Diese Übungen sollen am **28. Jänner 2020** in den Übungsgruppen vorgestellt werden. Die Aufgaben, die Sie bereit sind in den Übungen an der Tafel zu präsentieren, müssen in TUWEL bis zum **27. Jänner 2020 um 23:30 Uhr** angekreuzt werden.

(1) **Kovarianz**
Es seien X, Y und Z Zufallsvariable mit endlicher Varianz und $a, b \in \mathbb{R}$ konstant. Die Kovarianz von X und Y ist gegeben durch $\mathbb{C}ov(X,Y) := \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$. Zeigen Sie

- (a) $\mathbb{C}ov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- (b) $\mathbb{C}ov(X,Y) = Cov(Y,X)$ (Symmetrie)
- (c) $\mathbb{C}ov(X,X) = \mathbb{V}ar(X)$
- (d) $\mathbb{C}ov(X,a) = 0$
- (e) $\mathbb{C}ov(aX+bZ,Y) = a\mathbb{C}ov(X,Y) + b\mathbb{C}ov(Z,Y)$ (Linearität)
- (f) Gilt die Linearität auch für die zweite Komponente?
- (g) $\mathbb{V}ar(X+Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y) + 2\mathbb{C}ov(X,Y)$
- (h) Falls X und Y unabhängig, so ist $\mathbb{C}ov(X,Y) = 0$. (Tipp: Hier dürfen Sie nutzen, dass für X und Y unabhängig gilt, dass $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.)

(2) **Korrelation (Teil 1)**
Es seien X und Y Zufallsvariable mit $\mathbb{V}ar(X), \mathbb{V}ar(Y) \in (0,\infty)$. Dann ist deren Korrelation gegeben durch

$$\mathbb{C}orr(X,Y) := \frac{\mathbb{C}ov(X,Y)}{\sqrt{\mathbb{V}ar(X)} \cdot \sqrt{\mathbb{V}ar(Y)}}.$$

- (a) Es seien X und Z unkorrelierte Zufallsvariable mit $\mathbb{V}ar(X) = \mathbb{V}ar(Z) = 1$, und es sei $k \in [-1,1]$. Was ist die Korrelation von X und $Y := kX + \sqrt{1-k^2}\,Z$?
  Wir bemerken, dass Y linear von X (und linear von Z) abhängt.
- (b) Für X gelte $\mathbb{E}[X] = \mu_x$ und $\mathbb{V}ar(X) = \sigma_x^2$. Konstruieren Sie in Anlehnung an Teil (a) eine Zufallsvariable Y, welche ebenfalls linear von X und linear von Z abhängt, aber vorgegebenes $\mu_y := \mathbb{E}[Y] \in \mathbb{R}$, $\sigma_y^2 := \mathbb{V}ar(Y) > 0$ und $\rho := \mathbb{C}orr(X,Y) \in [-1,1]$ besitzt.
  Tipp: Standardisieren Sie X.
- (c) Da Y aus (b) linear von X und linear von Z abhängt, ist es von der Form $Y = \beta_0 + \beta_1 X + \sigma Z$. Zeigen Sie, dass
  $$\beta_1 = \rho \cdot \frac{\sigma_y}{\sigma_x} \qquad \text{und} \qquad \beta_0 = \mu_y - \beta_1 \mu_x,$$
  und dass sich σ durch ρ und σ_y ausdrücken lässt.

(3) **Korrelation (Teil 2)**
Es seien $X \sim N(3,1)$ und $Y \sim N(2,4)$. Plotten Sie 100 Realisierungen von (X,Y) (y_i gegen x_i aufgetragen, $i = 1,\dots,100$) wobei

- (a) $\mathbb{C}orr(X,Y) = 0$
- (b) $\mathbb{C}orr(X,Y) = 0.8$
- (c) $\mathbb{C}orr(X,Y) = -0.5$
- (d) $\mathbb{C}orr(X,Y) = -1$

Fügen Sie zudem Geraden der Form $\{(x,y) \in \mathbb{R}^2 | y = \beta_0 + \beta_1 x\}$ mit β_0 und β_1 gemäß Aufgabe 2 Teil c ein. Berechnen Sie jeweils die empirische Korrelation r der Realisierungen und fügen Sie diese den plots bei.

(4) **Korrelation (Teil 3)**
Es sei $X \sim N(\mu_x, \sigma_x^2)$.
(a) Was ist die Korrelation von X und $Y := (X - \mu_x)^2$? (Tipp: Nutzen Sie, dass bei der Normalverteilung die ungeraden zentrierten Moment verschwinden, d.h., $\mathbb{E}[(X - \mathbb{E}[X])^k] = 0$ für $k = 1, 3, 5, \ldots$)
(b) Es sei $\mu_x = 2$ und $\sigma_x^2 = 4$. Generieren Sie 100 Realisierungen $(x_i)_{i=1,\ldots,100}$ von X und plotten Sie sämtliche $y_i = (x_i - \mu_z)^2$ gegen x_i. Geben Sie die empirische Korrelation r der $(x_i)_i$ und $(y_i)_i$ an. Passt ihr Ergebnis zu Teil (a)?
(c) Fügen Sie zudem die Gerade der Form $\{(x, y) \in \mathbb{R}^2 | y = \beta_0 + \beta_1 x\}$ mit β_0 und β_1 gemäß Aufgabe 2 Teil c ein. Würden Sie Sagen, dass die Unkorreliertheit die Unabhängigkeit impliziert? Was misst die Korrelation? Ergibt es Sinn diese Gerade hier einzuzeichnen?

(5) **Regression (Teil 1)**
Eine Vorlesung wurde evaluiert. In der Datei `Evaluation.Rdata` finden Sie Daten von $n = 25$ Studierenden. Erfasst wurden erstens, die in den angegliederten Übungen erzielten Übungspunkte (zwischen 0 und 200 möglich), und zweitens das Klausurergebnis (in %). Lässt sich das Klausurergebnis durch die erzielten Übungspunkte erklären?
(a) Plotten Sie die Klausurergebnisse (y_i) gegen die Übungspunkte (x_i). Erkennen Sie einen Zusammenhang?
(b) Berechnen Sie den y-Achsenabschnitt b_0 und die Steigung b_1 der Regressionsgeraden (ohne `lm()`) und zeichnen Sie Letztere ein. Kommentieren Sie die Bedeutung der Steigung.
(c) Würden Sie sagen, dass der Zusammenhang kausal ist? Das heißt, dass z.B. viele Übungspunkte der Grund für gutes Abschneiden in der Klausur sind?

(6) **Regression (Teil 2)**
Fortsetzung von Aufgabe 5: Nehmen Sie das lineare Regressionsmodell $Y_i = \beta_0 + \beta_1 x_i + \sigma Z_i$ $(i = 1, \ldots, n)$ an, mit $Z_1, \ldots, Z_n$ unabhängig und identisch verteilt und $Z_1 \sim N(0, 1)$, und testen Sie die Nullhypothese $H_0 : \beta_1 = 0$ (ohne `lm()`) zum Signifikanzniveau von 5%. Berechnen Sie dafür
(a) den Standardfehler der Regression,
(b) den Standardfehler von β_1,
(c) den Wert der t-Statistik,
(d) die mit H_0 assoziierte t-Verteilung,
(e) und den P-Wert.
(f) Formulieren Sie ein Ergebnis.

(7) **Regression (Teil 3)**
Führen Sie die Analyse aus Aufgabe 6 nun mit Hilfe des Befehls `lm()` durch.
(a) Passen Sie die Daten mit Hilfe des Befehls `lm()` an das Modell an. Lesen Sie den Ausgang des Tests der Nullhypothese $H_0 : \beta_1 = 0$ aus der Ausgabe von `summary()` ab.
(b) Plotten Sie (erneut) die Datenpunkte und die Regressionsgerade und diskutieren Sie, inwieweit die Modellannahmen plausibel sind.
(c) Welches Klausurergebnis prognostizieren Sie Studierenden, die 140 Übungspunkte erhalten haben? Markieren Sie die Prognose in der Grafik.
(d) Mit welchem Fehler müssen Sie bei der Prognose rechnen? Fügen Sie der Grafik zwei Geraden hinzu, welche parallel der Regressionsgeraden verlaufen, aber deren y-Achsenabschnitt dem der Regressionsgeraden um einen Standardfehler der Regression erhöht bzw. erniedrigt ist.