Statistische Methoden der Datenanalyse
Beispielsammlung

R. Frühwirth

Institut für Hochenergiephysik
der Österreichischen Akademie der Wissenschaften
A-1050 Wien, Nikolsdorfer Gasse 18

Wintersemester 2013/2014

Übung 1

Beispiel 1.1

Es sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ ein Datensatz. Zeigen Sie, dass das Mittel aus dem 1. und dem 3. Quartil (midhinge) sowie der Mittelwert aller Daten zwischen dem 1. und dem 3. Quartil (interquartile mean, IQM) Lagemaße sind.

Beispiel 1.2

Es sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ ein Datensatz. Eine L-Statistik ist eine Linearkombination der *geordneten* Stichprobe. Zeigen Sie, dass eine L-Statistik ein Lagemaß ist, wenn die Koeffizienten symmetrisch und nichtnegativ sind und auf 1 summieren.

Beispiel 1.3

Es sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ ein Datensatz. Die Größe

$$y = \mathrm{med}_i \, |x_i - \mathrm{med}_j(x_j)|$$

wird MAD (median absolute deviation) genannt. Zeigen Sie, dass MAD ein Streuungsmaß ist.

Beispiel 1.4

Es sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ ein Datensatz. Zeigen sie, dass die Größe

$$S_n = \mathrm{med}_i(\mathrm{med}_j \, |x_i - x_j|)$$

ein Streuungsmaß ist.

Beispiel 1.5

Es sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ ein Datensatz. Zeigen sie, dass die Größe

$$z = (\bar{x} - \tilde{x})/\mathrm{IQR}$$

ein Schiefemaß ist, wobei $\bar{x}$ das Mittel und $\tilde{x}$ der Median des Datensatzes ist.

Beispiel 1.6 (Prog)

Bestimmen Sie die empirische Verteilung der folgenden Lagemaße anhand von 1000 simulierten Stichproben vom Umfang 100 aus einer Normalverteilung und vergleichen Sie die Verteilungen mit Hilfe von Boxplots:

- Arithmetisches Mittel
- Median
- Midhinge = Mittel aus dem 1. und dem 3. Quartil
- IQM = Mittelwert aller Daten zwischen dem 1. und dem 3. Quartil

Wiederholen Sie die Simulation mit einer kontaminierten Normalverteilung.

Beispiel 1.7 (Prog)

Bestimmen Sie die empirische Verteilung der folgenden Streuungsmaße anhand von 1000 simulierten Stichproben vom Umfang 100 aus einer Normalverteilung und vergleichen Sie die Verteilungen mit Hilfe von Boxplots:

- Standardabweichung
- IQR/1.349
- MAD $\cdot$ 1.4826
- $S_n \cdot 1.1926$

Wiederholen Sie die Simulation mit einer kontaminierten Normalverteilung.

Übung 2

Beispiel 2.1

Eine Münze wird dreimal geworfen. Was ist die Ergebnismenge des Experiments? Welches Ereignis beschreibt die Aussage „Es wird öfter Kopf als Zahl geworfen"?

Beispiel 2.2

A und B seien zwei Ereignisse mit $W(A) = 3/4, W(B) = 2/3$. Wie groß muß $W(A \cap B)$ mindestens sein? Wie groß ist $W(A \cap B)$ unter der Annahme der Unabhängigkeit von A und B? Berechnen Sie für diesen Fall die Wahrscheinlichkeiten der folgenden Ereignisse:

a) Keines der beiden Ereignisse tritt ein

b) Genau eines der beiden Ereignisse trit ein

c) Beide Ereignisse treten ein +

d) Mindestens eines der beiden Ereignisse tritt ein

e) Höchstens eines der beiden Ereignisse tritt ein

Beispiel 2.3

Beim Bau eines Gerätes werden 5 Widerstände und 4 Kondensatoren verwendet. Die Fehlerwahrscheinlichkeit der Widerstände sei 2%, die der Kondensatoren 3%. Berechnen Sie unter geeigneten Unabhängigkeitsannahmen die Wahrscheinlichkeit, dass mindestens zwei Bauteile fehlerhaft sind.

Beispiel 2.4

Sie werfen eine symmetrische Münze $2n$ mal. Wie groß ist die Wahrscheinlichkeit, genau n mal „Kopf" zu werfen? Wie verhält sich die Wahrscheinlichkeit für große n?

Beispiel 2.5

In einem Experiment wurde der Trigger so aufgesetzt, dass nur 1% der gewünschten Ereignisse verworfen wird. Die Analyse der aufgezeichneten Ereignisse ergab, dass davon 96% vom erwünschten Typ sind. Insgesamt wurden 92% aller Ereignisse aufgezeichnet.

a) Wie groß ist der ursprüngliche Anteil an erwünschten Ereignissen?

b) Wieviel % der Untergrundereignisse wurden vom Trigger verworfen?

c) Wie groß ist der Anteil an erwünschten Ereignissen an den nicht aufgezeichneten?

Beispiel 2.6

Ein Experiment verwendet ein große Zahl von ICs. Es bezieht diese von drei verschiedenen Herstellern A, B und C, und zwar von A und B je 25%, und von C 50%. Die Wahrscheinlichkeit, dass ein IC mindestens 40000 Stunden fehlerfrei arbeitet, beträgt für die drei Hersteller 0.92, 0.95 und 0.97.

a) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählter IC mindestens 40000 Stunden arbeitet?

b) Eine IC fällt vor Ablauf der 40000 Stunden aus. Mit welcher Wahrscheinlichkeit stammt er von A, B oder C?

Beispiel 2.7 (Prog)

Die Datei `trunclandau.txt` enthält einen Datensatz aus einer abgeschnittenen Landauverteilung. Bestimmen Sie Modus und FWHM (Breite bei halber maximaler Höhe) mittels Kernschätzer der Dichte. Stellen Sie die geschätzte Dichte und die empirische Verteilungsfunktion graphisch dar.

Beispiel 2.8 (Prog)

Für zwei Merkmale A und B gilt: $W(A) = 0.65$, $W(A|B) = 0.53$, $W(B|A) = 0.34$. Erzeugen Sie 1000 Vierfeldertafeln mit jeweils 250 Einträgen gemäß diesen Wahrscheinlichkeiten und untersuchen Sie die Verteilung der empirischen Vierfelderkorrelation. Wie ändert sich die Verteilung, wenn die Anzahl der Einträge verdoppelt wird?

Übung 3

Beispiel 3.1

Ein Messgerät zur Entfernungsmessung arbeitet mit Präzision σ^2, d.h. der Messwert streut um den wahren Wert x gemäß einer Normalverteilung $\mathcal{N}(x, \sigma^2)$. Wie groß darf σ maximal sein, damit der Messwert vom wahren Wert mit 99% Sicherheit nicht mehr als 1 mm abweicht?

Beispiel 3.2

Es sei $X \sim \mathrm{No}(\mu, \sigma^2)$. Bestimmen Sie die Dichte, den Erwartungswert und die Varianz von $Y = e^X$.

Beispiel 3.3

Es sei $X_1 \sim \mathrm{Ga}(a_1, b)$ und $X_2 \sim \mathrm{Ga}(a_2, b)$. Bestimmen Sie die Verteilung von $Y = X_1 + X_2$.

Beispiel 3.4

Es sei $X \sim \mathrm{Ga}(a, b)$. Bestimmen Sie die Dichte und den Erwartungswert von $Y = \ln X$.

Beispiel 3.5

Es sei $X \sim \mathrm{Ga}(5, 2)$. Bestimmen Sie den Erwartungswert und die Varianz von $Y = 1/X$ mit Fehlerfortpflanzung und vergleichen Sie mit den exakten Werten.

Beispiel 3.6

Die Lebensdauer eines Bauteils ist exponentialverteilt. Wie groß muß die mittlere Lebensdauer mindestens sein, damit ein Bauteil mit 50% Wahrscheinlichkeit nach einem Jahr noch funktioniert?

Beispiel 3.7 (Prog)

Die Lebensdauer eines elektrischen Bauteils ist exponentialverteilt mit dem Mittel τ. Sie schalten N gleichartige Bauteile gleichzeitig ein. Wie ist die Wartezeit bis zum ersten Ausfall verteilt? Wie ist die Wartezeit bis zum letzten Ausfall verteilt? Berechnen Sie jeweils den Mittelwert der Verteilung. Vergleichen Sie Ihre Rechnung mit dem Resultat einer Simulation.

Beispiel 3.8 (Prog)

Die Lebensdauer einer Glühbirne (in Stunden) ist normalverteilt mit dem Mittel $\mu = 1200$ und der Standardabweichung $\sigma = 25$. Sie schalten $N = 25$ gleichartige Birnen gleichzeitig ein. Wie ist die Wartezeit bis zum ersten Ausfall verteilt? Wie ist die Wartezeit bis zum letzten Ausfall verteilt? Vergleichen Sie Ihre Rechnung mit dem Resultat einer Simulation.

Übung 4

Beispiel 4.1

Es sei $X_i \sim \mathrm{Po}(\lambda_i),\ i = 1, 2$. Bestimmen Sie die Verteilung von $Y = X_1 + X_2$.

Beispiel 4.2

Es sei $X_i \sim \mathrm{Bi}(n_i, p),\ i = 1, 2$. Bestimmen Sie die Verteilung von $Y = X_1 + X_2$.

Beispiel 4.3

Sie messen die Aktivität einer Quelle mit einer mittleren Zerfallsrate λ. Die Ansprechwahrscheinlichkeit Ihres Detektors beträgt jedoch nur 80%. Wie ist die Zahl der beobachteten Zerfälle verteilt? Wie groß ist ihr Mittelwert? Wie ist die Wartezeit zwischen zwei beobachteten Zerfällen verteilt?

Beispiel 4.4

Sie wiederholen ein Bernoulli-Experiment mit Erfolgswahrscheinlichkeit p so lange, bis der erste Erfolg eintritt. Wie ist die Versuchsanzahl n verteilt? Bestimmen Sie Erwartungswert und Varianz von n mit Hilfe der charakteristischen Funktion. Berechnen Sie den Maximum-Likelihood-Schätzer von p und untersuchen Sie seine Eigenschaften.

Beispiel 4.5 (Prog)

Die Datei `truncexpo.txt` enthält einen Datensatz aus einer bei $x = 10$ abgeschnittenen Exponentialverteilung $\mathrm{Ex}(\tau)$. Schätzen Sie τ mit der Maximum-Likelihood-Methode und geben Sie den näherungsweisen Fehler der Schätzung an.

Beispiel 4.6 (Prog)

Schätzen Sie den Parameter a einer Gammaverteilung

$$f(x) = \frac{x^{a-1}\, e^{-x}}{\Gamma(a)}, \quad x \geq 0$$

mit der Maximum-Likelihood-Methode aus einer simulierten Stichprobe vom Umfang 250 (a=2). Bestimmen Sie den Fehler der Schätzung durch 1000-malige Wiederholung der Stichprobe. Vergleichen Sie die resultierende Verteilung des Schätzwertes mit einer individuellen Likelihoodfunktion.

Beispiel 4.7 (Prog)

Schätzen Sie den Skalenparameter s einer Cauchyverteilung

$$f(x) = \frac{1}{\pi s(1 + x^2/s^2)}, \quad x \in \mathbb{R}$$

mit der Maximum-Likelihood-Methode aus einer simulierten Stichprobe vom Umfang 500 (s=1.5). Bestimmen Sie den Fehler der Schätzung durch 1000-malige Wiederholung der Stichprobe. Vergleichen Sie die resultierende Verteilung des Schätzwertes mit einer individuellen Likelihoodfunktion.

Übung 5

Beispiel 5.1

Es sei $(X_1, \ldots, X_n)$ eine zufällige Stichprobe, wobei Beobachtung X_i aus der Normalverteilung $\mathrm{No}(\mu, \sigma_i^2)$ mit bekannter Varianz σ_i^2 stammt. Ferner sei

$$\tilde{X} = \sum_{i=1}^{n} w_i X_i$$

eine konvexe Kombination der Beobachtungen, d.h.

$$\sum_{i=1}^{n} w_i = 1, w_i \geq 0$$

Zeigen Sie, dass $\tilde{X}$ ein erwartungstreuer Schätzer von μ ist und bestimmen Sie die Faktoren ("Gewichte") w_i so, dass die Varianz von $\tilde{X}$ minimal ist.

Beispiel 5.2

Es sei $(X_1, \ldots, X_n)$ eine zufällige Stichprobe, wobei Beobachtung X_i aus der Normverteilung $\mathrm{No}(\mu, \sigma_i^2)$ mit bekannter Varianz σ_i^2 stammt. Bestimmen Sie den ML-Schätzer von μ und diskutieren Sie seine Eigenschaften.

Beispiel 5.3

Es sei $(X_1, \ldots, X_n)$ eine zufällige Stichprobe aus $\mathrm{No}(\mu, \sigma^2)$, mit unbekanntem μ und σ^2. Sie schätzen σ^2 durch

$$\Sigma = C \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Bestimmen Sie C so, dass

a) Σ unverzerrt ist;

b) Σ den minimalen quadratischen Fehler (MSE) hat.

Beispiel 5.4 (Prog)

Die Laplace-Verteilung $\text{La}(m, s)$ mit Lageparameter m und Skalenparameter s hat die Dichte

$$f(x; m, s) = \frac{1}{2s} \exp\left(-\frac{|x-m|}{s}\right)$$

a) Berechnen Sie Erwartung und Varianz der Verteilung.

b) Bestimmen sie die ML-Schätzer von m und s und diskutieren Sie ihre Eigenschaften. Untersuchen Sie Verzerrung und Varianz auch mittels Simulation.

c) Konstruieren Sie alternative Schätzer von m und s mit Hilfe des Stichprobenmittels bzw. der Stichprobenstandardabweichung und vergleichen Sie sie mit den ML-Schätzern.

Beispiel 5.5 (Prog)

Simulieren Sie eine Stichprobe vom Umfang $n = 200$ aus $\text{La}(m, s)$ mit $m = 1, s = 0.5$. Berechnen Sie den Maximum-Likelihood-Schätzer von (a, b) und ermitteln Sie eine näherungsweise Kovarianzmatrix aus der Log-Likelihoodfunktion.

Übung 6

Beispiel 6.1

Die Stichprobe $(X_1, \ldots, X_n)$ entstammt einer Gleichverteilung im Intervall $[a, b]$ mit unbekannten Grenzen a und b.

a) Berechnen Sie die ML-Schätzer von a und b und untersuchen Sie ihre Eigenschaften.

b) Korrigieren Sie die ML-Schätzer so, dass sie unverzerrt sind.

c) Bestimmen Sie den ML-Schätzer von $\mathsf{E}[X]$ und diskutieren Sie seine Eigenschaften.

d) Vergleichen Sie den ML-Schätzer von $\mathsf{E}[X]$ mit dem Stichprobenmittel und dem Stichprobenmedian.

Beispiel 6.2

Es sei $(X_1, \ldots, X_n)$ eine zufällige Stichprobe aus der Gammaverteilung $\text{Ga}(a, b)$. Berechnen Sie die Fisher-Informationsmatrix bezüglich a und b.

Beispiel 6.3 (Prog)

Simulieren Sie eine Stichprobe vom Umfang $n = 500$ aus $\text{Ga}(a, b)$ mit $a = 4, b = 2$. Berechnen Sie den Maximum-Likelihood-Schätzer von (a, b) und ermitteln Sie eine näherungsweise Kovarianzmatrix aus der Log-Likelihoodfunktion. Vergleichen Sie die Kovarianzmatrix mit der inversen Fisher-Informationsmatrix.

Beispiel 6.4 (Prog)

Der Datensatz `threeclusters.txt` enthält Beobachtungen aus drei normalverteilten Gruppen. Stellen Sie den Datensatz graphisch dar und schätzen Sie die Gruppenmittel und -varianzen mit dem EM-Algorithmus. Stellen Sie die Zuordnung der Datenpunkte zu den drei Gruppen graphisch dar. Vergleichen Sie die geschätzten Zuordnungen mit den wahren Zuordnungen.

Übung 7

Beispiel 7.1

In einer Stichprobe waren 27 von 500 erzeugten Produkte fehlerhaft. Bestimmen Sie das 95%-ige Konfidenzintervall für die Schadenswahrscheinlichkeit

a) nach Clopper-Pearson

b) mit Näherung durch Normalverteilung (Bootstrap, robust, Korrektur von Agresti-Coull).

Beispiel 7.2

Ein normalverteilter Datensatz hat Umfang $n = 200$. Es ist $\sum X_i = 559.65$ und $\sum X_i^2 = 1574.73$.

a) Schätzen Sie Mittelwert μ und Varianz σ^2 der Normalverteilung.

b) Geben Sie 95%-ige Konfidenzintervalle für beide Größen an.

Beispiel 7.3

Eine Messreihe vom Umfang $n = 250$ stammt aus einer Exponentialverteilung. Es ist $\sum X_i = 395.24$.

a) Schätzen Sie den Mittelwert τ der Exponentialverteilung.

b) Geben Sie 95%-ige Konfidenzintervalle für τ an (symmetrisch, linksseitig, rechtsseitig).

Beispiel 7.4 (Prog)

In einem Laboratorium wird mit einem Geigerzähler die Hintergrundstrahlung gemessen. Die Datenreihe in der Datei `poiss.txt` enthält die beobachteten Zählraten (pro Minute).

a) Schätzen Sie die mittlere Rate pro Minute λ mit der Maximum-Likelihood-Methode.

b) Geben Sie die Standardabweichung Ihrer Schätzung an.

c) Geben Sie 95%-ige Konfidenzintervalle für λ an (symmetrisch, linksseitig).

Beispiel 7.5 (Prog)

Konstruieren Sie zwei näherungsweise Konfidenzintervalle für den Lageparameter m einer Laplaceverteilung mit unbekanntem Skalenparameter s, unter Verwendung der asymptotischen Verteilung des Stichprobenmedians bzw. des Stichprobenmittels. Vergleichen Sie die Länge und die tatsächliche Überdeckungswahrscheinlichkeit der beiden Intervalle mittels Simulation für die Stichprobengrößen $n = 100, 200, 400, 800$.

Übung 8

Beispiel 8.1

Es sei Y geometrisch verteilt mit Erfolgswahrscheinlichkeit p.

a) Zeigen Sie, dass die Betaverteilung die konjugierte a-priori-Verteilung für p ist.

b) Berechnen Sie Jeffrey's prior für p, die entsprechende a-posteriori-Verteilung sowie deren Erwartung und Varianz.

Beispiel 8.2

Es sei X exponentialverteilt mit mittlerer Rate λ. Zeigen Sie, dass die Gammaverteilung die konjugierte a-priori-Verteilung für λ ist.

Beispiel 8.3 (Prog)

In 25 Wiederholungen eines Alternativversuchs erhalten Sie 8 Erfolge. Berechnen Sie die a-posteriori-Dichte der Erfolgswahrscheinlichkeit p und ihren Mittelwert

a) mit der a-priori-Dichte mit maximaler Entropie;

b) mit Jeffrey's prior.

Stellen Sie die a-posteriori-Dichte graphisch dar und berechnen Sie das symmetrische Vertrauensintervall und das HPD-Intervall (jeweils 95%).

Beispiel 8.4 (Prog)

In einem Laboratorium wird mit einem Geigerzähler die Hintergrundstrahlung gemessen. Die Datenreihe in der Datei `poiss1.txt` enthält die beobachteten Zählraten pro Sekunde. Berechnen Sie die a-posteriori-Dichte der mittleren Rate λ und ihren Mittelwert

a) mit der improperen konstanten a-priori-Dichte;

b) mit Jeffrey's prior.

Stellen Sie die a-posteriori-Dichte graphisch dar und berechnen Sie das linksseitige Vertrauensintervall und das HPD-Intervall (jeweils 95%). Wie muss die a-priori-Dichte aussehen, damit der a-posteriori-Mittelwert gleich dem ML-Schätzer von λ ist?

Beispiel 8.5 (Prog)

Simulieren Sie eine Stichprobe vom Umfang $n = 200$ aus der Exponentialverteilung mit Mittel τ, wobei τ zufällig aus einer a-priori-Gammaverteilung mit $a = 4, b = 1$ gezogen wird. Ermitteln Sie die a-posteriori-Dichte, ihren Mittelwert und das 95%-ige symmetrische Vertrauensintervall. Wiederholen Sie die Stichprobe 1000-mal und zählen Sie, wie oft der wahre Wert von τ im Vertrauensintervall liegt.

Übung 9

Beispiel 9.1

Bei einer Umfrage geben 171 von 500 befragtenPersonen an, Produkt X zu kennen. Testen Sie die Hypothese, dass der Bekanntheitsgrad mindestens 35% beträgt.

Beispiel 9.2

Ein normalverteilter Datensatz aus $\text{No}(\mu, \sigma^2)$ hat Umfang $n = 200$. Das Stichprobenmittel ist gleich $\bar{x} = 24.93$, die Stichprobenstandardbweichung ist gleich $S = 2.093$. Testen Sie

a) die Hypothese $\mu > 25$

b) die Hypothese $\sigma < 2$

mit dem Signifikanzniveau $\alpha = 0.01$.

Beispiel 9.3

Eine Messreihe vom Umfang $n = 250$ stammt aus einer Exponentialverteilung. Es ist $\sum X_i = 388.2$. Testen Sie die Hypothese $\tau < 1.5$ mit dem Signifikanzniveau $\alpha = 0.05$.

Beispiel 9.4

Das Ergebnis des 1000-maligen Würfelns eines Würfels sind folgende Häufigkeiten: 1: 170, 2: 152, 3: 164, 4: 179, 5: 158, 6: 177. Besteht Grund zu der Annahme, dass der Würfel nicht symmetrisch ist?

Beispiel 9.5

Bei einer Wahl vor zwei Jahren ergab sich für die Parteien A,B,C,D folgende Stimmenverteilung: A=33.4%, B=32.1%, C=18.5%, D=16%. Bei einer aktuellen Umfrage unter 500 Personen sieht die Parteipräferenz folgendermaßen aus: A=181, B=152, C=103, D=64. Besteht Grund zu der Annahme, dass sich das Wählerverhalten seit der Wahl geändert hat?

Beispiel 9.6 (Prog)

In einem Laboratorium wird mit einem Geigerzähler die Hintergrundstrahlung gemessen. Die Datenreihe in der Datei `poiss1.txt` enthält die beobachteten Zählraten pro Sekunde. Testen Sie die Hypothese, dass die mittlere Rate der Hintergrundstrahlung höchstens 2 pro Sekunde ist, mit dem Signifikanzniveau $\alpha = 0.01$.

Beispiel 9.7 (Prog)

Die Datei `truncexpo.txt` enthält einen Datensatz aus einer bei $x = 10$ abgeschnittenen Exponentialverteilung $\mathrm{Ex}(\tau)$. Schätzen Sie τ mit der Methode der kleinsten Fehlerquadrate und geben Sie den näherungsweisen Fehler der Schätzung an.

Beispiel 9.8 (Prog)

Die Datei `survival.txt` enthält die Lebenszeit (in Tagen) von 50 Labormäusen mit Krebstumoren an, die mit einer bestimmten Krebstherapie behandelt wurden.

a) Nehmen Sie an, dass die Daten log-normalverteilt sind, und schätzen Sie die Parameter μ und σ^2 mit der Maximum-Likelihood-Methode.

b) Entscheiden sie mit Hilfe des Kolmogorov-Smirnov-Tests auf einem Signifikanzniveau von 5%, ob die Daten aus einer Log-Normalverteilung mit $\mu = 3$ und $\sigma = 0.5$ stammen können.

Hinweis: Daten sind Log-normalverteilt mit den Parametern μ und σ^2, wenn ihre natürlichen Logarithmen normalverteilt mit Mittel μ und Varianz σ^2 sind.

Beispiel 9.9 (Prog)

Die Zufallsgröße s_n ist die standardisierte Summe von n unabhängigen in $(0, 1)$ gleichverteilten Zufallsgrößen. Simulieren Sie jeweils 10000 Werte von s_n und testen Sie, ob die Verteilung von s_n signifikant von einer Standardnormalverteilung abweicht.