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

9. Übungsblatt „Einführung in die Statistik"

Abgabe bis Mittwoch 15.6., 16 Uhr.

---

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 $1, 2, \ldots, 9$ 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 $\{1, 2, \ldots, 9\}$ gleichverteilt sind.

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

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

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.) Nun sei Z eine Zufallsvariable mit stetiger Verteilung auf $(0, \infty)$. Diese schreiben wir als Dezimalzahl, das heißt,

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

mit Ziffern $Z_0 \in \{1, \ldots, 9\}$, $Z_1, Z_2, Z_3, \ldots \in \{0, 1, \ldots, 9\}$ 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

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

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

2. (Bestzeiten von Sprintern) Die nachfolgende Tabelle enthält die Bestzeiten (in Sekunden) von zehn Sprintern aus Grossbritannien über 200 m bzw. 100 m für das Jahr 1988. Es handelt sich um alle Sprinter, die 1988 die 200 m in weniger als 21.20 Sekunden liefen und ausserdem eine Bestzeit über 100 m angaben. Wir vermuten, dass sich die Durchschnittsgeschwindigkeit der Sprinter über 200 m von jener der Sprinter über 100 m systematisch

unterscheidet, wir wissen jedoch nicht in welche Richtung: Man könnte vermuten, dass die Durchschnittsgeschwindigkeit der Sprinter über 200 m aufgrund des geringeren Einflusses der Start- und Beschleunigungsphase höher ist als jene der Sprinter über 100 m. Denkbar wäre auch, dass die längere Distanz zu einer niedrigeren Durchschnittsgeschwindigkeit führt.
Testen Sie (zum Beispiel unter Verwendung von R) die Nullhypothese, dass es keine systematischen Unterschiede gibt, zum Niveau $\alpha = 0.05$ auf zwei verschiedene Arten.

Athlet	Bestz. 200 m	Bestz. 100 m
L. Christie	20.09	9.97
J. Regis	20.32	10.31
M. Rosswess	20.51	10.40
A. Carrott	20.76	10.56
T. Bennett	20.90	10.92
A. Mafe	20.94	10.64
D. Reid	21.00	10.54
P. Snoddy	21.14	10.85
L. Stapleton	21.17	10.71
C. Jackson	21.19	10.56

3. (Kolmogorov-Smirnov-Bänder und Quantile) Mit einer Sicherheit von $1 - \alpha$ können wir davon ausgehen, dass $\|\hat{F} - F\|_{\sup} \leq \kappa_{n,\alpha}$ für einen gewissen kritischen Wert $\kappa_{n,\alpha} > 0$. Wie die folgenden Betrachtungen zeigen, ergeben sich hieraus automatisch Konfidenzschranken für beliebige Quantile q_γ, $\gamma \in (0,1)$:

a) Bestimmen Sie für $\kappa > 0$ und $\gamma \in (0,1)$ einen möglichst großen Index $k = k(n,\gamma,\kappa) \in \{0,1,\ldots,n\}$ und einen möglichst kleinen Index $\ell = \ell(n,\gamma,\kappa)$, so dass gilt:

$$q_\gamma \;\in\; [X_{(k)}, X_{(\ell)}] \quad \text{falls } \|\hat{F} - F\|_{\sup} \leq \kappa.$$

Illustrieren und prüfen Sie Ihr Ergebnis anhand einer Skizze mit $n = 4$ Beobachtungen, wobei $X_{(1)} < X_{(2)} < X_{(3)} < X_{(4)}$, $\kappa = 0.2$ und $\gamma \in \{0.1, 0.5, 0.7\}$.

b) Bestimmen Sie einen Stichprobenumfang n mit $\mathbb{P}\left[\|\hat{F}_n - F\|_{\sup} \geq 0.01\right] \leq 0.01$. Welches Konfidenzintervall ergibt sich auf diese Weise für den Median ?

4. (Konfidenzintervalle für den Median)

a) Angenommen man möchte für einen Datensatz mit $n = 30$ Beobachtungen ein Konfidenzintervall für den zugrundeliegenden Median $q_{0.5}$ berechnen. Bestimmen Sie *alle* minimalen Indexpaare (k,ℓ) derart, dass $[X_{(k)}, X_{(\ell)}]$ ein 95%-Konfidenzintervall für $q_{0.5}$ ist. Dabei bedeutet „minimal", dass weder $[X_{(k+1)}, X_{(\ell)}]$ noch $[X_{(k)}, X_{(\ell-1)}]$ ein 95%-Konfidenzintervall für $q_{0.5}$ wäre. Verwenden Sie hierzu nachfolgende Tabelle oder

ein eigenes Programm.

x	$F_{30,0.5}(x)$	x	$F_{30,0.5}(x)$	x	$F_{30,0.5}(x)$	x	$F_{30,0.5}(x)$
0	0.0000	8	0.0081	16	0.7077	24	0.9998
1	0.0000	9	0.0214	17	0.8192	25	1.0000
2	0.0000	10	0.0494	18	0.8998	26	1.0000
3	0.0000	11	0.1002	19	0.9506	27	1.0000
4	0.0000	12	0.1808	20	0.9786	28	1.0000
5	0.0002	13	0.2923	21	0.9919	29	1.0000
6	0.0007	14	0.4278	22	0.9974	30	1.0000
7	0.0026	15	0.5722	23	0.9993		

b) Nachfolgend sehen Sie die Lebensdauern in Monaten für eine Stichprobe von $n = 30$ Hauskatzen:

145.4	71.7	147.9	104.9	118.1	105.6	139.3	119.0	138.9	147.7
145.1	80.2	125.6	188.2	169.6	114.2	198.0	96.4	112.6	94.9
167.2	216.6	162.8	165.9	225.7	162.0	89.7	107.0	172.9	66.6

Berechnen Sie eine untere 95%-Vertrauensschranke für den Median der Lebensdauer von Hauskatzen. Formulieren Sie das Ergebnis auch in Worten.

5. (Ordnungsstatistiken) Seien $U_1, U_2, \ldots, U_n$ unabhängige, auf $[0,1]$ gleichverteilte Zufallsvariablen, und seien $U_{(1)}, U_{(2)} \cdots, U_{(n)}$ die entsprechenden Ordnungsstatistiken.

a) Zeigen Sie, dass die Verteilung von $U_{(k)}$ absolutstetig ist mit Dichte

$$f_{U_{(k)}}(u) \;=\; n\binom{n-1}{k-1} u^{k-1}(1-u)^{n-k}\, 1_{[0,1]}(u)\,.$$

b) Begründen Sie ohne lange Rechnung, dass

$$\int_0^1 u^a(1-u)^b du \;=\; (a+b+1)^{-1}\binom{a+b}{a}^{-1} \qquad \text{für alle } a, b \in \mathbb{N}_0.$$

c) Folgern Sie, dass $\mathbb{E}\left[U_{(k)}\right] = \frac{k}{n+1} =: t_k$ und $\mathrm{Var}\left[U_{(k)}\right] = \frac{t_k(1-t_k)}{n+2}$.