Klausur Statistik
Lösungshinweise

Prüfungsdatum: 1. Juli 2015 – Prüfer: Etschberger, Heiden, Jansen
Studiengang: IM und BW

Aufgabe 1 — 14 Punkte

Ein Freund von Ihnen hat über einen Teil seiner Daten, die er für seine Bachelorarbeit erhoben hat, Kaffee geschüttet. Die bereits sortierte Urliste wurde dadurch zum Teil unleserlich. Einige Einträge sowie einige Eigenschaften des kompletten Datensatzes sind aber noch zu entziffern:

x_1	x_2	x_3	x_4	x_5	x_6	$\cdots$
2	2	3	5	5	5	$\cdots$

Sie können erkennen, dass $\overline{x} = 5{,}5$, der $x_{\text{mod}} = 6$ ist, die Spannweite 9 und $F(6) = F(8)$ ist.

a) Rekonstruieren Sie die Urliste aus den Ihnen zur Verfügung stehenden Informationen.

Für die Teilaufgaben b) – d) ist eine zweite Urliste eines anderen Merkmals gegeben:

x_1	x_2	x_3	x_4	x_5	x_6	x_7
2	2	3	5	5	5	8

b) R: Geben Sie einen R-Befehl an, mit dem man die Daten in einer Variable x speichert.

c) R: Geben Sie jeweils R-Befehle an, um den Mittelwert, den Median, die mittlere quadratische Abweichung sowie die Spannweite der Daten zu berechnen.

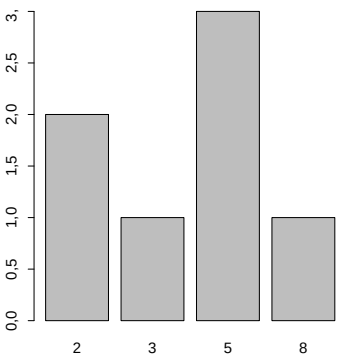
d) R: Welcher R-Befehl gibt ein Balkendiagramm aus, bei dem die absoluten Häufigkeiten aller Ausprägungen der x_i dargestellt sind?

Lösungshinweis:

a) 6 kommt mind. 4 mal vor, kein Wert bei $x \in (6;8]$, höchster Wert ist 11. Nimmt man einen Wert bei 9 noch dazu kommt man auf $\overline{x} = 5{,}5$. Damit ist Urliste $2, 2, 3, 5, 5, 5, 6, 6, 6, 6, 9, 11$.

b) , c), d) In R:

```
x = c(2, 2, 3, 5, 5, 5, 8) # b) Daten einlesen
mean(x)                    # c) arithm. Mittel
median(x)                  #    Median
mean((x-mean(x))^2)        #    mittl. quadr. Abweichung
max(x) - min(x)            #    Spannweite
barplot(table(x))          # d) Balkendiagramm
```

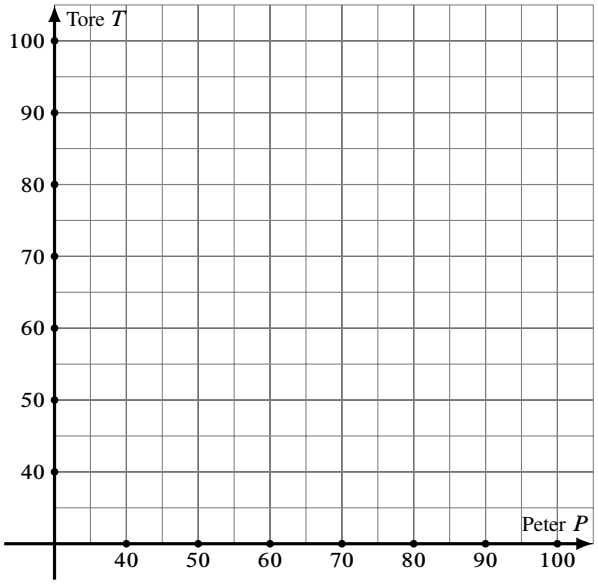


Aufgabe 2 — 24 Punkte

Boris interessiert sich eigentlich nicht für Fußball. Er hat aber neulich Barbara kennengelernt, die leidenschaftlich gerne Fußball kuckt. Um bei Ihr nicht als total ahnungslos dazustehen, möchte Boris das Wissen seines WG-Kumpels Peter nutzen, der sich als Fachmann bezeichnet. Peter hatte schon in der Vergangenheit immer Tipps über die Anzahl der Tore abgegeben, die ein bestimmter Verein in der kommenden Saison insgesamt erzielen wird.

Boris findet eine Tabelle zur vergangenen Saison mit Peters damaligen Prognosen und den dann tatsächlich gefallenen Toren von 10 Vereinen. Er liest:

Verein	Peters Prognose	tatsächliche Tore
1	40	81
2	76	55
3	94	36
4	46	82
5	33	87
6	78	48
7	65	63
8	86	55
9	97	39
10	33	99



a) Tragen Sie die beiden Merkmale Peters Prognose P und die tatsächlich gefallenen Tore T als Streuplot in das nebenstehende Koordinatensystem ein.

b) Berechnen Sie einen geeigneten Korrelationskoeffizienten der beiden Variablen.

c) Die Prognosen von Peter scheinen ziemlich schlecht zu sein. Warum kann man basierend auf diesen Daten trotzdem Peters Prognosen vermutlich als Ausgangspunkt einer neuen, eigenen Prognose nutzen?

d) R: Geben Sie zwei Zeilen R-Code an, mit denen die beiden Merkmale in den Variablen P und T in R eingegeben werden können (Die Datenwerte können Sie dabei abkürzen).

e) R: Wie berechnet man die Korrelation von P und T in R?

f) Boris möchte das „Wissen“ von Peter ausnutzen und berechnet zu diesem Zweck ein lineares Regressionsmodell der Toranzahl in Abhängigkeit von Peters Prognosewerten. Berechnen Sie auch dieses Modell und geben Sie die Modellgleichung an.

g) Angenommen Peter prognostiziert für einen Verein in der kommenden Saison 45 Tore: Wieviel Tore würde Boris (basierend auf dem Regressionsmodell) schätzen?

h) R: Geben Sie R-Befehle an, mit denen man
- den Streuplot zeichnet,
- das Regressionsmodell in R berechnet,
- und die Regressionsgerade in R in den Streuplot einzeichnet.

Lösungshinweis:

a) Streuplot: siehe h)

b) Bravais-Pearson: $r = -0{,}9736118$

c) Auch die negative Korrelation kann man ausnutzen, vorausgesetzt, sie setzt sich in der Zukunft so fort...

d) R:

```
P = c(p1, p2, ...)
T = c(t1, t2, ...) # jeweils 10 Werte
```

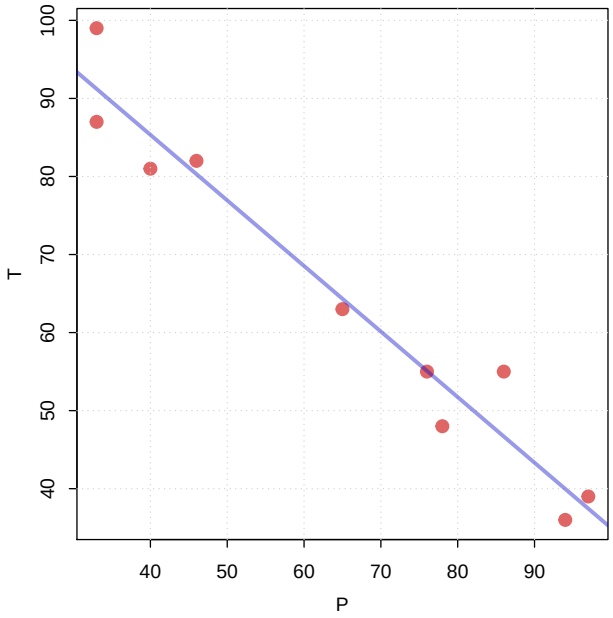
e)
```
cor(P, T)

## [1] -0,9736118
```

f) $T(P) = 118{,}945076 - 0{,}8402018 \cdot P$

g) $T(45) = 118{,}945076 - 0{,}8402018 \cdot 45 \approx 81{,}1359955$

h)
```
plot(P, T)
Modell = lm( T ~ P )
abline(Modell)
```



Aufgabe 3 — 5 Punkte

a) Bestimmen Sie die Anzahl der verschiedenen Permutationen, die aus allen Buchstaben des Wortes SEEWEG gebildet werden können.

b) Wie viele von den Wörtern beginnen und enden mit einem E?

c) Wie viele von den Wörtern beginnen mit E und enden mit einem G?

d) In wie vielen Wörtern stehen alle drei E hintereinander?

Lösungshinweis:

a) $\frac{6!}{3!} = 120$

b) $1 \cdot 4! \cdot 1 = 24$

c) $1 \cdot \frac{4!}{2!} \cdot 1 = 12$

d) $4 \cdot 3! = 24$

Aufgabe 4 — 15 Punkte

Berechnen Sie für die Zufallsvariable Z und die Fälle a) ... e) jeweils die Wahrscheinlichkeit

$$P(2 \leqq Z < 4.5)\,.$$

Hinweis: Bitte geben Sie alle Zahlenergebnisse (auch Zwischenergebnisse) mit 4 Nachkommastellen an

a) Z ist binomialverteilt nach $B(20; 0.10)$.

b) (R): Geben Sie einen R-Befehl an, der die gesuchte Wahrscheinlichkeit aus Teilaufgabe a) ausgibt.

c) Z ist hypergeometrisch verteilt mit $N = 50$, $M = 2$ und $n = 5$.

d) Z ist gleichverteilt im Intervall [1;b] und es gelte $F(2) = \frac{1}{20}$. Bitte berechnen Sie hier zunächst die Intervallgrenze b und geben Sie diese an.

e) Z ist poissonverteilt und es gilt $P(Z = 0) = 0.1353$. Berechnen Sie hier zunächst den Parameter λ der Poissonverteilung.

f) (R): Benutzen Sie λ aus Teilaufgabe e) und geben Sie einen R-Befehl an, der die gesuchte Wahrscheinlichkeit aus e) ausgibt.

Lösungshinweis:

a) $P(2 \leqq Z < 4.5) = f(2) + f(3) + f(4)$
$= \binom{20}{2} \cdot 0.10^2 \cdot 0.90^{18} + \binom{20}{3} \cdot 0.10^3 \cdot 0.90^{17} + \binom{20}{4} \cdot 0.10^4 \cdot 0.90^{16} \approx 0{,}5651$

b)
```
P = pbinom(4, size = 20, prob = 0.10) - pbinom(1, size = 20, prob = 0.10)
```

c) $P(2 \leqq Z < 4.5) = P(Z = 2)$, da nur 2 Treffer möglich
$P(2 \leqq Z < 4.5) = \frac{\binom{2}{2}\binom{50-2}{5-2}}{\binom{50}{5}} \approx 0{,}0082$

d) $F(2) = \frac{1}{20}$ und $a = 1 \quad \Rightarrow \quad b = 21$
$P(2 \leqq Z < 4.5) = F(4.5) - F(2) = 0.175 - 0.05 = 0.125$

e) $P(Z = 0) = 0.1353 \Rightarrow f(0) = \frac{\lambda^0}{0!} e^{-\lambda} = 0.1353$
Nach Lambda auflösen oder in Verteilungsfunktion suchen ergibt $\lambda = -\ln\ 0.1353 \approx 2$
Mit $\lambda = 2$ dann Wahrscheinlichkeit berechnen bzw. aus Verteilungstabelle ablesen.
$P(2 \leqq Z < 4.5) = F(4) - F(1) = 0{,}9473 - 0{,}406 = 0{,}5413$

f)
```
ppois(4, lambda = 2) - ppois(1, lambda = 2)
```

Aufgabe 5 — 16 Punkte

Für die zweidimensionale Zufallsvariable (X,Y) sei folgendes bekannt:

- X hat den Wertebereich $\{-2; 0; 2\}$
- Y hat den Wertebereich $\{0; 1\}$
- Es gilt: $P(X = -2) = P(X = 0) = 0.3$, $P(X = -2, Y = 0) = 0.3$, $P(X = 2, Y = 0) = 0.1$, $P(Y = 0) = 0.6$

a) Berechnen Sie die fehlenden gemeinsamen Wahrscheinlichkeiten und Randwahrscheinlichkeiten von (X,Y) und tragen Sie diese in eine passende Tabelle ein.

Gehen Sie für die Teilaufgaben b) ... d) von der zweidimensionalen Zufallsvariablen (A,B) mit

$\downarrow A \backslash B \rightarrow$	0	3	
-1	0.05	0.05	0.10
0	0.20	0.00	0.20
1	0.40	0.30	0.70
	0.65	0.35	1

aus und berechnen Sie bitte folgende Größen:

b) Den Erwartungswert und die Varianz von A.

c) Den Erwartungswert und die Varianz von B.

d) Den Erwartungswert der Zufallsvariablen $C = A \cdot B$ sowie $\mathrm{Cov}[A,B]$.

Lösungshinweis:

a) Lösungstabelle:

$\downarrow X \backslash Y \rightarrow$	0	1	
-2	0.3	0.0	0.3
0	0.2	0.1	0.3
2	0.1	0.3	0.4
	0.6	0.4	1

b) $\mathrm{E}(A) = -1 \cdot 0.1 + 0 \cdot 0.2 + 1 \cdot 0.7 = 0.6$
$\mathrm{E}(A^2) = (-1)^2 \cdot 0.1 + 1^2 \cdot 0.7 = 0.8$
$\mathrm{Var}(A) = \mathrm{E}(A^2) - [\mathrm{E}(A)]^2 = 0.8 - (0.6)^2 = 0.44$

c) $\mathrm{E}(B) = 0 \cdot 0.65 + 3 \cdot 0.35 = 1.05$
$\mathrm{Var}(B) = 3^2 \cdot 0.35 - (1.05)^2 = 3.15 - 1.1025 = 2.0475$

d) $\mathrm{E}(C) = \mathrm{E}(A \cdot B)$. Mit

$A \cdot B$	0	-3	3
$f(A \cdot B)$	0.65	0.05	0.3

folgt: $\mathrm{E}(A \cdot B) = -3 \cdot 0.05 + 3 \cdot 0.3 = 0.75$ und damit
$\mathrm{Cov}[A, B] = \mathrm{E}(A \cdot B) - \mathrm{E}(A) \cdot \mathrm{E}(B) = 0.75 - 0.6 \cdot 1.05 = 0{,}12$

Aufgabe 6 — 16 Punkte

Eine Hochschule interessiert sich für das Einkommen ihrer Absolventen. Dazu werden 25 berufstätige Alumni 10 Jahre nach dem Abschluss zu ihrem aktuellen Einkommen (Merkmal X, in Tausend Euro pro Jahr) befragt. Die Beobachtungen können als Ergebnis einer einfachen Stichprobe aus einer normalverteilten Grundgesamtheit angesehen werden. Es ergeben sich für die Ausprägungen a_i bzw. für die Häufigkeiten h_i in der Stichprobe:

a_i	20	25	30	35	40	45	50	55	60	65	85	90	95
h_i	1	3	2	1	1	2	3	3	1	3	3	1	1

a) Bestimmen Sie ein Konfidenzintervall für den Mittelwert der Einkommen in der Grundgesamtheit (unabhängig vom Studiengang) zu einem Konfidenzniveau von 90 %.

b) R: Angenommen, die Urliste ist in R in der Variable x gespeichert. Geben Sie *einen* R-Befehl an, mit dem man das Konfidenzintervall aus a) berechnen kann.

c) Wie müsste die Nullhypothese H_0 und die Gegenhypothese H_1 lauten, wenn die Hochschulleitung statistisch bestätigen möchte, dass das durchschnittliche Einkommen in der Grundgesamtheit (Gehalt aller Absolventen 10 Jahre nach dem Abschluss) höher als 40.000 € ist?

d) Würden Sie eher ein hohes oder ein niedriges Signifikanzniveau wählen, wenn Sie diese Vermutung statistisch bestätigen wollen?

e) Was bedeutet der Fehler 1. Art hier?

f) R: Beim Ausführen eines Tests in R ergibt sich mit den Einkommensdaten als Ausgabe:

```
t.test(x, mu = 40, alternative = "greater")

##
##  One Sample t-test
##
## data:  x
## t = 2,971, df = 24, p-value = 0,003324
## alternative hypothesis: true mean is greater than 40
## 95 percent confidence interval:
##  45,5986     Inf
## sample estimates:
## mean of x
##      53,2
```

Was bedeutet in dieser Ausgabe $t = \ldots$, $\mathrm{df} = \ldots$?
Würden Sie anhand der Ausgabe zu einem Signifikanzniveau von 5% die Aussage bestätigen, dass das durchschnittliche Einkommen in der Grundgesamtheit (Gehalt aller Absolventen 10 Jahre nach dem Abschluss) höher als 40.000 € ist? Woran haben Sie Ihre Entscheidung abgelesen?

Lösungshinweis:

a) $c = x_{0,95} = 1{,}711, \bar{x} = 53{,}2, s = 22{,}215 \quad \Rightarrow \quad \left[\bar{x} \pm \frac{sc}{\sqrt{n}}\right] = [45{,}599, 60{,}801]$

b)
```r
t.test(x, conf.level = 0.9)$conf.int # alternativ: ohne $conf.int und
                                     # Intervall aus Ausgabe ablesen

## [1] 45,6 60,8
## attr(,"conf.level")
## [1] 0,9
```

c) $H_0 : \mu = 40$ gegen $H_1 : \mu > 40$.

d) Je größer α, desto eher wird H_0 abgelehnt, also sollte ein hohes Signifikanzniveau gewählt werden (dafür: höheres Risiko für Fehler 1. Art)

e) Fehler 1. Art: Die Stichprobe führt zu einer Ablehnung der Nullhypothese ($\mu = 40$), obwohl H_0 stimmt.

f) t entspricht dem Testfunktionswert $v = (\bar{x}-\mu)\sqrt{n}/s$
df steht für „degrees of freedom“, also die Anzahl der Freiheitsgrade, hier $n - 1 = 25 - 1 = 24$
H_0 würde hier verworfen, da p-value $< \alpha$.