Stochastik

Serie 10

Aufgabe 10.1

Mit folgender Simulationsaufgabe sollen Sie mit dem Normalplot vertraut werden.

a) Simulieren Sie $n = 10, 20, 50$ und 100 (standard-) normalverteilte Zufallszahlen und betrachten Sie die jeweilige Folge mit einem Normalplot. Wiederholen Sie diese Simulationen einige Male, bis Sie abschätzen können, wie weit zufällige Abweichungen von einer Geraden im Normalplot üblich sind.

   R-Hinweise:

```
x <- rnorm(10)
qqnorm(x)
```

b) **Langschwänzige Verteilung**: Simulieren Sie je $n = 20$ und 100 t-verteilte Zufallszahlen mit $\nu = 20, 7$ und 3 Freiheitsgraden. Wiederholen Sie diese Simulationen einige Male, bis Sie abschätzen können, wie gross Abweichungen von einer Geraden im Normalplot üblich sind.

   R-Hinweise:

```
x <- rt(10, df = 20)
```

c) **Schiefe Verteilung**: Simulieren Sie je $n = 20$ und 100 chiquadrat-verteilte Zufallszahlen mit $\nu = 20$ und 1 Freiheitsgraden. Wiederholen Sie diese Simulationen einige Male, bis Sie abschätzen können, wie gross Abweichungen von einer Geraden im Normalplot üblich sind.

   R-Hinweise:

```
x <- rchisq(10, df = 20)
```

Aufgabe 10.2

In dieser Aufgabe untersuchen Sie die Wirkung des Zentralen Grenzwertsatzes mittels Simulation. Gehen Sie von einer Zufallsvariablen X aus, die folgendermassen verteilt ist: die Werte $0, 10$ und 11 werden je mit einer Wahrscheinlichkeit $\frac{1}{3}$ angenommen. Simulieren Sie nun die Verteilung von X sowie die Verteilung des Mittelwerts $\overline{X}_n$ von mehreren X.

a) Simulieren Sie X. Stellen Sie die Verteilung von X mittels eines Histogramms von 1000 Realisierungen von X dar, und vergleichen Sie sie mittels des Normalplots mit der Normalverteilung.

```
# Mehrere Grafiken neben- und untereinander
par(mfrow = c(4, 2))
# moegliche Werte von X
werte <- c(0, 10, 11)
# X simulieren
sim <- sample(werte, 1000, replace = TRUE)
# Histogramm erstellen
hist(sim, main = paste("Original"))
# Normalplot erstellen
qqnorm(sim)
```

b) Simulieren Sie nun $\overline{X}_5 = \frac{X_1+X_2+X_3+X_4+X_5}{5}$, wobei die X_i die gleiche Verteilung haben wie X und unabhängig sind. Stellen Sie die Verteilung von $\overline{X}_5$ anhand von 1000 Realisierungen von $\overline{X}_5$ dar, und vergleichen Sie mit der Normalverteilung.

```
n <- 5
# X_1,...,X_n simulieren und in einer n-spaltigen
# Matrix (mit 1000 Zeilen) anordnen
sim <- matrix(sample(werte, n * 1000, replace = TRUE),
    ncol = n)
# In jeder Matrixzeile Mittelwert berechnen
sim.mean <- apply(sim, 1, "mean")
hist(sim.mean)
title(paste("Mittelwerte von", n, "Beobachtungen"))
qqnorm(sim.mean)
```

c) Simulieren Sie nun die Verteilung von $\overline{X}_n$ auch für die Fälle, wo $\overline{X}_n$ das Mittel von $n = 10$ resp. $n = 200$ X_i ist.

d) Geben Sie die Verteilung von $\overline{X}_{200}$ an, zusammen mit den Werten der Verteilungsparameter.

Aufgabe 10.3

In einer Studie wurde untersucht, wie bei Mäusen die Aufnahme von Eisen (Fe^{3+}) von der Dosis abhängt. Dazu wurden 54 Mäuse zufällig in 3 Gruppen zu je 18 Mäusen eingeteilt und jeweils mit Dosis hoch, mittel und tief gefüttert (hoch = 10.2 millimolar, mittel=1.2 millimolar, tief=0.3 millimolar). Mittels radioaktiver Markierung wurde der Anteil des zurückgehaltenen Eisens in Prozent nach einer gewissen Zeit bestimmt.

Die Daten sind auf Ilias in der Datei `ironF3.dat` abgelegt; Sie können sie einlesen mit dem Befehl

```
iron <- read.table("./Daten/ironF3.dat", header = TRUE)
```

a) Erstellen Sie für jede der 3 Versuchsbedingungen einen Boxplot, am besten gerade nebeneinander. Wie unterscheiden sich die Daten der verschiedenen Versuchsbedingungen?

b) Transformieren Sie alle Werte mit dem Logarithmus und erstellen Sie wieder die 3 Boxplots wie bei Aufgabe a). Was hat sich durch die Transformation geändert?

c) Erstellen Sie einen Normalplot der Daten bei mittlerer Dosis vor und nach dem Logarithmieren. Wann passt die Normalverteilung besser? Verwenden Sie die R-Funktion

   ```
   qqnorm(...)
   ```

d) Unter der Annahme, dass die Daten bei mittlerer Dosis normalverteilt sind, schätzen Sie die Parameter μ und σ^2. Wie gross ist die Wahrscheinlichkeit, dass eine Maus mehr als 10 % Eisen zurückhält.

e) (**Zusatzaufgabe**) Unter der Annahme, dass die Daten bei mittlerer Dosis log-normalverteilt sind, schätzen Sie die Parameter μ und σ^2. Wie gross ist die Wahrscheinlichkeit, dass eine Maus mehr als 10 % Eisen zurückhält.

   Hinweis: Ist $Y = \log(X)$ normalverteilt ist, so heisst X log-normalverteilt.

Aufgabe 10.4

Ein Statistiker beobachtet, dass ein Angler innerhalb von 2 Stunden 15 Fische fängt. Er nimmt an, dass es sich um einen Poissonprozess handelt und überlegt sich:

a) Mit welcher Wahrscheinlichkeit dauert es länger als 12 Minuten, bis der nächste Fisch anbeisst?

   Hinweis: Benützen Sie die Momentenmethode, um Parameter zu schätzen.

b) Mit welcher Wahrscheinlichkeit beissen innerhalb der nächsten 12 Minuten genau 2 Fische an?

c) Ein Fischer hat die Wartezeiten zwischen zwei Fischfängen aufgeschrieben. Erstellen Sie einen QQ-Plot für die angegebenen Zeitdifferenzen. Tragen Sie dazu die empirischen Quantile der Messungen gegen die theoretischen Quantile der Exp(1)-Verteilung auf. Passt die Exponentialverteilung zu den Messdaten? Bestimmen Sie die Steigung der Regressionsgeraden im QQ-Plot. Was ist die Bedeutung von der Steigung der Regressionsgeraden in diesem QQ-Plot? Schätzen Sie aufgrund der Steigung den Parameter λ der Exp(λ)-Verteilung.

i	$x_{(i)}$ in Minuten
1	16.9
2	4.20
3	6.70
4	8.83
5	10.7
6	22.4
7	1.37
8	3.00
9	4.82
10	4.53
11	6.77
12	4.81

Aufgabe 10.5

In dieser Aufgabe geht es um Parameterschätzung. Wir betrachten eine stetige Verteilung mit folgender Dichte:

$$f(x) = \begin{cases} \dfrac{\alpha}{x^{\alpha+1}} & x \geq 1 \\ 0 & x < 1 \end{cases}$$

wobei $\alpha > 0$ ein unbekannter Parameter ist. Wir wollen den Parameter α aus einer Stichprobe schätzen.

a) Bestimmen Sie die Likelihood- und die Log-Likelihood-Funktion basierend auf n unabhängigen identisch verteilten Beobachtungen $x_1, \ldots, x_n$ einer Zufallsvariablen mit obiger Dichte.

b) Bestimmen Sie den zugehörigen Maximum-Likelihood-Schätzer für α. Schreiben Sie zuerst die allgemeine Formel für n Beobachtungen hin und berechnen Sie den Schätzer dann für die folgende konkrete Stichprobe:

x_1	x_2	x_3	x_4	x_5
12.0	4.0	6.9	27.9	15.4

c) Bestimmen Sie den Momentenschätzer für α, wieder zuerst allgemein basierend auf n unabhängigen Beobachtungen $x_1, \ldots, x_n$ und dann für obige Stichprobe.

Hinweise: Der Erwartungswert einer Zufallsvariablen X mit obiger Dichte f ist

$$\mathrm{E}[X] = \frac{\alpha}{\alpha - 1} \text{ für } \alpha > 1$$

Für $\alpha \leq 1$ ist der Erwartungswert gleich ∞ und der Momentenschätzer ist nicht definiert. Sie müssen für diese Teilaufgabe also annehmen, dass $\alpha > 1$.

d) Vergleichen Sie den Maximum-Likelihood und den Momentenschätzer für obige Stichprobe. Ist der Momentenschätzer hier sinnvoll?

Aufgabe 10.6

(*Zusatzaufgabe*)

Die Inkubationszeit eines Virus ist definiert als die Zeitspanne (gemessen in Tagen) von der Ansteckung bis zum Ausbruch der Krankheit. Diese Inkubationszeit wird mit folgender Dichte beschrieben:

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \dfrac{1}{x\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(\log(x)-1)^2}{2\sigma^2}} & x > 0 \end{cases}$$

wobei wir mit $\log(\cdot)$ den natürlichen Logarithmus bezeichnen. Hier ist $\sigma > 0$ der unbekannte Parameter, welchen es zu schätzen gilt. Wir möchten nun den Parameter der Verteilung kennen, dazu nehmen wir eine Stichprobe. Das Ziel ist, diesen unbekannten Parameter daraus zu schätzen, um somit zum Beispiel die Dauer von Quarantänemassnahmen für ein neuartiges Virus abschätzen zu können. Wir betrachten die Daten (in Tagen):

x_1	x_2	x_3	x_4	x_5
1.62	5.29	3.93	1.49	1.57

Hinweis: Für eine Zufallsvariable X mit obiger Dichte gilt $\mathrm{E}[X] = \mathrm{e}^{1+\sigma^2/2}$.

a) Bestimmen Sie die Likelihood- und die log-Likelihood-Funktion basierend auf n unabhängigen identisch verteilten Beobachtungen $x_1, \ldots, x_n$ einer Zufallsvariable mit obiger Dichte. Berechnen Sie daraus den Maximum-Likelihood Schätzer für den unbekannten Parameter σ. Die konkreten Werte brauchen Sie nicht einzusetzen.

b) Bestimmen Sie den Momentenschätzer für σ, sowohl basierend auf n unabhängigen Beobachtungen $x_1, \ldots, x_n$, wie auch den realisierten Wert für die gegebene Stichprobe.

c) Nehmen wir für diese Teilaufgabe an, dass wir folgende Daten beobachtet haben:

x_1	x_2	x_3
2.12	1.63	4.21

Welcher Schätzer (Momentenschätzer oder Maximum-Likelihood Schätzer) ist bei diesen Daten zu bevorzugen? Begründen Sie.

Aufgabe 10.7

Ein Weinhändler behauptet, dass die von ihm gefüllten Weinflaschen mindestens 70 Zentiliter enthalten. Ein skeptischer Konsument vermutet aber, dass der Weinhändler zu wenig Wein abfüllt und will diese Behauptung überprüfen. Deshalb kauft er 12 Weinflaschen und misst ihren Inhalt. Er findet:

$$71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72 \quad \text{(in Zentiliter)}$$

Nehmen Sie zunächst an, dass die Standardabweichung der Abfüllung im Voraus bekannt ist. Sie beträgt $\sigma = 1.5$ Zentiliter. Da die Standardabweichung der Messungen bekannt ist, können wir einen z-Test durchführen. Führen Sie den (einseitigen; in welche Richtung?) Test auf dem 5 %- Signifikanzniveau durch. Geben Sie die Modellannahmen, H_0, H_A, den Verwerfungsbereich, den Wert der Teststatistik und das Testergebnis explizit an. Formulieren Sie in einem Satz die Schlussfolgerung für den kritischen Konsumenten.

Kurzlösungen vereinzelter Aufgaben

A 10.3:

d) 0.371

e) 0.271

A 10.4:

a) 0.223

b) 0.251

c) Aus dem QQ-Plot folgt $\lambda = 1/7.4$.

A 10.6:

a) $\log l(\sigma; x_1, \ldots, x_n) = \log(1) - \frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum\limits_{i=1}^{n}\log(x_i) - \frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(\log(x_i) - 1)^2$

und $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(\log(x_i) - 1)^2}$

b) $\hat{\sigma}_{\text{MoM}} = \sqrt{\log\left(\left(\frac{1}{n}\sum\limits_{i=1}^{n}x_i\right)^2\right) - 2} = 0.2119$