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

Statistik II — Aufgabenzettel 4

(Abgabe: 29.04.2024)

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 29.04.2024, 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 ⇝ 0 Punkte
  - sich an der Aufgabe versucht, aber sehr wenig richtig gemacht ⇝ 1 Punkt
  - teilweise richtige Lösungen/Lösungsansätze vorgelegt ⇝ 2 Punkte
  - die Aufgabe gut bearbeitet, mit kleinen Schönheitsfehlern ⇝ 3 Punkte
  - die Aufgabe vollständig zufriedenstellend bearbeitet ⇝ 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: Zweidimensionale Verteilung

Die folgende Kontingenztafel gibt die gemeinsame Verteilung von zwei diskreten Zufallsvariablen X und Y an:

		Y		
		1	2	3
X	0	0.0	0.1	0.4
	1	0.3	0.2	0.0

Bestimmen Sie den Korrelationskoeffizienten $\rho(X, Y)$. Dokumentieren Sie dabei Ihren Rechenweg.

Aufgabe 2: Normalverteilung

In dieser Aufgabe betrachten wir die täglichen Renditen (d.h. die prozentualen Kursänderungen) des Aktienindex S&P 500, welcher Anteile der 500 größten US-amerikanischen Unternehmen an der Börse enthält. Die Renditen der letzten 20 Jahre erhalten Sie in R durch Aufruf des folgenden Befehls:

```
SP500 <- read.csv("http://www.rolandlangrock.com//Daten//SP500.csv")
```

Schauen Sie sich mit `head(SP500)` und `dim(SP500)` zunächst die Struktur der Daten an. In der zweiten Spalte sind die Renditen enthalten.

a) Ermitteln Sie für die gegebenen Daten die zwei Tage, an denen die größten negativen Renditen, also die größten prozentualen Verluste aufgetreten sind. Was war an diesen beiden Tagen bzw. allgemein in den Zeiträumen, in denen diese hohen negativen Renditen auftraten, jeweils der Auslöser für den Kursrutsch?

b) Wir gehen im Folgenden davon aus, dass sich die täglichen Renditen gut durch eine Normalverteilung beschreiben lassen. Schätzen Sie in R die entsprechenden Parameter μ und σ als das arithmetische Mittel bzw. die korrigierte Stichprobenstandardabweichung der vorliegenden Daten. Erstellen Sie dann in R ein Histogramm der Daten — die Optionen `prob=TRUE` und `breaks=50` nutzend — und ergänzen Sie die Dichte der angepassten Normalverteilung. In welcher Hinsicht passt die Verteilung nicht optimal zu den Daten?

c) Der S&P 500 steht aktuell bei 5000 Punkten. Berechnen Sie unter der in b) identifizierten Normalverteilung für die prozentuale Kursänderung an einem Tag die Wahrscheinlichkeit folgender Ereignisse:

   (i) der S&P 500 steigt am nächsten Tag um 1% oder mehr;

   (i) der S&P 500 fällt am nächsten Tag um 2% oder mehr;

   (iii) der S&P 500 wird am Ende des nächsten Tages zwischen 4950 und 5050 Punkten liegen.

Aufgabe 3: Monte-Carlo-Simulation

Hinweis: Möglicherweise werden einige von Ihnen die Umsetzung dieser Aufgabe in R schwierig finden, da keine ähnlichen Aufgaben in den Tutorien besprochen wurden. Das Beispiel zur Monte-Carlo-Simulation in diesem Skript könnte hilfreich sein, da es einem ähnlichen Muster folgt, insbesondere was die Nutzung von Loops angeht.

Wir betrachten weiterhin die Situation aus Aufgabe 1. Sie sollen nun basierend auf der in Aufgabe 1b) angepassten Normalverteilung in R simulieren, wie der weitere Kursverlauf des S&P 500 aussehen könnte, von dem aktuellen Kurs von 5000 Punkten ausgehend. Gehen Sie dabei wie folgt vor:

a) Simulieren Sie zunächst eine Realisation aus der Normalverteilung (d.h. eine prozentuale Änderung) und geben Sie an, was der daraus resultierende Kurs des S&P 500 nach einem Tag wäre.

b) Simulieren Sie nun 250 Realisationen aus der angepassten Normalverteilung, entsprechend den (ca.) 250 Börsentagen in einem Jahr. Berechnen Sie den Kursverlauf des S&P 500 über ein Börsenjahr, der sich gemäß dieser Simulation der täglichen prozentualen Änderungen ergeben würde, und geben Sie den daraus resultierenden Stand des S&P 500 am Ende des Börsenjahres an.

c) Wiederholen Sie die Simulation aus b) 1000 Mal und zeichnen Sie ein Histogramm der 1000 sich am Ende des simulierten Börsenjahres ergebenden Kurse.

d) *(ohne Wertung)* Halten Sie die eben durchgeführte Simulation des Verlaufs des S&P 500 für realistisch? Wo sehen Sie gegebenenfalls Schwachstellen des Modells?

Aufgabe 4: ZGWS

Als Vorbereitung für die Mathe-Klausur haben Sie sich vorgenommen, 20 Ihnen zur Verfügung gestellte Übungsaufgaben durchzurechnen. Aus Erfahrung wissen Sie, dass Ihre persönliche Bearbeitungszeit einer einzelnen Übungsaufgabe in Mathe einer $Exp(0.04)$–Verteilung (Einheit: Minuten) folgt und dass die einzelnen Bearbeitungszeiten voneinander unabhängig sind. Wir interessieren uns im Folgenden für

$$X = \text{Gesamtdauer, bis Sie alle Übungsaufgaben durchgerechnet haben.}$$

a) Geben Sie die approximative Verteilung von X gemäß des ZGWS an.

b) Bestimmen Sie gemäß Ihres Ergebnisses aus a) (approximativ) die Wahrscheinlichkeit, dass Sie in weniger als 10 Stunden alle Übungsaufgaben durchgerechnet haben werden.

c) Für eine Summe von exponentialverteilten Zufallsvariablen kennt man tatsächlich die exakte sich ergebende Verteilung: Es handelt sich um eine sogenannte Gammaverteilung. Sie können diese exakte Verteilung von X in R wie folgt zeichnen:

```
curve(dgamma(x,25,1/20),from=0,to=1000,bty="n",lwd=2,ylab="Dichte")
```

Ergänzen Sie — z.B. mit `curve(...,add=TRUE)` — die in a) basierend auf dem ZGWS bestimmte approximative Normalverteilung von X in dem Plot. Kommentieren Sie dann basierend auf dem Plot, ob die in b) erhaltene Approximation der gesuchten Wahrscheinlichkeit zu niedrig oder zu hoch ist.

Aufgabe 5: Schätzfunktionen

Wir betrachten iid Realisationen $X_1, \ldots, X_5$ aus einer $N(\mu, \sigma^2)$–Verteilung, basierend auf denen wir den Parameter μ schätzen wollen. Hierzu betrachten wir die folgenden möglichen Schätzer:

$$\hat{\mu}_1 = \frac{1}{2}(X_1 + X_2)$$

$$\hat{\mu}_2 = \frac{1}{6}(X_1 + X_2 + X_3) + \frac{1}{2}X_4$$

$$\hat{\mu}_3 = \frac{1}{5}X_1 + \frac{1}{6}X_2 + \frac{1}{5}X_3 + \frac{1}{6}X_4 + \frac{1}{4}X_5$$

a) Berechnen Sie Bias und Varianz aller angegebenen Schätzer. Welchen Schätzer würden Sie verwenden, wenn vorgegeben ist, dass der Schätzer unbedingt erwartungstreu sein muss?

b) Wir wissen, dass die Summe von normalverteilten Zufallsvariablen wieder normalverteilt ist (Slide 392). Entsprechend sind alle drei Schätzer normalverteilt. Die jeweiligen Erwartungswerte und Varianzen haben Sie gerade eben ausgerechnet, d.h. Sie kennen jetzt die exakte Verteilung der drei Schätzer. Zeichnen Sie nun die drei Normalverteilungen der Schätzer in R für den Fall, dass die wahren Parameterwerte $\mu = 3$ und $\sigma^2 = 4$ sind. Sie können sich hierfür an dem folgenden, noch unvollständigen Code orientieren:

```
curve(dnorm(x,...,...),from=-1,to=7,bty="n",ylim=c(0,0.45),ylab="Dichte")
curve(dnorm(x,...,...),from=-1,to=7,bty="n",col="red",add=TRUE)
curve(dnorm(x,...,...),from=-1,to=7,bty="n",col="blue",add=TRUE)
```

(der hier vorgegebene Wertebereich ist sinnvoll, sofern Sie in a) die Varianz korrekt berechnet haben — bei falscher Berechnung der Varianz kann es sein, dass sie mit dem hier angegebenen Wertebereich keine sinnvolle Grafik erhalten; dann sollten Sie idealerweise Ihren Fehler in a) finden, oder alternativ zumindest hier den Wertebereich anpassen)

Allein basierend auf dieser Grafik: Welcher der drei Schätzer scheint am besten zu sein?