Prof. Dr. Roland Fried
M. Sc. Jana König

Wintersemester 2019/2020

Übungen zur Vorlesung **Multivariate Verfahren**

— Blatt Nr. 4 —

Aufgabe P8

Seien $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim N_d(\boldsymbol{\mu}, \Sigma)$ unabhängig und identisch d-dimensional normalverteilt. Bezeichne mit $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ den Schätzer des Erwartungswertvektors $\boldsymbol{\mu}$. Der Schätzer für die Kovarianzmatrix Σ wird mit $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T$ bezeichnet.

Zeigen Sie, dass $\bar{\boldsymbol{X}}$ und $\boldsymbol{S}$ stochastisch unabhängig sind. Gehen Sie dabei wie folgt vor:

i) Überlegen Sie, wie $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_n')'$ verteilt ist.

ii) Prüfen Sie die Unabhängigkeit von $\bar{\boldsymbol{X}}$ und $\boldsymbol{X}_1 - \bar{\boldsymbol{X}}$.

iii) Begründen Sie, warum es ausreicht zu zeigen, dass $\boldsymbol{X}_1 - \bar{\boldsymbol{X}}$ und $\bar{\boldsymbol{X}}$ unabhängig sind.

Aufgabe P9

Betrachten Sie erneut den Zufallsvektor $\boldsymbol{X}$ aus Aufgabe P5: Sei also

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}, \begin{pmatrix} 7 & 0 & -2 \\ 0 & 5 & -3 \\ -2 & -3 & 5 \end{pmatrix} \right).$$

Seien $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots$ unabhängig und identisch wie $\boldsymbol{X}$ verteilt, ferner sei

$$g : \mathbb{R}^3 \to \mathbb{R} : (x_1, x_2, x_3)' \mapsto \ln(1 + x_1^3 + x_2^2 + x_3).$$

Bestimmen Sie mit Hilfe von Satz 2.35 ein asymptotisches Prognoseintervall zum Niveau 0.95 für $g(\bar{\boldsymbol{X}}_n)$, d.h. ein Intervall, das mit einer Wahrscheinlichkeit von etwa 95% den Wert $g(\bar{\boldsymbol{X}}_n)$ enthält.

Hinweis: Das asymptotische Prognoseintervall hängt von n ab (und wird mit wachsendem n kleiner). Das Attribut *asymptotisch* bezieht sich hierbei darauf, dass es aus der asymptotischen Verteilung von $\sqrt{n}(g(\bar{\boldsymbol{X}}_n) - g(\boldsymbol{\mu}))$ hergeleitet wird.

Aufgabe H8 (10/10)

Gegeben ist folgender zweidimensionale Datensatz ($n = 10$) mit emp. Korrelation $\hat{\rho}_n = 0.88$:

i	1	2	3	4	5	6	7	8	9	10
X_i	-0.6	1.4	1.0	-0.3	-0.8	1.2	0.4	0.5	-0.6	0.3
Y_i	-0.5	0.9	1.5	-0.8	-0.7	1.0	0.3	1.2	-1.4	0.0

Wir gehen davon aus, dass die Daten unabhängige Realisierungen einer zweidimensionalen Normalverteilung mit Korrelation ρ sind. Nun soll ein Konfidenzintervall für ρ bestimmt werden, welches auf dem empirischen Korrelationskoeffizienten $\hat{\rho}_n$ basiert. Dazu können wir unter anderem die folgenden beiden Grenzwertaussagen verwenden (Bemerkung 2.37):

i) $\sqrt{n}\,(\hat{\rho}_n - \rho)(1-\rho^2)^{-1} \xrightarrow{d} N(0,1)$,

ii) $\sqrt{n-3}\,(g(\hat{\rho}_n) - g(\rho) - \frac{\rho}{2(n-1)}) \xrightarrow{d} N(0,1)$, mit $g(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right)$ Fishers Z-Transformation.

Da aus diese Formen Konfidenzintervalle für ρ analytisch schwer bestimmbar sind, können diese numerisch berechnet werden. Alternativ kann man nach Bemerkung 2.33 (Slutsky) auch die folgenden Aussagen verwenden:

i') $\sqrt{n}\,(\hat{\rho}_n - \rho)(1-\hat{\rho}_n^2)^{-1} \xrightarrow{d} N(0,1)$,

ii') $\sqrt{n-3}\,(g(\hat{\rho}_n) - g(\rho) - \frac{\hat{\rho}_n}{2(n-1)}) \xrightarrow{d} N(0,1)$, mit $g(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right)$ Fishers Z-Transformation.

Bearbeiten Sie folgende Aufgaben:

a) Bestimmen Sie mit Hilfe von i) und ii), bzw. i') und ii'), jeweils ein approximatives 95%-Konfidenzintervall für ρ für die gegebenen Daten.

b) Gehen Sie davon aus, dass die wahre Korrelation $\rho = 0.9$ beträgt. Vergleichen Sie die beiden in a) konstruierten Konfidenzintervalle numerisch. Generieren Sie dafür $N = 10\,000$ Stichproben der Größe $n = 10$ von einer bivariaten Normalverteilung mit Korrelation $\rho = 0.9$ und Varianzen von jeweils 1. Bestimmen Sie jeweils beide 95%-Konfidenzintervalle und ermitteln Sie, wie oft der wahre Parameter tatsächlich überdeckt wurde. Halten die beiden Verfahren die vorgeschriebene Überdeckungswahrscheinlichkeit ein? Welches von beiden liefert im Durchschnitt kürzere Intervalle? Welches Verfahren ist vorzuziehen?

c) Wiederholen Sie Ihre Simulation aus Aufgabenteil b) für Stichprobenumfänge von $n = 500$ und $n = 1000$. Was beobachten Sie jetzt?

Hinweis: Die R-Funktion `rmvnorm` aus dem Paket `mvtnorm` kann Beobachtungen aus einer multivariaten Normalverteilung generieren.

Abgabe der Hausaufgaben bis zum 11.11.2019 12 Uhr. Gruppenarbeit unter Angabe der Gruppenmitglieder möglich, aber pro Person eine eigene handschriftliche Abgabe. Bei Software-Aufgaben schicken Sie bitte **zusätzlich** zum ausgedruckten Blatt (R-Code inkl. Ergebnisse) den zugehörigen Programmcode an ihren Übungsleiter.

Übung	Mailadresse	Briefkasten
Mittwoch 10:15 - 11:45 Uhr	jkoenig@statistik.tu-dortmund.de	131
Donnerstag 08:30 - 10:00 Uhr	nilsjannik.schuessler@tu-dortmund.de	132