Klausur Statistik
Lösungshinweise

Prüfungsdatum: 20. Juli 2016 – Prüfer: Etschberger, Jansen, Schneller, Wins
Studiengang: IM, BW, Inf und W-Inf
Punkte: 15, 18, 9, 23, 16, 9 ; Summe der Punkte: 90

Aufgabe 1 — 15 Punkte

Für ein metrisches Merkmal X wurden 30 Beobachtungen erfasst. Für X ergibt sich die empirische Verteilungsfunktion F mit

$$F(x) = \begin{cases} 0 & \text{für } x < 4 \\ 0.2 & \text{für } 4 \leqq x < 8 \\ 0.4 & \text{für } 8 \leqq x < 12 \\ 0.4 & \text{für } 12 \leqq x < 15 \\ 0.7 & \text{für } 15 \leqq x < 22 \\ 0.9 & \text{für } 22 \leqq x < 24 \\ 1 & \text{für } x \geq 24 \end{cases}$$

a) Erstellen Sie eine Tabelle der absoluten Häufigkeiten.

b) Berechnen Sie mit Hilfe der angegebenen empirischen Verteilungsfunktion
   (1) den Modus des Merkmals X.
   (2) die relative Häufigkeit der Ausprägung 21.
   (3) die absolute Häufigkeit der Ausprägung 15.

Für die Teilaufaben c) bis e) sei ein weiteres metrisches Merkmals Y mit ebenfalls $n = 30$ Beobachtungen gegeben. Für Y sind die Ausprägungen a_i und die relativen Häufigkeiten $f(a_i)$ in der folgenden Tabelle aufgeführt:

a_i	3	6	16	22	25
$f(a_i)$	0.1	0.3	0.2	0.2	0.2

c) Bestimmen Sie den Median von Y.

d) Bestimmen Sie die kumulierte relative Häufigkeit für die Ausprägung 17.

e) Berechnen Sie den Anteil der Beobachtungen von Y, an denen eine Ausprägung von mindestens 12, aber weniger als 23 vorliegt.

R

Nehmen Sie für die Teilaufgaben f) bis h) an, dass eine Urliste $x_1, \ldots, x_n$ in einem R-Vektor `data` gespeichert ist. Geben Sie das (die) R-Kommando(s) an, mit dem (denen) Sie

f) einen horizontal dargestellten Boxplot der Daten erstellen.

g) ein Histogramm mit den Klassengrenzen 75, 80, 95 und 105 erstellen.

h) eine Tabelle der kumulierten absoluten Häufigkeiten erstellen.

Für Teilaufgabe i) ist folgende Tabelle zu den Daten der Urliste $x_1, \ldots, x_7$ gegeben.

k	1	2	3	4	5	6	7
x_k	2	2	4	8	8	10	10
p_k	$2/44$	$2/44$	$4/44$	$8/44$	$8/44$	$10/44$	$10/44$
v_k	$2/44$	$4/44$	$8/44$	$16/44$	$24/44$	$34/44$	1
u_k	$1/7$	$2/7$	$3/7$	$4/7$	$5/7$	$6/7$	1

i) Bestimmen Sie die Knickstellen der zugehörigen Lorenzkurve.
(*Hinweis*: Die Lorenzkurve muss nicht gezeichnet werden)

Lösungshinweis:

a)

a_k	4	8	15	22	24
f_k	0.2	0.2	0.3	0.2	0.1
h_k	6	6	9	6	3

b) (1) $x_{\text{Mod}} = 15$, da maximale relative Hfk.
(2) $h(21) = 0$, (3) $h(15) = 0.3$

c) Sortierte Urliste x_i: 3 3 3 6 6 6 6 6 6 6 6 6 16 16 16 16 16 16 22 22 22 22 22 22 25 25 25 25 25 25
n gerade $\Rightarrow y_{\text{Med}} = \frac{1}{2}(y_{(15)} + y_{(16)}) = 16$

d) $F(17) = 0.6$

e) $F(22) - F(11) = F(22) - F(8) = 0.9 - 0.4 = 0.5$
oder: $h(22) + h(15) = 0.2 + 0.3 = 0.5$

f) `boxplot(data, horizontal=TRUE)`

g) `hist(data, breaks=c(75,80,95,105))`

h)
```
table1=table(data)
cumsum(table1)
```

i) Knickstellen bei k=2,3,5

Aufgabe 2 — 18 Punkte

Stefan Jumper nimmt seit 8 Jahren am gleichen Marathon teil und hat jedes Jahr seine Trainings- und Ergebnisdaten dokumentiert. Er hat dazu pro Jahr jeweils die Ergebniszeit im Marathon (Merkmal *Ergebnis Marathon*, in Minuten) sowie die durchschnittliche Anzahl gelaufener Trainingskilometer in den 8 bzw. 16 Wochen vor dem Marathon (Merkmal *8-Wochen-Trainings-Durchschnitt* bzw. *16-Wochen-Trainings-Durchschnitt*, jeweils in km pro Woche) in der folgenden Tabelle festgehalten:

Jahr	1	2	3	4	5	6	7	8
Ergebnis Marathon	172	161	156	147	152	167	157	168
8-Wochen-Trainings-Durchschnitt	85	105	125	150	130	90	110	95
16-Wochen-Trainings-Durchschnitt	70	80	120	125	125	100	105	75

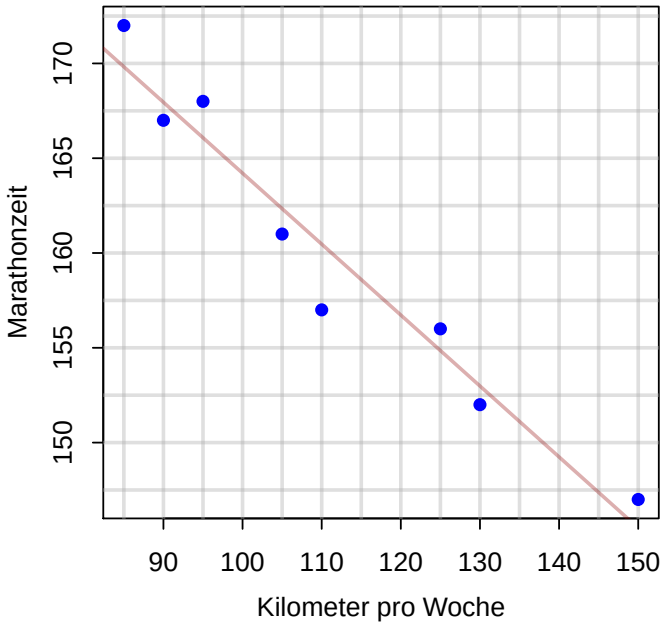
Sie haben Stefan laufend von den interessanten Themen der Statistik-Vorlesung berichtet, die Sie in diesem Semester besucht haben. Stefan wünscht sich daraufhin von Ihnen eine statistische Auswertung seiner Ergebnisse.

a) Stefan vermutet, dass der 8-Wochen-Trainings-Durchschnitt eine bessere Prognose für seine gelaufene Marathonzeit darstellt als der 16-Wochen-Trainings-Durchschnitt.

Berechnen Sie jeweils einen geeigneten Korrelationskoeffizienten, vergleichen Sie die beiden Ergebnisse und geben Sie an, ob Stefans Vermutung durch das Ergebnis gestützt wird.

b) Bestimmen Sie den Funktionsterm eines linearen Regressionsmodells für das Ergebnis im Marathon in Abhängigkeit vom 8-Wochen-Trainings-Durchschnitt.

c) Zeichnen Sie die Datenpunkte sowie die Regressionsgerade in das nebenstehende Koordinatensystem ein. Beschriften Sie dazu auch die Achsen geeignet. (Hinweis: Der Schnittpunkt der Koordinatenachsen soll nicht bei (0, 0) liegen)



d) Welche Marathonzeit erwartet Stefan auf Grundlage des linearen Modells für einen 8-Wochen-Trainings-Durchschnitt von 175 km pro Woche?

R e) Korrigieren Sie den folgenden nachlässig aufgeschriebenen R-Code, mit dem ein Student einen Teil der obigen Aufgabe lösen wollte:

```
Zeit = c[172:161:156:147:152:167:157:168]
8Wochendurchschnitt == seq(85,105,125,150,130,90,110,95)
corr(Zeit, 8Wochendurchschnitt, method = "spearman")
z <- LM(8Wochendurchschnitt~Zeit)
plotter(Zeit,8Wochendurchschnitt, color = "blau",
        ylab = Kilometer pro Woche, ylab == "Zeit")
abline(z)
```

Lösungshinweis:

```
Marathon    = c(172, 161, 156, 147, 152, 167, 157, 168)
WochenKM.8  = c(85, 105, 125, 150, 130, 90, 110, 95)
WochenKM.16 = c(70, 80, 120, 125, 125, 100, 105, 75)
```

```
cor(Marathon, WochenKM.8,  method = "pearson")
cor(Marathon, WochenKM.16, method = "pearson")

Marathon.Regression = lm(Marathon ~ WochenKM.8)
Marathon.Regression

plot(WochenKM.8, Marathon, col = "blue", pch=16, cex=1.5,
     xlab = "Kilometer pro Woche", ylab = "Marathonzeit")
grid()
abline(Marathon.Regression, lwd=2, col="#90000050")
```

Aufgabe 3 — 9 Punkte

Zwei faire Würfel werden geworfen. Folgende Ereignisse werden definiert:

- A: Die Augensumme beider Würfel beträgt 3
- B: Die Augensumme beider Würfel beträgt 7
- C: Mindestens einer der beiden Würfel zeigt eine 1

a) Tragen Sie die folgenden gesuchten Wahrscheinlichkeiten in die Tabelle ein. Geben Sie auch jeweils eine kurze Begründung oder Berechnung für das Ergebnis an.

Gesucht	Ergebnis	Begründung
$P(A)$	$2/36 \approx 5.56\,\%$	$A = \{(1,2),(2,1)\}$
$P(B)$	$6/36 \approx 16.67\,\%$	$A = \{(1,6),(2,5),\ldots,(6,1)\}$
$P(C)$	$11/36 \approx 30.56\,\%$	$C = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,1)\}$
$P(C \cap A)$	$2/36 \approx 5.56\,\%$	$= P(A)$
$P(C \cup A)$	$11/36 \approx 30.56\,\%$	$= P(C)$
$P(A\|C)$	$2/11 \approx 18.18\,\%$	$P(A \cap C)/P(C) = P(A)/P(C)$
$P(C\|A)$	1	$P(A)/P(A)$
$P\big((A \cup B)\|C\big)$	$4/11 \approx 36.36\,\%$	$(A \cup B) \cap C = \{(1,2),(2,1),(1,6),(6,1)\}$

b) Sind A und C unabhängig?

c) Sind B und C unabhängig?

Lösungshinweis:

a) siehe oben

b) Nein, da $P(A|C) = 2/11 \neq 2/36 = P(A)$

c) Nein, da $P(B|C) = 2/11 \neq 6/36 = P(B)$

Aufgabe 4 — 23 Punkte

Ein Mensch zwinkert durchschnittlich alle 5 Sekunden 1 mal. Gehen Sie davon aus, dass die Augen in einer Stunde durch Zwinkern durchschnittlich 4 Minuten und 48 Sekunden geschlossen sind.

a) Wie lange dauert ein Zwinkern durchschnittlich?

Fotograf Felix Fix fotografiert oft Menschen und denkt darüber nach, dass er gelegentlich Leute mit geschlossenen Augen auf seinen Bildern hat.

(*Hinweis*: Vernachlässigen Sie im Folgenden die Belichtungszeit der Fotos und gehen Sie davon aus, dass alle fotografierten Personen wach und ihre Gesichter auf dem Foto sichtbar sind)

b) Es wird ein Foto einer Person zufällig geschossen. Berechnen Sie die Wahrscheinlichkeit dafür, dass diese Person die Augen auf dem Foto nicht geschlossen hat.

c) Berechnen Sie die Wahrscheinlichkeiten dafür, dass auf einem Foto mit 50 Leuten

i) keiner

ii) höchstens 3

iii) 5 bis 10 Leute

die Augen geschlossen haben.

d) Wie groß ist die Wahrscheinlichkeit dafür, dass bei 5 Fotos mit 50 Leuten mindestens eins dabei ist auf dem niemand die Augen geschlossen hat?

e) Wie viele Fotos müssen mindestens geschossen werden, so dass mit einer Wahrscheinlichkeit von mindestens 99 % auf mindestens einem Foto niemand der 50 Leute die Augen geschlossen hat.

R f) Geben Sie die Lösungen zu den Teilaufgaben c) und d) jeweils mit Hilfe einer R-Funktion an.

Lösungshinweis:

a) Anzahl Zwinkern pro Stunde $= 3600/5 = 720$.
Dauer für ein Zwinkern $= (4 \cdot 60 + 48)/720 = 288/720 = 0.4$ Sekunden.

b) $p = 1 - (288/3600) = 1 - 0.08 = 0.92$

c) X : Anzahl der Leute mit geschlossenen Augen auf dem Foto.
$X \sim B(n = 50, p = 0.08)$.

i) keiner: $P(X = 0) = \binom{50}{0} \cdot 0.08^0 \cdot 0.92^{50} = 0.0154665 := p_{ci} \approx 1.55\,\%$

ii) höchstens 3: $P(X \leqq 3) = F(3) = 0.4252957 \approx 42.53\,\%$

iii) 5 bis 10 Leute: $P(5 \leqq X \leqq 10) = F(10) - F(4) = 0.3693422 \approx 36.93\,\%$

d) Y : Anzahl Fotos von 5 Fotos (mit jeweils 50 Leuten drauf), jeweils alle mit offenen Augen.
$Y \sim B(n = 5, p_{ci} \approx 0.0154665)$

$P(Y \geqq 1) = 1 - P(Y = 0) = 1 - \binom{5}{0} \cdot p_{ci}^0 \cdot (1 - p_{ci})^5 = 0.074977$

e) Z : Anzahl Fotos von n Fotos (mit jeweils 50 Leuten drauf), jeweils alle mit offenen Augen.
$Z \sim B(n, p_{ci} \approx 0.0154665)$

$$P(Z \geqq 1) = 1 - P(Z = 0) \geqq 0.99 \quad ,$$

$$\Leftrightarrow P(Z = 0) \leqq 0.01 \Leftrightarrow \binom{n}{0} \cdot p_{ci}^{0} \cdot (1 - p_{ci})^{n} \leqq 0.01$$

$$\Leftrightarrow n \geqq \frac{\ln 0.01}{\ln(1 - p_{ci})} \approx 295.4431792$$

also sind mind. 296 Fotos nötig.

f)
```
p.ci = dbinom(x = 0, size = 50, prob = 0.08) # Teilaufgabe c, i)
p.ci
## [1] 0.01546648
pbinom(q = 3, size = 50, prob = 0.08)        # Teilaufgabe c, ii)
## [1] 0.4252957
pbinom(10, 50, 0.08) - pbinom(4, 50, 0.08)   # Teilaufgabe c, iii)
## [1] 0.3693422
1-dbinom(x=0, size=5, prob = p.ci)           # Teilaufgabe d
## [1] 0.07497697
```

Aufgabe 5 — 16 Punkte

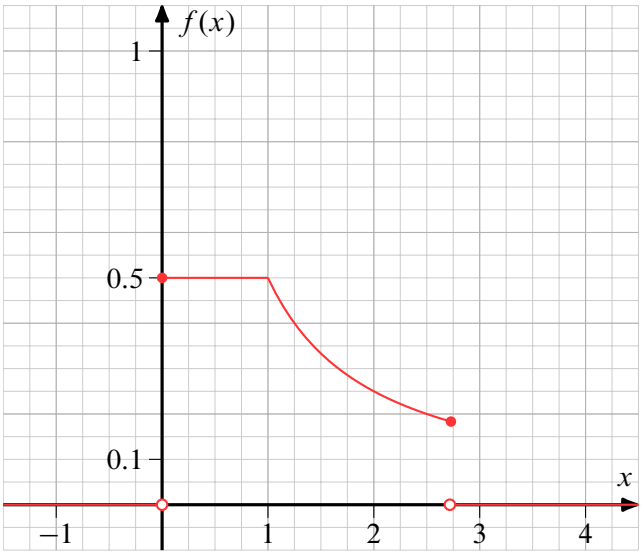
Gegeben sei zur Zufallsvariablen X die Dichtefunktion f gemäß:

$$f(x) = \begin{cases} 0.5 & \text{falls } 0 \leqq x < 1 \\ \dfrac{1}{2x} & \text{falls } 1 \leqq x \leqq a \\ 0 & \text{sonst} \end{cases}$$

a) Zeigen Sie, dass f genau dann eine Dichtefunktion ist, wenn $a = \mathrm{e}$.

b) Skizzieren Sie den Graph von f mit $a = \mathrm{e}$ in nebenstehendes Koordinatensystem.

c) Bestimmen Sie die Verteilungsfunktion $F(x)$.



Für die folgenden beiden Teilaufgaben sei eine Zufallsvariable Y stetig gleichverteilt auf dem Intervall [1; 3].

d) Berechnen Sie die Wahrscheinlichkeit $P(|Y - 2| < 0.5)$.

R e) Geben Sie einen R-Befehl an, der die Wahrscheinlichkeit aus Teilaufgabe d) ausgibt. Benutzen Sie dazu die Verteilungsfunktion der Gleichverteilung.

Lösungshinweis:

a) $\displaystyle\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 0.5 + \int_1^a \frac{1}{2x}\mathrm{d}x = 0.5 + \frac{1}{2}\Big[\ln(x)\Big]_1^a = 0.5 + 0.5 \cdot \ln(a) = 1$
$\Leftrightarrow \ln(a) = 1 \Leftrightarrow a = \mathrm{e}$

b) s.o.

c) $$F(x) = \begin{cases} 0 & \text{falls } x < 0 \\ 0.5x & \text{falls } 0 \leqq x < 1 \\ 0.5 + 0.5 \cdot \ln(x) & \text{falls } 1 \leqq x \leqq \mathrm{e} \\ 1 & \text{falls } x > \mathrm{e} \end{cases}$$

d) $P = 0.5$

e)
```
punif(2.5, min=1, max=3) - punif(1.5, min=1, max=3)
## [1] 0.5
```

Aufgabe 6 — 9 Punkte

Ein Sprachprofessor Ihrer Hochschule interessiert sich dafür, ob in einem seiner Kurse die Präsenzquote der Studenten und Studentinnen mit der in diesem Fach erreichten Noten zusammenhängt. Dazu erfasst er von 100 Teilnehmern jeweils neben dem Anteil der besuchten Veranstaltungen (in Prozent der Termine) auch die in der Abschlussprüfung erreichte Note. Er fasst die erhobenen Daten in folgender Kontingenztabelle zusammen:

	Anwesenheit (in Prozent der Veranstaltungen)			
Note	[0; 50)	[50; 80)	[80; 95)	[95; 100]
1	0	1	4	2
2	2	5	11	5
3	3	5	19	1
4	0	4	3	0
5	16	12	4	3

Die erhobenen Daten sollen im Folgenden als einfache Stichprobe der Grundgesamtheit aller Studenten des betreffenden Studiengangs angesehen werden.

Für die Tabelle der bei Unabhängigkeit erwarteten Häufigkeiten ergibt sich

	[0; 50)	[50; 80)	[80; 95)	[95; 100]
1	1.47	1.89	2.87	0.77
2	4.83	6.21	9.43	2.53
3	5.88	7.56	11.48	3.08
4	1.47	1.89	2.87	0.77
5	7.35	9.45	14.35	3.85

Für die Anteile an χ^2 erhält man damit

	[0; 50)	[50; 80)	[80; 95)	[95; 100]
1	1.47	0.419	0.445	1.965
2	1.658	0.236	0.261	2.411
3	1.411	0.867	4.926	1.405
4	1.47	2.356	0.006	0.77
5	10.18	0.688	7.465	0.188

a) Ergänzen Sie die fehlenden Einträge in den Tabellen.

b) Für den Testwert ergibt sich $v \approx 40.6$ (muss nicht nachgerechnet werden).
Testen Sie zum Signifikanzniveau $\alpha = 0.05$, ob die beiden Merkmale in der Grundgesamtheit unabhängig sind.

c) Kann man nur durch das Ergebnis des Tests sagen, dass viel Anwesenheit in diesem Kurs tendentiell eine gute Note verspricht?

Lösungshinweis:

a) s.o.

b) Daraus ergibt sich: Testwert $v \approx 40.6$
χ^2-Verteilung mit 12 Freiheitsgraden: $B = (21.03; \infty)$
Also: Nullhypothese ablehnen, Einkommen und Geschlecht sind abhängig.

c) Nein, denn wir wissen aus dem Testergebnis nur, dass es wahrscheinlich eine Abhängigkeit gibt, aber nicht in welcher Richtung (Könnte z.B. sein, dass unerwartet viele Leute bei hoher Anwesenheit sehr gut oder sehr schlecht in der Prüfung abschneiden)