Langrock, Vienken, Hopfengärtner, Müller-Hanssen, Lorch Sommersemester 2024

Statistik II — Aufgabenzettel 5

(Abgabe: 27.05.2024)

Hinweise zur Abgabe dieses Aufgabenzettels

- es handelt sich um eine Gruppenabgabe, wobei Gruppen mit 2-3 Mitgliedern zulässig sind — **Einzelabgaben sind nicht mehr zulässig**
- Gruppen können tutorienübergreifend gebildet und bei späteren Aufgabenzetteln jederzeit neu zusammengesetzt werden
- jede Person kann in nur einer Gruppe mitwirken und jede Gruppe nur eine Lösung abgeben
- die Abgabe ist bis zum 27.05.2024, 23:59 Uhr über Moodle möglich und von nur *einer* (beliebigen) Person aus jeder Gruppe durchzuführen
- geben Sie in Ihrer Abgabedatei auf der ersten Seite der Lösungen klar und deutlich alle Mitglieder der Gruppe an (Vorname, Name, Matrikelnummer)
- wir empfehlen, dass die abgebende Person sowohl die finale Abgabedatei als auch einen Screenshot der Abgabe in Moodle an die anderen Gruppenmitglieder schickt
- Lösungen können u.a. aus folgenden Komponenten bestehen: R Code und R Output, Screenshots, abfotografierte oder eingescannte handschriftliche Lösungen, mit einem Editor wie MS Word erstellte Lösungen — wir sind hier flexibel, alles was lesbar ist wird akzeptiert
- die einzelnen Komponenten sind für die Abgabe jedoch zu einer *einzigen* **PDF Datei** zusammenzufügen — hierfür gibt es kostenlose Online-Tools, z.B. Smallpdf
- der Lösungsweg muss immer klar ersichtlich und die Lösung vollständig sein — sollten beispielsweise Grafiken zu erstellen sein, so sollten diese auch Teil Ihrer Abgabe sein
- wir nutzen das folgende Punkteschema (pro Aufgabe):
  - gar nichts gemacht $\rightsquigarrow$ 0 Punkte
  - sich an der Aufgabe versucht, aber sehr wenig richtig gemacht $\rightsquigarrow$ 1 Punkt
  - teilweise richtige Lösungen/Lösungsansätze vorgelegt $\rightsquigarrow$ 2 Punkte
  - die Aufgabe gut bearbeitet, mit kleinen Schönheitsfehlern $\rightsquigarrow$ 3 Punkte
  - die Aufgabe vollständig zufriedenstellend bearbeitet $\rightsquigarrow$ 4 Punkte

  (bewertet wird nicht kleinteilig jede Teilaufgabe, sondern nach Gesamteindruck)
- nach Ende der Abgabefrist werden Musterlösungen bei Moodle bereitgestellt, eine Besprechung in den Tutorien wird es nicht geben
- falls während der Bearbeitungszeit Fragen auftreten sollten, dann melden Sie sich jederzeit gerne über die Pinnwand (eine Antwort kommt in der Regel innerhalb einiger Stunden)

Aufgabe 1: Maximum-Likelihood-Schätzung (analytisch)

Laden Sie in R wie folgt Daten zu 100 Tauchgängen von Kegelrobben in der Ostsee:

```
dauer <- read.csv("http://www.rolandlangrock.com//Daten//kegelrobbe.csv")$x
```

Konkret wird hier für jeden der 100 Tauchgänge die Dauer (in Sekunden) angegeben. Wir nehmen nun an, dass die Tauchgänge unabhängig voneinander sind und die jeweilige Dauer aus einer Verteilung mit folgender Dichte generiert wurde:

$$f_\theta(x) = \frac{\theta^6 x^5}{120} \exp\big(-\theta x\big), \quad x > 0$$

Bestimmen Sie den Maximum-Likelihood-Schätzer $\hat{\theta}$, zunächst für allgemeine $x_1, \ldots, x_n$ und dann konkret für die gegebenen Daten.

Optional (gibt keine Punkte, nur Beifall): Erstellen Sie ein Histogramm der Beobachtungen und ergänzen Sie die an die Daten angepasste Dichtefunktion.

Aufgabe 2: Maximum-Likelihood-Schätzung (numerisch)

Laden Sie in R wie folgt die Körpergewichte (in Gramm) von 439 Neugeborenen:

```
gewicht <- read.csv("http://www.rolandlangrock.com//Daten//gewicht.csv")$x
```

Wir nehmen an, dass die Körpergewichte der Babys unabhängig voneinander einer Verteilung mit folgender Dichte entstammen:

$$f_\theta(x) = 5\theta^5 x^4 \exp\big(-(\theta x)^5\big), \quad x > 0$$

Bestimmen Sie mit Newton-Raphson den Maximum-Likelihood-Schätzer $\hat{\theta}$ für die gegebenen Daten (Startwert $\theta^{(0)} = 0.0002$, mindestens drei Iterationen).

Optional (gibt keine Punkte): Erstellen Sie ein Histogramm der Beobachtungen und ergänzen Sie die an die Daten angepasste Dichtefunktion.

Aufgabe 3: Lineare Regression

In dieser Aufgabe betrachten wir Daten zu $n = 275$ Airbnb-Unterkünften in Berlin. Nutzen Sie folgende Befehle, um die Daten einzulesen:

```
airbnb <- read.csv("http://www.rolandlangrock.com/Daten/airbnb.csv")
attach(airbnb)
```

Die im Datensatz enthaltenen Variablen sind:

- 'Preis': der Preis (in Euro) für eine Übernachtung in der Unterkunft;
- 'Distanz': die Entfernung (in Km) zum Stadtzentrum von Berlin.

Diese Variablen können Sie durch Eingabe von `Preis` bzw. `Distanz` direkt in R abrufen.

a) Erstellen Sie ein Streudiagramm mit den Distanzen auf der x– und den Preisen auf der y–Achse.

b) Nutzen Sie den `lm`–Befehl in R, um das Regressionsmodell

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

mit

- Y_i: Preis (in Euro) für eine Übernachtung in der i–ten Unterkunft
- x_i: Distanz der i–ten Unterkunft zum Stadtzentrum,
- ϵ_i: zufälliger Fehler

an die Daten anzupassen. Geben Sie die Modellgleichung mit den geschätzten Parametern $\hat{\beta}_0$ und $\hat{\beta}_1$ an. Um wie viel Euro verringert sich gemäß des Modells der mittlere Preis einer Übernachtung, wenn man bereit ist, drei zusätzliche Kilometer Entfernung zum Stadtzentrum in Kauf zu nehmen?

c) Ergänzen Sie die Regressionsgerade in dem oben erstellten Streudiagramm.

d) Ermitteln Sie rechnerisch, für welche Distanz zum Stadtzentrum der mittlere Preis für eine Übernachtung gemäß des Modells exakt 60 Euro wäre.

Aufgabe 4: Breite von Konfidenzintervallen

Bestimmen Sie für jede der folgenden Situationen die minimale Stichprobengröße n, die *garantiert*, dass die Breite des resultierenden Konfidenzintervalls maximal $B = 0.05$ beträgt.

a) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Bern(p)$, bestimmt werden soll ein 90%–KI für p

b) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Bern(p)$, Zusatzinfo $\hat{p} \leq 0.2$, bestimmt werden soll ein 95%–KI für p

c) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $\sigma = 1$ ist bekannt, bestimmt werden soll ein 90%–KI für μ

Warum kann man die Bestimmung der benötigten Stichprobengröße n analog zur Situation c) im Fall σ *unbekannt* nicht vornehmen?