Andreas Bender, Philipp Kopper
Institut für Statistik

Einführung in die Statistische Software (R) – Hausarbeit 2

Winter Semester 2021/22, 21.12.2021 - 06.01.2022

Name: _____

Immatrikulationsnummer: _____

Studiengang: _____

Hiermit bestätige ich, dass ich die Anweisungen auf diesem Blatt gelesen und verstanden habe. Ich bestätige, dass die abgegebene Lösung vollständig und alleinig von mir bearbeitet und erstellt worden ist, ohne Hilfe von anderen in Anspruch zu nehmen. Ich bestätige, dass ich über die Vorlesungsmaterialien hinausgehende Quellen wie Bücher oder Internetseiten im Code angegeben und falls zutreffend verlinkt sind.

Unterschrift: _____

Prüfungshinweise:

1. Überprüfen sie ob die heruntergeladene Angabe vollständig ist. Sie sollte 3 Aufgabenblöcke enthalten. Einzelne Aufgabeblöcke können aus mehreren Teilaufgaben bestehen.
2. Insgesamt können (ohne Bonuspunkte) 20 Punkte erreicht werden. Die Aufteilung der Punkte auf die einzelnen Aufgabenblöcke kann der Angabe entnommen werden.
3. Die Lösungen sollen in Form von `.Rmd` Dateien abgegeben werden. Nutzen Sie Markdown um Beginn und Ende einzelner Aufgaben und Teilaufgaben zu kennzeichnen. Ist die Zugehörigkeit von Code zu einer der (Teil-)Aufgaben nicht eindeutig deklariert, kann es passieren, dass Sie dafür keine Punkte bekommen.
4. Jede Aufgabe soll in einer getrennten `.Rmd` bearbeitet werden. Wenn Sie nicht 3 separate `.Rmd` Dateien abgeben, müssen Sie mit Punktabzug bis hin zu einer Bewertung mit Null Punkten rechnen.
5. Es liegt in Ihrer Verantwortung, dass Ihre Ergebnisse lokal von den Prüfern repliziert werden können. Fügen Sie daher Ihrer Lösung alle notwendigen Dateien zum Kompilieren hinzu, verwenden Sie keine lokalen Pfade und laden Sie verwendete Pakete (diese Liste ist nicht vollständig.). Falls Ihre Ergebnisse nicht ohne Weiteres repliziert werden können müssen Sie mit Punktabzug rechnen hin bis zu einer Bewertung mit 0 Punkten.
6. Achten Sie darauf, dass alle Funktionen nach der Vorgabe in den Übungen dokumentiert und dass bei allen Funktionen grundlegende Input-Checks durchgeführt werden sollen.
7. Beachten Sie auf die weiteren formale Bewertungskriterien, die zu Beginn der Fragestellung erläutert werden.
8. Sie dürfen neben den beim Starten von R vorhandenen Paketen und Funktionen ausschließlich die Pakete `purrr` und `ggplot2` verwenden.

9. Sollten Sie technische oder andere Schwierigkeiten haben, kontaktieren Sie Bitte **alle Kursleiter gemeinsam in einer Email**. E-mail: andreas.bender@stat.uni-muenchen.de, philipp.kopper@stat.uni-muenchen.de, sven.lorenz@campus.lmu.de. (Bitte die Emails an alle gelisteten Personen schicken!)
10. Die Aufgaben müssen alle eigenständig bearbeitet werden. Vor allem sind keine Arbeitsgruppen erlaubt. Außerdem sind sonstige Diskussion der Aufgaben und Lösungen mit anderen Personen (egal ob diese Statistik studieren oder nicht) nicht zulässig.
11. Das Internet kann passiv genutzt werden. D.h. es dürfen Internetseiten oder Foren aufgerufen und gelesen werden. Das aktive Stellen von Fragen, die relevant zur Lösung der Aufgaben sind, ist allerdings nicht zulässig. Ebenso dürfen keine Aufgaben oder Lösungsvorschläge und anderen Hinweise im Internet gepostet oder per Chat, Email und anderen Kommunikationswegen diskutiert oder verteilt werden.
12. Sollte der Verdacht auf Plagiat, Betrug oder anderweitig unzulässiges Verhalten bestehen, können zusätzliche (mündliche) Prüfungen einberufen werden um die eigenständige Bearbeitung der Aufgaben zu prüfen.
13. Zweifel an der eigenständigen Bearbeitung ihrer Abgabe führen zum nicht-bestehen der Prüfung und der Benachrichtigung des Prüfungsausschusses.
14. Die Abgabe erfolgt bis Mitternacht (23:59 Uhr) am 06.01.2022.

Achten Sie bei der Bearbeitung insgesamt darauf, dass alle top-level Funktionen gut dokumentiert sind und zumindest Basis-Checks für alle Inputs der Funktionen durchzuführen. Achten Sie bei Ihren Outputs darauf, dass diese gut leserlich sind, nicht über den Rand hinausgehen (wenn man die `.Rmd` Datei zu einer PDF kompiliert) und dass Graphiken gute und gut leserliche Beschriftungen und Legenden haben. Sollte dies nicht der Fall sein, kann es zu Punktabzügen kommen.

Aufgabe 1	Bonus Punkte

Sie können bei dieser Hausarbeit folgende Bonuspunkte sammeln:

(a) Abgabe via Github Classroom (BONUS: 2P)

(b) Rmarkdown Datei kompiliert ohne Fehler und Output wohl formatiert (BONUS: 2P)

Aufgabe 2	7 Punkte

Das Gesetz der großen Zahlen besagt, dass wenn ein Experiment mehrfach wiederholt wird, der Mittelwert der Experimentausgänge sich dem Erwartungswert des Experiments annähert. Formal:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{für } n \to \infty,$$

wobei $X_i, i = 1, \ldots, n$ einzelne Ausgänge eines Experiments beschreibt, $\bar{X}_n$ den Mittelwert über n Experimente und μ den Erwartungswert des Experiments.

Sie möchten das Gesetz der Großen Zahlen empirisch anhand eines Münzwurf-Experiments mit eintausend Würfen untersuchen. Allerdings möchten Sie die Münze nicht tatsächlich eintausend Mal werfen und entscheiden sich deshalb das Experiment in `R` durchzuführen. Stellen Sie sicher, dass ihr Zufallsexperiment genau reproduzierbar ist.

(a) Schreiben Sie eine Funktion `toss_coin` die das einmalige werfen einer Münze simuliert. Das Argument `prob` gibt dabei an, mit welcher Wahrscheinlichkeit Kopf geworfen wird:

```r
toss_coin <- function(prob) {
  # TODO
}
```

Als Output soll die Funktion entweder "Kopf" oder "Zahl" zurückgeben.

(b) Unter Verwendung der Funktion aus (a) und Annahme einer fairen Münze, führen Sie das Experiment "Münzwurf" 1000 mal durch und speichern Sie Folgende Informationen in einem Objekt der Klasse `data.frame` (Spaltennamen in Klammern):

- Iterationszähler (`iteration`)
- Ausgang des jeweiligen Experiments (`outcome`)
- Gesamte Zahl der Experimente mit Ausgang "Kopf" in der jeweiligen Iteration (`number_heads`).
- Die relative Häufigkeit der Experimente mit Ausgang "Kopf" in der jeweiligen Iteration (`mean`).

Die ersten vier Zeilen des resultierende Datensatzes könnten z.B. wie folgt aussehen:

```
##   iteration outcome number_heads mean
## 1         1    Kopf            1 1.00
## 2         2    Kopf            2 1.00
## 3         3    Kopf            3 1.00
## 4         4    Zahl            3 0.75
```

(c) Stellen Sie nun die Entwicklung des Mittelwert graphisch dar. Zeichnen Sie hierzu die *Differenz* der relativen Häufigkeit von "Kopf" ($\bar{X}_n$) und dem Erwartungswert des Experiments (μ) als Linie (y-Achse) über die Iterationen (x-Achse). Zeichnen Sie zusätzlich eine gestrichelte horizontale Linie bei 0 ein.

(d) Erstellen Sie eine zweite Graphik identisch zu (c), allerdings soll die Differenz zwischen der absoluten Anzahl der Experimente mit Ausgang "Kopf" zur erwarteten Anzahl Experimente mit Ausgang "Kopf" auf der y-Achse dargestellt werden.

(e) Schreiben Sie nun eine Funktion, die das Experiment `n` mal durchführt und die Graphik aus (c) erzeugt. Die Signatur ist unten gegeben.

```r
visualize_coin_toss_experiment <- function(prob, n = 1000) {
  # TODO
}
```

(f) Wenden Sie Ihre Funktion aus (e) an um ein Münzwurf-Experiment mit einer gezinkten Münze durchzuführen, bei der die Wahrscheinlichkeit für Kopf 0.6 beträgt.

Aufgabe 3	5 Punkte

Betrachten Sie wieder den Datensatz der CO2 Emissionen für verschiedene Länder (`co2data.Rds`). Lesen Sie die Daten in R ein. Verwenden Sie im Folgenden die Funktion `split` und die `map*`-Familie von Funktionen aus dem `purrr` package.

(a) Berechnen Sie die durchschnittlichen CO2 Emissionen pro Jahr. Der Output soll ein benamter Vektor sein, wobei die Namen das Jahr und die Werte die durchschnittlichen CO2 Emissionen angeben.

(b) Wie (a), allerdings pro Kontinent (`continent`) statt Jahr.

(c) Wie (b), es sollen aber die durchschnittlichen CO2 emissionen pro Einwohner berechnet werden.

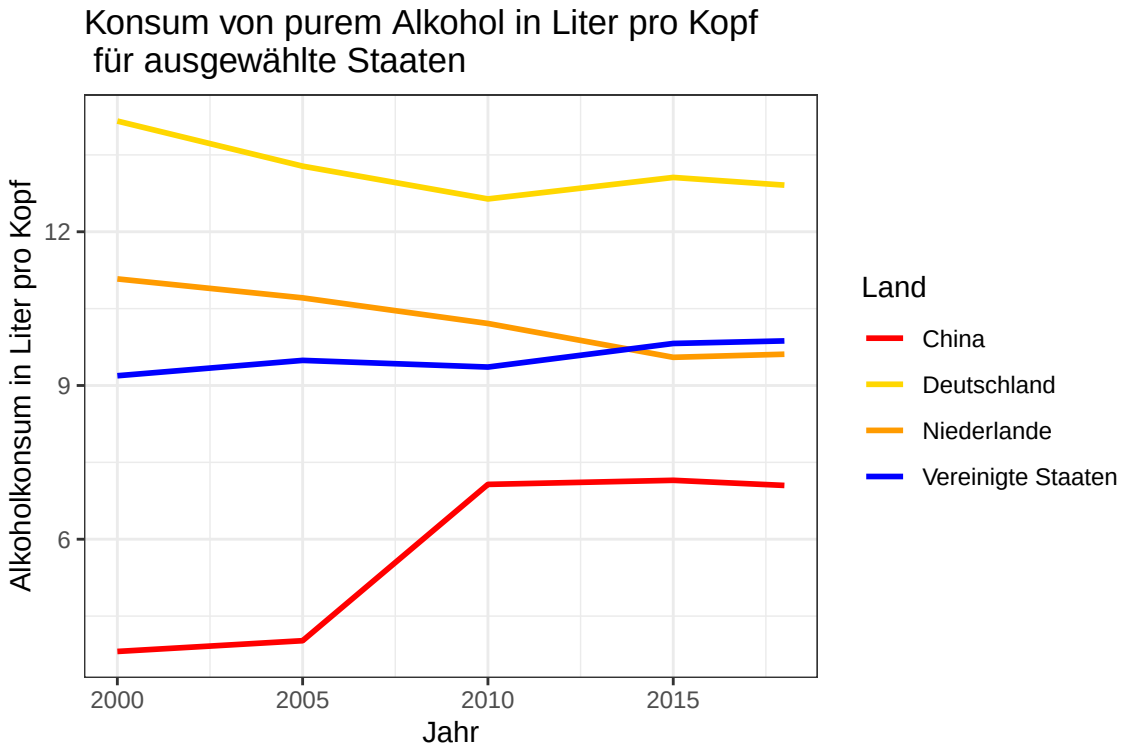
(d) Berechnen Sie pro Region (`region`) die durchschnittlichen CO2 Emissionen pro Einwohner und das durchschnittliche GDP pro Einwohner. Das Ergebnis soll *ein* `data.frame` sein, mit Spalten `region`, `gdp_per_capita` und `co2_per_capita` und einer Zeile pro Region.

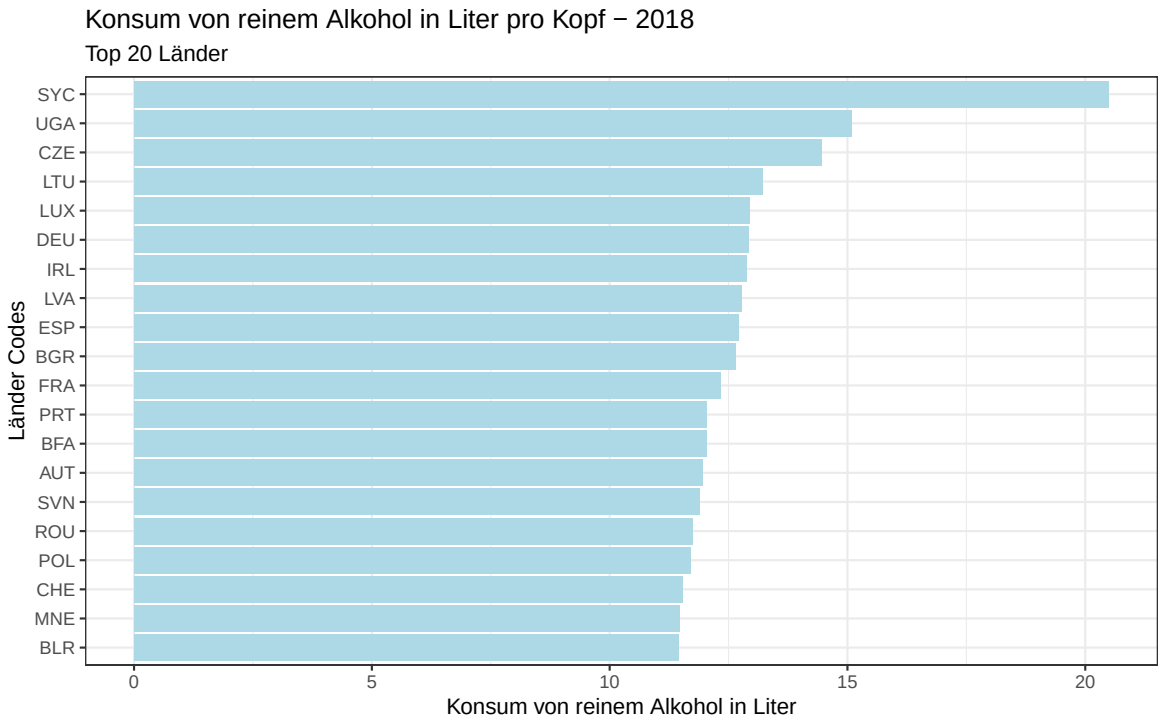
Aufgabe 4	8 Punkte

In dieser Aufgabe sollen Sie mit ihren bisherigen `ggplot2` Kenntnissen Graphiken zum Alkoholkonsum in verschiedenen Ländern bauen. Laden Sie hierfür den Datensatz `alcohol-consumption.Rds` von der Moodle Seite herunter.

Verwenden Sie für die Visualisierung das Paket `ggplot2`. Teilweise ist Pre-processing der Daten notwendig. Danach sollen die Graphiken mit einem zusammenhängenden `ggplot2` Befehl erstellt werden. Achten Sie bei allen Teilaufgaben darauf ordentliche und sauber beschriftete Plots zu erstellen.

Bauen Sie die zwei Nachfolgenden Graphiken nach. Die dabei verwendeten Farben sind für Graphik 1 `"red"`,`"gold"`,`"#FF9B00"` und `"blue"` und für Graphik 2 `"lightblue"`.




Konsum von reinem Alkohol in Liter pro Kopf − 2018
Top 20 Länder
Länder Codes
SYC
UGA
CZE
LTU
LUX
DEU
IRL
LVA
ESP
BGR
FRA
PRT
BFA
AUT
SVN
ROU
POL
CHE
MNE
BLR
0
5
10
15
20
Konsum von reinem Alkohol in Liter