Institut für Angewandte Mathematik
Sommersemester 2024
Andreas Eberle, Stefan Oberdörster

9. Übungsblatt „Einführung in die Statistik"

Abgabe bis Dienstag 18.6.

---

1. (Benfordsches Gesetz) Welcher Verteilung gehorcht die erste Ziffer einer Zahl?

a) Wir betrachten folgende Stichprobe: Aus einem Ortsverzeichnis wurde zufällig eine Seite aufgeschlagen. Diese enthält 305 Orte. In der Tabelle unten ist nun aufgeführt, in wie vielen Orten die Einwohnerzahl mit den Ziffern $1, 2, \dots, 9$ beginnt.

Erste Ziffer	1	2	3	4	5	6	7	8	9
Häufigkeit	107	55	39	22	13	18	13	23	15

(i) Testen Sie die Hypothese, dass diese Ziffern auf der Menge $\{1, 2, \dots, 9\}$ gleichverteilt sind.

(ii) Testen Sie die Hypothese, dass die Ziffern der Benford-Verteilung gehorchen, d.h.

$$\mathbb{P}\,(\text{Erste Ziffer} = k) = \log_{10}(1 + 1/k) \quad \text{für } k = 1, 2, \dots, 9.$$

b) Hinter der Benford-Verteilung steht ein allgemeines Phänomen: Ist X eine Zufallsvariable mit stetiger Verteilungsfunktion F auf $\mathbb{R}$, und ist diese Verteilung „recht diffus", dann ist die Zufallsvariable $Y = X - \lfloor X \rfloor$ „näherungsweise" gleichverteilt auf $[0, 1)$. Diese vage Aussage lässt sich mathematisch präzisieren. Sei Z eine Zufallsvariable mit stetiger Verteilung auf $(0, \infty)$. Diese schreiben wir als Dezimalzahl

$$Z \;=\; Z_0, Z_1 Z_2 Z_3 \ldots \cdot 10^W \;=\; (Z_0 + 10^{-1} Z_1 + 10^{-2} Z_2 + 10^{-3} Z_3 + \dots) \cdot 10^W$$

mit Ziffern $Z_0 \in \{1, \dots, 9\}$, $Z_1, Z_2, Z_3, \dots \in \{0, 1, \dots, 9\}$ und einem ganzzahligen Exponenten W. Wir gehen davon aus, dass $X = \log_{10}(Z)$ „recht diffus" verteilt ist. Wie kann man nun aus dem oben beschriebenen Phänomen ableiten, dass

$$\mathbb{P}(Z_0 = k) \approx \log_{10}(1 + 1/k) \quad \text{für } k = 1, 2, \dots, 9\ ?$$

Anmerkung: Das Benfordsche Gesetz wird beispielsweise bei Steuerprüfungen verwendet, um Manipulationen von Datenmaterial aufzuspüren.

2. (Monte-Carlo-p-Werte) Für eine Teststatistik $t = T(\text{Daten})$ betrachten wir den p-Wert

$$p \;:=\; 1 - F_0(t-)$$

für eine gegebene Verteilungsfunktion F_0 sowie den Monte-Carlo-p-Wert

$$\hat{p} \;:=\; \frac{\#\bigl\{s \in \{1, \ldots, B\} : T_s \geq t\bigr\} + 1}{B+1}.$$

Dabei sind $T_1, T_2, \ldots, T_B$ untereinander (und von den Daten) unabhängige, nach F_0 verteilte Zufallsvariablen. Nun vergleichen wir p und $\hat{p}$ bei gegebenen Daten, berücksichtigen also nur den Zufall in den (simulierten) Variablen $T_1, \ldots, T_B$. Zeigen Sie, dass

$$\mathbb{E}\left[(\hat{p} - p)^2\right] \;\leq\; \frac{1}{4B+1} \quad \text{falls } B \geq 2.$$

3. (Bootstrap-Konfidenzintervalle für Korrelationskoeffizienten) Auf der Vorlesungshomepage finden Sie unter "Vergleich Abiturnoten Bachelornoten" einen Datensatz, in dem die Abiturnoten und die Bachelorabschlussnoten von 15 Studierenden verglichen werden.

a) Berechnen Sie einen Plug-in-Schätzer $\hat{\rho}$ für die Korrelation ρ zwischen Abitur- und Bachelornoten.

b) Geben Sie in Pseudocode einen Algorithmus zur Simulation von $B = 1000$ Bootstrap-Replikationen des Datensatzes an.

c) Implementieren Sie den Algorithmus und berechnen Sie die entsprechenden Bootstrap-Schätzwerte für die Korrelation. Stellen Sie die empirische Verteilung der Bootstrap-Stichprobe graphisch als Histogramm dar.

d) Berechnen Sie einen Bootstrap-Schätzer für die Standardabweichung von $\hat{\rho}$, und bestimmen Sie Bootstrap-Konfidenzintervalle für ρ zum Niveau 95% mit und ohne Verwendung einer Normalapproximation.

e) Diskutieren Sie die Ergebnisse.

4. (Computer-Experiment zum Bootstrap-Verfahren) Seien $X_1, \ldots, X_n \sim \mathcal{N}(m, 1)$. Zum Schätzen von $g(m) = e^m$ verwenden wir den Plug-in-Schätzer $\hat{g} = e^{\bar{X}_n}$

a) Simulieren Sie für $m = 5$ eine Stichprobe $x = (x_1, \ldots, x_n)$ mit $n = 100$ Stichprobenwerten, und berechnen Sie den Wert von $\hat{g}$.

b) Schätzen Sie mithilfe eines Bootstrap-Verfahrens die Standardabweichung von $\hat{g}$ und berechnen Sie ein Bootstrap-Konfidenzintervall für e^m zum Niveau 95%.

c) Erstellen Sie ein Histogramm der Bootstrap-Replikationen, und vergleichen Sie dieses mit der tatsächlichen Verteilung von $\hat{g}$.

d) Diskutieren Sie die Ergebnisse.