Bachelorprüfung: Statistik Musterlösung

1. a) **(Total 1P)** Der Test muss gepaart sein, da zu jedem Läufer genau ein Wert mit jedem Schuhtyp vorliegt und die beiden Werte eines Läufers von dessen Lauffähigkeiten abhängen.

b) **(Total 1P)** Nullhypothese: $H_0 : \mu_x = \mu_y$ **(1/2P)**
Alternativhypothese: $H_A : \mu_x < \mu_y$ **(1/2P)**

c) **(Total 3.5 P)**

1. **Modell**: Differenzen D_i i.i.d. $\sim \mathcal{N}(\mu = \mu_X - \mu_Y, \sigma^2)$, $i = 1, \ldots, n$.
2. **Nullhypothese** $H_0 : \mu_X = \mu_Y$, d.h. $\mu = 0$;
   Alternativhypothese $H_A : \mu_X < \mu_Y$, d.h. $\mu < 0$.
3. **Teststatistik:** $T = \dfrac{\sqrt{n}\hat{\mu}}{\hat{\sigma}}$ **(1 P)**
   Verteilung der Teststatistik unter H_0**:** $T \sim t_9$
4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich:**
   $K = (-\infty, -t_{9,1-\alpha}) = (-\infty, -1.833)$ **(1 P)**
6. **Testentscheid:** $t = \dfrac{\sqrt{10} \cdot (-0.22)}{0.26} = -2.68 \in K$ **(0.5 P)**, also wird H_0 verworfen **(1 P)**

d) **(Total 2.5 P)** Das einseitige Vertrauensintervall ist gegeben durch

$$\left(-\infty, \bar{x} - \bar{y} + t_{9,1-\alpha}\frac{\hat{\sigma}}{\sqrt{n}}\right) = \left(-\infty, -0.22 + 1.833 \cdot \frac{0.26}{\sqrt{10}}\right) = (-\infty, -0.069).$$

2. a) **(1 Punkt)** Anzahl gewonnener Pokerrunden durch Markus X ist $\mathrm{Bin}(n, \pi)$.

b) **(1 Punkt)** Daher ist

$$\mathbf{E}[X] = n\pi = 20 * 0.2 = 4 \ (\mathbf{1/2P})$$

und

$$\mathrm{Var}(X) = n\pi(1-\pi) = 20 * 0.2 * 0.8 = 3.2, \ (\mathbf{1/2P})$$

d.h. es werden 4 Siege von Markus erwartet.

c) **(2 Punkte)** Verwende Zentralen Grenzwertsatz:

$$Y = \sum_{k=1}^{100} X_i \ (\mathbf{1/2P})$$

$$Y \sim \mathcal{N}\left(20, 4^2\right) \ (\mathbf{1/2P})$$

Daher ist

$$\begin{aligned} \mathrm{P}[Y >= 28] &= \mathrm{P}\left[\frac{Y-20}{4} >= 2\right] = 1 - \mathrm{P}[Z < 2] \ (\mathbf{1/2P}) \\ &= 1 - \mathrm{P}[Z \leq 2] = 1 - 0.977 = 0.022 = 2.2\% \ (\mathbf{1/2P}) \end{aligned} \quad (1)$$

d)
1. **Modell**: Y: Anzahl Siege von Markus; $Y \sim \mathrm{Binomial}(n, \pi)$ mit $n = 100$
2. **Nullhypothese** $H_0 : \pi = \pi_0 = 0.2$ (Markus spielt gleich gut wie die andern); **Alternativhypothese** $H_A : \pi > 0.2$ (Markus spielt besser). **(0.5 P)**
3. **Teststatistik:** $Z = \dfrac{Y - n\pi}{\sqrt{n\pi(1-\pi)}}$ **(0.5 P)**
   Verteilung der Teststatistik unter H_0**:** $Z \approx \mathcal{N}(0, 1)$
4. **Signifikanzniveau:** $\alpha = 0.01$
5. **Verwerfungsbereich:**
   $K = [\Phi^{-1}(1-\alpha), \infty) = [2.326, \infty)$ **(0.5 P)**
6. **Testentscheid:** $Z = 2.000 \notin K$, daher wird die Nullhypothese *nicht* verworfen. **(0.5 P)**

Das 99%-Quantil der Standard-Normalverteilung ist 2.326. Daher ist das 99%-Quantil der Verteilung von Y unter der Nullhypothese gleich $20 + 2.3 \cdot 4 = 29.3$: Markus müsste also mindestens 30 Mal gewinnen, damit die Nullhypothese verworfen würde. **(0.5 P)**

e) **(2 Punkte)** Die geschätzte Gewinnwahrscheinlichkeit von Markus ist

$$\hat{\pi} = \frac{X}{n} = 0.28 \ \textbf{(1P)}$$

Das zweiseitige 95% Vertrauensintervall von π ist

$$[\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}] = [0.28 \pm 1.96\sqrt{0.28 * (1 - 0.28)/100}] = [0.192, 0.368] \ (\mathbf{1P})$$

3.

1) **(1 Punkt)** Die Anzahl Beobachtungen n ist gegeben durch die Summe der Anzahl Freiheitsgrade und der Anzahl Parameter $n = n - p + p = 9 + 2 = 11$. Somit ist Antwort c) korrekt.
2) **(1 Punkt)** Der Intercept lässt sich schnell aus den Werten im angegebenen R-Output berechen. Man multipliziert die t-Teststatistik mit der Standardabweichung und erhält so den Intercept. Dass heisst, $51.25 \cdot -1.831 \approx -93.8$. Demnach ist Antwort b) richtig.
3) **(1 Punkt)** Um diese Frage zu beantworten betrachtet man den entsprechenden P-Wert. Ist der P-Wert kleiner als 0.05 wird die Nullhypothese auf dem 5% Niveau verworfen. Der P-Wert für $H_0 : \beta = 0$ und $H_A : \beta \neq 0$ ist, wie man aus dem R-Output erkennt, $0.00249 < 0.05$. Somit ist Antwort a) richtig.
4) **(1 Punkt)** Die t-Teststatistik lässt sich schnell aus den Werten im angegebenen R-Output berechen. Die t-Teststatistik erhält man nämlich, indem man den geschätzten Wert durch seine Standardabweichung teilt. Somit ergiebt sich $\frac{59.05}{14.24} \approx 4.15$. Richtig ist also Antwort d).
5) **(1 Punkt)** Das zweiseitige Vertrauensintervall für β ist gegeben durch $\hat{\beta} \pm t_{9;0.975} \cdot \hat{\sigma}_\beta$. Das Quantil $t_{9;0.975}$ lässt sich in der Tabelle der t-Verteilung nachschauen, es gilt $t_{9;0.975} = 2.262$. Somit ist Antwort d) korrekt.
6) **(1 Punkt)** Der Tukey-Anscombe Plot zeigt, dass die Annahme der konstanten Varianz der Fehler ϵ_i verletzt ist. Somit ist ist Aussage c) zutreffend.
7) **(1 Punkt)** Durch das hinzufügen einer Beobachtung an der Stelle $(2.5, 150)$ wird die Regressionsgerade flacher, dass bedeutet $\hat{\beta}$ wird kleiner. Weiter wird $\hat{\sigma}$ grösser, da durch das Hinzufügen der Beobachtung die Residuen mehrheitlich dem Betrage nach grösser werden und somit auch $\hat{\sigma}$. Das heisst Antwort a) ist korrekt.
8) **(1 Punkt)** Bei der einfachen linearen Regression erhält man die empirische Korrelation zwischen erklärender Variable und Zielvariable aus der Wurzel von R^2. Somit ist $\hat{\rho} = \sqrt{R^2} = \sqrt{0.6566} \approx 0.81$, also ist Antwort e) korrekt.
9) **(1 Punkt)** Die Nullhypotesen $H_{0,1} : \beta_1 = 0$ und $H_{0,2} : \beta_2 = 0$ werden nicht verworfen. Aber der F-Test hat einen kleinen P-Wert. Somit ist Antwort a) richtig.
10) **(1 Punkt)** In der multiplen Regression lässt sich R^2 nicht mehr als quadratische Stichprobenkorrelation zwischen erklärenden Variablen und Zielvariablen interpretieren. Somit ist Antwort b) richtig.

Übersicht

1) c)
2) b)
3) a)
4) d)
5) d)
6) c)
7) a)
8) e)

9) a)
10) b)

4. 1) b)

2) f)
$X + Y = -2 + \frac{5}{4}X$. Also gilt $\mathbf{E}[X + Y] = -2 + \frac{5}{4}2 = 0.5$.

3) b)
Siehe 2: $X + Y = -2 + \frac{5}{4}X$. Also ist $\mathrm{Var}(X + Y) = \frac{25}{16}\,\mathrm{Var}(X) = \frac{25}{16}$. Bem.: X und Y sind nicht unabhängig, daher darf man die Varianzen nicht addieren!

4) c)
Zentraler Grenzwertsatz: Summe von vielen Zufallsvariablen ist (approx.) normalverteilt.

5) a)
Da die Teststatistik grösser 0 ist (d.h. sie "liegt auf der richtigen Seite") ist der p-Wert 0.01.