Übungsfragen

Aufgabe 1

Beschreiben Sie was man unter einem **Merkmal** versteht.

Aufgabe 2

Skizzieren Sie den Unterschied zwischen **deskriptiver** und **induktiver** Statistik.

Aufgabe 3

Beschreiben Sie den Unterschied zwischen einer **diskreten** und einer **stetigen** Variable anhand von zwei Beispielen.

Aufgabe 4

Nennen Sie die **zwei R-Befehle** mit denen sich Zusatzpakete installieren und laden lassen.

Aufgabe 5

Skizzieren Sie den Unterschied zwischen den beiden Datentypen unten:

```
participant.no <- c(1, 2, 3, 4, 5)
class(participant.no)
```

```
## [1] "numeric"
```

```
country <- "Russia"
class(country)
```

```
## [1] "character"
```

Aufgabe 6

In der Grafik unten (Figure 1) sind die zwei Variablen y = **Demokratiequalität** (gemessen von 0 bis 100, wobei 100 die theoretisch bestmöglichste Demokratie beschreibt) und x = **Bruttoinlandsprodukt** (gemessen in Tausend US-Dollar, d.h. 1 entspricht 1000$) abgetragen. Zu der Grafik hat eine Forscherin diese Gleichung berechnet:

$$y = 46 + 0.86 * x$$

a. Wie *nennt* man diese Grafik?
b. *Diskutieren Sie:* Was bedeutet die berechnete Gleichung im Zusammenhang mit der Grafik, welche Teile der Grafik beschreibt sie?
c. *Beschreiben* Sie den Zusammenhang zwischen den beiden Variablen. Wie können Sie diesen interpretieren? Wo liegen die Grenzen dieser Interpretation?

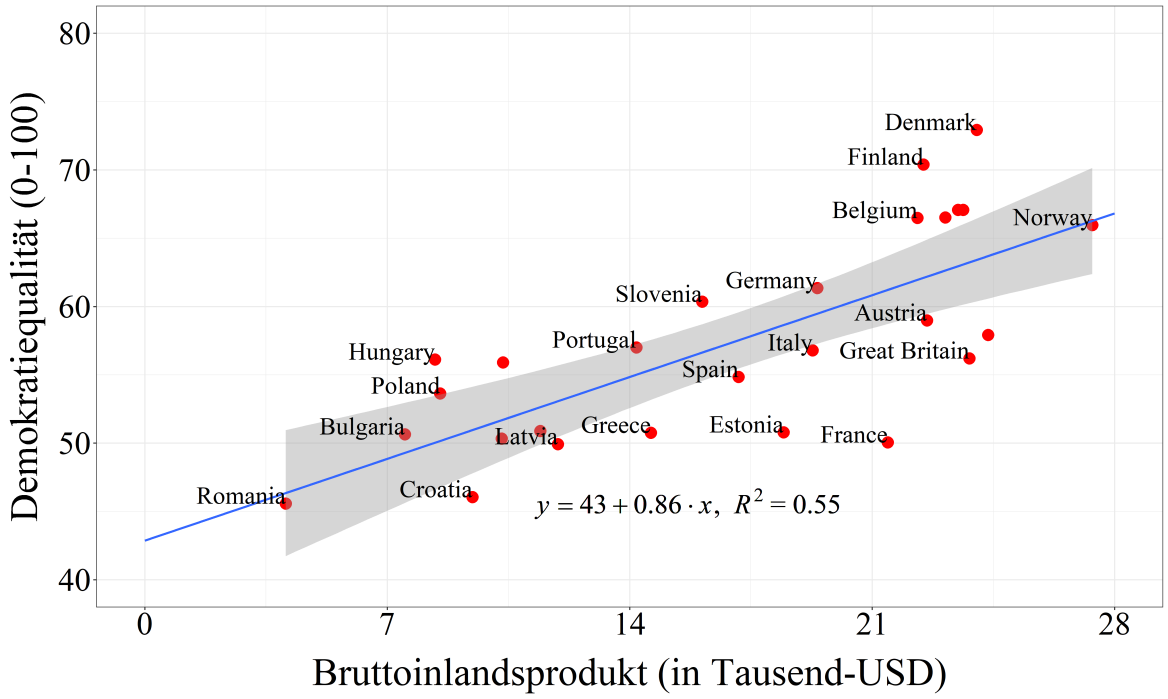


Figure 1:

Aufgabe 7

Beschreiben Sie kurz was man unter diesen Begriffen versteht:

- **einem unstandardisierten Regressionskoeffizient**?
- **Residuen**?
- **metrischen** Merkmalen?

Aufgabe 8

Nennen Sie das Skalenniveau der folgenden Merkmale (Ausprägungen in Klammern) an und begründen sie kurz, warum Sie sich für dieses Skalenniveau entschieden haben:

a. **Temperatur** (in Grad Celsius)
b. **Berufsgruppe** (1 = Arbeiter, 2 = Angestellter, 3 = Selbstständig, 4 = Andere)
c. **Jährliche Ausgaben für Berufskleidung** (in Euro)
d. **Konsum von Tageszeitungen** (0 = 'nie', 1 = 'monatlich', 2 = 'wöchentlich', 3 = 'mehrmals wöchentlich', 4 = 'täglich')

Aufgabe 9

Beschreiben Sie den Unterschied zwischen **Maßen der zentralen Tendenz** und **Maßen der Variabilität**.

Aufgabe 10

Bei der Erhebung der Intelligenz von 20 Studierenden fallen folgende IQ-Werte an:

`109 92 93 94 96 96 97 98 100 101 101 102 103 103 103 104 105 105 107 91`

Berechnen Sie den **Median** und den **Modus**.

Aufgabe 11

5 Personen bearbeiten einen psychologischen Test. Es treten folgende Testergebnisse auf:

$x_1 = 80, x_2 = 70, x_3 = 60, x_4 = 50, x_5 = 40$

a. *Berechnen* Sie die Stichprobenvarianz s^2. Notieren sie ihren Rechenweg.
b. *Notieren* Sie den genauen `R-Code` um aus diesen Werten die Stichprobenvarianz zu berechnen.

Aufgabe 12

Eine Schülerin hat sich 5 verschiedene USB-Sticks zu folgenden Preisen gekauft:

`10 Euro, 10 Euro, 13 Euro, 16 Euro, 16 Euro`

a. *Berechnen Sie*: Wie viel kostet die Schülerin ein USB-Stick im Durchschnitt?
b. *Berechnen Sie* die Standardabweichung s der Preise. Notieren sie ihren Rechenweg.

Aufgabe 13

In einem Fragebogen wurde den Befragten folgende Aussage vorgelesen:

"Ich möchte lieber ein Bürger/eine Bürgerin Deutschlands als irgendeines anderen Landes auf der Welt sein."

Die Antworten der Befragten finden Sie in der Tabelle unten. Die Antworten wurden im Datensatz `Allbus` zusammengefasst, die Antworten auf die Frage oben in der Variable `natio2` abgespeichert.

Codierung	Ausprägung	n_i
1	Stimme voll und ganz zu	60
2	Stimme zu	20
3	Weder noch	10
4	Stimme nicht zu	70
5	Stimme überhaupt nicht zu	40

a. Notieren Sie den `R-Code` mit dem man sich die absoluten Häufigkeiten der Variable `natio2` (ähnlich der Tabelle unten) anzeigen lassen kann.

b. In der Tabelle unten sind die absoluten Häufigkeiten der Antworten abgetragen. Berechnen Sie die **kumulierten Häufigkeiten** $F_{1,..,i}$ und bestimmen Sie den **Median**.

Aufgabe 14

Die Wahlbehörde von Minneapolis veröffentlichte einen Datensatz (im Excel-Format) mit Wahlergebnissen auf Stimmbezirksebene. Unten finden Sie ein Bild dieses Datensatzes. *Diskutieren Sie*, was an diesem Datensatz nicht dem Konzept von **tidy data** entspricht und skizzieren Sie kurz im Datensatz, wo Sie Verbesserungen vornehmen würden.

City of Minneapolis Statistics
General Election November 5, 2013

Ward	Precinct	Registered Voters at 7am	Voters Registering at Polls	Voters Registering by Absentee	Total Registrations	Voters at Polls	Absentee Voters	Total Ballots Cast	Total Turnout	Percentage Absentee	% Registered to Total (Election Day)	Spoiled Ballots
City-Wide Total		**233,351**	**5,926**	**708**	**6,634**	**75,145**	**4,954**	**80,099**	**33.38%**	**6.18%**	**7.89%**	**3,358**
1	1	1,878	25	3	28	492	27	519	27.23%	5.20%	5.08%	14
1	2	2,769	43	1	44	836	56	892	31.71%	6.28%	5.14%	22
1	3	2,337	40	0	40	905	19	924	38.87%	2.06%	4.42%	34
1	4	2,139	24	5	29	768	26	794	36.62%	3.27%	3.13%	19
1	5	1,875	31	0	31	683	31	714	37.46%	4.34%	4.54%	14
1	6	2,258	69	0	69	739	20	759	32.62%	2.64%	9.34%	32
1	7	1,847	47	0	47	291	8	299	15.79%	2.68%	16.15%	17
1	8	1,332	43	0	43	415	5	420	30.55%	1.19%	10.36%	22
1	9	2,401	42	0	42	596	25	621	25.42%	4.03%	7.05%	15
Ward 1 Subtotal		**18,836**	**364**	**9**	**373**	**5,725**	**217**	**5,942**	**30.93%**	**3.65%**	**6.36%**	**189**
2	1	2,820	62	1	63	1,011	39	1,050	36.42%	3.71%	6.13%	42
2	2	1,377	39	5	44	679	37	716	50.39%	5.17%	5.74%	28
2	3	1,763	44	4	48	324	18	342	18.88%	5.26%	13.58%	19
2	4	1,582	53	0	53	117	3	120	7.34%	2.50%	45.30%	3
2	5	1,994	48	2	50	495	26	521	25.49%	4.99%	9.70%	26
2	6	1,120	35	1	36	433	19	452	39.10%	4.20%	8.08%	22
2	7	1,013	39	0	39	138	7	145	13.78%	4.83%	28.26%	4
2	8	2,543	49	1	50	1,206	36	1,242	47.90%	2.90%	4.06%	30
2	9	1,162	37	2	39	351	16	367	30.56%	4.36%	10.54%	15
2	10	2,822	87	0	87	196	5	201	6.91%	2.49%	44.39%	7
Ward 2 Subtotal		**18,196**	**493**	**16**	**509**	**4,950**	**206**	**5,156**	**27.56%**	**4.00%**	**9.96%**	**196**

Figure 2:

Aufgabe 15

Skizzieren Was macht man mit den folgenden Zeichen in **R** deutlich?

a. `#`
b. `<-`
c. `%>%`

Aufgabe 16

Im Datensatz `BabyNames` sind die Namen (`name`) von Neugeborenen in den USA gelistet sowie die Häufigkeit, wie oft die Namen in einem Jahr vergeben wurden (`count`). Der Datensatz erfasst das Geschlecht der Neugeborenen (`sex`) und deckt einen Zeitraum von 1880 bis 2015 ab (`year`).

Sechs zufällige Zeilen des Datensatzes:

```
##              name sex count year
## 838547  Nathanial   M   179 1978
## 13548       Missie   F     9 1886
## 1355827      Isiac   M    12 2000
## 380914       Angel   M   108 1941
## 70324        Naoma   F    19 1905
## 551151       Karel   F    31 1958
```

a. Erklären Sie zeilenweise die Bedeutung dieses R-Code Abschnitts:

```
1  library(tidyverse)
2  R-Code:
3
4  MySum <-
5
6    BabyNames %>%
7
```

```
  filter(year == "2005") %>%

  group_by(sex) %>%

  summarise( total = sum(count) )
```

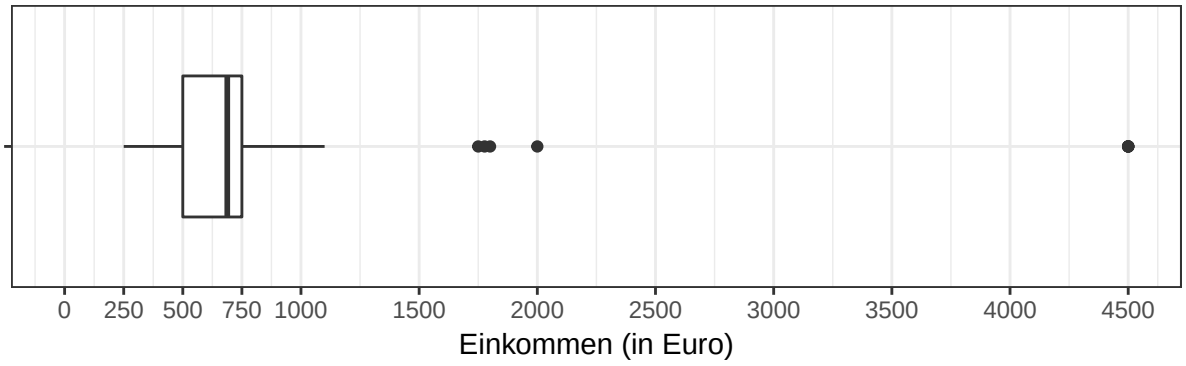
b. Unten sind vier verschiedene R-Code Abschnitte notiert. Drei Abschnitte machen das gleiche, einer berechnet etwas anderes. Welcher ist das und warum?

```
#1
BabyNames %>%
  group_by(year, sex) %>%
  summarise( totalBirths = sum(count) )

#2
group_by(BabyNames, year, sex) %>%
  summarise( totalBirths = mean(count) )

#3
group_by(BabyNames, year, sex) %>%
  summarise( totalBirths = sum(count) )

#4
Tmp <- group_by(BabyNames, year, sex)
summarise( Tmp, totalBirths = sum(count) )
```

Aufgabe 17

Ein Student hat ein Problem. Er möchte gerne etwas über die **finanzielle Situation von Studierenden** am Institut für Astronomie sagen. Dazu hat er 200 Studierende befragt wie viel Geld ihnen monatlich zur Verfügung steht.

Um das durchschnittlich verfügbare Geld anzugeben, hat er das **arithmetische Mittel** berechnet, das bei **847 Euro** liegt. Das erscheint dem Student viel zu hoch. Seine Freundin gibt ihm den Tipp, doch den **Median** zu verwenden, denn der sei *"besser"*. Tatsächlich liegt der Median nur bei **688 Euro**. Der Student ist verwirrt. Welches Maß soll er denn nun nehmen? Außerdem hat er von seinem Statistikprogramm noch diese Grafik angezeigt bekommen:



a. Wie nennt man die Grafik oben? **Beschriften Sie diese allgemein** und zeichnen sie das **arithme-**

tische Mittel ein.

b. Geben Sie die Grenzen des **oberen Quartils** der Einkommensverteilung an.

c. *Diskutieren* Sie ob der Student das arithmetische Mittel oder den Median verwenden soll. *Beschreiben* Sie dazu den Unterschied zwischen den beiden Maßen.

Aufgabe 18

Ein Wissenschaftler geht der Frage nach, ob die **Zufriedenheit von Studierenden** mit den Studienbedingungen davon abhängt, ob die Studierenden eine Studienberatung besucht haben. Er möchte gerne zeigen, dass die Teilnahme an der Beratung die Zufriedenheit beeinflusst. Dazu hat er Erstsemester-Studierende befragt und deren Daten analysiert.

- Zufriedenheit mit den Studienbedingungen hat er auf einer 11-stufigen Skala von 1 = "sehr unzufrieden" bis 11 = "sehr zufrieden" abgefragt.
- Teilnahme an der Studienberatung hat die Ausprägungen "ja" und "nein".

Verbalisieren und *formulieren sie mathematisch* eine Null- und die dazu passende Alternativhypothese.

Aufgabe 19

Zur Europawahl 2014 wurden im Rahmen der *European Election Study (EES)* 950 Personen in Frankreich zu ihren politischen Einstellungen befragt.

Eine Forscherin interessiert sich für die Gründe, mit denen sich eine mögliche Wahl der Partei **Front National** erklären lässt. Sie nimmt an, dass die Wahrscheinlichkeit den Front National zu wählen mit inhaltlichen Motiven, aber auch mit der eigenen ökonomischen Lage zusammenhängt.

Im Fragebogen der Studie hat sie folgende Fragen gefunden:

1. *"Wenn Sie an den FN denken, welcher Wert von 0 bis 10 beschreibt am besten, wie wahrscheinlich es ist, dass Sie jemals diese Partei wählen werden?"*, kodiert als `wahlw.fn`, mit den Ausprägungen $0 - 10$, wobei "0" bedeutet, dass dies überhaupt nicht wahrscheinlich und "10", dass dies sehr wahrscheinlich ist.

2. *"Und wie wird sich Ihrer Meinung nach die allgemeine Wirtschaftslage in Deutschland in den nächsten 12 Monaten entwickeln? Wird sie . . . ?"*, kodiert als `oekonomische.lage`, mit den Ausprägungen

- "Besser" (0),
- "Unverändert bleiben" (1),
- "Besser werden" (3)

3. *"Ich möchte Sie bitten sich zum Thema Einwanderung zu positionieren und zwar auf einer Skala von 0 bis 10"* wobei (0) bedeutet *Sie sind absolut für eine restriktive Einwanderungspolitik* und (10) *Sie sind absolut gegen eine restriktive Einwanderungspolitik.*, kodiert als `einwanderung`.

4. *Sind Sie in den letzten 12 Monaten arbeitslos geworden/gewesen?*, kodiert als `job.verloren`, mit den Ausprägungen nein = 1 und ja = 0.

5. *Geschlecht?*, kodiert als `geschlecht`, mit den Ausprägungen weiblich = 1 und männlich = 0.

6. Stellung in der Gesellschaft: *"Auf der folgenden Skala entspricht die Stufe 1 der niedrigsten, die Stufe 10 der höchsten Stellung in der Gesellschaft. Können Sie mir sagen, wo Sie sich selbst einordnen würden?"*, kodiert als `stellung.gesellschaft`.

Im Anschluss hat die Forscherin ein Regressionsmodell gerechnet, dessen R-Code und Output unten abgebildet sind.

a. Analysieren Sie den Output: Was ist die **abhängige Variable** und welche sind die **unabhängigen Variablen** in ihrem Regressions-Modell?

b. Interpretieren Sie das **Regressionsmodell**. Beurteilen Sie dazu die **Effekte** bezüglich ihrer inhaltlichen Bedeutung, ihrer Richtung, und Signifikanz. Nehmen Sie ein Signifikanzniveau von $\alpha = 0,05\%$ an.
c. Was ändert sich, wenn Sie ein Signifikanzniveau von $\alpha = 0,10\%$ annehmen? *Erklären Sie*, was das inhaltlich bedeutet!

R-Code:

```
library(tidyverse)
load("./EESFrance.RData")

lm(wahlw.fn ~ oekonomische.lage + einwanderung + job.verloren +
  geschlecht + stellung.gesellschaft, data=EESFrance, weight=wexpol) %>% summary()
```

Output:

```
##
## Call:
## lm(formula = wahlw.fn ~ oekonomische.lage + einwanderung + job.verloren +
##     geschlecht + stellung.gesellschaft, data = EESFrance, weights = wexpol)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1839.4  -469.8  -163.2   109.7  3502.1
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.07076    0.61326   9.899  < 2e-16 ***
## oekonomische.lageGleich       0.34161    0.35767   0.955 0.339818
## oekonomische.lageSchlechter   1.40982    0.36630   3.849 0.000128 ***
## einwanderung                 -0.33439    0.04210  -7.944 6.86e-15 ***
## job.verlorenNo               -0.37063    0.29684  -1.249 0.212185
## geschlechtFemale             -0.89790    0.25533  -3.517 0.000462 ***
## stellung.gesellschaft        -0.15299    0.08685  -1.762 0.078543 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 771.3 on 780 degrees of freedom
##   (91 observations deleted due to missingness)
## Multiple R-squared:  0.1492, Adjusted R-squared:  0.1426
## F-statistic:  22.8 on 6 and 780 DF,  p-value: < 2.2e-16
```

Aufgabe 20

Erklären Sie abstrakt mit eigenen Worten, was man im Allgemeinen mit R^2 beschreibt. Welches verhältnis beschreibt der Ausdruck? Welche nützliche Information können Sie diesem Wert für ihr Regressionsmodell entnehmen?

Aufgabe 21

Wie *nennt* man die abgebildete Grafik in Figure 3? *Beschreiben* Sie die Verteilung.

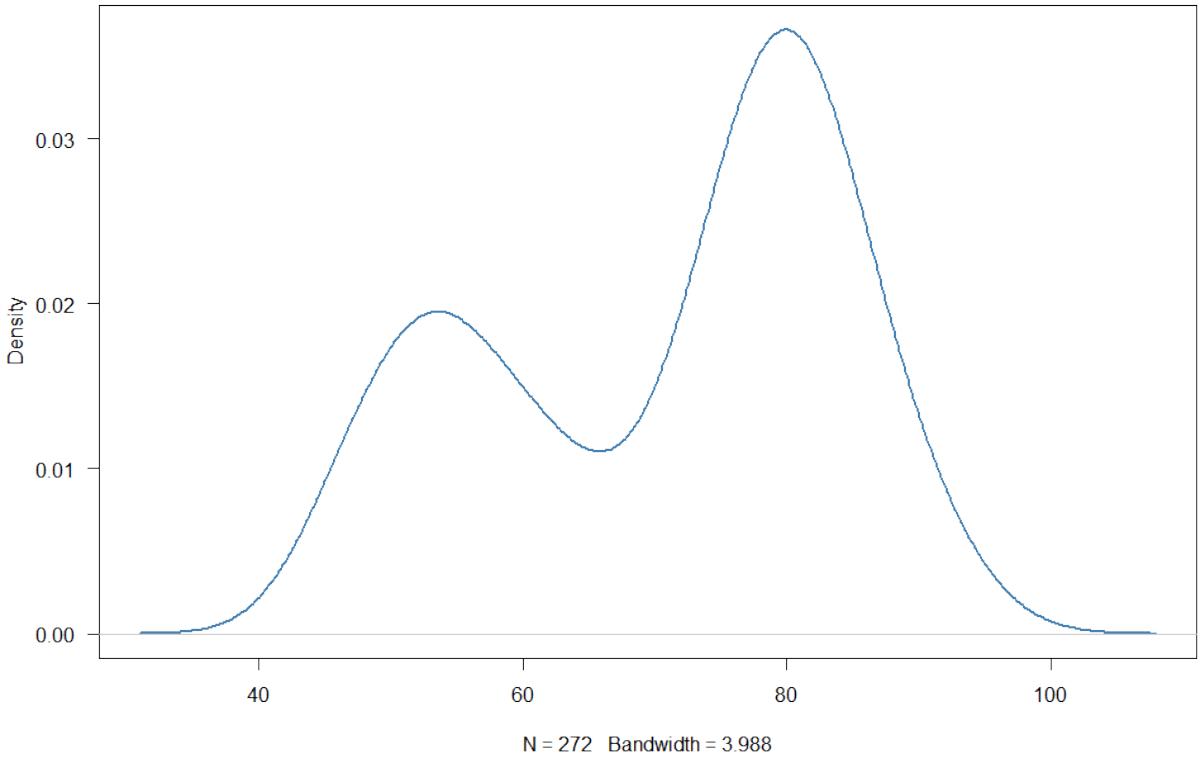


Figure 3: