Übungen zur Vorlesung

Datenanalyse - Dr. Terveer, Vogt, Pohl

Sommersemester 2022 Blatt 7 31.05.2022

Aufgabe 18 (Klausur WS1718) Ein Marktforschungsinstitut hat das Alter von je 10 Kunden der Social-Media-Dienste „Fakebook“ und „Glitter“ erhoben:

Fakebook	13	15	17	18	22	23	24	28	30	31
Glitter	16	19	21	25	26	32	33	34	35	37

Testen Sie - zum Niveau 5% statistisch gesichert - ob die Kundengruppen im Mittel gleich alt sind, d.h. geben Sie insbesondere an:

1) Modellannahmen für die Daten, die Hypothesen und die Testbezeichnung,
2) die allgemeine Form/konkreten Wert der Prüfgröße (Rechenweg!),
3) den zur Entscheidungsfindung notwendigen konkreten Schwellenwert,
4) die konkrete Entscheidung, die der Test trifft (mit Begründung).

Aufgabe 19 (Wilcoxon-Test) Bei der olympischen Disziplin „Schießen Luftpistole 10 Meter“ haben Final-Wettbewerbe für Frauen und Männer gleiche Bedingungen[2].

Eine Kommission des IOC überlegt, ob in dieser Disziplin künftig ein gemeinsamer Wettbewerb von Frauen und Männern durchgeführt werden sollte, was nur dann als sinnvoll erachtet wird, wenn die Leistungen von Frauen und Männern in dieser Disziplin grundsätzlich vergleichbar sind. Die Kommission zieht zu diesem Zweck die Leistungen der Finalrunden der Olympiade 2012 heran[3]:

Nr	Finale (M)	Finale (F)
1	101.80	101.00
2	100.60	100.25
3	100.28	100.10
4	100.22	99.65
5	99.30	99.64
6	98.30	98.60
7	97.70	98.50
8	96.10	97.60

a) Formulieren Sie das Hypothesenpaar, welches der IOC-Überlegung zugrunde liegt. Benennen Sie auch das für die Hypothesen verwendete Verteilungsmodell.

b) Berechnen Sie die Teststatistik $W = \sum_{i=1}^{n_1} rg(X_i)$ für den vorliegenden Datensatz.

c) Entscheiden Sie anhand der Quantiltabellen aus der Vorlesung zum Signifikanzniveau 0,05 das Testproblem (d.h. mit dem von Ihnen formulierten Hypothesenpaar).

Aufgabe 20 (Approximativer Test) $X_1, \ldots, X_N$ seien Sequenzlängen zu N Tweets eines bestimmten Themas[4]. Es gelte $\mathbb{P}(X_n = k) = \pi^{k-1}(1-\pi) \quad$ für $k \in \mathbb{N}$, $n = 1, \ldots, N$. Ein Thema gelte als wichtig, wenn $\pi > \frac{1}{3}$.

Zum Thema „lebenswerteste Stadt der Welt“ findet man 2523 Tweets, die nicht gleichzeitig Retweets sind, mit einer durchschnittlichen Sequenzlänge $\bar{X} = 1,552$. Prüfen Sie mit einem approximativen **statistischen Test** zum Niveau $\alpha = 0.01$ die Hypothese $\mathbb{H} : \pi \leq \frac{1}{3}$ auf und geben Sie auch die **Signifikanz** an. Hinweis: $EX_1 = \frac{1}{1-\pi}$, $var(X_1) = \frac{\pi}{(1-\pi)^2}$.

---

[2] Es mussten von jedem/jeder der acht Finalteilnehmer/innnen insgesamt 10 Schüsse abgegeben werden, jeweils mit Wertung zwischen 0 und 10.9 in 0.1-Punkte-Schritten.

[3] Quelle: Wikipedia; die Ausgangsdaten liegen mit nur einer Nachkommastelle vor. die zweite Nachkommastelle wurde zum Auflösen von Bindungen simuliert.

[4] Ein „Retweet“ T' ist die - unveränderte - Wiederholung des Tweets eines Nutzers, in Zeichen $T \succeq T'$. Eine Retweetsequenz ist eine Folge $T = T(1) \succeq \cdots \succeq T(n)$ mit (maximaler) Sequenzlänge n.