Data Science in R (DSR) – Testat 2

Markus Gumbel

03. November 2026

Inhaltsverzeichnis

Hinweise **1**

1 Zentrale Grenzwertsatz **3**
1.1 Verteilung . . . 3
1.2 Plot . . . 3
1.3 Abweichung . . . 3

2 Untersuchungen zur koronaren Herzkrankheit **4**
2.1 Einlesen der Herz-Daten . . . 5
2.2 Cholesterin im Vergleich Männer/Frauen . . . 5
2.3 Cholesterin im Vergleich zur Erkrankung . . . 5
2.4 Systolischer Ruheblutdruck . . . 6

3 Clustering und PCA auf die Herzdaten **7**
3.1 Einlesen der Herz-Daten . . . 7
3.2 Bedeutet "ähnliche Merkmale" auch "gleiche Diagnose"? . . . 7
3.3 PCA . . . 8

Hinweise

Aktion		Datum	Ablauf
Ausgabe	Di.,	03.11.26	über Moodle
Abgabe	Mo.,	23.11.26, 09:00 Uhr	über Moodle
Testat	Mo.,	23.11.26	als Präsentation

- Die Abnahme der Testate gilt als Prüfungsleistung. Bei einer Verhinderung durch Krankheit ist eine ärztliche Bescheinigung der Arbeitsunfähigkeit vorzulegen.
- Geben Sie Ihre Lösungen fristgerecht via Moodle ab. Es werden dort folgende Abgaben erwartet (Ausarbeitungen in einem anderen Format werden nicht berücksichtigt):
  - Für jede Aufgabe ein R-Notebook (als *.Rmd-Datei) und etwaige benötigten R-Skripte (*.R).
  - Genau ein PDF mit allen Lösungen, das außerdem die Gruppennummer und alle Namen enthält.
- Nutzen Sie das R-Notebook, das es für jede Aufgabe gibt und **fügen Sie dort Ihre Lösungen ein**, d.h. Source-Code und normaler Text.
- **Kommentieren Sie Ihre Lösungen ausführlich** – entweder als normalen Text im Notebook oder als Kommentar in R-Anweisungen. Ohne Kommentare gibt es Punktabzug.

- Jedes abgegebene R-Notebook muss ohne Fehler auf einem anderen Rechner mit R (4.1 oder höher) ausgeführt werden können und ein HTML-Dokument erzeugen. Installationsanweisungen für die benutzen Packages sind nicht nötig – laden Sie aber die Packages im Notebook.
- Während der Abnahme sind die Ergebnisse am Rechner live zu demonstrieren.
- Bei der Abnahme der Übung ist der Studentenausweis vorzulegen.

1 Zentrale Grenzwertsatz

Für jedes $n = 1, 2, \ldots$ seien die Zufallsvariablen $X_1, X_2, \ldots, X_n$ unabhängig und besitzen die gleiche Verteilung mit dem Erwartungswert $\mu = E(X_i)$ und der Varianz $\sigma^2 = Var(X_i)$. Dann gilt für die Verteilungsfunktion der standardisierten Summen $G_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$:

$$\lim_{n\to\infty} P\left(G_n \leq x\right) = \Phi(x)$$

Dabei ist Φ die Verteilungsfunktion der Standard-Normalverteilung $N(0, 1)$.

Überprüfen Sie mit R den zentralen Grenzwertsatz.

1.1 Verteilung

Erzeugen Sie einen Data Frame, der die gezogenen Zufallszahlen G_n enthält. Die Stichprobe soll einen Umfang von 50000 haben (Anzahl Zeilen der Tabelle). Der Data Frame hat sechs Spalten, die die Werte für die Zahlen $n = 1, 2, 3, 10, 100, 1000$ enthalten. Geben Sie die ersten 10 Zeilen des Data Frames aus.

Welche Verteilung Sie für X_i nehmen, bleibt Ihnen überlassen. Es soll nur keine Normalverteilung sein.

```
# Ihre Lösung:
```

1.2 Plot

Plotten Sie sechs Histogramme, die je für $n = 1, 2, 3, 10, 100, 1000$ die Verteilung im Vergleich zu einer $(0, 1)$-Normalverteilung zeigen. Die Intervallbreite soll 0,25 sein.

Tipp: Platzieren Sie die sechs Plots auf einem 3×2-Gitter.

```
# Ihre Lösung:
```

1.3 Abweichung

Nun soll die Abweichung der Standardisierten G von der Normalverteilung für alle n quantifiziert werden. Hierzu soll für alle Balken (bins) der Histogramme aus der vorigen Aufgabe die Differenz von G zur Normalverteilung gebildet und quadriert werden. Diese Werte werden aufaddiert und durch die Anzahl der Intervalle geteilt. Daraus wird die Wurzel gezogen.

Sie können sich auch ein anderes Maß zur Bestimmung der Abweichung überlegen.

Geben Sie die Abweichungen aus. Stimmt es, dass die Abweichungen mit größerem n kleiner werden?

Tipp: `hist(plot = FALSE)` erzeugt ein Histogramm, ohne es zu plotten. Gerne können Sie auch Ihr eigenes Histogramm nutzen.

```
# Ihre Lösung:
```

2 Untersuchungen zur koronaren Herzkrankheit

In diesem Abschnitt sollen Daten von Probanden bzw. Patienten auf das Risiko für koronare Herzkrankheit untersucht werden. Dies ist eine Erkrankung der Herzkranzgefäße (Koronararterien), die sich durch Ablagerungen in den Gefäßwänden verengen. Der Original-Herz-Datensatz ist unter

- https://archive.ics.uci.edu/ml/datasets/Heart+Disease

beschrieben. Wir nutzen eine konsolidierte CSV-Datei, die bereits Header enthält. Download unter:

- https://oc.informatik.hs-mannheim.de/s/wyzFq34K9HiNjXR/download

Die Datei enthält 13 Merkmale, die einen Einfluss auf eine koronare Herzkrankheit haben können. Das 14. Merkmal `goal` (im Original auch `num`) ist die Diagnose (Klassifizierung). Der Wert ist 0, falls keine krankhafte Verengung der Gefäße vorliegt, oder 1, 2, 3 oder 4, falls – je nach Stärke – eine krankhafte Verengung der Gefäße vorliegt. Wir unterscheiden im Folgenden nur die Zustände "gesund" (0) und "krank" (1, 2, 3, 4). Unter

- https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

finden Sie eine Beschreibung aller Attribute. Hier ist eine Zusammenfassung. Wir benötigen insbesondere die Merkmale `sex`, `trestbps`, `chol` und `goal`.

Feld	Bedeutung
age	age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
trestbps	resting systolic blood pressure (in mmHg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest.[1]
slope	slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
ca	number of major vessels (0-3) colored by flourosopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect
goal	diagnosis of heart disease (0: < 50% diameter narrowing ; 1, 2, 3, 4: > 50% diameter narrowing)

[1] ST depression refers to a finding on an electrocardiogram, wherein the trace in the ST segment is abnormally low below the baseline.

2.1 Einlesen der Herz-Daten

Lesen Sie die Datei aus der URL als Data Frame zur weiteren Bearbeitung ein. Überlegen Sie, ob sie Faktoren sinnvoll einsetzen können. Geben Sie die ersten drei Zeilen und fünf Spalten aus[2]:

```
# Ihre Lösung:
```

2.2 Cholesterin im Vergleich Männer/Frauen

Nun sollen die Cholerstin-Werte untersucht werden – zunächst im Vergleich Männer zu Frauen.

2.2.1 Überblick über Cholesterin-Daten

Verschaffen Sie sich einen Überblick, indem Sie ein Boxplot für das Cholesterin gruppiert nach dem Geschlecht plotten.

```
# Ihre Lösung:
```

2.2.2 Konfidenz-Intervall

Berechnen Sie das Konfidenz-Intervall (Niveau 95%) für den Cholesterin-Level jeweils für Männer und Frauen.

Tabelle Geben Sie das Ergebnis als `kable`-Tabelle aus:

```
# Ihre Lösung:
```

Überlappung? Überlappen sich die Bereiche?

2.3 Cholesterin im Vergleich zur Erkrankung

Nun sollen die Cholesterin-Werte in Abhängigkeit der Diagnose untersucht werden.

2.3.1 Überblick über Cholesterin-Daten

Verschaffen Sie sich einen Überblick, indem Sie ein Boxplot für das Cholesterin gruppiert nach der Diagnose plotten.

```
# Ihre Lösung:
```

2.3.2 Konfidenz-Intervall

Berechnen Sie die Konfidenzintervalle für beide Gruppen und geben Sie das Ergebnis als `kable`-Tabelle aus:

```
# Ihre Lösung:
```

2.3.3 Test

Es sieht so aus, als ob der Cholesterin-Wert bei den erkrankten Patienten höher ist als bei den nicht erkrankten. Überprüfen Sie das mit einem Hypothesen-Test.

Wie lauten die Hypothesen? Formulieren Sie die Hypothesen (H_0 und H_1).

[2] Möglicherweise kommt es zu einem Fehler beim Einlesen des ersten Attributs (`age`). Manuelles Umbenennen hilft.

Testanwendung Wenden Sie den Test mit R an. Was ist das Ergebnis?

```
# Ihre Lösung:
```

2.4 Systolischer Ruheblutdruck

Der systolische Blutdruck liegt beim gesunden Menschen bei ca. 120 mmHg.

2.4.1 Überblick über Blutdruck

Plot Verschaffen Sie sich einen Überblick, indem Sie ein Boxplot für den Blutdruck in Ruhe gruppiert nach der Diagnose plotten.

```
# Ihre Lösung:
```

Normalverteilt? Kann überhaupt davon ausgegangen werden, dass die Daten normalverteilt sind?

2.4.2 Konfidenzintervalle nach Erkrankung

Berechnen Sie die Konfidenzintervalle für den Ruheblutdruck aufgeschlüsselt nach der Diagnose (erkrankt/nicht erkrankt) und geben Sie das Ergebnis als `kable`-Tabelle aus:

```
# Ihre Lösung:
```

2.4.3 Test, ob Kranke höheren Ruhe-Blutdruck haben

Überprüfen Sie mit einem Hypothesen-Test, ob Erkrankte einen höheren Ruhe-Blutdruck haben als gesunde Probanden.

Wie lauten die Hypothesen? Formulieren Sie die Hypothesen (H_0 und H_1).

Testanwendung Wenden Sie den Test mit R an. Was ist das Ergebnis?

```
# Ihre Lösung:
```

3 Clustering und PCA auf die Herzdaten

3.1 Einlesen der Herz-Daten

Es werden wieder die Herzdaten aus der letzten Aufgabe genutzt. Lesen Sie diese als Data Frame ein.

```
# Ihre Lösung:
```

3.2 Bedeutet "ähnliche Merkmale" auch "gleiche Diagnose"?

Für jeden Datensatz ist bekannt, zu welcher Klasse er gehört: 0 (gesund) und 1 (erkrankt). Wir wollen untersuchen, wie gut *ähnliche* Datensätze zur gleichen Klasse gehören. Dafür soll mit dem k-means-Clusterverfahren der Datensatz in zwei Cluster eingeteilt werden.

3.2.1 Nur reelle Merkmale

Zunächst sollen **nur die nummerischen Merkmale** benutzt werden und nicht jene, die Faktoren sind.

Clustering Clustern Sie diese Daten. Überlegen Sie, ob Sie die Daten standardisieren wollen.

```
# Ihre Lösung:
```

Richtig? Berechnen Sie, wie viel Prozent der Datensätze richtig einem Cluster eingeordnet wurden und geben Sie die Zahl auf zwei Nachkommastellen gerundet aus.

Hinweis: Berücksichtigen Sie, dass die Vergabe der Clusternummern zufällig ist. D.h. sowohl die Cluster (1, 2) wie auch (2, 1) sind möglich.

```
# Ihre Lösung:
```

Scatterplot age vs. thalach Plotten Sie die Merkmale `age` und `thalach` als Scatterplot. Färben Sie die Punkte gemäß ihrer Clusterzuordnung ein. Die Form (`shape`) eines Punkts soll zeigen, ob die Klassifikation (d.h. der Cluster) richtig oder falsch ist.

```
# Ihre Lösung:
```

3.2.2 Mit Dummy-Variablen

Nun sollen **alle Merkmale** benutzt werden.

Clustering Clustern Sie diese Daten. Überlegen Sie, wie die Faktoren zu Zahlen werden.

```
# Ihre Lösung:
```

Richtig? Berechnen Sie für diesen Fall, wie viel Prozent der Datensätze richtig einem Cluster eingeordnet wurden und geben Sie die Zahl auf zwei Nachkommastellen gerundet aus. Wie hat sich der Wert verändert? Warum ist dies so?

```
# Ihre Lösung:
```

Scatterplot age vs. thalach Plotten Sie erneut und schauen Sie, wie die richtigen nun Punkte verteilt sind.

```
# Ihre Lösung:
```

3.3 PCA

Wenden Sie eine PCA auf diesen Datensatz an. Es sollen alle Merkmale berücksichtigt werden.

3.3.1 Wichtige Merkmale

Welche Merkmale der ersten Hauptkomponente tragen am meisten zur Varianz bei? Geben Sie die TOP-10-Merkmale an.

```
# Ihre Lösung:
```

3.3.2 Erste und zweite Hauptkomponente

Plotten Sie die erste und zweite Hauptkomponente als Scatterplot. Färben Sie die Punkte gemäß ihrer Klasse (Disease) ein.

```
# Ihre Lösung:
```

3.3.3 PVE

Plot Plotten Sie die Proportion of Variance explained (PVE) für jede Hauptkomponente sowie die akkumulierte PVE.

Wichtige Hauptkomponenten Wie viele Hauptkomponenten erklären mehr als 50% der Varianz?

Möglicherweise tragen bei Ihrem Ergebnis die letzten Hauptkomponenten keine Varianz mehr bei. Überlegen Sie, woran das liegen könnte.