Nachholklausur Statistik
Lösungshinweise

Prüfungsdatum: 28. Januar 2020 – Prüfer: Etschberger, Henle, Wesp
Studiengang: BW, IM
Punkte: 19, 20, 18, 10, 11, 12 ; Summe der Punkte: 90

Aufgabe 1 — 19 Punkte

Ein Datensatz zu Bevölkerungsmerkmalen einiger Ländern liegt in der R-Variable `Laendermerkmale` vor. Die Ausgabe dieser Daten ergibt:

Land	Kontinent	Zufriedenheit	Lebenserwartung	Einkommen	Bevoelkerung
Luxembourg	Europe	253	82	47 716	562 958
Estonia	Europe	170	77	12 421	1 313 271
Slovenia	Europe	220	80	24 750	2 068 959
Ireland	Europe	253	81	48 073	4 635 400
Norway	Europe	247	82	31 208	5 205 434
Finland	Europe	257	81	24 931	5 494 379
Denmark	Europe	273	80	34 797	5 699 220
Switzerland	Europe	273	83	48 414	8 279 700
Israel	Asia	223	82	24 084	8 434 600
Austria	Europe	260	81	28 051	8 623 073
Sweden	Europe	257	82	25 196	9 838 480
Hungary	Europe	190	75	13 419	9 849 000
Portugal	Europe	203	81	16 664	10 374 822
Czech Republic	Europe	213	78	16 626	10 541 466
Greece	Europe	210	81	15 061	10 846 979
Belgium	Europe	243	80	28 015	11 280 134
Netherlands	Europe	250	81	32 120	16 976 600
Australia	Australia	243	83	42 617	23 991 100
Canada	Americas	253	82	37 469	35 851 774
Poland	Europe	197	77	15 768	38 484 000
Spain	Europe	233	83	22 866	46 439 864
Italy	Europe	230	83	21 096	60 725 000
United Kingdom	Europe	237	81	30 064	64 800 000
France	Europe	220	82	22 718	67 128 000
Germany	Europe	240	81	28 636	81 292 400
Japan	Asia	207	84	26 122	126 890 000
United States	Americas	247	79	45 582	322 425 000

a) Mit dem folgenden R Kommando wird diese Datentabelle modifiziert:

```
L.Auswahl = Laendermerkmale %>%
  mutate(Bevoelkerung = round(Bevoelkerung / 1000000, 1)) %>%
  filter(Kontinent == "Europe" & Bevoelkerung > 15) %>%
  select(Land, Lebenserwartung, Bevoelkerung) %>%
  arrange(Lebenserwartung)
```

Schreiben Sie die resultierende Tabelle vollständig auf.

Hinweis: Die folgenden Teilaufgaben beziehen sich auf das Ergebnis aus Teilaufgabe a). Falls Sie a) nicht lösen können, rechnen Sie bitte mit dem nebenstehenden (falschen) Ergebnis weiter.

Land	*Zufriedenheit*	*Lebenserwartung*	*Bevoelkerung*
Australia	*243*	*83*	24 000
Canada	*253*	*82*	35 900
Czech Republic	*213*	*78*	10 500
Denmark	*273*	*80*	5700
Estonia	*170*	*77*	1300
Germany	*240*	*81*	81 300
Ireland	*253*	*81*	4600
Luxembourg	*253*	*82*	600
Sweden	*257*	*82*	9800

b) Geben Sie eine Tabelle mit den Ausprägungen a_i und den relativen Häufigkeiten f_i der `Lebenserwartung` an.

c) Bestimmen Sie zur empirischen Verteilungsfunktion F des Merkmals `Lebenserwartung` den Wert $F(81)$.

d) Bestimmen Sie zum Merkmal `Lebenserwartung` das empirische $^3/_7$-Quantil sowie das 80 %-Quantil.

(Hinweis: In e), f) wird das Merkmal Bevoelkerung betrachtet.)

e) Berechnen Sie den Median, das arithmetische Mittel sowie die Spannweite des Merkmals `Bevoelkerung`.

f) Geben Sie R-Sequenzen an, mit dem Sie aus dem tibble `L.Auswahl` aus Teilaufgabe a) die drei Ergebnisse aus Teilaufgabe e) berechnen können.

Lösungshinweis:

a)

Land	Lebenserwartung	Bevoelkerung
Poland	77	38.5
Netherlands	81	17.0
United Kingdom	81	64.8
Germany	81	81.3
France	82	67.1
Spain	83	46.4
Italy	83	60.7

b)
```
## # A tibble: 4 x 3
##      ai    hi    fi
##   <int> <int> <dbl>
## 1    77     1 0.143
## 2    81     3 0.429
## 3    82     1 0.143
## 4    83     2 0.286
```

c) $F(81) = 0.5714286$

d)
```
x.L = L.Auswahl %>% select(Lebenserwartung) %>% as_vector()
quantile(x.L, probs=c(3/7, 0.8), type=2)
```

```
## 42.85714%       80%
##        81        83
```

e) und f)

```
L.Auswahl %>%
  summarize(Median=median(Bevoelkerung),
            arithm.Mittel=mean(Bevoelkerung),
            SP=max(Bevoelkerung) - min(Bevoelkerung) )
```

```
## # A tibble: 1 x 3
##   Median arithm.Mittel    SP
##    <dbl>         <dbl> <dbl>
## 1   60.7          53.7  64.3
```

```
# alternativ:
x = L.Auswahl %>% select(Bevoelkerung) %>% as_vector()

# Median
median(x)
```

```
## [1] 60.7
```

```
# arithm. Mittel:
mean(x)
```

```
## [1] 53.68571
```

```
# Spannweite:
max(x)-min(x)
```

```
## [1] 64.3
```

Aufgabe 2 — 20 Punkte

In der Statistikvorlesung wurden zufällig 6 Studenten befragt nach

- der Anzahl der erreichten Punkte in der Matheklausur (Merkmal `Punkte`),
- der Anzahl der Stunden, die Sie in den 4 Wochen vor der Klausur für die Vorbereitung investiert haben (Merkmal `Lernzeit`) sowie
- einer persönlichen Einschätzung der Qualität der Mathevorlesung (Merkmal `Qualitaet`) auf einer Skala von 1 bis 5
  (5 ≡ phänomenal, 1 ≡ unterirdisch).

Student	Punkte	Lernzeit	Qualitaet
1	67	70	1
2	82	50	4
3	41	30	2
4	63	55	5
5	12	15	4
6	59	45	4

Es ergibt sich nebenstehende Tabelle.

Gehen Sie im Folgenden davon aus, dass die Merkmale Punkte, Lernzeit jeweils metrisch (kardinal) und das Merkmal Qualitaet ordinal skaliert sind.

a) Berechnen Sie einen geeigneten Korrelationskoeffizienten zwischen den beiden Merkmalen `Punkte` und `Qualitaet`. Begründen Sie, warum Sie sich für diesen Koeffizienten entschieden haben und interpretieren Sie den Zahlenwert des Ergebnisses.

b) Berechnen Sie ein geeignetes Maß für die Korrelation der `Punkte` und der `Lernzeit`. Begründen Sie auch hier Ihre Wahl und interpretieren Sie den Wert.

R c) Geben Sie R-Befehle an, um die Teilaufgabe a) und b) zu lösen.

d) Stellen Sie ein lineares Modell auf, mit dem die Anzahl der in der Klausur erreichten Punkte in Abhängigkeit von der Lernzeit prognostiziert werden kann.

e) Mit wieviel Punkten rechnen Sie gemäß diesem Modell bei einer Lernzeit von 40 Stunden?

f) Wieviele Punkte erwarten Sie gemäß Modell bei 250 Stunden Lernzeit? Warum ist dieses Ergebnis unrealistisch?

g) Wieviele Stunden müsste man gemäß diesem Modell in die Vorbereitung investieren, um die Klausur gerade so (mit 45 Punkten) zu bestehen?

Lösungshinweis:

```
Daten = tribble(~Student, ~Punkte, ~Lernzeit, ~Qualitaet,
                1, 67, 70, 1,
                2, 82, 50, 4,
                3, 41, 30, 2,
                4, 63, 55, 5,
                5, 12, 15, 4,
                6, 59, 45, 4)

# a) c)
cor(Daten$Punkte, Daten$Lernzeit)

## [1] 0.8495672
```

```r
# Zusammenhang stark: Mehr lernen -> Mehr Punkte

# b) c)
cor(Daten$Punkte, Daten$Qualitaet, method="spearman")
```

```
## [1] -0.03035884
```

```r
# Fast kein Zus.:
# Mehr Punkte -> keine Auswirkung auf Qulitätsbeurteilung der VL

# d)
D.lm = lm(Punkte~Lernzeit, data=Daten)$coefficients
a = D.lm[1]
b = D.lm[2]
D.lm
```

```
## (Intercept)    Lernzeit
##    6.547884    1.074388
```

```r
# Punkte = a + b * Lernzeit

# e)
a + b * 40
```

```
## (Intercept)
##    49.52339
```

```r
# f)
a + b * 250
```

```
## (Intercept)
##    275.1448
```

```r
# Macht keinen Sinn, da Modell bei Extrapolation nicht
# linear, vor allem außerhalb möglicher Werte

# g)
(45 - a)/b
```

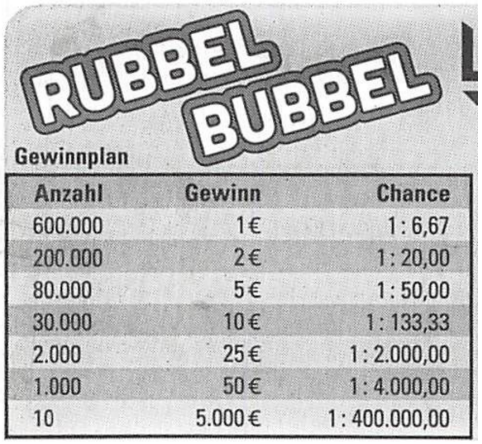
```
## (Intercept)
##     35.7898
```

Aufgabe 3 — 18 Punkte

Sebastian Schanze hat sich zum Geburtstag ein Rubbellos gekauft, für das er 1 € bezahlt hat. Nachdem er die Gewinnfelder freigerubbelt hat stellt er fest, dass er leider nicht gewonnen hat. Auf der Rückseite des Loses findet Sebastian neben der Angabe, dass insgesamt 4 Millionen Lose in dieser Serie verkauft werden auch einen Gewinnplan (siehe Abbildung 1).

Sebastian möchte sich einen Eindruck von der Fairness des Spiels verschaffen und bittet Sie um Ihre Hilfe.

Gehen Sie im Folgenden davon aus, dass jedes Los zufällig aus einer Grundgesamtheit von allen 4 Millionen Losen gezogen wird (also mit Zurücklegen).



Anzahl	Gewinn	Chance
600.000	1 €	1 : 6,67
200.000	2 €	1 : 20,00
80.000	5 €	1 : 50,00
30.000	10 €	1 : 133,33
2.000	25 €	1 : 2.000,00
1.000	50 €	1 : 4.000,00
10	5.000 €	1 : 400.000,00

Abbildung 1: Gewinnplan des Loses

a) Wie hoch ist die Wahrscheinlichkeit bei einem zufällig gezogenen Los irgendeinen Gewinn zu erzielen?

(Hinweis: Falls Sie die Teilaufgabe a) nicht lösen können, rechnen Sie bitte mit dem (falschen) Wert $p = 0.35$ *weiter.)*

b) Wie hoch ist die Wahrscheinlichkeit bei 10 zufällig gezogenen Losen mindestens einmal zu gewinnen?

c) Wieviele Lose muss Sebastian mindestens kaufen, um mit mindestens einer Wahrscheinlichkeit von 99 % mindestens einmal zu gewinnen?

d) Wie hoch sind Erwartungswert und Standardabweichung des Gewinns der Betreiberfirma bei einem zufällig verkauften Los?

e) Betrachten Sie den durchschnittlichen Gewinns der Betreiberfirma pro Los bei 10 000 verkauften Losen mit jeweils einem Gewinn G_i, die zufällig und unabhängig voneinander gezogen wurden, also

$$\overline{G} = \frac{1}{10\,000} \sum_{i=1}^{10\,000} G_i \,.$$

Wie hoch sind der Erwartungswert und die Standardabweichung von $\overline{G}$?

f) Geben Sie den 3-σ-Bereich von $\overline{G}$ an und interpretieren Sie diesen Wert.

Lösungshinweis:

a) Gesamtanzahl der Gewinne $n = 600000 + 200000 + 80000 + 30000 + 2000 + 1000 + 10 = 913\,010$.
Gewinnwahrscheinlichkeit $p = \frac{913\,010}{4\,000\,000} = 0.228\,252\,5$.

b) X : Anzahl Gewinne bei 10 gezogenen Losen. $X \sim B(10; p)$.
Gesucht: $P(X \geqq 0) = 1 - P(X = 0) = 1 - \binom{10}{0} p^0 (1-p)^{10} = 0.925\,053\,3$.

c) Jetzt: X: Anzahl Gewinne bei n gezogenen Losen. Also:

$$\begin{aligned} & P(X \geqq 0) = 1 - (1-p)^n \geqq 0.99 \\ \Leftrightarrow\ & (1-p)^n \leqq 0.01 \\ \Leftrightarrow\ & n \geqq \log_{1-p} 0.01 \approx 17.77 \end{aligned}$$

also müssen mindestens 18 Lose gezogen werden.

d) Gewinn G der Firma pro Los:
$\mathrm{E}[G] = 1 - \left(1 \cdot \frac{600\,000}{4\,000\,000} + 2 \cdot \frac{200\,000}{4\,000\,000} + \ldots\right) \approx 0.5375.$
$\mathrm{Sta}[G] = \sqrt{\left(1^2 \cdot \frac{600\,000}{4\,000\,000} + 2^2 \cdot \frac{200\,000}{4\,000\,000} + \ldots\right) - \mathrm{E}^2[G]} \approx 8.0513111.$

e) $\mathrm{E}[\overline{G}] = \mathrm{E}[G]$. $\sigma = \mathrm{Sta}[\overline{G}] = \mathrm{Sta}[G]/\sqrt{10\,000} \approx 0.0805131$

f) $[\mathrm{E}[\overline{G}] \pm 3\sigma] = [0.296, 0.779]$. Der Gewinn pro Los liegt bei 10 000 verkauften Losen mit sehr hoher Wahrscheinlichkeit (ca. 99.8 %) in diesem Intervall.

Aufgabe 4 — 10 Punkte

Noch 20 Minuten bis zur Statistikklausur. Norbert steht vor dem Prüfungsraum und ist furchtbar aufgeregt. Er spürt bei sich einen erhöhten Pulsschlag. Um herauszufinden, ob er durch seine Nervosität einen Nachteil hat, möchte er wissen, wie der Puls seiner Mitprüflinge sich verhält. Er fragt 10 Kommilitonen nach ihrer Pulsfrequenz und erhält folgende Antworten in Pulsschlägen pro Minute:

63	108	55	81	114	68	71	67	74	83

(Gehen Sie im Folgenden davon aus, dass es sich bei den Daten um eine einfache Stichprobe aus einer normalverteilten Grundgesamtheit aller Prüflinge dieser Klausur handelt.)

Bestimmen Sie zum Konfidenzniveau von 95 % jeweils ein Konfidenzintervall für

a) den durchschnittlichen Puls sowie für

b) die Standardabweichung des Pulses

in der Grundgesamtheit.

Lösungshinweis:

a)
```
## [1] 64.8 92.0
```

b)
```
## [1] 13.1 34.8
```

Aufgabe 5 — 11 Punkte

Gegeben ist zu einer diskreten Zufallsvariable X der Graph der Verteilungsfunktion F in Abb. 2.

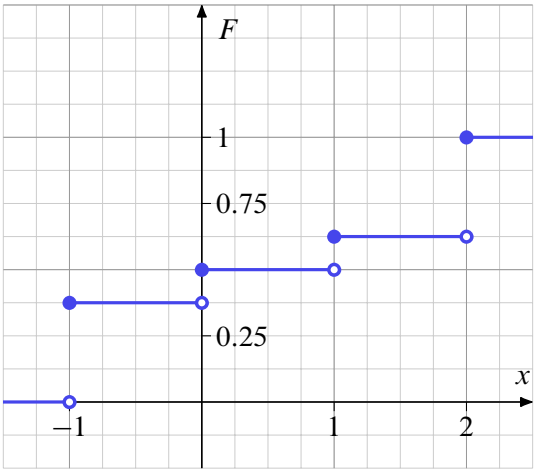


Abbildung 2: Graph von F

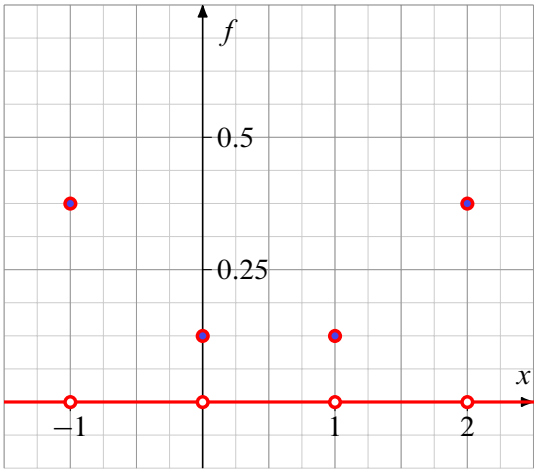


Abbildung 3: Graph von f

a) Zeichnen Sie zu X den Graph der Wahrscheinlichkeitsfunktion f in Abbildung 3 ein.

b) Nutzen Sie die Information in den Graphen und bestimmen Sie, wenn möglich, den Wert folgender Wahrscheinlichkeiten:

(1) $P(X < -1)$
(2) $P(X = 0)$
(3) $P(X \geqq 0)$
(4) $P(-1 \leqq X < 2)$
(5) $P(X \leqq 1 | X \geqq 2)$
(6) $P(X < 0 | X \leqq 0)$

Gegeben ist jetzt eine stetige Zufallsvariable Y mit der Dichtefunktion g. Von g ist bekannt:

- g ist achsensymmetrisch, also gilt $g(y) = g(-y)$ für alle $y \in \mathbb{R}$.
- $g(y) = 0$ für $|y| > 2$.
- Der Verlauf des Graphen von g für $x \in (-1.5, 1.5)$ ist in Abbildung 4 dargestellt.

g ist für $[-2, -1.5] \cup [1.5, 2]$ unbekannt (schwarze Balken).

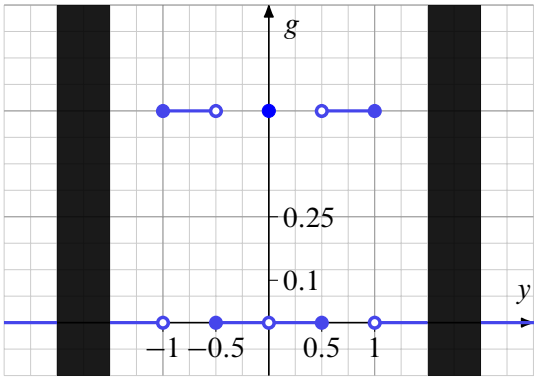


Abbildung 4: Dichtefunktion g zu Y

c) Kreuzen Sie bei den folgenden Vorschlägen für den Verlauf von g im unbekannten Bereich jeweils an, ob er möglich oder unmöglich ist. Eine korrekte Begründung (die in das Begründungsfeld passt) ist Voraussetzung für Punkte.

$g(x) =$	möglich	unmöglich	Begründung
$3/8$ für $x \in [-2, -1.5] \cup [1.5, 2]$		X	$2 \cdot 3/8 + 2 \cdot (1/2 \cdot 1/2) < 1$
$-1/8$ für $x \in [-2, -1.5] \cup [1.5, 2]$		X	Dichte < 0
$1/2$ für $x \in [-2, -1.5] \cup [1.5, 2]$	X		Fläche unter Dichte $= 1$
0 für $x \in [-2, -1.5] \cup [1.5, 2]$		X	Fläche unter Dichte $= 0.5 < 1$
$\begin{cases} 1 & \text{für } x \in [-2, -1.5] \\ 0 & \text{für } x \in [1.5, 2] \end{cases}$		X	Dichte nicht symmetrisch

Lösungshinweis:

a) siehe oben

b) (1) $P(X < -1) = 0$

(2) $P(X = 0) = 1/8 = 0.125$

(3) $P(X \geqq 0) = 0.125 + 0.125 + 0.375 = 0.625$

(4) $P(-1 \leqq X < 2) = 3/8 + 1/8 + 1/8 = 5/8 = 0.625$

(5) $P(X \leqq 1 | X \geqq 2) = P(\emptyset)/P(X \geqq 2) = 0$

(6) $P(X < 0 | X \leqq 0) = P(X < 0)/P(X \leqq 0) = \frac{3/8}{3/8+1/8} = 3/4 = 0.75$

c) s.o.

Aufgabe 6 — 12 Punkte

Betrachtet wird eine Funktion $g : \mathbb{R}^2 \to \mathbb{R}$. Gegeben ist dazu der Gradient ∇g sowie die Hessematrix $H_g(x,y)$ mit

$$\nabla g(x,y) = \begin{pmatrix} x^3 - 3x \\ y^3 - 12y \end{pmatrix}, \quad H_g(x,y) = \begin{pmatrix} 3\left(x^2-1\right) & 0 \\ 0 & 3\left(y^2-4\right) \end{pmatrix}$$

Die 9 kritischen Punkte von g sind in folgender Tabelle gegeben (Das müssen Sie nicht nachrechnen).

a) Kreuzen Sie jeweils an, um welche Art von kritischem Punkt es sich dabei handelt. Tragen Sie auch jeweils eine Begründung ein. Ohne (richtige) Begründung gibt es jeweils keine Punkte.

Punkt	Art: Minimum	Art: Maximum	Art: Sattelpunkt	Begründung
$(0,0)$		X		ad-bc>0, a<0
$(\pm\sqrt{3},0)$			X	ad-bc<0
$(0,\pm\sqrt{12})$			X	ad-bc<0
$(\pm\sqrt{3},\pm\sqrt{12})$	X			ad-bc>0, a>0
$(\mp\sqrt{3},\pm\sqrt{12})$	X			ad-bc>0, a>0

Nun wird eine anderen Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ betrachtet, von der Folgendes bekannt ist:

- Die 1. partielle Ableitung von f nach x ist gegeben mit

$$f_x(x,y) = x^2 - x - 6\,.$$

- Die 1. partielle Ableitung von f nach y ist ein Polynom 2. Grades von der Form

$$f_y(x,y) = y^2 + ay + b\,.$$

Die Koeffizienten $a, b \in \mathbb{R}$ sind unbekannt.

b) Bestimmen Sie die Nullstellen von $f_x(x,y)$.

c) In Abbildung 5 sind alle kritischen Punkte von f mit ● markiert. Bestimmen Sie die unbekannten Koeffizienten a, b von f_y.

d) Geben Sie die Hessematrix $H_f(x,y)$ von f an.

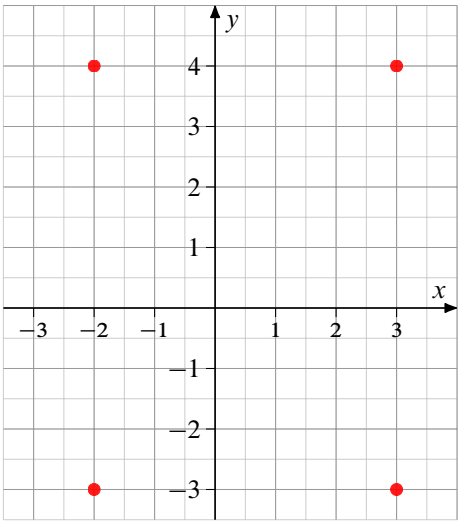


Abbildung 5: Kritische Punkte von f

Lösungshinweis:

a) $x_{1/2} = 3, -2$

b) Setze $y = 4, -3$ ein: $16 + 4a + b = 0$

$$9 - 3a + b = 0$$

$a = -1, b = -12$, also $f_y(x, y) = y^2 - y - 12$

c) $H_f(x, y) = \begin{pmatrix} 2x - 1 & 0 \\ 0 & 2y - 1 \end{pmatrix}$

d) s.o.