Institut für Angewandte Mathematik
Sommersemester 2022
Andreas Eberle, Luis Aniello La Rocca

1. Klausur „Einführung in die Statistik"

Musterlösung zur Klausur

---

1. (Einige kurze Fragen)

a)

$$(i) \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad (ii) \quad \hat{\rho}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

b) Der Getrimmte Mittelwert, der Median und der Interquartilsabstand sind robust.

c) Sei $\Theta \subseteq \mathbb{R}$ die Menge aller möglichen Parameter, dann ist

$$f(\vartheta|x) = \frac{f(\vartheta)f(x|\vartheta)}{\int_{\Theta} f(\vartheta')f(x|\vartheta')\mathrm{d}\vartheta'}$$

die Dichte der a posteriori Verteilung von ϑ gegeben x.

d) Die Funktion $p \mapsto F_{n,p}(c)$ ist monoton fallend und $p \mapsto G_{n,p}(u)$ ist monoton wachsend.

2. (Gauß-Modell)

a) Ein Konfidenzintervall (KI) für m zum Konfidenzniveau $\alpha \in (0,1)$ ist ein reelles Intervall $I(X_1,\ldots,X_n) = [a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]$ (oder offen) mit $a, b\colon \mathbb{R}^n \to \mathbb{R}$ messbar, für das gilt:

$$\mathbb{P}_{m,v}\left[m \in I(X_1,\ldots,X_n)\right] \geq \alpha$$

für alle $m \in \mathbb{R}$ (und für alle $v > 0$ falls v unbekannt ist).

b) Es gilt

$$\mathbb{E}_{m,v}\left[\bar{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{m,v}\left[X_i\right] = m \quad \text{und} \quad \mathrm{Var}_{m,v}\left[\bar{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{m,v}\left[X_i\right] = \frac{v}{n}.$$

Also gilt $\bar{X}_n \sim \mathcal{N}(m, \frac{v}{n})$ unter $\mathbb{P}_{m,v}$ als Linearkombination von unabhängigen Gaußschen Zufallsvariablen. Beziehungsweise $\frac{\bar{X}_n - m}{\sqrt{v/n}} \sim \mathcal{N}(0,1)$. Damit gilt weiter

$$\begin{aligned}
\mathbb{P}_{m,v}\left[m \notin \left(\bar{X}_n \pm \sqrt{\frac{8}{n}}\right)\right] &= \mathbb{P}_{m,v}\left[\left|\bar{X}_n - m\right| \geq \sqrt{\frac{8}{n}}\right] \\
&= \mathbb{P}_{m,v}\left[\Big|\underbrace{\frac{\bar{X}_n - m}{\sqrt{v/n}}}_{\sim\mathcal{N}(0,1)}\Big| \geq \sqrt{\frac{8}{v}}\right] \\
&= 2\left(1 - \Phi\left(\sqrt{\frac{8}{v}}\right)\right) \\
&= 2\left(1 - \Phi(2)\right) < 5\%
\end{aligned}$$

wobei im vorletzten Schritt $v = 2$ benutzt wurde, und für die letzte Abschätzung verwendet wurde, dass bei der Normalverteilung 95% der Masse in einer 2σ-Umgebung um den Erwartungswert liegt.

c) Definiere die Statistik T als

$$T(X) = \frac{\bar{X}_n - m}{\sqrt{V/n}} \quad \text{, mit } V = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Dann ist $T(X) \sim t(n-1)$. Also ist

$$\mathbb{P}_{m,v}\left[\left|\bar{X}_n - m\right| \geq c\sqrt{\frac{V}{n}}\right] = \mathbb{P}_{m,v}\left[\left|T(X)\right| \geq c\right] = 2\mathbb{P}[T(X) \geq c] \leq 0,1$$

falls $c = q_{t(n-1);0,95}$. Also ist das KI gegeben durch

$$\left(\bar{X}_n - q_{t(n-1);0,95}\sqrt{V/n}\,,\ \bar{X}_n + q_{t(n-1);0,95}\sqrt{V/n}\right).$$

d) Wir haben $n = 8$ und berechnen

$$\begin{aligned}
\bar{x} &= \frac{0 + 3 - 4 + 1 + 1 + 4 - 2 - 3}{8} = 0 \\
s_x^2 &= \frac{1}{7}\underbrace{(3^2 + 4^2 + 1^2 + 1^2 + 4^2 + 2^2 + 3^2)}_{32+18+4+2} = \frac{56}{7} = 8
\end{aligned}$$

Dann ist das KI aus (b) gegeben durch

$$\left(\bar{x} - \sqrt{\frac{8}{n}}, \bar{x} + \sqrt{\frac{8}{n}}\right) = (\bar{x} - 1, \bar{x} + 1) = (-1, 1)$$

und das KI aus (c) ist gegeben durch

$$\begin{aligned}
\left(\bar{x} - q_{t(7);0,95}\sqrt{s_x^2/8}, \bar{x} + q_{t(7);0,95}\sqrt{s_x^2/8}\right) &= \left(\bar{x} - 1,895\sqrt{8/8}, \bar{x} + 1,895\sqrt{8/8}\right) \\
&= (-1,895, 1,895).
\end{aligned}$$

e) Das Intervall ist größer, weil die tatsächliche Stichprobenvarianz $\sqrt{8} = 2\sqrt{2} \approx 2,83$ größer als die in b) angenommene Varianz 2 ist. Außerdem vergrößert das Schätzen der Varianz das KI, da der Schätzfehler berücksichtigt werden muss.

3. (Pferderennen)

a) Das Modell ist gegeben durch

$$(H_1, H_2, \ldots, H_8) \sim \text{Mult}(n; p_1, \ldots, p_8) \quad \text{mit } n = 144$$
$$\Theta = WV(\{1, \ldots, 8\}) = \left\{ p \in \mathbb{R}^8 \colon p_i \geq 0 \, \forall i = 1, \ldots, 8, \sum_{i=1}^{8} p_i = 1 \right\}.$$

Die Nullhypothese ist, dass die Startposition keinen Einfluss auf die Gewinnchancen hat. Also

$$H_0 \colon p = \left(\frac{1}{8}, \frac{1}{8}, \ldots, \frac{1}{8} \right).$$

b) Wir testen die Hypothese mit Hilfe des χ^2-Tests. Die Teststatistik ist gegeben durch

$$T(H) = \sum_{i=1}^{8} \frac{(H_i - np_i^0)^2}{np_i^0}.$$

Dann ist für große n die Statistik $T(H) \approx \chi^2(8-1)$. Entsprechend verwerfen wir H_0 falls $T(H) > q_{1-\alpha,\chi^2(8-1)}$. Hier ist

$$\begin{aligned} q_{0,95;\chi^2(7)} &= 14,07 \\ np_i^0 &= \frac{144}{8} = 18 \quad \forall i = 1, \ldots, 8 \\ T(H) &= \frac{1}{18} \sum_{i=1}^{8} (H_i - 18)^2 = \frac{11^2 + 1^2 + 0^2 + 7^2 + 1^2 + 8^2 + 3^2 + 7^2}{18} \\ &= \frac{294}{18} = \frac{147}{9} = \frac{49}{3} = 16,\bar{3} > 14,07 \end{aligned}$$

Der χ^2-Test verwirft also H_0 zum Niveau 5%.

c) In einem linearen Regressionsmodell sind die Schätzwerte für die Koeffizienten der Regressionsgerade $y_i = a + bx_i + \xi_i$ gegeben durch $\hat{a} = 27,5357$ und $\hat{b} = -2,119$. Der Schätzwert für die Steigung ist mit p-Wert 1,69% signifikant kleiner als 0. Also können wir die Nullhypothese $H_0 \colon b = 0$ bei Annahme einer Normalverteilung z.B. zum Niveau 5% (oder sogar 2%) verwerfen und davon ausgehen, dass Startposition und Siegchance negativ korreliert sind.

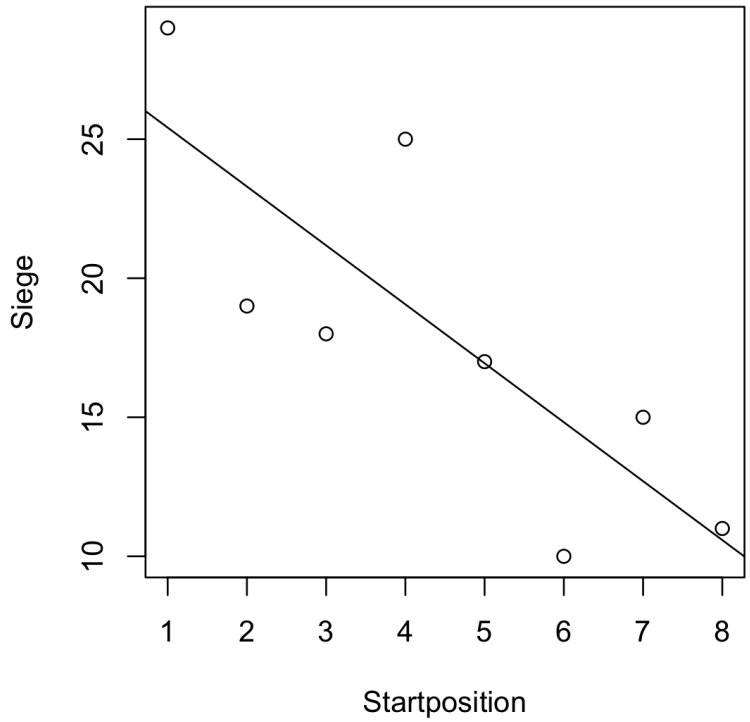

Abbildung 1: Plot aus Aufgabe 3c) Teil (i)

4. (Chancenquotienten)
Variante 1:

a) Seien $H_{1,+}, H_{2,+}$ fest. Dann ist ein statistisches Modell gegeben durch

$$\begin{aligned} H_{1,1} &\sim \mathrm{Bin}(H_{1,+}, p_1) \quad \text{und} \quad H_{1,2} = H_{1,+} - H_{1,1} \\ H_{2,1} &\sim \mathrm{Bin}(H_{2,+}, p_2) \quad \text{und} \quad H_{2,2} = H_{2,+} - H_{2,1} \end{aligned}$$

und $H_{1,1}, H_{2,1}$ unabhängig. Der Chanchenquotient ist gegeben durch

$$\rho = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

b)

$$\hat{\rho} = \frac{H_{1,1}H_{2,2}}{H_{1,2}H_{2,1}} = \frac{4 \cdot 77}{4 \cdot 13} = \frac{77}{13} \approx 5,9$$

c) $H_0 \colon p_1 = p_2$ und $H_1 \colon p_1 > p_2$.
Beziehungsweise $H_0 \colon \rho = 1$ und $H_1 \colon \rho > 1$.

Variante 2:

a) Es seien $(H_{1,1}, H_{12}, H_{2,1}, H_{2,2}) \sim \mathrm{Mult}(N, p_{11}, p_{12}, p_{21}, p_{22})$. Der Chanchenquotient ist gegeben durch

$$\rho = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

b)
$$\hat{\rho} = \frac{H_{1,1}H_{2,2}}{H_{1,2}H_{2,1}} = \frac{4 \cdot 77}{4 \cdot 13} = \frac{77}{13} \approx 5,9$$

c) $H_0 \colon \rho = 1$ und $H_1 \colon \rho > 1$.

d) Unter H_0 gilt $H_{1,1} \sim \mathrm{Hyp}(N, H_{+,1}, H_{1,+})$ gegeben $H_{+,1}$ und $H_{1,+}$. Dementsprechend: Unter H_0 hat jeder Arbeitnehmer dieselbe Wahrscheinlichkeit für Discushernien. $H_{1,1}$ erhalten wie folgt: Von den N befragten Arbeitnhmern haben $H_{+,1}$ Discushernien. Wir wählen nun $H_{1,+}$ von den N aus (die Kraftfahrer). Dann ist $H_{1,1}$ die Anzahl von Discushernien unter den ausgewählten, also
$$H_{1,1} \Big| (H_{1,+}, H_{+,1}) \sim \mathrm{Hyp}(N, H_{+,1}, H_{1,+})$$

e) Test: Verwerfe H_0 falls $H_{11} > c$
$$\begin{aligned} p\text{-Wert} &= \mathbb{P}_0[H_{1,1} \geq 4 | H_{1,+} = 8, H_{+,1} = 17] \\ &= \mathrm{Hyp}(98, 17, 8)[\{4, 5, 6, 7, 8\}] \end{aligned}$$

5. (Likelihood)

a)
$$L(p; x) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$$

b) Die Statistik $S(X) = \sum_{i=1}^{n} X_i$ ist suffizent, da die Likelihood eine Funktion von dieser Statistik und p ist. Ebenso ist $\bar{X}_n = \frac{1}{n} S(X)$ suffizient.

c) Bestimme die log-Likelihood-Funktion
$$\log L(p; x) = \sum_{i=1}^{n} x_i \cdot \log(p) + \left(n - \sum_{i=1}^{n} x_i \right) \cdot \log(1-p)$$
Die erste Ableitung nach p ist gegeben durch
$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}p} \log L(p; x) &= \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} x_i \right) \\ &= \left(\frac{1}{p} - \frac{1}{1-p} \right) \sum_{i=1}^{n} x_i - \frac{n}{1-p} \\ &= \frac{n}{p(1-p)} \bar{x}_n - \frac{n}{1-p} \end{aligned}$$
Die Ableitung verschwindet für $p = \bar{x}_n$. Außerdem ist die Ableitung für $p < \bar{x}_n$ strikt positiv und für $p > \bar{x}_n$ strikt negativ. Also ist $\bar{x}_n$ eindeutiges Maximum der Likelihood-Funktion.

d) Berechne für $p \in [0,1]$

$$\mathbb{E}_p[\hat{p}] = \mathbb{E}_p[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}_p[X_i]}_{=p} = p$$

$$\mathrm{Var}_p[\hat{p}] = \mathrm{Var}_p[\bar{X}_n] = \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{\mathrm{Var}_p[X_i]}_{p(1-p)} = \frac{p(1-p)}{n}$$

wobei zur Berechnung der Varianz die Unabhängigkeit der Zufallsvariablen genutzt wurde. Also ist der Schätzer erwartungstreu.

e) Informationsungleichung: Für jeden erwartungstreuen Schätzer $T(X)$ für p gilt unter Regularitätsvoraussetzungen (s.u.)

$$\mathrm{Var}_p\left[T(X)\right] \geq \frac{1}{I(p)}$$

Dabei ist

$$\begin{aligned} I(p) &= \mathbb{E}_p\left[\Big| \underbrace{\frac{\partial}{\partial p}\log L(p;X)}_{\frac{n}{p(1-p)}(\bar{X}_n - p)} \Big|^2\right] \\ &= \left(\frac{n}{p(1-p)}\right)^2 \underbrace{\mathbb{E}_p\left[(\bar{X}_n - p)^2\right]}_{=\mathrm{Var}[\bar{X}_n - p] = \frac{p(1-p)}{n}} \\ &= \frac{n}{p(1-p)}. \end{aligned}$$

Also gilt

$$\mathrm{Var}_p\left[T(X)\right] \geq \frac{p(1-p)}{n} = \mathrm{Var}_p[\hat{p}].$$

Geeignete Regularitätsannahmen sind zum Beispiel (mit $\theta = p$ und $f_\theta(x) = L(p;x)$):

Annahme (Regularität): Die folgenden Bedingungen sind erfüllt:

- *Nicht-degeneriert:* Für alle $x \in S$ und $\theta \in \Theta$ ist $f_\theta(x) > 0$.
- *Glattheit der Likelihood:* Für jedes $x \in S$ ist $\theta \mapsto f_\theta(x)$ stetig differenzierbar.
- *Vertauschen von Ableitung und Integral:* Es gilt

$$\begin{aligned} \int \frac{\partial}{\partial\theta} f_\theta(x)\,dx &= \frac{\partial}{\partial\theta}\int f_\theta(x)\,dx = 0, \\ \int T(x)\,\frac{\partial}{\partial\theta} f_\theta(x)\,dx &= \frac{\partial}{\partial\theta}\int T(x)\,f_\theta(x)\,dx. \end{aligned}$$