Klausur Statistik
Lösungshinweise

Prüfungsdatum: 30. Juni 2021 – Prüfer: Etschberger, Henle, Henle
Studiengang: BW, IM
Punkte: 14, 18, 11, 22, 11, 14 ; Summe der Punkte: 90

Aufgabe 1 — 14 Punkte

Gegeben ist die Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x,y) = x^4 y + xy^3 - \tfrac{5}{2}x^2 - 2y^2\,.$$

a) Berechnen Sie zu f den Gradienten $\nabla f(x,y)$ sowie die Hesse-Matrix $H_f(x,y)$.

b) f hat die kritischen Punkte $(0,0)$ und $(1,1)$. Bestimmen Sie für diese Punkte jeweils, ob es sich dabei um ein lokales Minimum, ein lokales Maximum oder einen Sattelpunkt handelt.

Jetzt werden drei andere Funktionen $g_i : \mathbb{R}^2 \to \mathbb{R}$, für $i = 1,2,3$ betrachtet. Von diesen ist jeweils der Gradient gegeben:

$$\nabla g_1(x,y) = \begin{pmatrix}1\\1\end{pmatrix}, \qquad \nabla g_2(x,y) = \begin{pmatrix}x\\2y\end{pmatrix}, \qquad \nabla g_3(x,y) = \begin{pmatrix}4xy^2\\4x^2y\end{pmatrix}$$

Bestimmen Sie jeweils einen möglichen Funktionsterm zu g_1, g_2, g_3, also

c) $g_1(x,y)$,

d) $g_2(x,y)$,

e) $g_3(x,y)$.

R f) Skizzieren Sie die Graphik, die durch folgende R-Befehle ausgegeben wird:

```
f = function(x, y) {x^4 * y + x * y^3 - 2.5 * x^2 - 2 * y^2}

tibble(x = c(-1, 0, 1), y = c(1, 1, 1) ) %>%
  mutate(f = f(x, y)) %>%
  ggplot(aes(x, f)) +
  geom_point()
```

Lösungshinweis:

a) $\nabla f(x,y) = \begin{pmatrix}4x^3y + y^3 - 5x\\ x^4 + 3xy^2 - 4y\end{pmatrix}$, $H(x,y) = \begin{pmatrix}12x^2y - 5 & 4x^3 + 3y^2\\ 4x^3 + 3y^2 & 6xy - 4\end{pmatrix}$.

b) $H(0,0) = \begin{pmatrix}-5 & 0\\ 0 & -4\end{pmatrix} \Rightarrow (0,0)$ ist ein Maximum.

c) $H(1,1) = \begin{pmatrix}7 & 7\\ 7 & 2\end{pmatrix} \Rightarrow (1,1)$ ist ein Sattelpunkt.

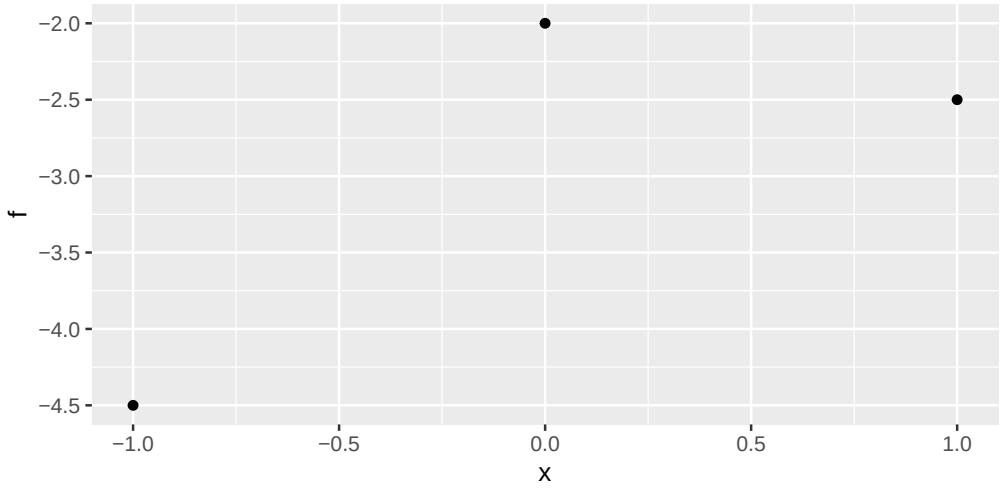
d) $g_1(x,y) = x + y + C$,

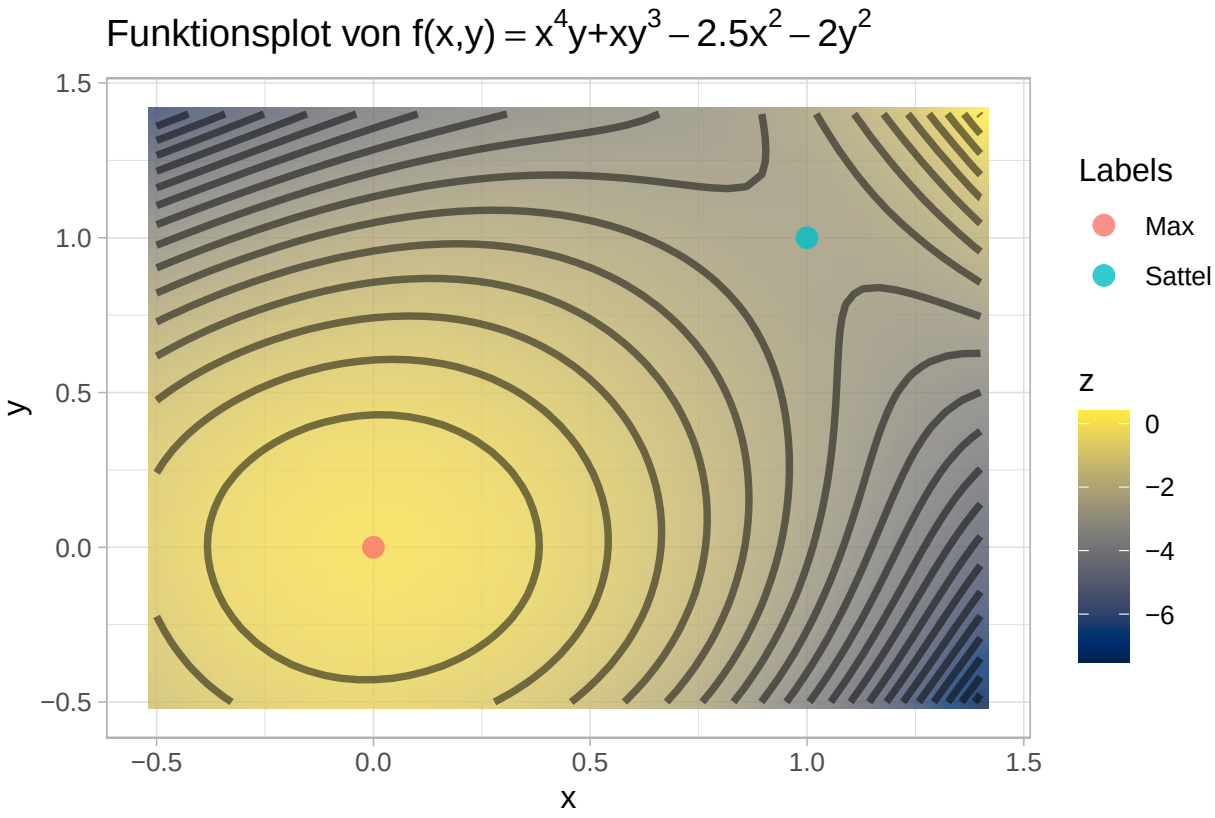
e) $g_2(x,y) = 0.5x^2 + y^2 + C$,

f) $g_3(x,y) = 2x^2y^2 + C$ (mit jeweils $C \in \mathbb{R}$).

g)
```
f = function(x, y) {x^4 * y + x * y^3 - 2.5 * x^2 - 2 * y^2}

tibble(x = c(-1, 0, 1), y = c(1, 1, 1) ) %>%
  mutate(f = f(x, y)) %>%
  ggplot(aes(x, f)) +
  geom_point()
```





Aufgabe 2 — 18 Punkte

Peter Planlos sitzt in der Statistikklausur neben Ihnen. Er ist leider nicht besonders gut in Statistik. Sie sind gut vorbereitet und Peter weiß das. Unbeobachtet von der Aufsicht schiebt er Ihnen seine Aufgabe zu mit der nonverbalen Bitte, seinen Versuch zu verbessern.

a) Zur Urliste $x = (20, 0, 1, 3, 1, 20, 3, 10, 8, 20)$ sollten die absoluten Häufigkeiten h_j, die relativen Häufigkeiten f_j, die kumulierten absoluten Häufigkeiten H_j und die kumulierten relativen Häufigkeiten F_j bestimmt werden.

Peters Lösung sieht folgendermaßen aus:

j	1	2	3	4	5	6
a_j	0	1	3	8	10	20
h_j	0	2	6	8	10	60
H_j	1	3	5	7	8	10
f_j	0.1	0.2	0.2	0.1	0.2	0.8
F_j	0.1	0.3	0.4	0.6	0.7	0.9

Korrigieren Sie die Lösung von Peter, indem Sie die falschen Werte in der Tabelle durchstreichen und durch die richtigen ersetzen.

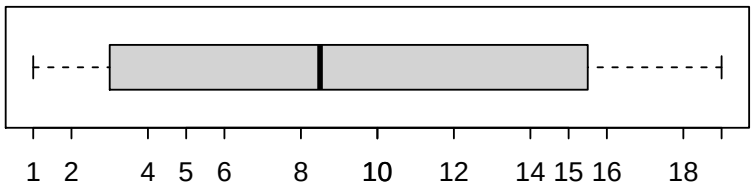
R b) Zur selben Urliste wie in a) sollte in dieser Aufgabe jeweils R Code angegeben werden, der

- Das 49 %-Quantil $\tilde{x}_{0.49}$ sowie
- einen plot der empirischen Verteilungsfunktion F graphisch ausgibt.

Die Urliste ist laut Aufgabenstellung im R-Vektor x gespeichert. Hier hat Peter gar nichts aufgeschrieben. Schreiben Sie die nötigen R-Befehle auf.

c) In der nächsten Aufgabe soll Peter zu einer anderen Urliste mit $n = 10$ Werten den Gini-Koeffizient G berechnen. Sein Ergebnis ist $G = 0.98$. Begründen Sie, ohne die Urliste zu kennen, warum dieses Ergebnis nicht stimmen kann.

d) Zuletzt soll Peter zur Urliste $y = (10, 7, 2, 2, 1, 19, 4, 7)$ einen Boxplot zeichnen. Peters Lösung sieht so aus:



Zeichnen Sie bitte den korrekten Boxplot für Peter.

R e) Mit welchem R-Code könnte Peter Teilaufgabe d) lösen?

Lösungshinweis:

a)

a	0	1	3	8	10	20
h	1	2	2	1	1	3
H	1	3	5	6	7	10
f	0.1	0.2	0.2	0.1	0.1	0.3
F	0.1	0.3	0.5	0.6	0.7	1.0

b)
```
x=c(0, 1, 1, 3, 3, 8, 10, 20, 20, 20)
quantile(x, probs=0.49, type=2)
ecdf(x) %>% plot
```

c) Mit $n = 10$ gilt: $G_{\max} = 9/10 < 0.98$

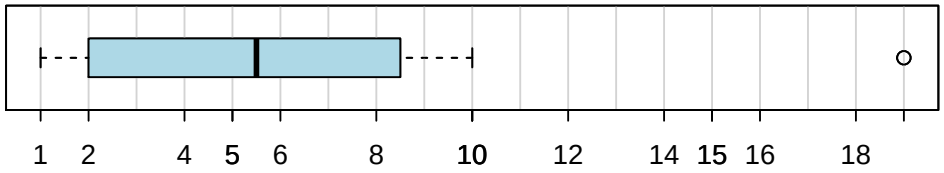
d)
```
y = c(10, 7, 2, 2, 1, 19, 4, 7)

Q1 = y %>% quantile(0.25, type=2)
Q2 = y %>% quantile(0.5, type=2)
Q3 = y %>% quantile(0.75, type=2)
IQR = Q3 - Q1

tibble(Q1, Q2, Q3,
       "Fence unten" = Q1 - 1.5 * IQR,
       "Fence oben"  = Q3 + 1.5 * IQR) %>%
  kbl(booktabs=TRUE)
```

Q1	Q2	Q3	Fence unten	Fence oben
2	5.5	8.5	-7.75	18.25

Also: Unterer Whisker bis 1, keine Ausreißer nach unten, oberer whisker bis 10, Ausreißer ist bei 19.



e)
```
y %>% boxplot()
```

Aufgabe 3 — 11 Punkte

Bernd und Ines sind auf der Suche nach einer gemeinsamen Wohnung mit einem guten Preis-Leistungsverhältnis. Um ein Gefühl für den Markt zu bekommen haben sie schon 10 Wohnungen besichtigt und für gut befunden. Von diesen Wohnungen haben sie sich jeweils die Grundfläche (in m^2) und die Kaltmiete (in €) notiert. Die beiden wollen ein lineares Regressionsodell erstellen, in dem die Miete (y) in Abhängigkeit von der Grundfläche (x) beschrieben wird. Zu den Daten haben Bernd und Ines die empirische Kovarianz Cov, sowie jeweils die Standardabweichung s_x, s_y und die arithmetischen Mittelwerte $\overline{x}, \overline{y}$ mit dem Taschenrechner ermittelt:

$$\text{Cov[x, y]} = \frac{1}{10}\sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y}) = 1844.5\,, \quad \overline{x} = 58.5\,, \quad \overline{y} = 603\,,$$

$$s_x = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(x_i - \overline{x})^2} = 14.86\,, \quad s_y = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(y_i - \overline{y})^2} = 129.23$$

Der Zettel mit den Originaldaten ist leider seit einem Unfall mit einer Kaffeetasse unleserlich.

a) Berechnen Sie den Determinationskoeffizienten des Regressionsmodells. Interpretieren Sie diesen Wert.

b) Schätzwerte für die Parameter eines Regressionsmodells $\hat{y} = \hat{a} + \hat{b}x$ kann man laut Vorlesung berechnen mit

$$\hat{b} = \frac{\sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{10}(x_i - \overline{x})^2}\,, \quad \hat{a} = \overline{y} - \hat{b}\overline{x}$$

Bestimmen Sie $\hat{a}, \hat{b}$.

Hinweis: Sollten Sie Teilaufgabe b) nicht lösen können (und nur dann) rechnen Sie mit den (falschen) Werten $\hat{a} = 182.962, \hat{b} = 13.36$ weiter.

c) Welche Grundfläche hätte eine Wohnung für 600 € Kaltmiete nach diesem Modell?

d) Wie viel mehr müssten die beiden nach diesem Modell ausgeben, wenn sie 10 m^2 Grundfläche mehr haben wollen?

e) Wieviel Miete sind nach dem Modell für eine Wohnung mit 0 m^2 zu erwarten? Erklären Sie, warum dieser Wert keinen Sinn ergibt.

R f) Bernd und Ines haben die Originaldaten doch wieder gefunden. Sie waren in einer R-Session in einem tidyverse tibble mit der Bezeichnung `Wohnungen` mit den beiden Spalten `Miete` und `Flaeche` gespeichert. Geben Sie R-Befehle an, so dass ein Streuplot mit den Daten der Wohnungen sowie der eingezeichneten Regressionsgerade ausgegeben wird.

Lösungshinweis:

a) $r = \text{Cov}/(s_x s_y) \Rightarrow R^2 = r^2 \approx 0.9226$,
d.h. ca. 92.3 % der Information in den Daten ist im Modell abgebildet.

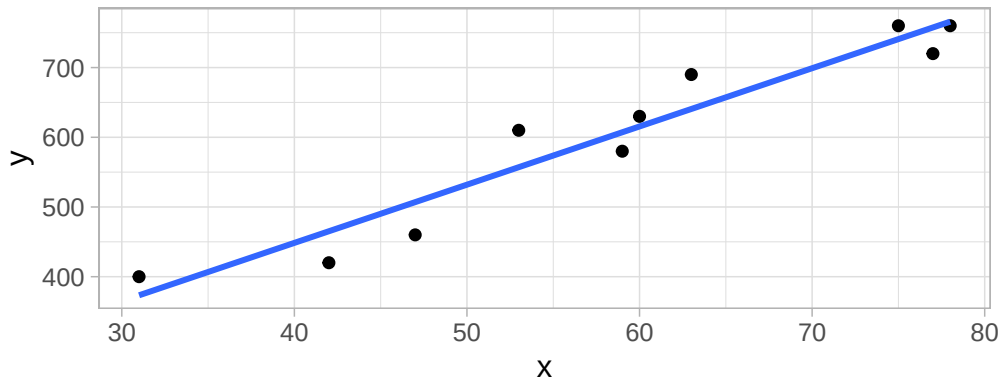
b) $\hat{b} = \frac{1844.5}{14.86^2} \approx 8.353$, $\hat{a} = 603 - 8.353 \cdot 58.5 \approx 114.35$

c) $x = \dfrac{y - a}{b} = \dfrac{600 - 114.351}{8.35} \approx 58.14$

d) $10 \cdot \hat{b} = 83.53$

e) $\hat{a} = 114.35$. Regressionsmodell funktioniert vermutlich außerhalb der beobachteten Werte schlechter (Extrapolation) und Kaltmiete skaliert eventuell nicht direkt proportional mit der Grundfläche (Sehr kleine Wohnungen haben höhere Kaltmiete pro qm).

f)
```
D %>%  ggplot(aes(x, y)) +
  geom_point() +
  theme_light() +
  geom_smooth(method = "lm", se = FALSE)
```



Aufgabe 4 — 22 Punkte

Gegeben sind zwei normalverteilte Zufallsvariablen X, Y mit

$$X \sim N(\mu_X, \sigma_X)\,, \quad Y \sim N(\mu_Y, \sigma_Y)\,.$$

Zu X ist gegeben, dass $\mu_X = 600, \sigma_X = 200$. Berechnen Sie damit für die Zufallsvariable X

a) $P(X = \sigma_X)$,

b) x, wenn gilt $P(X \leqq x) = 0.11314$,

c) $P(X \leqq 400 \vee X \geqq 800)$,

d) $P(\mu_X - 3\sigma_X \leqq X \leqq \mu_X + 3\sigma_X)$.

R e) Geben Sie jeweils einen R-Befehl an, der die Lösungen der Teilaufgaben b) und d) ausgibt.

Von Y sind zusätzlich die beiden folgenden Tatsachen bekannt:

$$P(Y > 9507) = 0.40\,, \quad P(Y \leqq 4347) = 0.01\,.$$

Berechnen Sie damit für die Zufallsvariable Y:

f) μ_Y,

g) σ_Y.

Lösungshinweis:

a) $P(X = \sigma_X) = 0$,

b) x, wenn gilt $P(X \leqq x) = 0.11314$; $x = 358$.

c) $P(X \leqq 400 \vee X \geqq 800) = 0.317$,

d) $P(\mu_X - 3\sigma_X \leqq X \leqq \mu_X + 3\sigma_X) \approx 0.997$,

e)
```
qnorm(0.11314, mean=600, sd=200) # Teilaufgabe b)
pnorm(3) - pnorm(-3)             # Teilaufgabe d)
```

f) $\mu_Y = 9004$,

g) $\sigma_Y = 1998$.

Aufgabe 5 — 11 Punkte

Es werden eine einfache Stichprobe mit Stichprobenumfang $n = 3$ sowie vier Stichprobenfunktionen betrachtet mit

$$\hat{\Theta}_1 = \frac{1}{3}(X_1 + X_2 + X_3)\,, \qquad \hat{\Theta}_2 = \frac{1}{6}(X_1 + 2X_2 + 3X_3)\,,$$
$$\hat{\Theta}_3 = 0.01 \cdot X_1 + 0.01 \cdot X_2 + 0.98 \cdot X_3\,, \qquad \hat{\Theta}_4 = \alpha(X_1 + X_2) + (1 - 2\alpha)X_3$$

Dabei ist $\alpha \in [0, 1]$.

a) Prüfen Sie für alle vier Stichprobenfunktionen, ob diese jeweils erwartungstreue Schätzfunktionen für den Erwartungswert μ der Grundgesamtheit sind.

b) Betrachten Sie die Stichpobenfunktionen $\hat{\Theta}_1, \hat{\Theta}_2, \hat{\Theta}_3$. Welche davon ist die wirksamste, welche davon ist am wenigsten wirksam?

c) Bestimmen Sie den Parameter α so, dass Sie die wirksamste Variante von $\hat{\Theta}_4$ erhalten.

Lösungshinweis:

a) $\mathrm{E}(\hat{\Theta}_1) = \mu$ (Stichprobenmittel, also erwartungstreu),
$\mathrm{E}(\hat{\Theta}_2) = \frac{1}{6}(\mu + 2\mu + 3\mu) = \frac{1}{6} \cdot 6\mu = \mu$, also erwartungstreu.
$\mathrm{E}(\hat{\Theta}_3) = 0.01\mu + 0.01\mu + 0.98\mu = \mu$, also erwartungstreu.
$\mathrm{E}(\hat{\Theta}_4) = \alpha\mu + \alpha\mu + (1 - 2\alpha)\mu = (2\alpha + 1 - 2\alpha)\mu = \mu$, also auch erwartungstreu.

b) $\mathrm{Var}(\hat{\Theta}_1) = \frac{1}{3}\sigma^2 \approx 0.33 \cdot \sigma^2$
$\mathrm{Var}(\hat{\Theta}_2) = \left(\frac{1}{6}\right)^2 (\sigma^2 + 2^2\sigma^2 + 3^2\sigma^2) = \frac{14}{36}\sigma^2 \approx 0.39 \cdot \sigma^2$
$\mathrm{Var}(\hat{\Theta}_3) = \left(0.01^2 + 0.01^2 + 0.98^2\right)\sigma^2 \approx 0.9606 \cdot \sigma^2$.
Damit gilt: $\hat{\Theta}_1$ ist am wirksamsten, $\hat{\Theta}_3$ am wenigsten wirksam von den drei Funktionen.

c) $\mathrm{Var}(\hat{\Theta}_4) = 2 \cdot \alpha^2\sigma^2 + (1 - 2\alpha)^2\sigma^2 = (6\alpha^2 - 4\alpha + 1) \cdot \sigma^2$;
damit $(6\alpha^2 - 4\alpha + 1)$ minimal wird: Differenziere und setze gleich 0:

$$(6\alpha^2 - 4\alpha + 1)' = 12\alpha - 4 = 0 \Leftrightarrow \alpha = \tfrac{1}{3}$$

Das ist ein Minimum, denn $(6\alpha^2 - 4\alpha + 1)'' = 12$, also ist $\mathrm{Var}(\hat{\Theta}_4)$ bzgl. α streng konvex und damit ist $\hat{\Theta}_4$ am wirksamsten für $\alpha = 1/3$.

Aufgabe 6 — 14 Punkte

Bertram Beier ist Mystery-Restaurant-Tester für einen internationalen Reiseführer. Heute testet er die Temperatur des servierten Bieres im Augsburger Biergarten *zum goldenen Vollhirsch*. Er bestellt im Laufe seines Arbeitstages 5 Halbe Helles und misst jeweils die Temperatur (Merkmal X) im Glas:

k	1	2	3	4	5
x_k	13	15	18	16	12

Gehen Sie im folgenden davon aus, dass es sich bei x um eine einfache Stichprobe einer normalverteilten Grundgesamtheit handelt.

a) Bestimmen Sie ein Konfidenzintervall zum Konfidenzniveau 99 % für die Varianz der Biertemperatur in der Grundgesamtheit. Warum ist das Ergebnis so schlecht und was könnte man dagegen machen?

b) Schreiben Sie die komplette Ausgabe folgender R Befehle auf:

```
biertemperatur = c(13, 15, 18, 16, 12)
konf.niveau = 0.99
alpha = 1 - konf.niveau

Ausgabe = tibble(x=biertemperatur) %>%
  summarize(x.var.mal.n.minus.1   = var(x) *( length(x) - 1),
            x.FG    = length(x) - 1 ) %>%
  mutate(Fraktil.u = qchisq(alpha/2,   df=x.FG),
         Fraktil.o = qchisq(1-alpha/2, df=x.FG) ) %>%
  mutate(KI.links   = x.var.mal.n.minus.1/Fraktil.o,
         KI.rechts  = x.var.mal.n.minus.1/Fraktil.u )
```

```
Ausgabe
```

c) Bertram belauscht ein Gespräch von zwei Servicekräften und erfährt, dass die Standardabweichung der Biertemperatur heute 4 Grad beträgt. Wie viele Biere müsste er demnach mindestens bestellen, so dass er ein Konfidenzintervall zur mittleren Biertemperatur in der Grundgesamtheit zu einem Konfidenzniveau von 99 % mit einer Genauigkeit (Breite des Konfidenzintervalls) von ± 1 Grad angeben könnte?

d) Bertram weiß, dass die im goldenen Vollhirsch servierte Biersorte am besten bei einer Temperatur von 10 Grad getrunken wird. Er testet die Nullhypothese: *Die Durchschnittstemperatur des servierten Bieres ist 10 Grad* und verwirft nachdem er den Test anhand einer Stichprobe durchgeführt hat diese Nullhypothese zu einem Signifikanzniveau von 1 %. Schreiben Sie in maximal 15 Worten auf, was das bedeutet. Erklären Sie (innerhalb der maximal 15 Worte) auch die Bedeutung des Signifikanzniveaus in diesem konkreten Zusammenhang.

Lösungshinweis:

a) K-Intervall für σ^2 in der GG: KI $= [1.53, 108.6]$; die Breite des Konfidenzintervalls ist so breit, dass die Aussage ziemlich nutzlos ist; liegt vor allem am kleinen Stichprobenumfang, aber auch am hohen Konfidenzniveau.

b) Ausgabe mit Fraktilswerten aus R:

x.var.mal.n.minus.1	x.FG	Fraktil.u	Fraktil.o	KI.links	KI.rechts
22.8	4	0.2069891	14.86026	1.534294	110.1507

Ausgabe mit Fraktilswerten aus der (Papier-)Tabelle

x.var.mal.n.minus.1	x.FG	Fraktil.u	Fraktil.o	KI.links	KI.rechts
22.8	4	0.21	14.86	1.53	108.6

c) $L = 2 \cdot \frac{\sigma c}{\sqrt{n}} \leqq 2 \Leftrightarrow n \geqq \sigma^2 c^2 = 4^2 \cdot 4.604^2 \approx 339.149056$.
Er müste also mindestens 340 Biere bestellen.

d) Statistischer Nachweis, dass Nullhypothese falsch, mit 1 % Wahrscheinlichkeit, dass Nullhypothese doch stimmt.