

Musterlösung

1. (10 Punkte) Ein Wissenschaftler behauptet, dass er ein einzelnes Gen gefunden hat, das einen kausalen Einfluss darauf hat, wie lange die Pflanzenart Arabidopsis blüht. Um dies zu überprüfen werden ihm fünf Paare von jeweils zwei genetisch identischen Pflanzen (Replikate) zur Verfügung gestellt. Jeweils bei einer der beiden schaltet er das besagte Gen aus ("Gen-Knockout") und vergleicht die Blütezeiten der Pflanzen (in Tagen).

Pflanze	1	2	3	4	5
Replikat 1 ohne Knockout	4.1	4.8	4.0	4.5	4.0
Replikat 2 mit Knockout	3.1	4.3	4.5	3.0	3.5

Nehmen Sie an, dass die Differenzen Blütezeit Replikat 1 minus Blütezeit Replikat 2 normalverteilt mit Erwartungswert μ und Varianz σ^2 sind.

- a) Es handelt sich um einen gepaarten Test. Warum?
 Jeder Pflanze mit Gen-Knockout kann eindeutig eine ohne Gen-Knockout zugeordnet werden: das jeweils andere Replikat. (1 Punkt)
- b) Geben Sie die Null- und die Alternativhypothese an und begründen Sie kurz Ihre Wahl.

$$H_0 : \mu = 0 \quad (0,5 \text{ Punkte})$$

$$H_A : \mu \neq 0 \quad (0,5 \text{ Punkte})$$

wobei μ der Erwartungswert von $D_i = X_i - Y_i$ ist mit X_i : Blütezeit der Pflanze ohne Gen-Knockout, Y_i : Blütezeit der Pflanze mit Gen-Knockout.

Das Blühverhalten einer Pflanze ist ein komplexer Vorgang, die Pflanze Arabidopsis hat zudem sehr viele verschiedene Gene. Wenn ein Wissenschaftler behauptet, er habe ein Gen mit kausalem Einfluss gefunden, möchte man sicher sein, dass seine Aussage auch stimmt. Die Wahrscheinlichkeit des Fehlers 'Das Gen hat keinen (messbaren) Einfluss, aber wir behaupten, es hat einen.' ist also zu beschränken. Daher wird dies der Fehler erster Art. Da der Wissenschaftler nicht behauptet, ob das Gen die Blütezeit verlängert oder verkürzt, benutzen wir einen zweiseitigen Test. (1 Punkt)

- c) Geben Sie eine Schätzung $\hat{\sigma}^2$ für die Varianz σ^2 der Differenz an (mit Lösungsweg). Mit $n = 5$ erhalten wir

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i = \frac{1}{5} (1 + 0.5 - 0.5 + 1.5 + 0.5) = 0.6$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2 \quad (0,5 \text{ Punkte}) \\ &= \frac{1}{4} ((0.4)^2 + (-0.1)^2 + (-1.1)^2 + (0.9)^2 + (-0.1)^2) \\ &= \frac{1}{4} (0.16 + 2 \cdot 0.01 + 1.21 + 0.81) = \frac{1}{4} 2.2 = 0.55 \quad (0,5 \text{ Punkte}) \end{aligned}$$

- d) Führen Sie den geeigneten t -Test zum Niveau 0.05 durch: Bestimmen Sie den Wert der Teststatistik T und dessen Verteilung unter der Nullhypothese, den Verwerfungsbereich für T und den Testentscheid. (Wenn Sie obige Aufgabe c) nicht lösen konnten, benutzen Sie im Folgenden $\hat{\sigma}^2 = 0.55$.)

1. Modell: $D_i \sim \mathcal{N}(\mu, \sigma^2)$ iid
2. Nullhypothese: $H_0 : \mu = 0$
 Alternative: $H_A : \mu \neq 0$

3. Teststatistik:

$$T = \frac{\sqrt{n}\bar{D}}{\hat{\sigma}} \quad (0,5 \text{ Punkte})$$

$$t = \frac{\sqrt{5} \cdot (0.6)}{\sqrt{0.55}} \approx 1.81 \quad (0,5 \text{ Punkte})$$

Verteilung von T unter H_0 : $T \sim t_{n-1} = t_4$. (0,5 Punkte)

4. Signifikanzniveau: $\alpha = 0.05$

5. Verwerfungsbereich:

$$K_1 = (-\infty, -t_{n-1;1-\alpha/2}] \cup [t_{n-1;1-\alpha/2}; \infty) \quad (0,5 \text{ Punkte})$$

$$= (-\infty, -t_{4;0.975}] \cup [t_{4;0.975}; \infty) = (-\infty, -2.776] \cup [2.776; \infty) \quad (0,5 \text{ Punkte})$$

6. Testentscheid: $t \notin K_1 \Rightarrow H_0$ wird nicht verworfen. (0,5 Punkte)

e) Bestimmen Sie ein zweiseitiges 95%-Vertrauensintervall für μ .

$$\begin{aligned} 0.95 &\leq \mathbf{P} \left(-t_{n-1;1-\alpha/2} \leq \sqrt{n} \frac{\bar{D} - \mu}{\hat{\sigma}} \leq t_{n-1;1-\alpha/2} \right) \\ &= \mathbf{P} \left(\bar{D} - \frac{\hat{\sigma} t_{n-1;1-\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{D} + \frac{\hat{\sigma} t_{n-1;1-\alpha/2}}{\sqrt{n}} \right) \\ \Rightarrow \text{VI} &= \left[\bar{D} - \frac{\hat{\sigma} t_{n-1;1-\alpha/2}}{\sqrt{n}} ; \bar{D} + \frac{\hat{\sigma} t_{n-1;1-\alpha/2}}{\sqrt{n}} \right] \quad (0,5 \text{ Punkte}) \\ &\approx [-0.32 ; 1.52] \quad (0,5 \text{ Punkte}) \end{aligned}$$

f) Kann der Wissenschaftler das Testergebnis durch mehrfaches Kopieren der Daten verändern, d.h. dadurch, dass er vorgibt insgesamt $5 \cdot k$ unabhängige Pflanzenpaare untersucht zu haben und jeden Datenpunkt k -fach notiert?

(Geben Sie eine kurze Erklärung für Ihre Antwort, in der Sie ohne Beweis erwähnen, was mit dem Mittelwert und der Standardabweichung geschieht. Sie dürfen benutzen, dass $t_{f;1-\alpha/2}$ mit steigendem f gegen $z_{1-\alpha/2}$ konvergiert.)

Der Mittelwert \bar{D}_k der Daten verändert sich nicht, (0,5 Punkte)

die Standardabweichung $\hat{\sigma}_k$ wird mit steigendem k leicht kleiner. (0,5 Punkte)

Die (nichtbenötigte) Begründung hierfür ist:

$$\hat{\sigma}_k^2 = \frac{1}{nk-1} k(n-1) \hat{\sigma}_1^2 \leq \hat{\sigma}_1^2$$

(Nebenbei: $\hat{\sigma}_k^2 \rightarrow \hat{\sigma}_1^2$ für $k \rightarrow \infty$.)

Die Teststatistik $T = \sqrt{kn} \frac{\bar{D}_k}{\hat{\sigma}_k}$ folgt einer t -Verteilung mit $k \cdot n - 1$ Freiheitsgraden, die demnach mit wachsendem k gegen $N(0,1)$ konvergiert. Dadurch wird die Teststatistik mit wachsendem k irgendwann grösser als $t_{k \cdot n - 1; 1 - \alpha/2}$ (0,5 Punkte)

und der Test wird verworfen. (0,5 Punkte)

(Nebenbei: Die Annahme, dass die Samples i.i.d. sind, ist beim k -fachen Kopieren der Daten sicherlich verletzt.)

2. a) Geben Sie die Verteilung von X inklusive Parameter. Wie gross ist die Wahrscheinlichkeit, dass das Lied von keinem der ersten fünf Kunden gekauft wurde?

$X \sim \text{Bin}(10, 0.4)$ (0.5 Punkte) und $0.6^5 = 0.07776$ (0.5 Punkte)

b) Wie gross ist die Wahrscheinlichkeit, dass das Lied genau von 3 Kunden gekauft wurde? Wie gross ist die Wahrscheinlichkeit, dass das Lied von höchstens 3 Kunden gekauft wurde?

$$P(X = 3) = \binom{10}{3} 0.4^3 * 0.6^7 = 0.215 \quad (0.5 \text{ Punkte})$$

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 0.6^{10} + \binom{10}{1} 0.4 * 0.6^9 + \binom{10}{2} 0.4^2 * 0.6^8 + 0.215 = 0.382 \quad (0.5 \text{ Punkte})$$

c) Was sind die Null- und die Alternativhypothese?

$H_0 : \pi = 0.4$ und $H_A : \pi > 0.4$ (1 Punkt)

- d) Führen Sie einen geeigneten exakten Test durch. Geben Sie den Verwerfungsbereich für X und den Testentscheid an.

(2 Punkte)

1. Model: $X \sim \text{Bin}(10, \pi)$ ist die Anzahl der Kunden die das neue Lied des Rappers 20Rappen an einem Tag hören und kaufen

2. Nullhypothese: siehe oben

3. Die Teststatistik ist: $T = X$, Verteilung von T unter H_0 : $T \sim \text{Bin}(10, 0.4)$

4. Signifikanzniveau: $\alpha = 0.05$

5. Verwerfungsbereich:

Da $P(X = 10) = 0.4^{10} = 0.0001$, $P(X \geq 9) = 0.4^{10} + 10 \cdot 0.6 \cdot 0.4^9 = 0.00167$, $P(X \geq 8) = 0.0123$, $P(X \geq 7) = 0.0547$ gilt, dass der Verwerfungsbereich $\{8, 9, 10\}$ ist. **(1 Punkt)**

6. Testentscheid: Da $6 \notin \{8, 9, 10\}$ wird nicht verworfen **(1 Punkt, mit Folgefehler)**.

- e) Wie gross ist die Macht des Tests bei konkreter Alternative $\pi = 0.5$?

$Y \sim \text{Bin}(10, 0.5)$ und die Macht ist $P(Y \geq 8) = 0.0546$ **(1 Punkt)**.

- f) MeTunes ist ebenfalls an dem Experiment interessiert und lädt 250 Kunden ein, das neue Lied von 20Rappen und Pihanna zu hören. Das Lied wurde von 108 Kunden gekauft. Führen Sie erneut einen Hypothesentest durch und berechnen Sie anschliessend das zweiseitige 95%-Intervall für π .

(2 Punkte)

Hypothesentest (da n gross ist verwenden wir die Normalapproximation)

1. Model: $X \sim \text{Bin}(250, \pi)$ ist die Anzahl der Kunden die das neue Lied des Rappers 20Rappen an einem Tag hören und kaufen

2. Nullhypothese: $H_0 : \pi = 0.4$ $H_A : \pi > 0.4$, oder $H_A : \pi \neq 0.4$

3. Die Teststatistik ist: $T = X$, Verteilung von T unter H_0 : $T \sim \text{Bin}(250, 0.4)$

4. Signifikanzniveau: $\alpha = 0.05$

5. Verwerfungsbereich: mit Normalapproximation für $H_A : \pi > 0.4$

$K_1 = \{c, \dots, n\}$ mit $c = 250 \cdot 0.4 + 1.64 \cdot \sqrt{250 \cdot 0.4 \cdot 0.6} \approx 112.7$. Also ist $K_1 = \{113, \dots, 250\}$.

Oder $K_2 = [1.64, \infty)$ ($1.64 = \Phi^{-1}(0.95)$), oder $K_3 = [0.451, \infty)$ ($0.451 = 0.4 + 1.64 \cdot \sqrt{0.4 \cdot 0.6} / \sqrt{250}$)

(1 Punkt)

Verwerfungsbereich: mit Normalapproximation für $H_A : \pi \neq 0.4$

$K_1 = (-\infty, 84] \cup [115, \infty)$ ($c_i = 250 \cdot 0.4 \pm 1.96 \cdot \sqrt{250 \cdot 0.4 \cdot 0.6}$), oder $K_2 = (-\infty, -1.96] \cup [1.96, \infty)$, oder $K_3 = (-\infty, 0.34] \cup [0.461, \infty)$ ($c_i = 0.4 \pm 1.96 \cdot \sqrt{0.4 \cdot 0.6} / \sqrt{250}$) **(1 Punkt)**

6. Testentscheid: Da $t = 108 \notin K_1$ wird H_0 nicht verworfen, oder da $z = \frac{108 - 250 \cdot 0.4}{\sqrt{250 \cdot 0.4 \cdot 0.6}} = 1.032 \notin K_2$

wird H_0 nicht verworfen, oder da $z = \frac{108}{250} = 0.432 \notin K_3$ wird H_0 nicht verworfen **(0.5 Punkte)**.

Das zweiseitige 95%-Vertrauensintervall ist:

$$I \approx \left[\frac{108}{250} - 1.96 \sqrt{\frac{108}{250} \left(1 - \frac{108}{250}\right) \frac{1}{250}}, \frac{108}{250} + 1.96 \sqrt{\frac{108}{250} \left(1 - \frac{108}{250}\right) \frac{1}{250}} \right] \approx [0.37, 0.49]$$

(0.5 Punkte).

3. **(7 Punkte)** In Zürich wird untersucht, inwiefern die Anzahl Liter verkauften Glühweins (wine) durch die Aussentemperatur (temp) und den Preis (price) beeinflusst wird. Es wurden 30 Messungen gemacht und folgendes Modell angepasst:

$$\text{wine}_i = \beta_0 + \beta_1 \cdot \text{temp}_i + \beta_2 \cdot \text{price}_i + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, 30.$$

Der (unvollständige) Regressionsoutput sieht wie folgt aus:

Residuals:

Min	1Q	Median	3Q	Max
-1.6759	-0.1504	-0.0064	0.1704	1.3397

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.0122	???	-0.11	0.91
temp	-1.9600	0.1012	-19.37	<2e-16 ***
price	9.5989	11.0057	???	???

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.551 on ?? degrees of freedom

Multiple R-squared: 0.938, Adjusted R-squared: 0.934

F-statistic: 205 on 2 and ?? DF, p-value: <2e-16

- 1) Was ist der Standardfehler von $\hat{\beta}_0$? (auf zwei signifikante Stellen genau)
 - a) -0.013
 - b) 0.33
 - c) 0.11
 - d) 0.87

c (0.11)
- 2) Wie gross ist die Schätzung von σ^2 ?
 - a) 0.551
 - b) 0.3
 - c) 0.74
 - d) 0.88

b (0.3)
- 3) Mit wievielen Freiheitsgraden wurde der "residual standard error" berechnet?
 - a) 28
 - b) 3
 - c) 2
 - d) 27

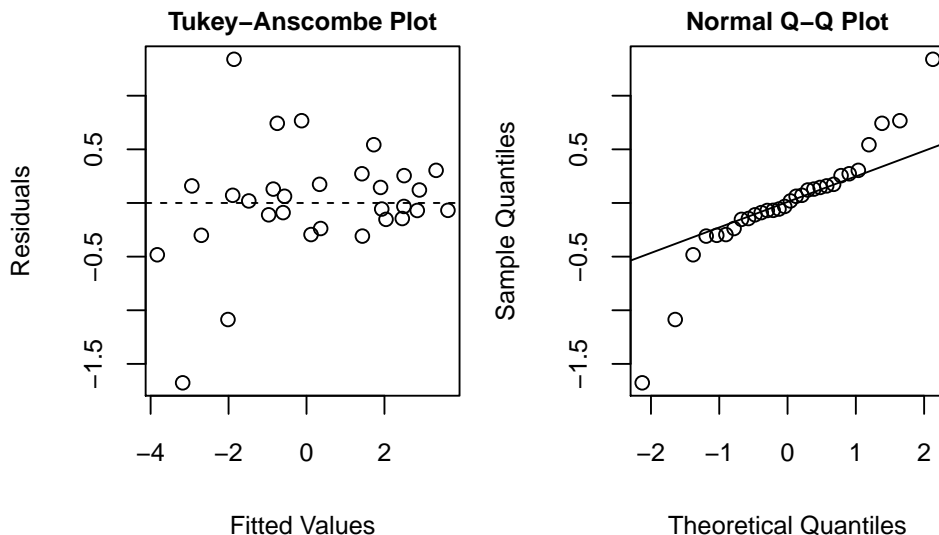
d (27)
- 4) Berechnen Sie das exakte zweiseitige 95%-Konfidenzintervall für β_1 . (auf zwei signifikante Stellen genau)
 - a) $[-2.13, -1.78]$
 - b) $[-12.98, 32.18]$
 - c) $[-2.17, 1.75]$
 - d) $[-2.17, -1.75]$

d ($[-2.17, -1.75]$)
- 5) Wird die Nullhypothese $H_0 : \beta_2 = 0$ auf dem 1%-Niveau verworfen?
 - a) Ja.
 - b) Nein.
 - c) Keine Angabe möglich.

b (Nein. Der Verwerfungsbereich für β_2 ist $K = (\infty, -2.771] \cup [2.771, \infty)$ und $T = \frac{9.59}{11} = 0.87$ ist nicht in K .)
- 6) Eine Person verkauft Glühwein zum Preis von 3 CHF. Wie soll die Aussentemperatur nach obigem Modell sein, wenn sie 18 Liter Glühwein zu verkaufen erwartet?
 - a) -6.5
 - b) 4.3
 - c) 5.5
 - d) 2.5

c (5.5)
- 7) Betrachten Sie die nachfolgenden Plots. Welche der folgenden Aussagen trifft zu?
 - a) Alle Modellannahmen sind erfüllt.
 - b) Die Fehlervarianz ist nicht konstant, aber die Normalverteilungsannahme ist plausibel.
 - c) Die Fehlervarianz ist konstant, aber die Normalverteilungsannahme trifft nicht zu.
 - d) Sowohl konstante Fehlervarianz als auch Normalverteilungsannahme treffen nicht zu.

d (Sowohl konstante Fehlervarianz als auch Normalverteilungsannahme treffen nicht zu.)



4. (8 Punkte) Die folgenden Aufgaben sind zufällig angeordnet und insbesondere nicht nach Schwierigkeitsgrad sortiert.

1) Es gilt für ein Ereignis A , dass $\text{odds}(A) = 3$. Was ist die Wahrscheinlichkeit für A ?

- a) $1/3$
- b) 0.5
- c) 0.25
- d) 0.75

d), denn $P(A) = \text{odds}(A)/(1 + \text{odds}(A))$.

2) Die geometrische Verteilung beschreibt die Anzahl Versuche x , die nötig sind, um einen Erfolg zu haben bei unabhängigen Bernoulliversuchen mit Erfolgswahrscheinlichkeit p (z.B. bei einem Münzwurf). Dies ist eine diskrete Verteilung mit Wahrscheinlichkeitsfunktion

$$P(X = x) = p(1 - p)^{x-1}, \quad x = 1, 2, 3, \dots$$

wobei $0 < p < 1$ die Erfolgswahrscheinlichkeit ist.

Zwei Personen ziehen unabhängig voneinander an einer Losbude so lange ein Los, bis sie den ersten Treffer haben. Dabei ist $x_1 = 5$ und $x_2 = 7$. Was ist der Maximum Likelihood Schätzer für p ?

- a) 0.083
- b) 0.833
- c) 0.171
- d) 0.166

d). Die Log-Likelihood ist

$$\log(p; x_1, \dots, x_n) = \sum_i \log(p) + (x - 1) \log(1 - p) = n \log(p) + \log(1 - p) \cdot \sum_i x_i - n \log(1 - p).$$

(1 Punkt) Ableiten und nullsetzen gibt

$$-\frac{\sum_i x_i - n}{1 - p} + \frac{n}{p} = 0,$$

d.h.

$$\hat{p}_{MLE} = \frac{n}{\sum_i x_i}.$$

- 3) Es gelten für zwei Ereignisse A und B , dass $P(A|B) = 0.2$, $P(A) = 0.5$ und $P(B) = 0.4$. Wie gross ist $P(B|A)$?
- 0.25
 - 0.16
 - 0.08
 - 0.04

b), denn $P(B|A) = P(A|B)P(B)/P(A)$.

- 4) Für zwei unabhängige Zufallsvariablen X und Y gelte $\text{Var}(X) = 3$ und $\text{Var}(Y) = 2$. Welchen Wert hat $\text{Var}(3X - 2Y + 5)$?
- 35
 - 60
 - 17
 - 22

a), denn $3^2 \cdot 3 + 2^2 \cdot 2 = 35$.

- 5) Ein Casino kauft 100 neue Spielautomaten. Pro Abend werden ca. 100 Spiele pro Automat gemacht. Sowohl einzelne Spiele als auch verschiedene Automaten sind unabhängig. Pro Spiel und Automat folgt der Gewinn folgender Verteilung:

Gewinn	-1	0	1
Wahrscheinlichkeit	0.3	0.2	0.5

Gib ein approximatives 95% Vertrauensintervall für den Gewinn, der pro Abend durch die neuen Maschinen entsteht, an.

- [2829.1, 3170.9]
- [1829.1, 2170.9]
- [1856.6, 2143.4]
- [1998.3, 2001.7]

b). Es ist $\mu = -1 \cdot 0.3 + 0 \cdot 0.2 + 1 \cdot 0.5 = 0.2$ und $\sigma^2 = E[X^2] - E[X]^2 = 0.8 - 0.2^2 = 0.76$. Es gilt $S_n \approx N(n\mu, n\sigma^2)$, also ist

$$[n\mu - \sqrt{n\sigma^2}\Phi^{-1}(0.975), n\mu + \sqrt{n\sigma^2}\Phi^{-1}(0.975)] \approx [1829.1, 2170.9]$$

ein approximatives 95% Vertrauensintervall.

- 6) Bei fixem α hat ein einseitiger Test verglichen mit dem entsprechenden zweiseitigen Test
- einen kleineren Fehler 1. Art
 - ein kleineres Signifikanzniveau
 - einen grösseren Fehler 2. Art
 - mehr Macht
- d).

- 7) Es sei $X \sim N(5, 3^2)$. Was ist $P(X \leq 8)$?

- 0.54
- 0.63
- 0.84
- 1

c).

- 8) Ein Test mit einer Nullhypothese H_0 ergibt einen p -Wert von 0.01. Welche der folgenden Aussagen trifft zu:
- Die Nullhypothese H_0 ist mit 1% Wahrscheinlichkeit richtig.
 - Die Alternative ist mit 99% Wahrscheinlichkeit richtig.
 - Die Alternative wird mit 1% Wahrscheinlichkeit fälschlicherweise angenommen.
 - Keine der obigen Aussagen ist richtig.

d). Der p -Wert ist die Wahrscheinlichkeit, eine Teststatistik zu erhalten, welche mindestens so extrem ist, unter der Annahme die Nullhypothese H_0 sei richtig.