Dr. W. Herff

RWTH Aachen, WS 2024/25

29. November 2024

# Einführung in die angewandte Statistik mit R

## 1. Test

## Allgemeine Hinweise:

- Der Test ist bestanden, wenn
  - mindestens **2** der insgesamt 3 Programme ohne Fehlermeldung „laufen“ (d.h. eine Ausgabe bzw. eine Graphik erzeugen) **und**
  - mindestens **28** von 70 möglichen Punkten erreicht werden.
- Erstellen Sie **zu jeder Aufgabe jeweils ein eigenes Programm**, und speichern Sie diese Programme bitte **auf dem Desktop** in der folgenden Form ab, wobei *Matrikelnummer* Ihre Matrikelnummer und *Nr.* die betreffende Aufgaben-Nr. angibt:

  *Matrikelnummer_Nr.R* (z.B. *123456_1.R*)

- Kopieren Sie abschließend die von Ihnen erstellten Test-Programme vom Desktop in das

  Unterverzeichnis Abgabe des Verzeichnisses 20241129-RPrak-Test1

  (auf dem Abgabe-Laufwerk der Computer des Zuselab).
- Fügen Sie bitte jedem Programm zu Beginn eine Kommentarzeile mit Ihrem Namen, Ihrer Matrikelnummer und der Aufgabennummer hinzu.
- Erstellen Sie bitte selbst zu jeder erzeugten Graphik einen geeigneten Titel und sinnvolle Achsenbezeichnungen, sofern diese nicht in der Aufgabenstellung vorgegeben werden.
- Gestalten Sie bitte Ihre Programme übersichtlich, und achten Sie darauf, dass die Programme gut ausgedruckt werden können (**maximal 80 Zeichen pro Zeile, ggf. Leerzeilen zwischen Programmteilen einfügen**).

  Überflüssige Programmpassagen sollten Sie aus dem Programm entfernen.
- Tragen Sie die gefragten Auswertungsergebnisse bitte jeweils in das Zusatzblatt ein.

**Viel Erfolg !**



Sie die gefragten Auswertungsergebnisse bitte jeweils in das Zu-
ein

## Aufgabe 1 (30 Punkte)

Im Rahmen einer Untersuchung von Verspätungen der Deutschen Bahn wurden 2023 die Verspätungszeiten verschiedener Zug-Kategorien auf bestimmten Verbindungsstrecken ermittelt. Die dieser Aufgabe zugrundeliegende Datenauswahl umfasst die Verpätungszeiten (in Minuten) zu 22 ICE-Zügen und 34 Nahverkehrs-Zügen. Diese Zeiten sind zusammen mit den zugehörigen Zug-Bezeichnungen in der Datei `verspaetung.dat` enthalten (im Unterverzeichnis Daten des Verzeichnisses `20241129-RPrak-Test1` auf dem `Abgabe`-Laufwerk).

Erstellen Sie zu diesen Daten **ein** R-Programm, mittels dessen die folgenden Teilaufgaben gelöst werden.

(i) Lesen Sie die Einträge der Datei `verspaetung.dat` in R ein, erzeugen Sie hieraus die Datentabelle mit der Bezeichnung `versp`, und geben Sie diese Datentabelle in der Konsole aus. Übernehmen Sie beim Einlesen die Merkmalbezeichnungen aus der Datei `verspaetung.dat`.

(ii) Berechnen Sie mittels R zu sämtlichen Verspätungszeiten (d.h. für beide Zug-Kategorien zusammen) den empirischen Median und den Quartilsabstand, und geben Sie die berechneten Kenngrößen jeweils in der Konsole aus. Verwenden Sie hierbei die Definition empirischer Quantile für metrische Daten aus der Vorlesung.

(iii) Erstellen Sie mittels R zu sämtlichen Verspätungszeiten (d.h. für beide Zug-Kategorien zusammen) ein Histogramm der absoluten Klassenhäufigkeiten zu folgender Klasseneinteilung:

$$[0,3]\ ,\ (3,6]\ ,\ (6,9]\ ,\ (9,12]\ ,\ (12,15]\ ,\ (15,18]\ ,\ (18,21]\ ,\ (21,24]\ .$$

Berücksichtigen Sie hierbei die folgenden Vorgaben:

- Verwenden Sie die (globale) Graphik-Option, mittels der die Zahlenwerte an der vertikalen Achse aufrecht (und damit besser ablesbar) dargestellt werden.
- Wählen Sie eine geeignete Skalierung für die horizontale Achse, so dass die einzelnen Klassengrenzen als Zahlenwerte an der horizontalen Achse angegeben werden.
- Erstellen Sie selbst passende Achsenbezeichnungen und einen geeigneten (aussagekräftigen) Titel.

(iv) Erzeugen Sie aus der Datentabelle `versp` die beiden Teil-Datentabellen zu den einzelnen Zug-Kategorien, in denen nur die Verspätungszeiten der ICE-Züge bzw. der Nahverkehrs-Züge zusammen mit der jeweils zugehörigen Zug-Bezeichnung enthalten sind (also jeweils beide Merkmale).

Geben Sie diese beiden Teil-Datentabellen ebenfalls jeweils in der Konsole aus.

(v) Berechnen Sie zu jeder Zug-Kategorie die Stichproben-Standardabweichung der betreffenden Verspätungszeiten, und geben Sie diese jeweils in der Konsole aus.

(vi) Erzeugen Sie eine Graphik, in der die beiden Box-Plots der Verspätungszeiten zu jeder Zug-Kategorie gemeinsam (senkrecht) dargestellt werden. Berücksichtigen Sie hierbei die folgenden Vorgaben:

- Wählen Sie unterschiedliche Darstellungsfarben für die Ränder (bzw. Linien) der beiden Box-Plots.
- Kennzeichnen Sie die beiden Box-Plots durch die zugehörigen Zug-Bezeichnungen (an der horizontalen Achse).
- Wählen Sie die Bezeichnung „Verspaetungszeiten (in Minuten)“ für die vertikale Achse, und erstellen Sie einen geeigneten (aussagekräftigen) Titel.

Tragen Sie bitte in das Zusatzblatt (auf Seite 5) stichwortartig ein, was Ihnen im Vergleich der beiden Box-Plots auffällt.

## Aufgabe 2 (16 Punkte)

Innerhalb von OECD-Studien wird unter anderem regelmäßig auch das Konsumverhalten Jugendlicher hinsichtlich Alkohol, Zigaretten und Drogen untersucht. Hierbei ergaben sich für das Jahr 2014 im Vergleich der drei Länder Belgien, Deutschland und Italien folgende Anzahlen zum Alkoholkonsum in der Altersgruppe 15 – 19 Jahre:

| | Alkoholkonsum | | | | |
|---|---|---|---|---|---|
| Land | taeglich | woechentlich | monatlich | selten | nie |
| Belgien | 1 | 62 | 51 | 16 | 42 |
| Deutschland | 5 | 309 | 435 | 199 | 293 |
| Italien | 3 | 121 | 177 | 83 | 524 |

(Hierbei bedeuten „selten“ weniger als einmal pro Monat und „nie“ keinen Alkoholkonsum innerhalb der letzten 12 Monate.)

Erstellen Sie zu diesen Daten **ein** R-Programm, mittels dessen die folgenden Teilaufgaben gelöst werden.

(i) Erzeugen Sie zu den beiden Merkmalen `Land` und `Alkoholkonsum` eine Kontingenztafel, deren Darstellung der oben angegebenen Tabelle entspricht, und geben Sie diese Kontingenztafel in der Konsole aus.

(ii) Ergänzen Sie die erzeugte Kontingenztafel um die zugehörigen (absoluten) Randhäufigkeiten, und geben Sie auch diese erweiterte Tabelle in der Konsole aus.

(iii) Erzeugen Sie weiter die Kontingenztafel der zugehörigen (bedingten) relativen Häufigkeiten des Alkoholkonsums bezogen auf die drei Länder. Geben Sie diese Kontingenztafel **mit einer Rundung auf vier Dezimalstellen** ebenfalls in der Konsole aus.

(iv) Berechnen Sie den zugehörigen korrigierten Kontingenzkoeffizienten nach Pearson, und geben Sie diesen Koeffizienten in der Konsole aus.

**Hinweis:** Die $\chi^2$-Größe zu den gegebenen Daten kann einfach berechnet werden mittels `chisq.test(Tabelle)$statistic`, wobei `Tabelle` die Kontingenztafel der absoluten Häufigkeiten (ohne Randhäufigkeiten) bezeichnet.
Ignorieren Sie hierbei die für die Berechnung der $\chi^2$-Größe unrelevante Warnmeldung.

Tragen Sie bitte in das Zusatzblatt (auf Seite 6) den Gesamtumfang der in dieser Aufgabe betrachteten Jugendlichen sowie den berechneten korrigierten Kontingenzkoeffizienten nach Pearson ein.
Bewerten Sie weiter die Stärke des Zusammenhangs zwischen den beiden Merkmalen `Land` und `Alkoholkonsum`, die sich aufgrund des von Ihnen berechneten korrigierten Kontingenzkoeffizienten nach Pearson ergibt.

## Aufgabe 3 (24 Punkte)

Zu einer 2014 installierten Photovoltaik-Anlage (Solar-Anlage) wurde am Ende eines jeden Jahres jeweils die gesamte bislang erzeugte Energie der Anlage seit der Inbetriebnahme erfasst. Die jeweilige Laufzeit (in Jahren) und die in den betreffenden Zeiträumen erzeugte Energie (in kWh) sind in der Datei solar.csv enthalten (im Unterverzeichnis Daten des Verzeichnisses 20241129-RPrak-Test1 auf dem Abgabe-Laufwerk).

Erstellen Sie hierzu **ein** R-Programm zur Lösung der folgenden Teilaufgaben.

(i) Lesen Sie die Einträge der Datei solar.csv in R ein, erzeugen Sie hieraus die Datentabelle mit der Bezeichnung `solar`, und geben Sie diese Datentabelle in der Konsole aus. Übernehmen Sie beim Einlesen die Merkmalbezeichnungen aus der Datei solar.csv.

(ii) Berechnen Sie mittels R den Korrelationskoeffizienten von *Bravais/Pearson* zu den beiden betrachteten Merkmalen Laufzeit und Energie.

(iii) Nehmen Sie an, dass sich die gesamte von der Photovoltaik-Anlage erzeugte Energie in Abhängigkeit von der Laufzeit durch den folgenden Regressions-Ansatz mit einem (unbekannten) Parameter $b \in \mathbb{R}$ geeignet beschreiben lässt:

$$\texttt{Energie} = b \cdot \texttt{Laufzeit} \ .$$

Berechnen Sie hierzu mittels R die Kleinste-Quadrate-Schätzung $\hat{b}$ des Regressionskoeffizienten $b$, und geben Sie diese Schätzung in der Konsole aus.

(iv) Erzeugen Sie mittels R ein (zweidimensionales) Streudiagramm zu den beiden Merkmalen der Datentabelle `solar`, mittels dessen die Energiewerte über den Laufzeiten dargestellt werden. Fügen Sie weiter dem Streudiagramm die durch die Kleinste-Quadrate-Schätzung $\hat{b}$ festgelegte Regressionsgerade hinzu. Berücksichtigen Sie hierbei die folgenden Vorgaben:

- Wählen Sie als Darstellungsbereiche das Intervall $[0, 11]$ für die horizontale und das Intervall $[0, 80\,000]$ für die vertikale Koordinatenachse.
- Setzen Sie die Offsets an den Koordinatenachsen jeweils auf $0$.
- Wählen Sie unterschiedliche Darstellungsfarben für die Datenpunkte und die Regressionsgerade.
- Erstellen Sie selbst geeignete Achsenbezeichnungen und einen geeigneten (aussagekräftigen) Titel.

Zu gegebenen Datenpaaren $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ werde die Anpassung einer Regressionsgeraden durch den Koordinatenursprung betrachtet (wie in (iii)).

In diesem Spezialfall des einfachen linearen Regressionsmodells sind die Kleinste-Quadrate-Schätzung $\hat{b}$ für den Steigungsparameter $b$ und das Bestimmtheitsmaß $B(\hat{b})$ gegeben wie folgt (unter der Voraussetzung $(x_1, \ldots, x_n) \neq (0, \ldots, 0)$ und $(y_1, \ldots, y_n) \neq (0, \ldots, 0)$):

$$(*) \qquad \hat{b} = \frac{\sum_{i=1}^{n} x_i \, y_i}{\sum_{i=1}^{n} x_i^2} \quad \text{und} \quad B(\hat{b}) = \hat{b}^2 \, \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$$

(v) Erzeugen Sie eine eigene R-Funktion `my.reg()`, die zu zwei eingegebenen Vektoren $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ und $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ die Kleinste-Quadrate-Schätzung $\hat{b}$ und das Bestimmtheitsmaß $B(\hat{b})$ gemäß der Darstellung $(*)$ berechnet und ausgibt. Beachten Sie hierbei die folgenden Vorgaben:

- Runden Sie die **Ausgabe-Ergebnisse** der von Ihnen erzeugten R-Funktion `my.reg()` auf vier Dezimalstellen.
- Ergänzen Sie den Ausgabe-Vektor um geeignete Komponenten-Bezeichnungen.

**Hinweis:** Warnmeldungen hierzu müssen *nicht* ergänzt werden und werden dementsprechend auch *nicht* bewertet.

(vi) Wenden Sie die von Ihnen erstellte R-Funktion `my.reg()` gemäß des in (iii) betrachteten Regressionsansatzes auf die beiden Merkmale der Datentabelle `solar` an.

Tragen Sie in das Zusatzblatt (auf Seite 6) bitte die berechnete Kleinste-Quadrate-Schätzung, das berechnete Bestimmtheitsmaß und die hieraus resultierende Anpassungsgüte ein.

## Ergebnisse der statistischen Auswertungen

**Aufgabe 1** **(Bewertungsanteil: 6 Punkte)**

Im Vergleich der beiden Box-Plots zeigen sich folgende Auffälligkeiten:
(Bitte jeweils nur **stichwortartig** angeben, aber gegebenenfalls begründen, worauf die Beobachtung basiert!)

Für ICE gibt es ein ~~Maximalpunkt~~ $x_{(6)} > x_0 = \max\{x_{(i)}\}$!
unklar, was gemeint ist

~~Deshalb ICE hat in diesem Fall~~
die Median von Nah ist größer als ICE, d.h
Nah immer ~~spätzer~~ mit Spät. (+1)

**Aufgabe 2** **(Bewertungsanteil: 3 Punkte)**

Der Gesamtumfang der in Aufgabe 2 betrachteten Jugendlichen lautet:

167,2845 J. 2321 (-1)

Der Wert des berechneten korrigierten Kontingenzkoeffizienten nach Pearson lautet:

NAN. fehlt (-1)

$C_K = 0{,}4089$

Aufgrund des berechneten korrigierten Kontingenzkoeffizienten nach Pearson ergibt sich zwischen den beiden Merkmalen `Land` und `Alkoholkonsum`

- ☒ kein
- ☐ ein sehr schwacher
- ☐ ein mittlerer
- ☐ ein starker
- ☐ ein sehr starker

Welches? (-1)

Zusammenhang.

**Aufgabe 3** **(Bewertungsanteil: 3 Punkte)**

Die berechnete Kleinste-Quadrate-Schätzung des unbekannten Regressionskoeffizienten $b$ lautet:

7621 (V) Ff. (+1)

Der Wert des zugehörigen Bestimmtheitsmaßes lautet:

1,0 (V) Ff. (+1)

Auf Grundlage dieses Wertes des Bestimmtheitsmaßes ergibt sich eine

Best.heitsmaß = 1 ⇒ ~~gute~~ sehr gute Anpassung

- ☒ sehr schlechte
- ☐ schlechte
- ☐ mittlere
- ☐ gute
- ☐ sehr gute

Anpassung der zugehörigen Regressionsgeraden an die Datenpunkte.

Bestimmtheitsmaßes = 1 ~~des~~

## Zusammenfassung der Bepunktung

| | Aufgabe 1 | Aufgabe 2 | Aufgabe 3 | Gesamt-Punktzahl |
|---|---|---|---|---|
| Programm | 18 /24 | M /13 | 16 /21 | |
| Auswertungen | 1 /6 | 0 /3 | 2 /3 | |
| Summe | 19 /30 | M /16 | 18 /24 | 48 /70 |