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

Aufgabe 1

- Das Gen für blaue Augen ist rezessiv, d. h. beide Gene einer Person müssen "blau" sein, damit sie blaue Augen hat. Andrea und Bernd haben beide braune Augen, obwohl ihre Väter beide blauäugig waren. Ihr Sohn Christian hat ebenfalls braune Augen und heiratet die blauäugige Dana. Mit welcher Wahrscheinlichkeit ist ihre Tochter Eva blauäugig?
- Annahmen:
  - es gibt nur die beiden Gen-Varianten für braun und blau
  - beide Gen-Varianten sind gleich häufig
  - man vererbt mit gleicher Wahrscheinlichkeit eins der beiden Gene an ein Kind

Aufgabe 2

- Wie groß ist die Wahrscheinlichkeit, beim Kniffel direkt eine kleine Straße (also mit 5 Würfeln genau vier aufeinanderfolgende Zahlen) zu würfeln?

Aufgabe 3

- Drei verschiedene Zulieferer A, B, C liefern eine Komponente, wobei A 40% der Gesamtmenge liefert, B 35% und C 25%. Die Fehlerrate bei A betrage 2%, bei B 3% und bei C 1%. Wie groß ist die Wahrscheinlichkeit, dass eine beliebig herausgegriffene Komponente defekt ist?

Aufgabe 4

- Zur Fertigung eines Werkstücks werden zwei Maschinen A und B eingesetzt:
  - Maschine A ist größer, sie erzeugt 80% der Teile und hat eine Fehlerquote von 0.2%,
  - Maschine B dient zur Ergänzung bei Anforderungsspitzen, sie erzeugt im Schnitt nur 20% der Teile und hat eine Fehlerquote von 0.4%.
- Am Ende wird ein defektes Teil entdeckt. Mit welcher Wahrscheinlichkeit stammt es von Maschine A?

Aufgabe 5

- Das folgende Problem hat unter dem Namen **Ziegenproblem** Furore gemacht:

  Nehmen Sie an, Sie wären in einer Spielshow und hätten die Wahl zwischen drei Toren. Hinter einem der Tore ist ein Auto, hinter den anderen sind Ziegen. Sie wählen ein Tor, sagen wir, Tor Nummer 1, und der Showmaster, der weiß, was hinter den Toren ist, öffnet ein anderes Tor, sagen wir, Nummer 3, hinter dem eine Ziege steht. Er fragt Sie nun: "Möchten Sie das Tor Nummer 2?" Ist es von Vorteil, die Wahl des Tores zu ändern? [1]

- Dabei sind folgende Voraussetzungen gemacht:

  Die Wahrscheinlichkeit, das Auto hinter einer bestimmten Tür zu finden, ist für alle Türen gleich.

  Der Showmaster muss eine andere Tür wählen, hinter der eine Ziege steht.

  Hat er dabei die Wahl zwischen zwei Möglichkeiten, entscheidet er sich mit gleicher Wahrscheinlichkeit für eine von beiden.

- Lösen Sie das Problem auf zwei verschiedene Weisen:

  a. Formulieren Sie geeignet Ereignisse und bedingte Wahrscheinlichkeiten und berechnen Sie daraus die Sieg-Wahrscheinlichkeiten ohne bzw. mit Wechsel.

  b. Schreiben Sie ein Matlab-Programm, das das Experiment häufig durchführt, und bestimmen Sie die Zahl der Gewinne in beiden Strategien. Die Matlab-Funktion `rand` könnte hier hilfreich sein.

Aufgabe 6

- Frank und Greta haben braune Augen, aber jeweils selbst ein blauäugiges Elternteil. Wie groß ist die Wahrscheinlichkeit, dass genau drei ihrer vier Kinder braune Augen haben (vgl. Aufgabe 1)?
- Tipp: Die Matlab-Funktion `nchoosek` könnte nützlich sein.

Aufgabe 7

- Ein Multiple-Choice-Test bestehe aus 20 Fragen mit jeweils 4 Möglichkeiten, von denen immer genau eine richtig ist. Zum Bestehen braucht man mindestens 10 richtige Antworten, eine 2 bekommt man ab 15, eine 1 ab 18 richtigen Antworten.
  a. Student A hat überhaupt keine Ahnung, so dass er die richtige Antwort immer nur rein zufällig - sprich: mit Wahrscheinlichkeit p = 1/4 - erwischt. Mit welcher Wahrscheinlichkeit besteht er die Klausur?
  b. Student B hat eifrig gelernt, so dass die Wahrscheinlichkeit, jeweils die richtige Antwort zu finden, p = 90% beträgt. Wie groß sind die Wahrscheinlichkeit, dass er durchfällt, dass er eine 2 bekommt oder dass er sogar eine 1 bekommt?
  c. Student C betreibt Optimierung: Er lernt so viel, dass die Wahrscheinlichkeit zu bestehen 75 % beträgt. Wie groß muss er die Erfolgswahrscheinlichkeit p einer Frage dafür machen?
- Lösen Sie die Aufgabe sowohl mit Matlab-Standardfunktionen als auch mit Hilfe der Funktionen `makedist` und `cdf` aus der Statistik-Toolbox (STB).

Aufgabe 8

- Eine Maschine produziere an einem Tag N = 500 Bauteile, von denen R = 50 defekt sind. Die Qualitätskontrolle entnehme n = 20 Stück. X beschreibe die Zahl der defekten Teile in der Stichprobe.
  a. Wie groß sind die Wahrscheinlichkeiten, dass
     - kein Teil defekt ist,
     - genau 2 Teile defekt sind,
     - mehr als 3 Teile defekt sind?
  b. Vernachlässigt man die Verringerung der Menge durch das Entnehmen der Stichprobe, vereinfacht sich die Verteilung von X. Bestimmen Sie damit die Wahrscheinlichkeiten aus a. näherungsweise.
  c. Plotten Sie die Wahrscheinlichkeitsfunktion und die kumulative Verteilungsfunktion von X in den Fällen a und b.
- Lösen Sie die Aufgabe sowohl mit Matlab-Standardfunktionen als auch mit Hilfe der Funktionen aus der STB. Sie enthält leider kein Objekt für die hypergeometrische Verteilung, aber Hilfsfunktionen `hygepdf` und `hygecdf`. Zum Plotten sind `stem` und `stairs` nützlich.

Aufgabe 9

- Um den Übergang von der Binomial- zur Poissonverteilung zu studieren, werden drei Verteilungen betrachtet:

  $X_1 \sim B(10, 0.5)$

  $X_2 \sim B(50, 0.1)$

  $X_3 \sim Po(5)$

- Lösen Sie folgende Aufgaben jeweils für alle drei Verteilungen:
  a. Plotten Sie die Wahrscheinlichkeitsfunktion für die Werte k = 0 ... 15.
  b. Berechnen Sie die Wahrscheinlichkeiten $P(X_i = 5)$, $P(X_i < 3)$ und $P(X_i \geq 10)$.
  c. Bestimmen Sie den Erwartungswert, die Varianz und die p-Quantile für p = 0.1, 0.5 und 0.9.

- Neben den Funktionen `pdf` und `cdf` sind auch `mean`, `var` und `icdf` nützlich.

Aufgabe 10

- Bei einer Fertigung von Kolbenstangen komme es zu Abweichungen von der Soll-Länge, die maximal 3 mm betragen. Werte um 0 herum sind dabei wahrscheinlicher als Maximalwerte.
- X beschreibe die Abweichung der Länge vom Sollwert in mm. Ihre Wahrscheinlichkeitsdichte soll durch eine Kosinusfunktion angenähert werden, also

$$f(x) = \begin{cases} A\cos(\frac{\pi}{6}x) & | \quad -3 < x < 3 \\ 0 & | \quad \text{sonst} \end{cases}$$

- Wie groß ist die Wahrscheinlichkeit, dass bei fünf aus der Produktion entnommenen Stangen alle eine Abweichung von weniger als 2 mm vom Sollwert haben?

Aufgabe 11

- An einer U-Bahn-Station komme alle 5 Minuten eine Bahn an. Herr K. hat flexible Arbeitszeit, verlässt sein Büro täglich, ohne auf die Uhr zu sehen, und nimmt die nächste U-Bahn.
  a. Die Zufallsvariable X beschreibe die Zeit in Minuten, die Herr K. am Gleis auf die Bahn wartet. Welcher Verteilung genügt sie?
  b. Wie lange wartet Herr K. im Mittel, wie groß ist die Standardabweichung von X?
  c. Herr K. hat den Eindruck, dass er oft länger als 4 Minuten warten muss. Wie groß ist die Wahrscheinlichkeit, dass ihm das mindestens dreimal in der (Arbeits-)Woche passiert?

Aufgabe 12

- Die Bearbeitungszeit X in einem zweistufigen Verarbeitungsprozess sei beschrieben durch eine Erlang-Verteilung mit $n = 2$ und $\lambda = 0.3/\text{min}$.
  a. Plotten Sie die Wahrscheinlichkeitsdichte und Verteilungsfunktion von X.
  b. Wie groß sind Erwartungswert und Standardabweichung der Bearbeitungszeit?
  c. Wie groß ist die Wahrscheinlichkeit, dass ein Bearbeitungsprozess länger als 10 Minuten dauert?

Aufgabe 13

- Bei elektrischen Widerständen der Normreihe E12 darf die Größe des Widerstandes um ±10% schwanken. In einem Herstellungsprozess sollen Widerstände der Größen 47 kΩ, 56 kΩ und 68 kΩ hergestellt werden. Die erzielten Werte sind normalverteilt, wobei die Erwartungswerte gleich den gewünschten Werten sind und die Standardabweichungen jeweils 3 kΩ betragen.

  a. Wieviel Prozent der hergestellten Widerstände liegen jeweils innerhalb des erlaubten Toleranzbereichs? Rechnen Sie mit und ohne Verwendung der Statistik-Toolbox.

  b. Wieviel Prozent der Widerstände kann man insgesamt gebrauchen, falls alle drei Prozesse gleich viele Widerstände erzeugen und man keine weiteren - kleineren oder größeren - Widerstände produzieren möchte?

Aufgabe 14

- Eine Zufallsvariable X heißt logarithmisch normalverteilt (mit Parametern μ und σ^2), kurz: $X \sim LN(\mu, \sigma^2)$, wenn ihr Logarithmus normalverteilt ist (mit Parametern μ und σ^2), also wenn

  $X = \exp(Y)$ mit $Y \sim N(\mu, \sigma^2)$

  Sie findet Anwendung z. B. zur Beschreibung von Aktienkursen im Black-Scholes-Modell.

  a. Berechnen Sie die Wahrscheinlichkeitsdichte von X

  b(*) Beweisen Sie folgende Formeln für Mittelwert und Varianz von X:

$$\begin{aligned} E(X) &= e^{\mu+\sigma^2/2} \\ \mathrm{Var}(X) &= e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right) \end{aligned}$$

- Der Kurs X einer Aktie zu verschiedenen Tagen gehorche einer logarithmischen Normalverteilung mit E(X) = 10 € und σ(X) = 2 €.

  c. Plotten Sie die Wahrscheinlichkeitsdichte und Verteilungsfunktion von X.

  d. Wie groß ist der Median von X?

  e. Wie groß ist die Wahrscheinlichkeit, dass der Wert der Aktie auf über 12 € steigt?

- Achtung: Dieses Modell ist voller (unrealistischer) Annahmen - damit sollte man nicht an der Börse spekulieren!

Aufgabe 15

- Eine stetige Zufallsvariable **X** = (X, Y) habe die gemeinsame Dichtefunktion

$$f(x,y) = \begin{cases} A(x+y)e^{-(x+y)} & \mid \quad x, y \geq 0 \\ 0 & \mid \quad \text{sonst} \end{cases}$$

a. Bestimmen Sie den Wert der Konstante A.

b. Berechnen Sie die Randverteilungen f_X, f_Y sowie deren Erwartungswerte und Varianzen.

c. Ermitteln Sie die Kovarianz und den Korrelationskoeffizienten von X und Y.

Aufgabe 16

- Wie oft darf man den Schmelzofen aus dem Beispiel maximal befüllen, damit die Kapazität mit einer Wahrscheinlichkeit von 99.9 % nicht überschritten wird? Wie groß ist dann die mittlere Ausnutzung?

Aufgabe 17

- Um einen Widerstand von R = 60 kΩ zu erreichen, werden mehrere Widerstände in Reihe geschaltet, nämlich folgende Nennwerte (alle Werte in kΩ):

  1. 27 + 33
  2. 15 + 18 + 27
  3. 10 + 10 + 18 + 22

  Die Werte eines Widerstand können um 10 % um den Nennwert R_N schwanken, sie seien jeweils im Intervall [0.9 R_N, 1.1 R_N] gleichverteilt.

  a. Berechnen Sie für alle drei Fälle die Standardabweichungen der einzelnen Widerstände und des Gesamtwiderstands.

  b. (*) Berechnen Sie für den ersten Fall die Dichtefunktion des Gesamtwiderstands. Plotten Sie diese und bestimmen Sie die Wahrscheinlichkeit, dass der Wert im 5%-Intervall [57, 63] liegt.

Aufgabe 18

- Die Lebensdauer einer Batteriesorte sei exponentiell verteilt mit einem Mittelwert von 10 Stunden. Ein Satz von n Batterien wird der Reihe nach verwendet, um eine bestimmte Gesamt-Laufzeit S_n zu erreichen.

  a. Plotten Sie die Dichtefunktion der standardisierten Verteilung von S_n für n = 3, 7, 50, zusammen mit der Standard-Normalverteilung.

  b. Berechnen Sie die Wahrscheinlichkeit $P(S_n \leq 10\, n)$ direkt bzw. über den zentralen Grenzwertsatz für die Werte von n aus a.

Aufgabe 19

a. Bei Twisted-GFSR-Generatoren wird für die Matrix A meistens folgende Form gewählt:

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & a_1 \\ 1 & 0 & 0 & \cdots & 0 & a_2 \\ 0 & 1 & 0 & \cdots & 0 & a_3 \\ 0 & 0 & 1 & \cdots & 0 & a_4 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{l-1} \\ 0 & 0 & 0 & \cdots & 1 & a_l \end{pmatrix}$$

A enthält also Einsen in der unteren Nebendiagonalen und einen Vektor a als letzte Spalte. Wie kann man dann die Berechnung von A · x in Matlab durchführen, wenn man x als einfache `uint`-Variable geeigneter Größe wählt?

b. Erzeugen Sie Zufallszahlen mit einem TGFSR-Generator mit folgenden Parametern

   i.) q = 5, r = 3, l = 4, a = 0xb

   ii.) q = 25, r = 14, l = 16, a = 0xa875

   Dabei bedeutet 0xNNNN (wie in der Informatik üblich), dass NNNN die Hexadezimaldarstellung einer Zahl ist.

c. Testen Sie die in ii. erhaltene Folge auf Gleichverteilung.

Aufgabe 20

- Berechnen Sie zwei Sätze normalverteilter Zufallszahlen mit dem Box-Muller-Verfahren bzw. mit der `randn`-Funktion. Vergleichen Sie die Ergebnisse, indem Sie beide als Histogramme darstellen.

Aufgabe 21 (*)

- Erstellen Sie mit Matlab ein Modell für eine Kassenschlange, bei der Kunden in zufälligen Abständen DT ankommen und die Abfertigungsdauer TK der einzelnen Kunden zufällig schwankt. Dabei sei DT exponentiell verteilt und TK (a) exponentiell bzw. (b) normalverteilt mit folgenden Parametern:

  $DT \sim Ex(\lambda)$, $\lambda = 0.9/\text{min}$

  $TK_a \sim Ex(\mu)$, $\mu = 1.0$ min

  $TK_b \sim N(\mu, \sigma^2)$, $\mu = 1.0$ min , $\sigma = 0.2$ min

- Simulieren Sie das Modell für N = 1000 Kunden und berechnen Sie die mittlere Wartezeit der Kunden und die mittlere Queuelänge. Erstellen Sie Plots der Queuelänge und der Wartezeiten.

- Testen Sie Ihr Modell, indem Sie alle Events für N = 10 ausgeben und per Hand mit den Ergebnisplots und Mittelwerten vergleichen.

Aufgabe 22

- Angenommen, die Frage, ob es heute regnet, hinge nur von den Wetterbedingungen von gestern und vorgestern ab, nämlich:
  - Wenn es zwei Tage hintereinander geregnet hat, regnet es am nächsten Tag mit einer Wahrscheinlichkeit $p = 0.7$,
  - wenn es gestern geregnet hat, vorgestern aber trocken blieb, regnet es heute mit $p = 0.5$,
  - wenn es gestern nicht geregnet hat, aber vorgestern, regnet es heute mit $p = 0.4$,
  - wenn es zwei Tage hintereinander trocken blieb, regnet es am nächsten Tag mit $p = 0.2$.
- Dies ist kein Markov-Prozess, kann aber mit einem Trick leicht zu einem gemacht werden:
  - Man ordnet jedem Tag einen Zustand (X,Y) zu, wobei
    - X = T bzw. X = R, wenn es am Vortag trocken war bzw. geregnet hat,
    - Y = T bzw. Y = R, wenn es am Tag selbst trocken war bzw. geregnet hat.
  - Damit hängt die Regen-Wahrscheinlichkeit an einem Tag nur vom Zustand des Vortags ab.
- Berechnen Sie die Übergangs-Matrix des entsprechenden Markov-Prozesses und zeichnen Sie den Übergangs-Graphen. Ist der Prozess reduzibel oder irreduzibel? Welche Zustände sind transient, welche rekurrent?
- Berechnen Sie die Gleichgewichtsverteilung auf drei Weisen:
  - per Hand
  - durch Lösen des Gleichungssystems mit Matlab
  - durch Bestimmen des Langzeitverhaltens
- An wievielen Tagen im Jahr regnet es durchschnittlich?

Aufgabe 23

- Berechnen Sie mit Matlab symmetrische Zufallspfade in 1, 2 und 3 Dimensionen, die jeweils am Ursprung beginnen. Erreicht man den Ursprung immer wieder?

Aufgabe 24

- Ein Webserver werde als einfaches M|M|1-System modelliert, wobei im Schnitt 800 Anfragen pro Sekunde ankommen, deren Beantwortung durchschnittlich 1.1 ms dauert.
  a. Wie lange muss ein Anwender im Schnitt auf eine Antwort warten? Wie lang ist die Warteschlange im Schnitt?
  b. Wie ändern sich die Werte, wenn die Zahl der Anfragen auf 900/s steigt?
  c. Wie viele Warteplätze für die Queue muss man vorsehen, damit eine Anfrage mit 90% bzw. 99% Wahrscheinlichkeit nicht abgewiesen wird (jeweils bei 800 oder 900 Anfragen pro Sekunde)?

Aufgabe 25

- Untersuchen Sie ein Bediensystem, bei dem die Zahl der Warteplätze beschränkt ist, also ein M|M|1|c-System mit gegebenem c. Im einzelnen:
  a. Erstellen Sie den Übergangs-Graphen.
  b. Berechnen Sie die GGV.
  c. Bestimmen Sie Beziehungen für die Auslastung und die Kennzahlen L, L_Q.
  d. Wie groß ist die tatsächliche Ankunftsrate λ_c?
  e. Geben Sie schließlich noch Formeln an für W und W_Q.
- Vergleichen Sie die numerischen Werte für c = 3, c = 10 und c = ∞ bei λ = 10/s, μ = 12/s.

Aufgabe 26

- Die Tabelle pkwzulassungen.xlsx enthält die Zahl der Neuzulassungen der einzelnen PKW-Marken für Juni 2017 und Januar bis Juni 2017 sowie jeweils Anzahlen für Fahrzeuge mit Diesel-, Hybrid- und Elektroantrieb.
    a. Stellen Sie die Anteile an den Neuzulassungen im Juni für die größten Marken (Anteil ≥ 2%) als Tortendiagramm dar und zeigen Sie die Änderungen der Anteile gegenüber den Werten von Januar bis Juni als Balkendiagramm.
    b. Erstellen Sie ein Tortendiagramm, das die Verteilung der Antriebsarten an den Neuzulassungen im Juni zeigt, und geben Sie die prozentualen Veränderungen der Anteile gegenüber den Halbjahreswerten als Balkendiagramm an.

Aufgabe 27

- Die Datei solaranlage.csv enthält für jeden Tag der Jahre 2009 - 2016 die von einer privaten Solaranlage erbrachte Energie.

  a. Lesen Sie die Daten ein, wobei Sie Matlabs `datetime`-Format für die Zeitangaben verwenden, und verschaffen Sie sich einen ersten Überblick, indem Sie die gesamte Zeitreihe plotten.

  b. Vergleichen Sie nun die Ergebnisse der verschiedenen Jahre. Plotten Sie dazu Mittelwerte und Standardabweichung für die einzelnen Jahre und einen Boxplot über die Jahre.

  c. Gewinnen Sie nun einen Überblick über die jahreszeitlichen Schwankungen, indem Sie Plots wie in b. für die über alle Jahre monatlich zusammengefassten Daten erstellen.

  d. Die Werte für Januar und Juni unterscheiden sich natürlich stark. Versuchen Sie, anhand der Ergebnisse aus c., eine Vorstellung der Verteilungen in diesen beiden Monaten zu gewinnen. Erstellen Sie entsprechende Histogramme, um diese zu überprüfen.

Aufgabe 28

- Die Tabelle arbeitslose.xlsx enthält die Arbeitslosenquoten in Deutschland im Jahresdurchschnitt von 1948 - 2016, die Tabelle wahlergebnisse.xlsx die Ergebnisse der Bundestagswahlen.
- Stellen Sie für einen ersten Überblick die Wahlergebnisse der CDU/CSU, der SPD, der "Linken" (Linke, KPD/DKP zusammen) und der "Rechten" (NPD, REP und AfD zusammen) gemeinsam mit der Arbeitslosenquote über die Jahre graphisch dar.
- Um den Einfluss der Arbeitslosigkeit auf die Wahlergebnisse zu untersuchen, berechnen Sie die Korrelationskoeffizienten jeweils der vier Parteiengruppen mit

  a. der Arbeitslosenquote im Wahljahr,

  b. dem Mittelwert der Arbeitslosenquote über die vier letzten Jahre,

  c. der Änderung der Arbeitslosenquote gegenüber dem Vorjahr

  und stellen die Werte graphisch dar.

Aufgabe 29

- Erstellen Sie Q-Q-Plots zum Vergleich der beiden Lebensdauer-Datensätze (lebensdauer.dat) jeweils mit der
    a. Exponential-Verteilung,
    b. Gamma-Verteilung,
    c. Weibull-Verteilung [Ross, Beucher].
- Berechnen Sie dazu zunächst Mittelwert und Varianz der Daten und bestimmen daraus die Parameter der jeweiligen Verteilung.
- Welche Verteilung erscheint Ihnen am plausibelsten?

Aufgabe 30

- Bestimmen Sie für das Beispiel 1 (defekte Werkstücke) den Maximum-Likelihood-Schätzer für die Ausfall-Wahrscheinlichkeit p. Ist er erwartungstreu? Ist er konsistent?

Aufgabe 31

- Bei der Fertigung von Akkus kommt es aufgrund von Fertigungstoleranzen zu schwankenden Kapazitäten, für die eine Normalverteilung angenommen wird. Die Datei akkus100.dat enthält Messwerte für die Kapazitäten einer Stichprobe von n = 100 Stück (in mAh).

  a. Bestimmen Sie 99%-Konfidenzintervalle für den Mittelwert und die Standardabweichung der Kapazität.

  b. Dass die Standardabweichung kleiner ist als verlangt, ist unproblematisch. Deswegen interessiert man sich hier für ein einseitiges Konfidenzintervall

     $$P(\sigma^2 \leq g_o) = 1 - \alpha$$

     Bestimmen Sie einen entsprechenden Schätzer G_o und ermitteln Sie die obere Grenze für die Standardabweichung der Akku-Kapazität mit 99% Konfidenz.

Aufgabe 32

- Ein Marktforschungsinstitut möchte herausfinden, wieviel Prozent der Haushalte eine bestimmte Werbung gesehen haben (und sich daran erinnern können). Dazu ruft man zufällig ausgewählte Haushalte an.

  a. Bei 140 Antworten erhielt man 39 positive (die sich an die Werbung erinnern konnten). Geben Sie ein 95%-Konfidenzintervall für den gesuchten Anteil an.

  b. Wie viele Antworten braucht man, damit man 95% sicher sein kann, dass die Schätzung im Bereich von ± 0.01 genau ist?

Aufgabe 33

- In dieser Aufgabe soll an einem Beispiel untersucht werden, wie gut die Berechnung von Konfidenzintervallen bei nicht normalverteilten Größen funktioniert. Dazu werde eine Zufallsgröße X ~ Ex(λ) simuliert mit λ = 2. Für sie gilt

$$\begin{aligned} E(X) &= \frac{1}{\lambda} = 0.5 \\ \sigma(X) &= \frac{1}{\lambda} = 0.5 \end{aligned}$$

a. Berechnen Sie mit Matlab Ex(2)-verteilte Zufallszahlen und erzeugen Sie daraus N = 1000 Stichproben mit je n = 100 Werten. Berechnen Sie für jede Stichprobe die Konfidenzintervalle für den Mittelwert bei einem Konfidenzlevel 1-α = 0.95. Wie viele der Intervalle umfassen den wahren Wert nicht? Entspricht das Ihrer Erwartung?

b. Wiederholen Sie a. für kleinere Stichproben n = 30 und n = 10. Deuten Sie das Ergebnis.

c. Was passiert, wenn man den für normalverteilte Größen hergeleiteten Schätzer für das Konfidenzintervall der Standardabweichung hier verwendet?

Aufgabe 34

- Ein Gerät zur Messung des pH-Werts einer Lösung hat eine vom Hersteller angegebene Standardabweichung von $\sigma = 0.02$. Für eine Synthese-Reaktion wird ein pH-Wert von 8.95 benötigt. In zwölf unabhängigen Messungen erhält man folgende Werte:

  8.928 8.944 8.894 8.884 8.964 8.947 8.949 8.948 8.875 8.935 8.994 8.896

a. Testen Sie den Mittelwert (H_0: $\mu = 8.95$) unter der Annahme, die Herstellerangabe sei richtig. Was schließen Sie bei einem Signifikanzniveau $\alpha = 0.05$? Wie groß ist der p-Wert des Ergebnisses?

b. Wiederholen Sie a., aber ohne der Angabe des Herstellers zu trauen.

c. Stimmt die Herstellerangabe für die Genauigkeit des Messgeräts? Es interessiert natürlich nur, ob das Gerät mindestens so genau ist wie angegeben.

d. Führen Sie alle Tests noch einmal durch, indem Sie die fertigen Test-Funktionen von Matlab/STB benutzen.

Aufgabe 35

- Berechnen Sie die Güte-Funktion $G(\theta)$ beim einseitigen Gauß-z-Test mit H_0: $\mu \leq \mu_0$. Erstellen Sie Plots für folgende Parameterwerte:
    a. $\mu_0 = 0$, $\sigma = 1$, $\alpha = 0.05$, $n = \{10, 30\}$
    b. $\mu_0 = 0$, $\sigma = 1$, $\alpha = \{0.05, 0.01\}$, $n = 10$

Aufgabe 36

- Die Daten aus solaranlage.csv zur Leistung einer Solaranlage sollen darauf untersucht werden, ob sich die Werte im Juli 2012 und im Juli 2013 unterscheiden. Dabei wird von einer Normalverteilung der Werte ausgegangen.

a. Erstellen Sie einen Boxplot für die beiden Messreihen. Was kann man daran erkennen?

b. Führen Sie einen Test durch, ob die Mittelwerte gleich sind. Gehen Sie dabei von gleicher Varianz aus. Was schließen Sie bei Signifikanzniveau $\alpha = 0.05$?

c. Bestimmen Sie Konfidenzintervalle zu $1-\alpha = 95\%$ für die Standardabweichungen σ_x, σ_y.

d. Überprüfen Sie mit einem Test, wieder bei $\alpha = 0.05$, ob die Annahme gleicher Varianzen gerechtfertigt ist.

Aufgabe 37

- Ziel der Aufgabe ist der Vergleich der verschiedenen Binomialtests durch Simulation vieler Experimente. Ausgangspunkt ist ein Bernoulli-Experiment mit $p_0 = 0.2$.
- Erzeugen Sie mit Matlab N = 1000 entsprechende Datensätze mit je n = 60 Stichproben und überprüfen Sie damit die Nullhypothese H_0: $p_0 = 0.2$, indem Sie zum Signifikanzniveau $\alpha = 0.05$

  a. exakte Binomialtests,
  b. genäherte Binomialtests,
  c. genäherte Binomialtests mit Stetigkeitskorrektur

  durchführen. Entsprechen die Ergebnisse Ihren Erwartungen?

Aufgabe 38

- In den Jahren 1966 - 69 erkrankten deutlich mehr sehr junge Frauen an Vaginalkrebs als früher. In [5] wurden diverse mögliche Einflussfaktoren an 8 erkrankten Frauen und einer Kontrollgruppe von 32 nichterkrankten Frauen untersucht. Es stellte sich heraus, dass 7 der Mütter der Erkrankten während der Schwangerschaft den Wirkstoff Diethylstilbestrol (DES) bekommen hatten, aber keine der Kontrollgruppe.
- Eine sinnvolle Nullhypothese lautet: "DES-Einnahme während der Schwangerschaft hat keinen Einfluss auf den Ausbruch von Vaginalkrebs bei der Tochter." Führen Sie mit diesen Daten einen Fisher-Test bzgl. der Nullhypothese durch und geben Sie den p-Wert an.
- Unter den untersuchten Faktoren war auch, ob die Mütter während der Schwangerschaft geraucht hatten. Es waren 7 rauchende Mütter in der Gruppe der Erkrankten, 21 in der Kontrollgruppe (das war Ende der 40er bis Anfang der 50er Jahre). Ist hier ein Zusammenhang zum Vaginalkarzinom der Töchter erkennbar?

Aufgabe 39

- Untersuchen Sie mit Hilfe des Wilcoxon-Rangsummen-Tests, ob die beiden Testreihen in lebensdauer.dat die gleiche Verteilung haben.
  a. Verwenden Sie nur die ersten 7 Werte beider Stichproben. Erstellen Sie zur Berechnung von $wr_{m,n}(k)$ eine rekursive Matlab-Funktion `ranksumcdf`.
  b. Verwenden Sie die komplette Stichprobe und berechnen Sie $wr_{m,n}(k)$ mit Hilfe einer geeigneten Approximation.
- Vergleichen Sie Ihre Ergebnisse mit denen der Testfunktion `ranksum`.

Aufgabe 40

- Untersuchen Sie, ob die Daten in akkus100.dat normalverteilt sind
  a. durch einen Chi-Quadrat-Anpassungstest,
  b. durch einen Kolmogorow-Smirnow-Test.
- Bestimmen Sie jeweils die p-Werte.

Aufgabe 41

- Graphen ist eine zweidimensionale Modifikation des Kohlenstoffs, die wegen ihrer ungewöhnlichen Eigenschaften zahllose Verwendungsmöglichkeit hat. Seine Zugfestigkeit z. B. ist ungefähr 125-mal so groß wie die von Stahl. Allerdings lassen sich einkristalline Proben schwer herstellen.
- In [10] untersuchen die Autoren die Festigkeit von Graphen-Proben verschiedener Körnigkeit, sie unterscheiden zwischen einkristallinem (pristine), grobkörnigem (LG) und feinkörnigem (SG) Graphen.
- Verwenden Sie die Daten aus graphen.dat (die künstlich sind, aber mit den Daten in [10, Fig. 2] und [10, Supp, Fig. S7] konsistent), um den Zusammenhang zwischen Korngröße und Biegesteifigkeit (in N/m) zu untersuchen.

Aufgabe 42

- Untersuchen Sie Gesamtprozentzahlen der Klausuren in klausuren.xslx: Sind sie in den einzelnen Jahren unterschiedlich verteilt? Da Klausurergebnisse in der Regel *nicht* normalverteilt sind, verwenden Sie den Kruskal-Wallis-Test ohne bzw. mit Berücksichtigung von Bindungen.
- Hinweis: Zum Berechnen der Ränge bei Bindungen gibt es die Funktion `tiedrank`.

Aufgabe 43

- Die (nichtlineare) Drehzahl-Drehmomenten-Kennlinie eines Asynchronmotors soll in der Umgebung des Arbeitspunkts linear angenähert werden. Man hat folgende Werte gemessen:

n [1000 U/min]	2.55	2.69	2.86	2.92	2.94	2.99	3.03	3.09	3.16	3.25	3.42
M [Nm]	24.69	22.98	15.90	9.93	6.62	1.83	-3.46	-9.89	-17.39	-22.75	-25.00

Der Arbeitspunkt liegt bei n_0 = 3000 Upm und M_0 = 0.

a. Bestimmen Sie die Regressionsgerade durch die Messwerte und geben Sie den Wert der Steigung an.

b. Um eine bessere Näherung für die Tangente zu erhalten, werden die Messwerte umso weniger gewichtet, je weiter sie vom Arbeitspunkt entfernt liegen. Definieren Sie dazu künstliche Standardabweichungen σ_i der Messwerte gemäß

$$\sigma_i = | n - n_0 |$$

und berechnen Sie die entsprechende Steigung der Regressionsgeraden.

c. Wiederholen Sie b. mit der Gewichtung

$$\sigma_i = (n - n_0)^2$$

- Plotten Sie alle drei Geraden zusammen mit den Daten. Welche würden Sie nehmen?

Aufgabe 44

- Bei der Messung des Sättigungsdampfdrucks p_S von Wasser als Funktion der Temperatur erhielt man folgende Werte:

T [°C]	80	90	100	110	120	130	140	150	160	170	180
p_S [bar]	0.53	0.64	0.86	1.48	2.19	2.57	4.34	5.21	6.15	7.57	9.19

a. lineares Modell
   - Nehmen Sie näherungsweise einen linearen Zusammenhang zwischen T und p_S an und berechnen Sie für die Ausgleichsgerade die Parameter α, β incl. ihrer Standardabweichungen sowie ein 95%-Konfidenzintervall für β.
   - Um zu überprüfen, ob die Annahme eines linearen Zusammenhangs gerechtfertigt ist, berechnen Sie das Bestimmtheitsmaß R^2 und erstellen für die standardisierten Residuen ein Streudiagramm sowie einen Q-Q-Plot.

b. nichtlineares Modell
   - Nach der Formel von Clausius und Clapeyron ist der Zusammenhang zwischen T und p_S nicht linear, vielmehr gilt:

     $$\ln(p/p_0) = A \cdot (1/T_0 - 1/T)$$

     Dabei sind die Werte in SI-Einheiten zu nehmen, und p_0 ist ein beliebiger Bezugsdruck - wir wählen p_0 = 1 bar.
   - Wiederholen Sie a. für die Ausgleichsgerade zwischen 1/T und $\ln(p/p_0)$.

Aufgabe 45

- Die Datei daten45a.txt enthält die Prozentergebnisse aller Physikklausuren seit 2006, zusammen mit der (geschätzten ☺) Körpergröße des Prüflings.

a. Erstellen Sie einen Scatterplot der Daten und berechnen Sie die Regressionsgerade. Testen Sie die Hypothese, dass $\beta = 0$ ist, also das Physikergebnis nicht von der Körpergröße abhängt.

b. Die Datei daten45b.txt erweitert die Daten jeweils um die Angabe des Geschlechts des Prüflings. Wiederholen Sie a. jeweils getrennt mit den Daten der männlichen bzw. weiblichen Kandidaten. Was schließen Sie daraus?

Aufgabe 46

- Man vermutet, dass die MTBF ("mean time between failures") einer Maschine abhängt von ihrer Versorgungsspannung, der Drehzahl und der Temperatur. Um dies zu überprüfen, werden Experimente durchgeführt, die folgende Ergebnisse liefern:

MTBF [min]	**U [V]**	**n [Upm]**	**T [°C]**
4076	220	578	40.0
4094	220	654	62.2
4218	220	770	40.0
4228	220	847	62.2
4203	230	647	45.6
4142	230	678	45.6
4294	240	578	40.0
4305	240	654	62.2
4435	240	770	40.0
4446	260	770	62.2

a. Bestimmen Sie die Parameter eines linearen Ausgleichsmodells und dessen Streuung $s_{y|x}$.

b. Wie groß ist die MTBF nach diesem Modell bei U_0 = 230 V, n_0 = 680 Upm, T_0 = 60 °C? Geben Sie ein 95%-Konfidenzintervall für eine Einzelmessung an.

- Lösen Sie die Aufgabe einmal durch direkte Anwendung der hergeleiteten Beziehungen, ein zweites Mal möglichst einfach mit Hilfe der Statistik-Toolbox.

Aufgabe 47

- Die Schätzer A und B für die Parameter der Regressiongeraden sind nicht unabhängig voneinander. In der Aufgabe soll die Idee des Konfidenzintervalls auf den zweidimensionalen Vektor (A,B) verallgemeinert werden.
- Die Beispieldaten werden mit dem Skript erzeugeZugversuchDaten.m generiert. Für n = 10 sind dies die Daten aus dem Zugversuch-Beispiel.

a. Verwenden Sie die Beispieldaten (bzw. das Skript mit n = 10). Die gemeinsame Verteilung von A und B ist eine bivariate Normalverteilung, deren Parameter $\mu_{1,2}$ und $\sigma_{1,2}$ wir schon berechnet haben. Bestimmen Sie den noch fehlenden Parameter ρ aus der Beziehung

   $\mathrm{Cov}(A,B) = \sigma_1\, \sigma_2\, \rho$

   für die konkreten Daten des Zugversuch-Beispiels und plotten Sie die entsprechende Verteilungsfunktion.

b. Berechnen Sie zunächst die 95%-Konfidenzintervalle für α und β. Verwenden Sie dann die STB-Funktion `mvncdf`, die die kumulative Verteilungsfunktion der multivariaten Normalverteilung berechnet, um folgende Wahrscheinlichkeiten zu bestimmen:

   P_A = P(α liegt im berechneten Konfidenzintervall)

   P_B = P(β liegt im berechneten Konfidenzintervall)

   P_{AB} = P(α und β liegen in den berechneten Konfidenzintervallen)

   Hätten Sie die Ergebnisse erwartet?

c. Wiederholen Sie b. mit n = 1000 Datensätzen. Was schließen Sie?

- Hinweis: Verwenden Sie die Funktion `fitlm` für die üblichen Berechnungen.

Anhang

- Literatur
- Nachweise
- Herleitungen
- Matlab-Beispiele

Literatur

1. A. Rooch: Statistik für Ingenieure
   Springer 2014, 244 S.
2. Chr. Maas: Stochastik für Dummies
   Wiley-VCH Verlag 2013, 343 S.
3. L. Fahrmeir, R., I., G. Tutz: Statistik: Der Weg zur Datenanalyse
   Springer 2012, 610 S.
4. G. Hübner: Stochastik
   Vieweg+Teubner, 5. Aufl. 2009, 220 S.
5. M. Sachs: Wahrscheinlichkeitsrechnung und Statistik
   Carl Hanser, 4. Aufl. 2013, 196 S.
6. S. M. Ross (1): Introduction to Probability Models
   Academic Press, 11. Aufl. 2014, 877 S.
7. S. M. Ross (2): Introduction to Probability and Statistics for Engineers and Scientists
   Academic Press; 5. Aufl. 2014, 686 S.
8. O. Beucher: Wahrscheinlichkeitsrechnung und Statistik mit MATLAB
   Springer, 2. Aufl. 2007, 548 S.
9. A. M. Law: Simulation Modeling and Analysis
   Mcgraw-Hill Publ.Comp., 4. Aufl. 2006, 776 S.
10. Z. Kabluchko, Skript zur Vorlesung "Mathematische Statistik"
    Uni Münster, WS 2015/16 (online)

Nachweise

Text

1. Seite "Ziegenproblem". In: Wikipedia, Die freie Enzyklopädie. Bearbeitungsstand: 2. Juni 2016, 07:09 UTC. (Abgerufen: 25. Juni 2016, 09:07 UTC)
   Originalquelle: Craig F. Whitaker: *Ask Marilyn.* In: *Parade Magazine.* 9. September 1990, S. 16.
2. S. Stidham Jr., A Last Word on L = λ W. Operations Research 22 (1974), p.417-421 (online).
3. L. D. Brown, T. T. Cai, A. DasGupta, Interval Estimation for a Binomial Proportion. Statistical Science, 2001, Vol. 16(2), p. 101-133 (online).
4. H. Ruben, A simple conservative and robust solution of the Behrens-Fisher problem: Sankhya Ind. J. Statist. Ser. A, 2002, Vol. 64, p. 139 – 155 (online).
5. A.L. Herbst, H. Ulfelder H, D. C. Poskanzer: Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. The New England Journal of Medicine, 1971, Vol. 284 (15), p. 878–81 (online).
6. Ch. E. Novitski: Revision of Fisher's analysis of Mendel's garden pea experiments. Genetics. 2004 Mar; 166(3): 1139–1140 (online).
7. D. L. Hartl, D. J. Fairbanks, Mud Sticks: On the Alleged Falsification of Mendel's Data. Genetics. 2007 Mar; 175(3): 975–979 (online).
8. G. Marsaglia, W. W. Tsang, J. Wang: Evaluating Kolmogorov's Distribution. Journal of Statistical Software. Vol. 8, Issue 18, 1-4, 2003 (online).
9. J. P. Meyer, M. A. Seaman: A comparison of the exact Kruskal-Wallis distribution to asymptotic approximations for all sample sizes up to 105. Journal of Experimental Education, 81(2), 139-156, 2014 (online).
10. G.-H. Lee et al: High-Strength Chemical-Vapor−Deposited Graphene and Grain Boundaries. Science 340, 1073-76, 2013 (online, supplement).

Daten

- lebensdauer.dat: künstliche Daten für die Lebensdauer von Festplatten
- klausuren.xlsx: historische anonymisierte Klausurdaten
- bundestagswahlen.xlsx: Ergebnisse der Bundestagswahlen nach Wahlkreisen vom Bundeswahlleiter
  Original als gezippte CSV-Dateien
- arbeitslose.xlsx: Jahresdurchschnitte, erzeugt aus zwei Datensätzen vom Statistischen Bundesamt
  Historische Daten zum Arbeitsmarkt, Daten ab 1950 (Excel-Tabelle)
  Unemployment statistics based on the ILO concept (Auswahl-Tabelle)
  sowie Werten für 1948/49 von der Bundesagentur für Arbeit
- wahl2013.xlsx: Ergebnisse der Bundestagswahl 2013, manuell aus den Werten in bundestagswahlen.xlsx zusammengestellt
- pkwzulassungen.xlsx: Zahl der Neuzulassungen der einzelnen PKW-Marken für Juni 2017 und Januar bis Juni 2017 sowie jeweils Anzahlen für Fahrzeuge mit Diesel-, Hybrid- und Elektroantrieb, Daten vom Kraftfahrt-Bundesamt (Excel-Tabelle)
- solaranlage.csv: von einer privaten Solaranlage erbrachte Energie, für jeden Tag der Jahre 2009 - 2016, mit freundlicher Genehmigung von P. Husmann
- wahlergebnisse.xlsx: Ergebnisse aller Bundestagswahlen für die größeren Parteien, aus den Werten in bundestagswahlen.xlsx zusammengestellt
- akkus100.dat: künstliche Daten für die Kapazität von Akkus
- schrauben100.dat: künstliche Daten für Schraubendurchmesser
- graphen.dat: künstliche Daten für die Festigkeit dreier Graphen-Probensorten. Werte für Probenklassen

"pristine", "LG" und "SG" jeweils in N/m. Konsistent mit den Werten aus [10].

- daten45a.txt, daten45b.txt: erzielte Prozentzahlen für alle Physikklausuren seit 2006, zusammen mit (künstlichen) Körpergrößen und Geschlecht (in b)

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 $\tilde{S}^2$ 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 $S^2_{y|x}$ bei mehrfacher linearer Regression

Erwartungswert bei der Binomialverteilung

- Den Erwartungswert einer nach B(n,p) verteilten Zufallsvariablen erhält man aus

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

- Nach dem Binomischen Lehrsatz gilt

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1$$

  Das ist natürlich auch deswegen klar, weil die Summe der Wahrscheinlichkeiten aller möglichen Ergebnisse einer binomialverteilten Größe 1 ist.

- Leitet man diese Gleichung nach p ab, erhält man:

$$\begin{aligned}
0 &= \sum_{k=0}^{n} \binom{n}{k} k p^{k-1} (1-p)^{n-k} - \sum_{k=0}^{n} \binom{n}{k} (n-k) p^k (1-p)^{n-k-1} \\
&= \frac{1}{p} \sum_{k=0}^{n} \binom{n}{k} k p^k (1-p)^{n-k} - \frac{1}{1-p} \sum_{k=0}^{n} \binom{n}{k} (n-k) p^k (1-p)^{n-k} \\
&= \frac{1}{p} E(X) - \frac{1}{1-p} \sum_{k=0}^{n} \binom{n}{k} n p^k (1-p)^{n-k} + \frac{1}{1-p} E(X) \\
&= \frac{1}{p} E(X) - \frac{1}{1-p} \cdot n + \frac{1}{1-p} E(X)
\end{aligned}$$

- Dies löst man leicht nach E(X) auf:

$$\begin{aligned}
\left(\frac{1}{p} + \frac{1}{1-p}\right) E(X) &= \frac{n}{1-p} \\
\Rightarrow \qquad \frac{1}{p(1-p)} E(X) &= \frac{n}{1-p} \\
\Rightarrow \qquad E(X) &= n \cdot p
\end{aligned}$$

Randverteilung der bivariaten Normalverteilung

- Gesucht ist

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} dy\, f(x,y) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{(x-\mu_1)^2}{2(1-\rho^2)\sigma_1^2}} \cdot \int_{-\infty}^{\infty} dy\, e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(y-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right)} \end{aligned}$$

- Die Substitution

$$z = \frac{y-\mu_2}{\sigma_2} \quad \Rightarrow \quad dz = \frac{dy}{\sigma_2}$$

liefert

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}}\, e^{-\frac{(x-\mu_1)^2}{2(1-\rho^2)\sigma_1^2}} \cdot \int_{-\infty}^{\infty} dz\, e^{-\frac{1}{2(1-\rho^2)}\left(z^2 - 2\rho\frac{(x-\mu_1)}{\sigma_1}z\right)} \\ &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}}\, e^{-\frac{(x-\mu_1)^2}{2(1-\rho^2)\sigma_1^2}} \cdot e^{\frac{\rho^2}{2(1-\rho^2)}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} \cdot \int_{-\infty}^{\infty} dz\, e^{-\frac{1}{2(1-\rho^2)}\left(z - \rho\frac{(x-\mu_1)}{\sigma_1}\right)^2} \\ &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} \cdot \int_{-\infty}^{\infty} dz\, e^{-\frac{1}{2(1-\rho^2)}\left(z - \rho\frac{(x-\mu_1)}{\sigma_1}\right)^2} \end{aligned}$$

- Mit der Substitution

$$t = \frac{z - \rho\frac{(x-\mu_1)}{\sigma_1}}{\sqrt{1-\rho^2}} \quad \Rightarrow \quad dt = \frac{dz}{\sqrt{1-\rho^2}}$$

ergibt dies

$$f_X(x) = \frac{1}{2\pi\sigma_1}\, e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} \cdot \int_{-\infty}^{\infty} dt\, e^{-\frac{1}{2}t^2}$$

- Wegen

$$\int_{-\infty}^{\infty} dt\, e^{-\frac{1}{2}t^2} = \sqrt{2\pi}$$

erhält man schließlich

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1}\, e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}$$

- Das bedeutet $X \sim N(\mu_1, \sigma_1^2)$

Erwartungswert E(X Y) der bivariaten Normalverteilung

- Auf geht's:

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, xy f(x,y) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, x\, y\, e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)} \end{aligned}$$

- Substitution

$$s = \frac{x-\mu_1}{\sigma_1},\; t = \frac{y-\mu_2}{\sigma_2} \quad \Rightarrow \quad ds = \frac{dx}{\sigma_1},\; dt = \frac{dy}{\sigma_2}$$

  ergibt

$$\begin{aligned} E(XY) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} ds \int_{-\infty}^{\infty} dt\, (\sigma_1 s + \mu_1)(\sigma_2 t + \mu_2) e^{-\frac{1}{2(1-\rho^2)}\left(s^2 - 2\rho s t + t^2\right)} \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} ds \int_{-\infty}^{\infty} dt\, (\sigma_1 s + \mu_1)(\sigma_2 t + \mu_2) e^{-\frac{1}{2(1-\rho^2)}(s-\rho t)^2} e^{-\frac{1}{2}t^2} \end{aligned}$$

- Weitere Substitution

$$z = \frac{s - \rho t}{\sqrt{1-\rho^2}} \quad \Rightarrow \quad dz = \frac{ds}{\sqrt{1-\rho^2}}$$

  liefert

$$\begin{aligned} E(XY) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} dt\, (\sigma_1\sqrt{1-\rho^2}\, z + \sigma_1 \rho\, t + \mu_1)(\sigma_2\, t + \mu_2)\, e^{-\frac{1}{2}z^2} e^{-\frac{1}{2}t^2} \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} dt\, (\sigma_2\, t + \mu_2)\, e^{-\frac{1}{2}t^2} \int_{-\infty}^{\infty} dz\, (\sigma_1\sqrt{1-\rho^2}\, z + \sigma_1 \rho\, t + \mu_1)\, e^{-\frac{1}{2}z^2} \end{aligned}$$

- Wegen

$$\int_{-\infty}^{\infty} dz\, e^{-\frac{1}{2}z^2} = \sqrt{2\pi}$$

  und

$$\int_{-\infty}^{\infty} dz\, z\, e^{-\frac{1}{2}z^2} = 0 \qquad \text{Antisymmetrie!}$$

  ist dies

$$\begin{aligned} E(XY) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} dt\, (\sigma_2\, t + \mu_2)\, e^{-\frac{1}{2}t^2} (\sigma_1 \rho\, t + \mu_1)\sqrt{2\pi} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dt\, (\sigma_1\sigma_2\rho\, t^2 + \mu_1\mu_2)\, e^{-\frac{1}{2}t^2} \end{aligned}$$

  wobei im letzten Schritt die gemischten Terme (mit Faktor t) wieder weggefallen sind.

- Benutzt man weiter

$$\int_{-\infty}^{\infty} dz\, z^2 e^{-\frac{1}{2}z^2} = \sqrt{2\pi}$$

  erhält man schließlich

$$
\begin{aligned}
E(XY) &= \frac{1}{\sqrt{2\pi}}(\sigma_1\sigma_2\rho\sqrt{2\pi} + \mu_1\mu_2\sqrt{2\pi}) \\
&= \sigma_1\sigma_2\rho + \mu_1\mu_2
\end{aligned}
$$

Herleitung des Box-Muller-Verfahrens

- Trick:

  berechne statt eines Zufallswerts gleich zwei auf einmal

  betrachte dazu zwei unabhängige Zufallsvariablen Y_1, Y_2 ~ N(0,1)

- Die gemeinsame Verteilungsfunktion lautet

$$F(y_1, y_2) = \frac{1}{2\pi} \int_{-\infty}^{y_1} dy_1' \int_{-\infty}^{y_2} dy_2' \, e^{-y_1'^2/2} e^{-y_2'^2/2}$$

- Zur Berechnung führt man Polarkoordinaten ein

$$\begin{aligned} y_1 &= r\cos\theta \qquad r \in [0, \infty) \\ y_2 &= r\sin\theta \qquad \theta \in [0, 2\pi) \end{aligned}$$

- und erhält als transformiertes Volumen-Element (Jacobi-Matrix nachrechnen!)

  $dy_1\, dy_2 = r\, dr\, d\theta$

- Das Integral wird dann

$$\begin{aligned} F(r, \theta) &= \frac{1}{2\pi} \int_0^r dr' \int_0^\theta d\theta' \, r' e^{-r'^2/2} \\ &= \left(\frac{1}{2\pi} \int_0^\theta d\theta' \right) \cdot \left(\int_0^r dr' \, r' e^{-r'^2/2} \right) \\ &= F_1(\theta) \cdot F_2(r) \end{aligned}$$

- Dies ist die Verteilungsfuntion zweier unabhängiger Zufallsvariablen Θ und R, wobei Θ gleichverteilt ist in $[0, 2\pi]$.

- Die Verteilungsfunktion von R und ihr Inverses berechnet man leicht zu

$$\begin{aligned} F_2(r) &= 1 - e^{-r^2/2} \\ F_2^{-1}(x) &= \sqrt{-2\ln(1-x)} \end{aligned}$$

- Benutzt man noch, dass mit X ~ U(0,1) auch 1 - X ~ U(0,1), kann man also folgendermaßen vorgehen:
  1. bestimme X_1, X_2 ~ U(0,1)
  2. berechne daraus

$$\begin{aligned} \Theta &= 2\pi X_1 \\ R &= \sqrt{-2\ln X} \end{aligned}$$

  3. und damit

$$\begin{aligned} Y_1 &= R\cos\Theta \\ Y_2 &= R\sin\Theta \end{aligned}$$

Poisson-Grenzwertsatz

- Satz:

  Für n = 1, 2, ... sei $X_n \sim B(n, p_n)$ mit

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

  Dann geht X_n gegen Po(λ), d.h.

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

- Beweis:

  Man rechnet direkt nach

  $$\begin{aligned}
  P(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
  &= \frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k (1-p_n)^{n-k} \\
  &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{(np_n)^k}{k!} \cdot \left(1 - \frac{np_n}{n}\right)^n \cdot (1-p_n)^{-k} \\
  &\overset{n\to\infty}{\longrightarrow} 1 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \cdot 1 \\
  &= e^{-\lambda} \cdot \frac{\lambda^k}{k!}
  \end{aligned}$$

  mit der bekannten Beziehung

  $$\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x$$

Eigenschaften der geometrischen Verteilung

- Kumulative Verteilungsfunktion von G(p):

  X sei G(p)-verteilte Zufallsvariable. Dann erhält man

$$\begin{aligned} P(X \le n) &= \sum_{k=1}^{n} P(X=k) \\ &= \sum_{k=1}^{n} (1-p)^{k-1} p \\ &= p \sum_{k=0}^{n-1} (1-p)^{k} \\ &= p \cdot \frac{(1-p)^n - 1}{(1-p)-1} \\ &= 1-(1-p)^n \end{aligned}$$

- Grenzwert-Eigenschaft von G(p):
  - Sei X Zufallsvariable mit X/h ~ G(λh). Für h → 0 gilt dann

    X ~ Ex(λ)
  - Beweis durch Berechnen der kumulativen Verteilungsfunktion

$$\begin{aligned} P(X \le a) &= P\left(\frac{X}{h} \le \frac{a}{h}\right) \\ &= 1-(1-\lambda h)^{a/h} \\ &= 1-\left(1+\frac{(-a\lambda)}{a/h}\right)^{a/h} \end{aligned}$$

    - Mit der bekannten Beziehung

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$$

    - folgt

$$P(X \le a) \xrightarrow{h\to 0} 1-e^{-a\lambda} = F_\lambda(a)$$

    - wobei F_λ die Verteilungsfunktion von Ex(λ) ist

Erwartungswert des ML-Schätzers $\tilde{S}^2$ für die Varianz

Berechnung des Erwartungswerts von $\tilde{S}^2$ nach [Ross2]:

- Zunächst gilt

  $$\begin{aligned}\sum_{i=1}^{n}(x_i-\overline{x})^2 &= \sum_{i=1}^{n}(x_i^2-2\overline{x}x_i+\overline{x}^2)\\ &= \sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2\\ &= \sum_{i=1}^{n}x_i^2-2\overline{x}(n\overline{x})+n\overline{x}^2\\ &= \sum_{i=1}^{n}x_i^2-n\overline{x}^2\end{aligned}$$

  daher ist

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

- Nimmt man den Erwartungswert beider Seiten und benutzt noch die bekannte Beziehung für die Varianz einer beliebigen Zufallsvariablen Y

  $$E(Y^2)=\mathrm{Var}(Y)+E(Y)^2$$

  erhält man

  $$\begin{aligned}nE(\tilde{S}^2) &= E\left(\sum_{i=1}^{n}X_i^2\right)-nE(\overline{X}^2)\\ &= nE(X_1^2)-nE(\overline{X}^2)\\ &= n\,\mathrm{Var}(X_1)+nE(X_1)^2-n\,\mathrm{Var}(\overline{X})-nE(\overline{X})^2\\ &= n\sigma^2+n\mu^2-n\frac{\sigma^2}{n}-n\mu^2\\ &= (n-1)\,\sigma^2\end{aligned}$$

  also

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

Aufteilung der Varianzen bei einfaktorieller ANOVA

- Gegeben seien Werte von n Stichproben jeweils vom Umfang n_i, also

  $$X_{ij}, \quad i = 1 \ldots n,\ j = 1 \ldots n_i$$

  Mit der Gesamtanzahl N, dem Mittelwert $\overline{X}_i$ der i-ten Stichprobe und dem Gesamt-Mittelwert $\overline{X}$ hat man dann

  $$\sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\overline{X}_i-\overline{X})^2$$

- Beweis:

  Man berechnet zunächst

  $$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}((X_{ij}-\overline{X}_i)+(\overline{X}_i-\overline{X}))^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\overline{X}_i-\overline{X})^2 \\ &\quad + 2\sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)(\overline{X}_i-\overline{X})\end{aligned}$$

  Es bleibt zu zeigen, dass der gemischte Term verschwindet:

  $$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)(\overline{X}_i-\overline{X}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\overline{X}_i X_{ij} - \overline{X}_i^2 - \overline{X}X_{ij} + \overline{X}\,\overline{X}_i) \\ &= \sum_{i=1}^{n}\overline{X}_i\sum_{j=1}^{n_i}X_{ij} - \sum_{i=1}^{n}\overline{X}_i^2 n_i - \overline{X}\sum_{i=1}^{n}\sum_{j=1}^{n_i}X_{ij} + \overline{X}\sum_{i=1}^{n}\overline{X}_i n_i \\ &= \sum_{i=1}^{n}\overline{X}_i n_i \overline{X}_i - \sum_{i=1}^{n}n_i\overline{X}_i^2 - \overline{X}N\overline{X} + \overline{X}\sum_{i=1}^{n}\sum_{j=1}^{n_i}X_{ij} \\ &= -N\overline{X}^2 + \overline{X}N\overline{X} \\ &= 0\end{aligned}$$

Berechnung der Varianz des ML-Schätzers A bei einfacher linearer Regression

- Setzt man die Formeln für $\overline{Y}$ und B ein, kann man A als Linearkombination der Y_i schreiben

$$\begin{aligned}
A &= \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\overline{x}}{\sum_{k=1}^{n}(x_k-\overline{x})^2}\sum_{i=1}^{n}(x_i-\overline{x})Y_i \\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\overline{x}(x_i-\overline{x})}{\sum_{k=1}^{n}(x_k-\overline{x})^2}\right)Y_i
\end{aligned}$$

- Da alle Y_i unabhängig sind und die gleiche Varianz σ^2 haben, erhält man

$$\begin{aligned}
\mathrm{Var}(A) &= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\overline{x}(x_i-\overline{x})}{\sum_{k=1}^{n}(x_k-\overline{x})^2}\right)^2\sigma^2 \\
&= \sigma^2\left(\frac{1}{n} - \frac{2}{n}\sum_{i=1}^{n}\frac{\overline{x}(x_i-\overline{x})}{\sum_{k=1}^{n}(x_k-\overline{x})^2} + \sum_{i=1}^{n}\frac{\overline{x}^2(x_i-\overline{x})^2}{\left(\sum_{k=1}^{n}(x_k-\overline{x})^2\right)^2}\right) \\
&= \sigma^2\left(\frac{1}{n} - 0 + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right) \\
&= \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2}\left(\sum_{i=1}^{n}(x_i-\overline{x})^2 + n\overline{x}^2\right) \\
&= \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2}\left(\sum_{i=1}^{n}x_i^2 - 2\overline{x}\cdot n\overline{x} + n\overline{x}^2 + n\overline{x}^2\right) \\
&= \frac{\sigma^2}{n}\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}
\end{aligned}$$

Berechnung der Kovarianzmatrix des ML-Schätzers B bei mehrfacher linearer Regression

- Wegen

$$\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

  kann man B_i als Linearkombination der Y_k schreiben

  Dazu führt man die Hilfsmatrix C ein

$$\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = (C_{ij})_{\substack{i=1\ldots p \\ j=1\ldots n}}$$

  und hat dann

$$B_{i-1} = \sum_{l=1}^{n} C_{il}Y_l, \qquad i = 1\ldots p$$

- Man rechnet los

$$\begin{aligned} \mathrm{Cov}(B_{i-1}, B_{j-1}) &= \mathrm{Cov}(\sum_{l=1}^{n} C_{il}Y_l, \sum_{r=1}^{n} C_{jr}Y_r) \\ &= \sum_{l=1}^{n}\sum_{r=1}^{n} C_{il}C_{jr}\,\mathrm{Cov}(Y_l, Y_r) \end{aligned}$$

  Wegen

$$\mathrm{Cov}(Y_l, Y_r) = \begin{cases} 0 & | \quad l \neq r \\ \mathrm{Var}(Y_l) = \sigma^2 & | \quad l = r \end{cases}$$

  erhält man

$$\begin{aligned} \mathrm{Cov}(B_{i-1}, B_{j-1}) &= \sigma^2 \sum_{r=1}^{n} C_{ir}C_{jr} \\ &= \sigma^2(\boldsymbol{C}\boldsymbol{C}^T)_{ij} \end{aligned}$$

  also

$$\mathrm{Cov}(\boldsymbol{B}) = \sigma^2\boldsymbol{C}\boldsymbol{C}^T$$

- Nun ist aber

$$\begin{aligned} \boldsymbol{C}\boldsymbol{C}^T &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned}$$

  also

$$\mathrm{Cov}(\boldsymbol{B}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Formel zur Berechnung des Schätzers $S^2_{y|x}$ bei mehrfacher linearer Regression

- Zunächst gilt

$$\begin{aligned}
(n-k-1)\, S^2_{y|x} &= \sum_{i=1}^{n}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik})^2 \\
&= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) \\
&= (\boldsymbol{Y}^T - \boldsymbol{B}^T\boldsymbol{X}^T)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) \\
&= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B} - \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}
\end{aligned}$$

- Mit der Normalengleichung

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B} = \boldsymbol{X}^T\boldsymbol{Y}$$

  erhält man

$$\begin{aligned}
(n-k-1)\, S^2_{y|x} &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B} - \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y} \\
&= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B}
\end{aligned}$$

- Nun ist Y^T X B ein Skalar, daher gleich seinem Transponierten

$$\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B} = (\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y}$$

  also folgt

$$S^2_{y|x} = \frac{1}{n-k-1}(\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y})$$

Matlab-Beipiele

- 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

bsp_1_3_2.m

```
%% Produktionsfehler
p = 0.1;
n = 15;
binomi = @(k) nchoosek(n, k)*p^k*(1-p)^(n-k);
pa = binomi(2)
pb = binomi(1) + binomi(2)
pcC = binomi(13) + binomi(14) + binomi(15)

%% Geometrische Verteilung
pd = (1 - p)^4 * p

%% Lottozahlen
pe = nchoosek(6, 4)*nchoosek(43, 2)/nchoosek(49, 6)

%% Versicherung
la = 3;
poiss = @(k) exp(-la)*la^k/factorial(k);
pva = poiss(3)
pvb = poiss(0) + poiss(1) + poiss(2)
pvc = 1 - pva - pvb
```

bsp_1_4_3.m

```
% Beispiel für Kap 1.4.3
%% Beispiel Produkt-Lebensdauer
la = 1/6;   % in 1/a
P1 = exp(-1)
p = 0.1;
T = -log(1-p)/la

%% Beispiel Größenverteilung
x = 20/9;
A = 1 - 0.5*(erf(x/sqrt(2)) + 1)
A = 1 - normcdf(200, 180, 9)

%% Quantile der Standard-Normalverteilung
p = [50, 75, 90, 95, 97.5, 99]/100;
xp = norminv(p,0,1)

x = 1:4;
P = normcdf(x) - normcdf(-x)
```

bsp_1_5_1.m

```
% Beispiel für Kap 1.5.1
%% Beispiel 3-2-Würfel
X = [2,3,4,5]';
Y = [1,2,3,4,6]';
pXY = [1,0,0,0,0; 0,2,0,0,0; 0,0,1,1,0; 0,0,0,0,1]/6
pY = sum(pXY)'
pX = sum(pXY,2)

EX = X'*pX
EY = Y'*pY

EXX = (X.*X)'*pX
EYY = (Y.*Y)'*pY
EXY = sum(sum((X*Y').*pXY))

CovXY = EXY - EX*EY
VarX = EXX - EX*EX
VarY = EYY - EY*EY
rhoXY = CovXY/sqrt(VarX*VarY)
```

bsp_1_5_2.m

```
% Beispiel für Kap 1.5.2
%% Beladung eines Schmelzofens
kap = 3000;
mu1 = 100;
var1 = 400;
n = 27;

mug = n*mu1
varg = n*var1
sigmag = sqrt(varg)

z1 = (kap - mug)/sigmag
P = 100*(0.5 - 0.5*erf(z1/sqrt(2)))

% oder direkt
pd = makedist('Normal', 'mu', mug, 'sigma', sigmag);
P = 100*(1 - cdf(pd, kap))
```

bsp_1_5_3.m

```
% Beispiel für Kap 1.5.3
% Beispiel Produktionsfehler
p = 0.1;
n = 100;
k = 10;

Pb = nchoosek(n, k)*p^k*(1-p)^(n-k)
Pn = erf(0.5/3/sqrt(2))
```

bsp_3_1.m

```
% Beispiel für Kap 3.1
%% Roulette
p = 19/37;
q = 18/37;
P = diag(q*ones(5,1), 1) + diag(p*ones(5,1), -1);
P(1,2) = 0;
P(6,5) = 0;
P(1,1) = 1;
P(6,6) = 1
P5 = P*P*P*P*P
P5(3,6)

%% GGV bei Autovermietung
P = [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];
% homogenes System lösen = Null space finden
x = null(P' - eye(4))';
x = x/sum(x)
x*P - x;    % Test klappt!

%% GGV bei Roulette
P = [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];
x = null(P' - eye(6))'    % es gibt zwei, sie sind schon normiert
```

bsp_4_1_2a.m

```
% Beispiel für Kap 4.1.2a
% Beispiel Bundestagswahl 2013, Häufigkeitstabelle
file = '../daten/wahl2013.xlsx';
[wahl, wahlText] = xlsread(file);
wahlText = wahlText(3,:);

h = wahl(5:end);
a = string(wahlText(5:end));

n = sum(h);     % = wahl(4), gültige Stimmen
f = h/n;

% sortieren
[fS, idx] = sort(f, 'descend');
aS = a(idx);
nL = find(fS >= 0.01, 1, 'last');   % alle bis 1%
aS = aS(1:nL);
fS = fS(1:nL);

% Klassenbildung
% CDU, CSU zusammenfassen, einfach per Hand
aS(1) = "CDU/CSU";
fS(1) = fS(1) + fS(5);
aS(5:end-1) = aS(6:end);
fS(5:end-1) = fS(6:end);

% sonstige
aS(end) = "Sonstige";
fS(end) = 1 - sum(fS(1:end-1));

% tabelle ausgeben
fprintf('%10s', aS);
fprintf('\n');
fprintf('%10.2f', 100*fS);
fprintf('\n');

% Plotten
figure('Position',[1 1 600 400], 'Color',[1 1 1]);

% bar graph
bar(100*fS);
set(gca(), 'XTickLabel', aS);
ylabel('Stimmenanteil [%]');

% stem graph
figure
stem(100*fS);
set(gca(),'XTick', 1:9, 'XTickLabel', aS);
xlim([0,10])
ylabel('Stimmenanteil [%]');

% pie graph, umsortiert
figure
set(gcf(), 'Position',[1 1 400 400]);
[aS, idx] = sort(aS);
pie(100*fS(idx), cellstr(aS));
```

bsp_4_1_2b.m

```
% Beispiel für Kap 4.1.2b
% Beispiel Klausuren, Kennzahlen

% Daten einlesen
file = '../daten/klausuren.xlsx';
kl15 = xlsread(file, '2015');
kl14 = xlsread(file, '2014');
kl13 = xlsread(file, '2013');
X13 = kl13(1:end-2,5);
X14 = kl14(1:end-2,5);
X15 = kl15(1:end-2,5);
X = [X13; X14; X15];
n = size(X,1);
fprintf('%d Datensätze gelesen\n', n);

edges = 0:10:100;
subplot(2,2,1)
histogram(X13, edges)
xlabel('2013')

subplot(2,2,2)
histogram(X14, edges)
xlabel('2014')

subplot(2,2,3)
histogram(X15, edges)
xlabel('2015')

subplot(2,2,4)
histogram(X, edges)
xlabel('gesamt')

xq = zeros(1,4);
xq(1) = mean(X13);
xq(2) = mean(X14);
xq(3) = mean(X15);
xq(4) = mean(X);

xm = zeros(1,4);
xm(1) = median(X13);
xm(2) = median(X14);
xm(3) = median(X15);
xm(4) = median(X);

xq1 = zeros(1,4);
xq1(1) = quantile(X13, 1/4);
xq1(2) = quantile(X14, 1/4);
xq1(3) = quantile(X15, 1/4);
xq1(4) = quantile(X, 1/4);

xq3 = zeros(1,4);
xq3(1) = quantile(X13, 3/4);
xq3(2) = quantile(X14, 3/4);
xq3(3) = quantile(X15, 3/4);
xq3(4) = quantile(X, 3/4);

xs = zeros(1,4);
xs(1) = std(X13);
xs(2) = std(X14);
xs(3) = std(X15);
xs(4) = std(X);

fprintf('Kennzahl 2013    2014    2015    gesamt\n');
fprintf('mean     %5.2f   %5.2f   %5.2f   %5.2f\n', xq);
fprintf('Q1       %5.2f   %5.2f   %5.2f   %5.2f\n', xq1);
fprintf('median   %5.2f   %5.2f   %5.2f   %5.2f\n', xm);
fprintf('Q3       %5.2f   %5.2f   %5.2f   %5.2f\n', xq3);
fprintf('s        %5.2f   %5.2f   %5.2f   %5.2f\n', xs);
```

bsp_4_1_2c.m

```
% Beispiel für Kap 4.1.2c
% Beispiel Klausuren, Boxplot, Verteilungsfunktionen

% Daten einlesen
file = '../daten/klausuren.xlsx';
kl15 = xlsread(file, '2015');
kl14 = xlsread(file, '2014');
kl13 = xlsread(file, '2013');
X13 = kl13(1:end-2,5);
X14 = kl14(1:end-2,5);
X15 = kl15(1:end-2,5);
Xall = [X13; X14; X15];
n = size(Xall,1);
fprintf('%d Datensätze gelesen\n', n);

% alle auf gleiche Länge bringen, mit NaNs füllen
Y = NaN*zeros(length(Xall),4);
Y(1:length(X13), 1) = X13;
Y(1:length(X14), 2) = X14;
Y(1:length(X15), 3) = X15;
Y(1:length(Xall), 4) = Xall;

boxplot(Y)
set(gca(),'XTickLabel', ["2013", "2014", "2015", "gesamt"]);

% Verteilungsfunktion
figure;
set(gcf(), 'Position',[1 1 700 400], 'Color',[1 1 1]);
subplot(1,2,1)
dx = 5;
edges = 0:dx:100;
N = histcounts(Xall, edges);
bar(edges(1:end-1)+dx/2, N, 1.0)
set(gca(),'XTick',0:20:100)
xlabel('empirische Dichtefunktion')

subplot(1,2,2)
edges = 0:1:100;
N = histcounts(Xall, edges);
stairs(edges(1:end-1)+0.5, cumsum(N)/n)
xlabel('empirische Verteilungsfunktion')
```

bsp_4_1_3a.m

```
% Beispiel für Kap 4.1.3a
% Beispiel Klausuren, 2d-Daten

% Daten einlesen
file = '../daten/klausuren.xlsx';
kl15 = xlsread(file, '2015');
kl14 = xlsread(file, '2014');
kl13 = xlsread(file, '2013');
% 1. Spalte: Jahr, 2. - 4. Spalte: Aufgabe 1 - 3
X13 = kl13(1:end-2,1:4);
X14 = kl14(1:end-2,1:4);
X15 = kl15(1:end-2,1:4);
Xall = [X13; X14; X15];
n = size(Xall,1);
fprintf('%d Datensätze gelesen\n', n);

% Plotten
figure('Position',[1 1 1100 300], 'Color',[1 1 1]);
subplot(1,3,1)
scatter(Xall(:,2), Xall(:,3))
xlabel('Aufgabe 1')
ylabel('Aufgabe 2')

subplot(1,3,2)
gscatter(Xall(:,2), Xall(:,3), Xall(:,1))
xlabel('Aufgabe 1')
ylabel('Aufgabe 2')

subplot(1,3,3)
hist3(Xall(:, 2:3), [5,5])
xlabel('Aufgabe 1')
ylabel('Aufgabe 2')

% Kennzahlen
r13 = corr(X13(:,2), X13(:,3));
r14 = corr(X14(:,2), X14(:,3));
r15 = corr(X15(:,2), X15(:,3));
rAll = corr(Xall(:,2), Xall(:,3));
fprintf('       2013     2014     2015     gesamt\n');
fprintf('rxy    %6.4f   %6.4f   %6.4f   %6.4f\n', r13, r14, r15, rAll);

% check: per Hand
X = Xall(:,2);
Y = Xall(:,3);
xq = mean(X);
yq = mean(Y);
sx = std(X);
sy = std(Y);
sxy = (X-xq)'*(Y-yq)/(length(X)-1);
rxy = sxy/(sx*sy)
```

bsp_4_1_3b.m

```
% Beispiel für Kap 4.1.3b
% Beispiel Lebensdauern
X = load('../daten/lebensdauer.dat');
X1 = X(:,1);
X2 = X(:,2);
edges = 0:16;

figure('Position',[1 1 700 400], 'Color',[1 1 1]);
subplot(1,2,1)
histogram(X1, edges)
xlim([0,16])
xlabel("Lebensdauern von Typ 1");
subplot(1,2,2)
histogram(X2, edges)
xlim([0,16])
xlabel("Lebensdauern von Typ 2");

% berechnen der Quantile für i/N
N = 20;
p = (1:N-1)/N;
x1p = quantile(X1, p);
x2p = quantile(X2, p);

% berechnen der Parameter und Quantile der Normalverteilung
m = mean(X1);
s = std(X1);
xnp = norminv(p, m, s);

figure('Position',[1 1 700 400], 'Color',[1 1 1]);
subplot(1,2,1)
plot(x2p, x1p, 'o', [0 12], [0, 12])
title('Q-Q-Plot Typ1 über Typ2')
xlabel('Typ 2')
ylabel('Typ 1')

subplot(1,2,2)
plot(xnp, x1p, 'o', [0 12], [0, 12])
title('Q-Q-Plot Typ1 über Normalverteilung')
xlabel('Normalverteilung')
ylabel('Typ 1')
```

bsp_4_3.m

```
% Beispiel für Kap 4.3
%% erzeuge Beispielwerte für Kap. 4-3
% Quantile der Standard-Normalverteilung
alpha = [10, 5, 2, 1, 0.5, 0.1]/100;
x_al = norminv(1-alpha/2);

% Beispiel Füllmengen
mu = 0.72;
sigma = 0.03;
n = 20;
alpha = 0.05;
rng('default');  % For reproducibility
xi = normrnd(mu, sigma, 1, n);
xq = mean(xi)
z = norminv(1 - alpha/2)
gu = xq - sigma*z/sqrt(n)
go = xq + sigma*z/sqrt(n)

s = std(xi)
s2 = s^2
t = tinv(1 - alpha/2, n-1)
gu = xq - s*t/sqrt(n)
go = xq + s*t/sqrt(n)

chiu = chi2inv(1 - alpha/2, n-1)
chio = chi2inv(alpha/2, n-1)
gu = (n-1)*s2/chiu
go = (n-1)*s2/chio
sigu = sqrt(gu)
sigo = sqrt(go)


%% Beispiel Würfel
n = 30;
rng(4);  % For reproducibility
xi = floor(6*rand(1,n)) + 1;
alpha = 0.05;
y = sum(xi == 6)
z = norminv(1 - alpha/2)
xq = y/n
gu = (xq + z^2/(2*n) - (z/sqrt(n))*sqrt(xq*(1-xq) + z^2/(4*n)))/(1+z^2/n)
go = (xq + z^2/(2*n) + (z/sqrt(n))*sqrt(xq*(1-xq) + z^2/(4*n)))/(1+z^2/n)
```

bsp_5_1.m

```
% Beispiel für Kap 5.1
%% W.keit bei Würfeltest über B(n,p)
n = 200;
p = 0.5;
pd = makedist('Binomial','N',n,'p',p);
c = [10, 13, 14]
y = 2*cdf(pd,99-c)
```

bsp_5_2_1f.m

```
% Beispiel für Kap 5.2.1f
%% Beispiel Schraubendurchmesser
al = 0.05;
n = 10;
sigma = 0.05;
mu = 4.95;

fprintf('\nTest auf my bei bekanntem sigma\n');
mu0 = 5;
rng(42);  % For reproducibility
xi = normrnd(mu, sigma, 1, n)
xq = mean(xi)
Tx = sqrt(n)/sigma*abs(xq - 5)
pW = 2*(1-normcdf(Tx));
fprintf('\npW = %8.6f\n', pW)

fprintf('\nTest auf my bei unbekanntem sigma\n');
S = std(xi)
c = tinv(1-al/2, n-1)
Tx = sqrt(n)/S*abs(xq - 5)

fprintf('\nTest auf sigma\n');
c1 = chi2inv(al/2, n-1)
c2 = chi2inv(1-al/2, n-1)
Tx = (n-1)/sigma^2 * S^2

fprintf('\nTest auf gleiche Schraubenlänge\n');
n = 10;
m = 12;
mux = 4.95;
muy = 5.00;
sigma = 0.05;

al = 0.05;
xi = normrnd(mux, sigma, 1, n)
yi = normrnd(muy, sigma, 1, m)
c = norminv(1-al/2)
xq = mean(xi)
yq = mean(yi)
Tx = (xq - yq)/(sigma*sqrt(1/n + 1/m))

fprintf('\nTest auf gleiche Schraubenstreuung\n');
al = 0.01;
n = 10;
m = 10;
mux = 4.95;
muy = 5.00;
sigmax = 0.05;
sigmay = 0.03;
xi = normrnd(mux, sigmax, 1, n)
yi = normrnd(muy, sigmay, 1, m)
c1 = finv(al/2, n-1, m-1)
c2 = finv(1-al/2, n-1, m-1)
sx = std(xi)
sy = std(yi)
Tx = (sx/sy)^2

% bei größerem alpha
al0 = 0.05;
c1a = finv(al0/2, n-1, m-1);
c2a = finv(1-al0/2, n-1, m-1);
fprintf('\nc-Werte bei al=0.05: %6.4f %6.4f\n', c1a, c2a)
```

bsp_5_2_2.m

```
% Beispiel für Kap 5.2.2
%% Auswertung größeres Schraubenbeispiel
daten = load('daten/schrauben100.dat');
xi = daten(:,1);
yi = daten(:,2);
n = length(xi);
m = length(yi);
sx = std(xi)
sy = std(yi)
al = 0.01;
c1 = finv(al/2, n-1, m-1)
c2 = finv(1-al/2, n-1, m-1)
Tx = (sx/sy)^2
pW = 1 - fcdf(Tx, n-1, m-1)
```

bsp_5_2_3.m

```
% Beispiel für Kap 5.2.3
%% Ausschussbeispiel
p0 = 0.005;
n = 2000;
Tx = 13;
al = 0.05;
ca = binoinv(1-al, n, p0)
pW = 1 - binocdf(Tx, n, p0)

test = n*p0*(1 - p0) % < 9?
z = norminv(1-al)
ca = n*p0 + z*sqrt(n*p0*(1 - p0))

%% Medikamentenbeispiel
n = 100;
m = 120;
x1 = 40;
y1 = 42;
k = x1 + y1;
p1 = hygecdf(x1, n+m, n, k)
p2 = 1 - hygecdf(x1-1, n+m, n, k)
```

bsp_5_3.m

```
% Beispiel für Kap 5.3
%% Beispiel Saurer Regen
xi = [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]
n = length(xi);
p = 0.5;
m0 = 5.2;

Tx = sum(xi < 5.2)
pW = 1 - binocdf(Tx-1, n, p)

%% Lebensdauer Festplatten
X = [6.9175, 7.1546, 8.0619, 6.7216, 7.6289, 8.2784];
Y = [7.0103, 6.1237, 6.0309, 7.4330, 6.8247];
n = length(X);
m = length(Y);

lebensdauer = [X, Y];
typ = [ones(1,n), 2*ones(1,m)];
[wert,idx] = sort(lebensdauer);
herst = typ(idx);
rang = 1:(n+m);
fprintf('Rang   Wert   Hersteller\n')
fprintf(' %2d   %6.4f      %1d\n', [rang; wert; herst])

T = sum(rang(herst == 1))

c1 = ranksumcdf(T,n,m)
c2 = ranksumcdf(T-1,n,m)
pW = 2*min(c1, 1-c2)

%% 5.3 Mendels Erbsen
pj = [9,3,3,1]/16;
Nj = [315, 101, 108, 32];
n = sum(Nj);
T = sum((Nj - n*pj).^2./(n*pj))
z = chi2inv(1-0.05, 3)
pW = 1 - chi2cdf(T, 3)

%% 5.3 Schicht und Maschine
Nij = [12    14    13    12;
       16    13    15    11;
       11    10    9     19]
mj = sum(Nij)
ni = sum(Nij')'
n = sum(ni)
pi = ni/n
qj = mj/n
T = sum(sum((Nij - n*pi*qj).^2./(n*pi*qj)))
r = length(ni);
s = length(mj);
z = chi2inv(1-0.05, (r-1)*(s-1))
pW = 1 - chi2cdf(T, (r-1)*(s-1))

%% 5.3 Kolmogorov-Smirnov-Test
X0 = [4.9795 4.9359 4.9858 4.9968 5.0500 4.9599 5.0046 5.0137 4.8933 ...
      5.0202 4.9521 4.9837];
X = sort(X0);
n = length(X);
Fn = (1:n)/n;
mu = 5;
sigma = 0.05;
F0 = normcdf(X, mu, sigma);
do = abs(F0 - Fn);
du = abs(F0 - [0, Fn(1:n-1)]);

OUT = [(1:n)', X', Fn', F0', do', du'];
fprintf('i     xi     Fn(xi)  F0(xi)  d_i,o  d_i,u\n')
fprintf('%2d  %6.4f  %6.4f  %6.4f  %6.4f  %6.4f\n', OUT');
```

```
T = max([do, du])

% oder direkt mit STB-Funktion
[verwerfen, pWert, T, KS_al_n] = kstest((X0 - mu)/sigma)
```

bsp_5_4.m

```
% Beispiel für Kap 5.4
%% Biegefestigkeit von Querlenkern
mu = [480, 470, 480, 490];
sigma = 20;
ni = [12, 10, 8, 9];
n = length(ni);
nmax = max(ni);
rng(42);  % For reproducibility
xij = normrnd(ones(nmax,1)*mu, sigma, nmax, n);
for I=1:n
  xij(ni(I)+1:end,I) = NaN;
end
fprintf('%6.2f %6.2f %6.2f %6.2f %6.2f %6.2f %6.2f %6.2f %6.2f %6.2f %6.2f %6.2f\n', xij);

%% test all pairs with al=5%
al = 0.05;
fprintf('\n2-sample t-tests:\n');
for I = 1:(n-1)
  for J = (I+1):n
    % direkt mit STB-Funktion
    [h, p] = ttest2(xij(:,I), xij(:,J), 'Alpha', al);
    fprintf('   (%1d,%1d): %6.4f\n', I,J,p);
  end
end

%% test al with Bonferroni-Holm
alloc = al/6    % zu klein selbst für (2,4)

%% jetzt mit Varianzanalyse
xqi = mean(xij, 'omitnan')
N = sum(ni);
xq = sum(sum(xij, 'omitnan'))/N
Y1 = sum(sum((xij - ones(nmax,1)*xqi).^2, 'omitnan'))/sigma^2
Y2 = ((xqi - xq).^2)*ni'/sigma^2
T = (Y2/(n-1))/(Y1/(N-n))
al = 0.05;
c = finv(1-al, n-1, N-n)
pW = 1-fcdf(T, n-1, N-n)

% komplett mit Matlab-Test
[p,tbl] = anova1(xij, [], 'off') % keine Plots, tbl/F = T, tbl/Prob>F = pW

%% und schließlich Kruskal-Willis-Test
N = sum(ni);

liste = xij(:);              % xij als 1d-Vektor
[~, idx] = sort(liste);      % sortierte Indizes
raenge = (1:length(liste))';
[~, idx2] = sort(idx);    % idx2 = inverse Sortierung
rangMat = reshape(raenge(idx2), size(xij));  % raenge sortieren und umordnen
rangMat(rangMat > N) = NaN

rqi = mean(rangMat, 'omitnan')
T = 12/(N*(N+1))*(rqi.^2)*ni' - 3*(N+1)
al = 0.05;
c = chi2inv(1-al, n-1)
pW = 1 - chi2cdf(T, n-1)

% komplett mit Matlab-Test
[p, tbl] = kruskalwallis(xij, [], 'off')
```

bsp_6_1ff.m

```
% Beispiel für Kap 6.1 - 6.3
%% Beispiel Zugversuch
n = 10;
[xi, yi] = erzeugeZugversuchDaten(n);
% gerundetes Ergebnis, i.F. verwendet:
xi = [0.0601 0.1437 0.2399 0.3468 0.4407 0.5386 0.6188 0.7471 ...
      0.8235 0.9344]';
yi = [9.68 11.92 28.61 32.42 44.63 51.47 60.91 75.12 79.79 93.80]';

xq = mean(xi);
yq = mean(yi);
B = ((xi - xq)'*(yi - yq))/((xi - xq)'*(xi - xq))
A = yq - B*xq
sYX = sqrt(1/(n-2)*sum((yi - A - B*xi).^2))

%% Varianzen
VA = (sYX^2/n)*sum(xi.^2)/sum((xi - xq).^2);
VB = sYX^2/sum((xi - xq).^2);
sigA = sqrt(VA)
sigB = sqrt(VB)
% relativ
sigrelA = sigA/A
sigrelB = sigB/B

%% Konfidenzintervalle
p = 0.05;
t = tinv(1 - p/2, n - 2)
GAu = A - t*sYX*sqrt(sum(xi.^2)/(n*sum((xi - xq).^2)))
GAo = A + t*sYX*sqrt(sum(xi.^2)/(n*sum((xi - xq).^2)))
GBu = B - t*sYX/sqrt(sum((xi - xq).^2))
GBo = B + t*sYX/sqrt(sum((xi - xq).^2))
chiu = chi2inv(p/2, n-2)
chio = chi2inv(1 - p/2, n-2)
GSu = (n-2)/chio*sYX^2
GSo = (n-2)/chiu*sYX^2
Gsigu = sqrt(GSu)
Gsigo = sqrt(GSo)

%% Test
b0 = 90;
p = 0.01;
c = tinv(1-p, n-2)
T = (B - b0)/sYX*sqrt(sum((xi - xq).^2))
pWert = 1 - tcdf(T, n-2)

%% Bestimmtheitsmaß
yhati = A + B*xi;
SQT = sum((yi - yq).^2)
SQE = sum((yhati - yq).^2)
SQR = sum((yhati - yi).^2)
R2 = SQE/SQT
```

bsp_6_4.m

```
% Beispiel für Kap 6.4
% Härte von Stahl
A = [
 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];

TC = A(:,1);     % Glühtemperatur in °C
qCu = A(:,2);    % Kupferanteil in %
hr = A(:,3);     % Rockwellhärte HR30T

% x0-Werte
qCu0 = 0.15;
TC0 = 620;

% multilinearer Fit
n = size(A,1);
k = size(A,2)-1;
Y = hr;
X = [ones(n,1), qCu, TC];
x0 = [1;qCu0; TC0];

B = (X'*X)\(X'*Y)
sYX = sqrt((Y'*Y - B'*X'*Y)/(n - k - 1))

% Schätzwerte und Konfidenzintervalle für Mittel- und Einzelwert
Y0 = x0'*B

p = 0.05;
t = tinv(1-p/2, n-k-1);
dGM = t*sYX*sqrt(x0'*((X'*X)\x0))
dG = t*sYX*sqrt(1 + x0'*((X'*X)\x0))
```

ranksumcdf.m

```
function p = ranksumcdf(k, n, m)
% computes the cdf needed for the Wilcoxon rank-sum test
% given two samples of size n and m with equal distribution
% and R1 := rank-sum of the first sample in the union of both samples,
%   ranksumcdf(k,n,m) = P(R1 <= k)
if n == 1 && m == 0
  if k <= 0
    p = 0;
  else
    p = 1;
  end
elseif n == 0 && m == 1
  if k < 0
    p = 0;
  else
    p = 1;
  end
elseif n == 0
  p = m/(n+m)*ranksumcdf(k, n, m-1);
elseif m == 0
  p = n/(n+m)*ranksumcdf(k-n-m, n-1, m);
else
  p = n/(n+m)*ranksumcdf(k-n-m, n-1, m) + m/(n+m)*ranksumcdf(k, n, m-1);
end
```

erzeugeZugversuchDaten.m

```
function [xi, yi] = erzeugeZugversuchDaten(n)
% erzeugt n Datensätze für das Regressionsbeispiel "Zugversuch"
alpha = 0;
beta = 100;
sigma = 2;
rng(59);                 % reproduzierbar machen
epsi = normrnd(0, sigma, n, 1);
xi = (1/(2*n):1/n:1)' + normrnd(0, 0.02, n, 1);
yi = alpha + beta*xi + epsi;
```