
    $$\begin{aligned} E(A) &= E(\overline{Y} - B\, \overline{x}) \\ &= E(\overline{Y}) - \overline{x} E(B) \\ &= \alpha + \beta\, \overline{x} - \overline{x}\, \beta \\ &= \alpha \end{aligned}$$

  also: A und B sind erwartungstreu

  Berechnung der Varianzen

  - Wegen $\mathrm{Var}(Y_i) = \sigma^2$ und Unabhängigkeit der Y_i hat man
    $$\begin{aligned} \mathrm{Var}(B) &= \frac{1}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} \mathrm{Var}\left(\sum_{i=1}^{n} (x_i - \overline{x}) Y_i \right) \\ &= \frac{1}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} \sum_{i=1}^{n} (x_i - \overline{x})^2\, \mathrm{Var}(Y_i) \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \end{aligned}$$
  - genauso einfach, aber etwas länger, berechnet man (s. Anhang)
    $$\mathrm{Var}(A) = \frac{\sigma^2}{n} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

  im Beispiel, mit $s_{y|x}$ als Schätzwert für das unbekannte σ

  $\sigma_\alpha = 1.6262$

  $\sigma_\beta = 2.8843$

  - relative Schwankungen

    $\sigma_\alpha / a = 1.1543$

    $\sigma_\beta / b = 0.0298$
  - Wert für α deutlich ungenauer als der für β

- Verteilungen des Schätzers $S^2_{y|x}$:

Man kann beweisen

$$\frac{n-2}{\sigma^2} S_{y|x}^2 \sim \chi_{n-2}^2$$

daher erhält man

$$\begin{aligned} E(S_{y|x}^2) &= \sigma^2 \\ \mathrm{Var}(S_{y|x}^2) &= \frac{2\sigma^4}{n-2} \end{aligned}$$

außerdem: $S_{y|x}$ ist unabhängig von A und B

Verhältnisse ähnlich wie bei $\overline{X}$ und S^2 für Normalverteilungen

- Konfidenzintervalle für α und β:

Verteilungen von A und B erfüllen Voraussetzungen von Satz über t-Verteilung →

$$\begin{aligned} \frac{A-\alpha}{\sqrt{S_{y|x}^2}} \sqrt{\frac{n \sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n x_i^2}} &\sim t_{n-2} \\ \frac{B-\beta}{\sqrt{S_{y|x}^2}} \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2} &\sim t_{n-2} \end{aligned}$$

bei gegebener Irrtums-Wahrscheinlichkeit p gilt daher

$$\begin{aligned} P\left(-t_{n-2,1-p/2} \le \frac{A-\alpha}{\sqrt{S_{y|x}^2}} \sqrt{\frac{n \sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n x_i^2}} \le t_{n-2,1-p/2}\right) &= 1-p \\ \Rightarrow P\left(A - t_{n-2,1-p/2}\sqrt{S_{y|x}^2}\sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}} \le \alpha \right. & \\ \left. \le A + t_{n-2,1-p/2}\sqrt{S_{y|x}^2}\sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}}\right) &= 1-p \end{aligned}$$

man erhält also als Schätzer für die Grenzen des zweiseitigen (1-p)-Konfidenzintervalls für α

$$G_{o/u}^{\alpha} = A \pm t_{n-2,1-p/2}\sqrt{S_{y|x}^2}\sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}}$$

entsprechend ergibt sich für β

$$G_{o/u}^{\beta} = B \pm t_{n-2,1-p/2}\sqrt{S_{y|x}^2}\frac{1}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

im Beispiel für p = 5%

- α: [-2.3411, 5.1588]
- β: [90.2635, 103.5659]

- Konfidenzintervall für σ^2:

aus der bekannten Verteilung von $S^2_{y|x}$ erhält man für den symmetrischen Fall

$$P\left(\chi_{n-2,p/2}^2 \le \frac{n-2}{\sigma^2} S_{y|x}^2 \le \chi_{n-2,1-p/2}^2\right) = 1-p$$

- also als Schätzer für das (1-p)-Konfidenzintervall

$$G_u = \frac{n-2}{\chi_{n-2,1-p/2}^2} S_{y|x}^2, \qquad G_o = \frac{n-2}{\chi_{n-2,p/2}^2} S_{y|x}^2$$

im Beispiel für p = 5%

- σ^2: [2.9756, 23.9371]
- σ: [1.7250, 4.8926]

- Tests für α und β:

  Verteilung interessierender Teststatistiken meistens bekannt

  hier nur exemplarisch für β am Standard-Beispiel

  Fragestellung

  - gesucht ist Material mit Elastizitätsmodul-Modul (≡ β) ≥ 90 N/mm^2
  - Ist E-Modul der Probe groß genug für die Anforderungen?
  - im Zweifelsfall lieber nicht verwenden, daher p = 1 %
  - H_0: E-Modul β der Probe ist ≤ β_0 = 90 N/mm^2

  als Teststatistik gut geeignet

$$T = \frac{B - \beta_0}{\sqrt{S^2_{y|x}}} \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

  unter H_0 ist T ~ t_{n-2}, daher kritischer Wert

$$\begin{aligned} P(T > c \mid H_0) &\stackrel{!}{=} p \\ \Rightarrow \qquad c &= t_{n-2,1-p} \end{aligned}$$

  Messwerte des Beispiels liefern

  - c = 2.8965
  - T(x) = 2.3974

  bei 1% Signifikanzniveau wird H_0 akzeptiert, d. h. die Probe wird abgelehnt

  p-Wert von T(x)

$$\begin{aligned} \text{pWert} &= P(T \geq T(x) \mid H_0) \\ &= 1 - P(T \leq T(x) \mid H_0) \\ &= 2.17\% \end{aligned}$$

  - bei Signifikanz-Niveau 5% würde man hier verwerfen, also die Probe verwenden
  - wichtig: Signifikanzniveau vor der Testdurchführung festlegen!

Ergänzende Analysen bei linearer Regression

- Bestimmtheitsmaß:

  besteht überhaupt ein linearer Zusammenhang zwischen x- und y-Werten?
  - bei Messwerten existiert häufig ein direkter Zusammenhang, aber nicht unbedingt linear
  - bei allgemeinen statistischen Daten ist x oft nur ein Einflussfaktor

  Idee: zerlege Streuung der y_i-Werte um ihren Mittelwert in die durch Geradengleichung und die durch Fehlerterm

  man definiert die "exakten" Modellwerte

  $$\hat{y}_i = a + bx_i$$

  - und damit

  $$\begin{aligned} \mathrm{SQT} &:= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\ \mathrm{SQE} &:= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \\ \mathrm{SQR} &:= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned}$$

  Bedeutung
  - SQT = Summe der Quadrate Total = gesamte Streuung der y_i
  - SQE = Summe der Quadrate Erklärt = Streuung wegen Geradengleichung
  - SQR = Summe der Quadrate Rest = Streuung wegen Fehler ε_i

  damit gilt (nachrechnen!)

  SQT = SQE + SQR

  Definition des Bestimmtheitsmaßes R^2

  $$R^2 = \frac{\mathrm{SQE}}{\mathrm{SQT}}$$

  Eigenschaften von R^2
  - $0 \le R^2 \le 1$
  - $R^2 = 1 \rightarrow$ SQR = 0 $\rightarrow$ y_i liegen genau auf einer Geraden
  - $R^2 = 0 \rightarrow$ SQE = 0 $\rightarrow$ alle $\hat{y}_i = \overline{y}$ $\rightarrow$ Gerade ist horizontal
    $\rightarrow$ Streuung beruht komplett auf Fehlerterm

  Interpretation
  - R^2 nahe 1 $\Leftrightarrow$ Modell passt gut zu den Daten

  im Beispiel
  - SQT = 7.4157e+03
  - SQE = 7.3636e+03
  - SQR = 52.1764
  - R^2 = 0.9930

- Residualanalyse:

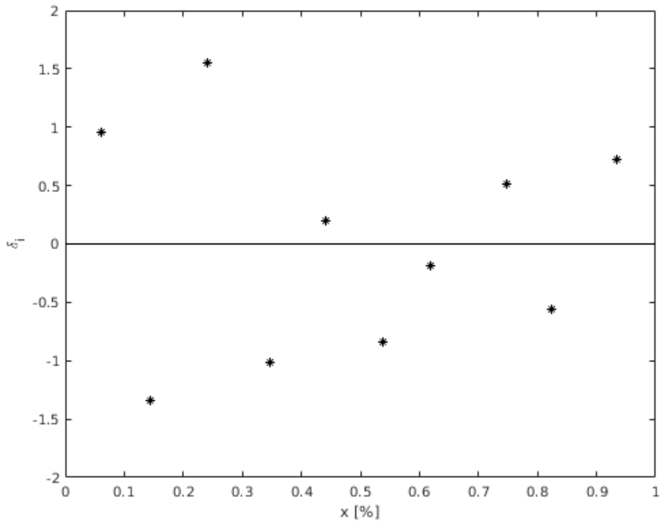
  wie gut ist das Fehlermodell?

  berechne standardisierte Residuen

  $$\delta_i := \frac{y_i - a - bx_i}{s_{y|x}}$$

  plotte Streudiagramm von δ_i über x_i

- im Beispiel alles ok

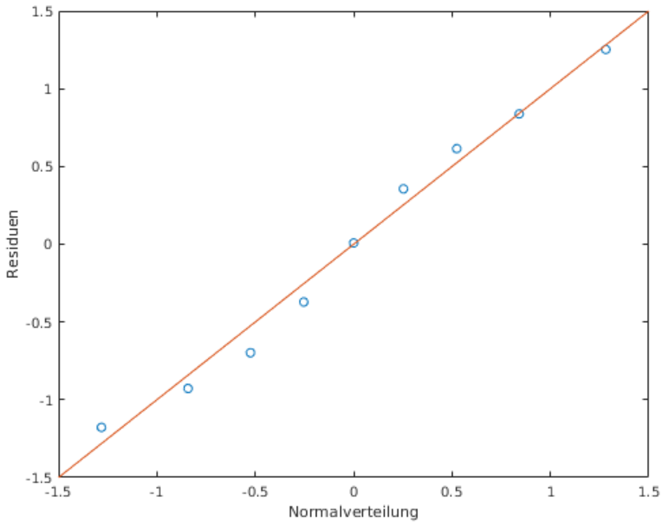


typische Probleme

- Punkte folgen grob einer Kurve → y hängt nichtlinear von x ab
- Abstand der Punkte von der x-Achse wächst mit x → ε_i haben mit x wachsendes σ

ε_i-Modell ok → $\delta_i \sim N(0,1)$

- überprüfen mit Q-Q-Plot
- im Beispiel



- Gewichtete Regression:

  Erweiterung auf Fehlerterme $\varepsilon_i \sim N(0, \sigma_i^2)$

  - Annahme: σ_i sind (näherungsweise) bekannt

  z. B. Messreihen mit mehreren Messungen pro x_i

  - → y_i wird als Mittelwert bestimmt
  - → σ_i wird aus Messreihe zu x_i abgeschätzt
  - oder man kennt die Genauigkeit des Messgeräts als Funktion von x

  Dichtefunktion der Y_i einfache Erweiterung von oben

$$f_{\alpha\beta\sigma}(y_1, \ldots, y_n) = \frac{1}{(2\pi)^{n/2} \prod_i \sigma_i} e^{-\frac{1}{2}\sum_{i=1}^n \frac{1}{\sigma_i^2}(y_i - \alpha - \beta x_i)^2}$$

  Maximum-Likelihood-Verfahren liefert lineares Gleichungssystem

$$\begin{pmatrix} \sum_{i=1}^n \frac{1}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{x_i}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \frac{y_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2} \end{pmatrix}$$

- Lösung ergibt Schätzer

Ergebnis

- Messwerte mit niedriger Genauigkeit werden schwächer berücksichtigt

weitere Anwendung

- σ_i manuell als Gewichte einführen, um bestimmte x-Bereiche zu unterdrücken oder hervorzuheben

- Aufgaben:

  Aufgabe 43

  Aufgabe 44

  Aufgabe 45

Mehrfache lineare Regression

- Problemstellung:

  Größe y hängt (vermutlich) von mehreren Einflussgrößen x_i, i = 1 ... k, ab

  Zusammenhang wird als linear angenommen

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_k x_k$

  man macht n Messungen, um die p = k+1 unbekannten Koeffizienten abzuschätzen (n > p) und erhält Werte y_i, x_{ij} (i = 1 ... n, j = 1 .. p-1)

  statistisches Modell

  $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_k x_k + \varepsilon_i$

  - mit unabhängigen normalverteilten Störungen $\varepsilon_i \sim N(0, \sigma^2)$

  Einführen von Vektoren und Matrizen

  $$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

  $$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

  - liefert Gleichung des Modells

  $$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Schätzer für die β_i und σ^2:

  Y_i sind i.i.d. mit $Y_i \sim N(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k, \sigma^2)$

  berechnen Maximum-Likelihood-Schätzer **B** = $(B_0, B_1, \dots, B_k)^T$ für **β** ähnlich wie oben

  längere Rechnung liefert die **Normalengleichung**

  $$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B} = \boldsymbol{X}^T\mathbf{Y}$$

  lineares Gleichungssytem mit p Gleichungen für p Unbekannte

  - $\mathbf{X}^T\mathbf{X}$ in der Regel invertierbar (sonst schlechte Messreihe!)
  - formale Lösung

  $$\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{Y}$$

  numerisch lösen

  - mit LU-Zerlegung (in Matlab: `B = (X'*X)\(X'*Y)`)
  - *nicht* mit Berechnung der Inversen (in Matlab: `B = inv(X'*X)*(X'*Y)`)

  Schätzer für σ^2 ähnlich wie oben

  $$S^2_{y|x} = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik})^2$$

- Verteilung des Schätzers B:

  Linearkombinationen normalverteilter Größen Y_i → (multivariat) normalverteilt

  Erwartungswert

$$\begin{aligned}E(\boldsymbol{B}) &= E((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y})\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y})\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\\ &= \boldsymbol{\beta}\end{aligned}$$

- → B ist erwartungstreu

Kovarianzmatrix

$$\mathrm{Cov}(\boldsymbol{B}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

- Beweis im Anhang

- Verteilung des Schätzers $S^2_{y|x}$:

  ähnlich wie oben ist

  $$\frac{n-k-1}{\sigma^2}S^2_{y|x} \sim \chi^2_{n-k-1}$$

  daher hat man

  $$\begin{aligned}E(S^2_{y|x}) &= \sigma^2\\ \mathrm{Var}(S^2_{y|x}) &= \frac{2\sigma^4}{n-k-1}\end{aligned}$$

  außerdem gilt folgende nützliche Beziehung zur Berechnung von $S^2_{y|x}$

  $$S^2_{y|x} = \frac{1}{n-k-1}(\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y})$$

  - Beweis im Anhang

- Prognose des Mittelwerts:

  Aufgabe: schätze aus den bekannten Daten **X**, **Y** den Mittelwert von Y_0 zu den Eingabewerten $\mathbf{x}_0 = (1\ x_{01}\ x_{02}\ \ldots\ x_{0k})^T$

  es ist

  $$\overline{Y}_0 = \boldsymbol{x}_0^T\boldsymbol{\beta}$$

  Schätzer natürlich

  $$\widehat{Y}_0 = \boldsymbol{x}_0^T\boldsymbol{B}$$

  offensichtlich erwartungstreu

  $$E(\widehat{Y}_0) = \boldsymbol{x}_0^T E(\boldsymbol{B}) = \boldsymbol{x}_0^T\boldsymbol{\beta} = \overline{Y}_0$$

  interessant sind Streuung und Konfidenzintervalle

  als Linearkombination der Y_i ist $\widehat{Y}_0$ normalverteilt

  Berechnung der Varianz

  $$\begin{aligned}\mathrm{Var}(\widehat{Y}_0) &= \mathrm{Var}(\sum_{i=1}^{n} x_{0i}B_i)\\ &= \mathrm{Cov}(\sum_{i=1}^{n} x_{0i}B_i, \sum_{j=1}^{n} x_{0j}B_j)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_{0i}x_{0j}\,\mathrm{Cov}(B_i, B_j)\\ &= \boldsymbol{x}_0^T\,\mathrm{Cov}(\boldsymbol{B})\boldsymbol{x}_0\\ &= \sigma^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0\end{aligned}$$

insbesondere gilt

$$\frac{\widehat{Y}_0 - \overline{Y}_0}{\sigma\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim N(0,1)$$

ersetzt man σ^2 durch seinen Schätzer $S^2_{y|x}$, hat man wie immer

$$\frac{\widehat{Y}_0 - \overline{Y}_0}{S_{y|x}\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t_{n-k-1}$$

daher erhält man Schätzer für ein (1-p)-Konfindenzintervall um $\overline{Y}_0$

$$G_{o/u} = \widehat{Y}_0 \pm t_{n-k-1,1-p/2}\, S_{y|x}\, \sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

- Prognose eines Werts:

  schätze den Wert Y($\mathbf{x}_0$) eines Experiments mit Eingabewerten $\mathbf{x}_0$, nicht den Mittelwert

  Schätzer bleibt gleich, aber Varianz ist größer (wegen ε_i)

  B_i stammen aus bisherigen Daten, Y($\mathbf{x}_0$) bezieht sich auf neue → B_i und Y($\mathbf{x}_0$) sind unabhängig

  daher

$$\begin{aligned} \mathrm{Var}(Y(\boldsymbol{x}_0) - \boldsymbol{x}_0^T\boldsymbol{B}) &= \mathrm{Var}(Y(\boldsymbol{x}_0)) + \mathrm{Var}(\widehat{Y}_0) \\ &= \sigma^2 + \sigma^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0 \\ &= \sigma^2(1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0) \end{aligned}$$

  Schätzer für ein (1-p)-Konfindenzintervall um Y($\mathbf{x}_0$)

$$G_{o/u} = \widehat{Y}_0 \pm t_{n-k-1,1-p/2}\, S_{y|x}\, \sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

- Beispiel Härte von Stahl (nach [Ross2]):

  Stahlhersteller möchte kaltgewalzte Stahlbleche mit Kupfergehalt qCu_0 = 0.15 % und Glühtemperatur T_0 = 620 °C herstellen

  gesucht ist Abschätzung der Härte hr_0 (genauer: Rockwellhärte HR30T)

  folgende Daten sind bekannt

T [°C]	qCu [%]	hr [HR30T]
540	0.15	84.30
565	0.02	79.20
590	0.15	70.40
650	0.03	64.00
650	0.09	61.30
675	0.03	55.70
705	0.04	56.30
705	0.10	58.60
760	0.09	49.80
760	0.13	51.30

  Ansatz: multilineares Modell

  hr = $\beta_0 + \beta_1$ T + β_2 qCu

  multilineares Regression liefert folgende Schätzwerte

  - B = [157.4; 16.60; -0.1450]
  - $s_{y|x}$ = 3.223

  damit Schätzwert für gesuchte Härte

- $hr_0 = 70.00$

95%-Konfidenzintervall für den Mittelwert

- $E(hr_0) = 70.00 \pm 4.2717$

95%-Konfidenzintervall für den Einzelwert

- $hr_0 = 70.00 \pm 8.7359$

- Aufgaben:

  Aufgabe 46

  Aufgabe 47