

Klausur Statistik

Lösungshinweise

Prüfungsdatum: 13. Januar 2017 – Prüfer: Etschberger, Jansen, Ivanov, Wins
 Studiengang: IM, BW, Inf und W-Inf
 Punkte: 21, 18, 12, 12, 11, 16 ; Summe der Punkte: 90

Aufgabe 1

21 Punkte

Für ein metrisches Merkmal X wurden 40 Beobachtungen mit ihren Ausprägungen a_i sowie den zugehörigen absoluten Häufigkeiten h_i erfasst:

i	1	2	3	4	5	6	7
a_i	4	6	8	9	11	12	14
h_i	8	4	8	5	6	4	5

- Berechnen Sie die relativen Häufigkeiten.
- Berechnen Sie den Modus, den Median, das arithmetische Mittel sowie die mittlere quadratische Abweichung des Merkmals X .
- Bestimmen Sie zur empirischen Verteilungsfunktion F von X den Wert $F(13)$.
- Berechnen Sie den Anteil der Beobachtungen von X , bei denen eine Ausprägung von mindestens 7, aber höchstens 12 vorliegt.
- Geben Sie die Koordinaten der ersten beiden Knickpunkte der Lorenzkurve zu X an.
Hinweis: Die Lorenzkurve soll nicht gezeichnet werden.

Nehmen Sie für die Teilaufgaben f) bis h) an, dass die Urliste für Merkmal X in einem R-Vektor `data` gespeichert ist. Geben Sie das (die) R-Kommando(s) an, mit dem (denen) Sie

- die Werte in Teilaufgaben d) berechnen,
- das empirische 20%-Quantil berechnen,
- die empirische Verteilungsfunktion graphisch darstellen.

Lösungshinweis:

```
a) n = sum(h)
    table(x)/n

## x
##  4  6  8  9 11 12 14
## 0.200 0.100 0.200 0.125 0.150 0.100 0.125
```

```
b) ##      Parameter  Wert
## 1      Modus      4
## 2      Median     8.5
## 3  arithm.Mittel 8.725
## 4 m.q.Abweichung 3.263
```

```
c) ecdf(x)(13)
## [1] 0.875
```

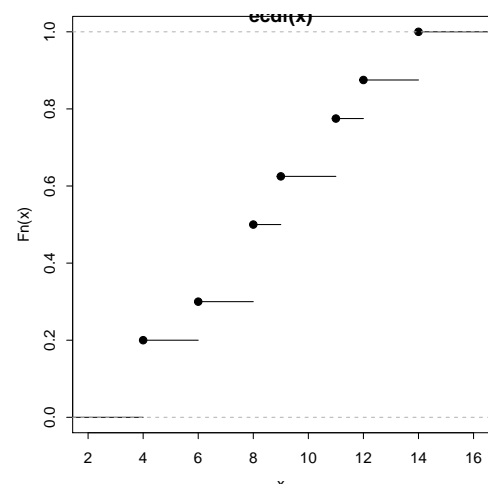
```
d) ecdf(x)(12) - ecdf(x)(6.9)
## [1] 0.575
```

```
e) ##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## uk      0 0.200 0.30 0.500 0.625 0.775 0.875 1
## vk      0 0.092 0.16 0.344 0.473 0.662 0.799 1
```

f) s.o.

```
g) quantile(x, probs = 0.2)
## 20%
## 5.6
```

```
h) plot(ecdf(x))
```



Aufgabe 2

18 Punkte

Der Läufer Stefan Jumper ist in die Jahre gekommen. Auch im Alter von 40 bis 55 nimmt er immer wieder an Läufen über 10 Kilometer teil. In der unten aufgeführten Tabelle sind neben seinem Alter die Durchschnittsgeschwindigkeit (in Kilometern pro Stunde) des besten Laufs im jeweiligen Jahr aufgeführt.

Alter	40	42	43	45	46	48	51	52	54	55
Geschwindigkeit	16.25	16.1	16.03	15.96	16.01	15.88	15.66	15.42	15.33	15.41

- Berechnen Sie eine geeignete Kenngröße für die Korrelation der Geschwindigkeit mit Stefans Alter.
 - Bestimmen Sie die Gleichung des linearen Regressionsmodells mit dem die Geschwindigkeit in Abhängigkeit des Alters geschätzt werden kann.
 - Prognostizieren Sie mit Hilfe des Modells die Geschwindigkeit im 65. Lebensjahr sowie die Geschwindigkeit, die Stefan zu seinen besten Zeiten (mit 30 Jahren) gelaufen ist.
- R** d) Geben Sie R-Befehle an, mit denen man das Ergebnis aus Teilaufgabe a) bestimmen kann.
Hinweise: die Eingabe der Daten kann sinnvoll verkürzt werden.
- R** e) Mit welchen R-Befehlen kann man die obige Regressionsgerade, sowie das zugehörige Streudiagramm der Daten graphisch darstellen?

Lösungshinweis:

- a) Berechnung des Bravais-Pearson Korrelationskoeffizient:

```
Alter = c(40, 42, 43, 45, 46, 48, 51, 52, 54, 55)
Geschwindigkeit = c(16.25, 16.10, 16.03, 15.96, 16.01, 15.88,
                    15.66, 15.42, 15.33, 15.41)
A = data.frame(Alter, Geschwindigkeit)
cor(A)

##           Alter Geschwindigkeit
## Alter      1.0000000      -0.9709773
## Geschwindigkeit -0.9709773      1.0000000
```

- b) Lineare Regression:

```
linReg = lm(Geschwindigkeit ~ Alter)
a = coef(linReg)[1] # Achsenabschnitt
b = coef(linReg)[2] # Steigung
```

Damit ist die Gleichung: $y = -0.0605519 \cdot x + 18.6872727$ und es ergibt sich für $x = 65$

```
a + b * c(65, 30)
## [1] 14.75140 16.87071
# oder
Prognose = data.frame(Alter=c(65, 30))
data.frame(Prognose, Prognose=predict(linReg, newdata=Prognose))

##   Alter Prognose
## 1    65 14.75140
## 2    30 16.87071
```

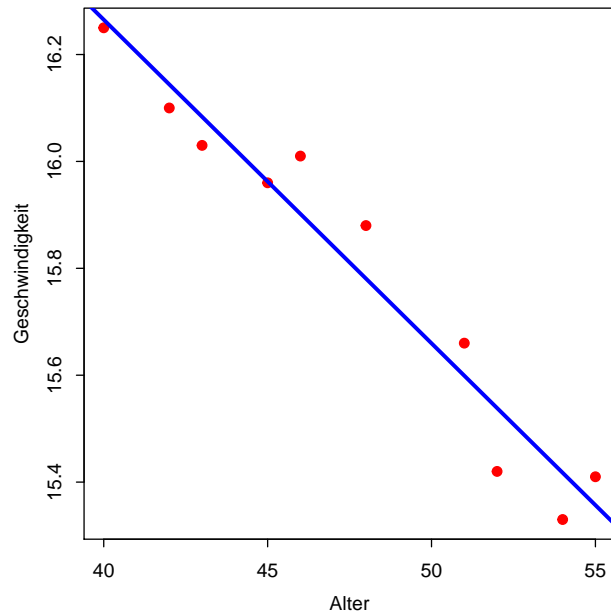
- e) `plot(Alter, Geschwindigkeit, pch=19, col="red")`
`abline(linReg, col = "blue", lwd=3)`

```
## Alter Prognose
## 1 65 14.75140
## 2 30 16.87071
```

c) s.o.

d) s.o.

```
e) plot(Alter, Geschwindigkeit, pch=19, col="red")
abline(linReg, col = "blue", lwd=3)
```



Aufgabe 3

12 Punkte

Der Nikolaus versteckt seit Neuestem in jedem zehnten Geschenk eine Autogrammkarte von sich. Hugo erfährt davon, ist passionierter Nikolausfan und möchte unbedingt eine Autogrammkarte haben.

Hinweis: Gehen Sie davon aus, dass die Wahrscheinlichkeit, eine Autogrammkarte zu enthalten, für jedes Geschenk gleich groß ist.

- Im Kaufhaus werden Geschenke „direkt vom Nikolaus“ verkauft. Hugo kauft sich davon für sich selbst 10, die er zufällig auswählt. Wie groß ist die Wahrscheinlichkeit, dass er mit diesem Einkauf mindestens eine Autogrammkarte erworben hat?
- Wie viele Geschenke müsste Hugo kaufen, wenn die Wahrscheinlichkeit, dass sich darunter mindestens eine Autogrammkarte befindet, mindestens 99 % sein soll?
- Wie viele Autogrammkarten müsste der Nikolaus pro 10 Geschenke verteilen, so dass Hugo, wenn er 10 Geschenke zufällig auswählt mindestens eine Autogrammkarte mit einer Wahrscheinlichkeit von mindestens 99 % erhält (es wird ein ganzzahliges Ergebnis gesucht)?

Lösungshinweis:

a) $P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{10}{0} \cdot 0.1^0 \cdot 0.9^{10} = 1 - 0.9^{10} \approx 0.6513$

b) $1 - 0.9^n \geq 0.99$. Also:

$$\frac{\ln 0.01}{\ln 0.9} \approx 43.709 \text{ also mind. 44 Geschenke müsste er kaufen}$$

- c) Sei k die (durchschn.) Anzahl von Autogrammkarten pro 10 Geschenken, also $X \sim B(n = 10, p = k/10)$. Dann folgt:

$$P(X \geq 1) \geq 0.99 \Leftrightarrow 1 - \left(\frac{k}{10}\right)^{10} \geq 0.99$$

$$\Leftrightarrow 0.01^{0.1} \geq 1 - \frac{k}{10} \Leftrightarrow k \geq 10 \cdot (1 - 0.01^{0.1}) \approx 3.69$$

Also sind mind. 4 Karten pro 10 Geschenke nötig.

Aufgabe 4

12 Punkte

Das kleine Unternehmen Hackintosh betreibt zehn Rechner als Server. In einem Monat fällt durchschnittlich einer der Rechner aus. Die Ausfälle der Rechner passieren unabhängig voneinander. Im Folgenden stehe die Zufallsvariable X für die Anzahl der Rechnerausfälle pro Monat. Gehen Sie davon aus, dass X poissonverteilt ist.

- a) Berechnen Sie für X die Wahrscheinlichkeit für
- (1) keinen Ausfall,
 - (2) einen Ausfall,
 - (3) maximal drei Ausfälle.
- R** b) Geben Sie einen R-Befehl an, mit dem man für X berechnet, wie hoch die Wahrscheinlichkeit für den Ausfall von höchstens drei Rechnern ist.
- c) Um einen reibungsfreien Betrieb zu garantieren müssen mindestens 7 der 10 Rechner immer laufen. Wenn ein Rechner in einem Monat ausfällt, so kann dieser erst wieder zu Beginn des nächsten Monats ersetzt werden. Wie groß ist die Wahrscheinlichkeit, dass innerhalb eines Jahres der reibungsfreie Betrieb aufgrund von Serverausfällen nicht garantiert werden kann?

Lösungshinweis:

- a) $\lambda = 1$ und damit ergibt sich
- (1) $P(x = 0, \lambda = 1) = \frac{1^0}{0!}e^{-1} = 0.3678794$
 - (2) $P(x = 1, \lambda = 1) = \frac{1^1}{1!}e^{-1} = 0.3678794$
 - (3) $P(x \leq 3, \lambda = 1) = \sum_{j=0}^3 \frac{1^j}{j!}e^{-1} = 0.9810118$

b) (1)

```
ppois(q = 3, lambda = 1)
## [1] 0.9810118
```

(2)

```
dbinom(x = 2, size = 10, prob = 0.1)
## [1] 0.1937102
```

- c) Es müssen mindestens vier Rechner in einem Monat ausfallen, damit ist die Wahrscheinlichkeit, dass in einem Monat der Betrieb nicht reibungsfrei garantiert werden kann

$$p = 1 - P(x \leq 3, \lambda = 1) = 0.0189882.$$

Daraus ergibt sich für keinen Ausfall im Jahr (Binomialverteilung):

$$P(X = 0) = \binom{12}{0} p^0 \cdot (1 - p)^{12} \approx 0.7944947$$

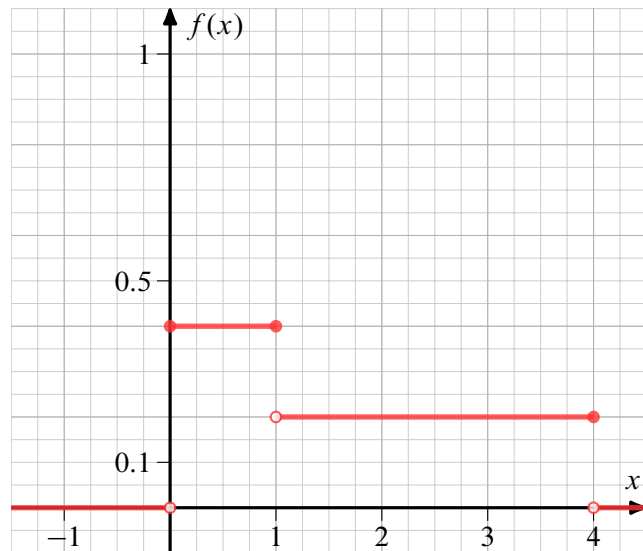
und damit ist $P(X \geq 1) = 0.2055053$.

Aufgabe 5

11 Punkte

Gegeben sei zur Zufallsvariablen X die Dichtefunktion $f : \mathbb{R} \rightarrow \mathbb{R}$ gemäß:

$$f(x) = \begin{cases} 2/5 & \text{falls } 0 \leq x \leq 1 \\ 1/5 & \text{falls } 1 < x \leq 4 \\ 0 & \text{sonst} \end{cases}$$



- Skizzieren Sie den Graph von f in das nebenstehende Koordinatensystem.
- Zeigen Sie, dass f eine Dichtefunktion ist.
- Geben Sie die Wahrscheinlichkeiten $P(X = 1)$, $P(X \leq 1)$, $P(X \geq 3)$, $P(0.5 < X \leq 1.5)$ an.
- Ergänzen Sie die fehlenden drei Terme zur Definition von $F(x)$:

$$F(x) = \begin{cases} 0 & \text{falls } x < 0 \\ 0.4x & \text{falls } 0 \leq x \leq 1 \\ 0.4 + 0.2(x - 1) & \text{falls } 1 < x \leq 4 \\ 1 & \text{falls } x > 4 \end{cases}$$

Im Folgenden sei $Y \sim N(2,2)$ und X sei gleichverteilt mit $a = 0$ und $b = 4$, also $X \sim U(0, 4)$. Bestimmen Sie damit

- R**
- die Wahrscheinlichkeit $P(1 \leq Y \leq 3)$ und geben Sie den zugehörigen R-Befehl an,
 - den Erwartungswert der Zufallsvariablen Z mit $Z = 2Y - X + 3$.

Lösungshinweis:

- s.o.
- $f \geq 0$ und Fläche = $1 \cdot 0.4 + 3 \cdot 0.2 = 1$
- $P(X = 1) = 0$, $P(X \leq 1) = 1 \cdot 0.4 = 0.4$, $P(X \geq 3) = 1 \cdot 0.2 = 0.2$,
 $P(0.5 < X \leq 1.5) = 0.5 \cdot 0.4 + 0.5 \cdot 0.2 = 0.3$
- s.o.
- $P(1 \leq Y \leq 3) = F(3) - F(1) = \Phi\left(\frac{3-2}{\sqrt{2}}\right) - \Phi\left(\frac{1-2}{\sqrt{2}}\right) = \Phi(1/2) - \Phi(-1/2) = 2\Phi(1/2) - 1 = 2 \cdot 0.6914625 - 1 = 0.3829249$

```
pnorm(3, mean = 2, sd = 2) - pnorm(1, mean=2, sd=2)
## [1] 0.3829249
```

- $E[Z] = 2 \cdot E[Y] - E[X] + 3 = 2 \cdot 2 - 2 + 3 = 5$

Eine Hochschule interessiert sich für die Anzahl (Merkmal X) der Bewerbungsgespräche, die nach dem Abschluss geführt werden müssen, bis der erste Job angenommen wird. Dazu werden 20 Absolventen/-innen 5 Jahre nach Ihrem Abschluss befragt. Es ergibt sich die folgende Tabelle des Umfrageergebnisses mit der Anzahl der Bewerbungsgespräche (a_i) und der absoluten Häufigkeit der jeweiligen Antworten (h_i):

i	1	2	3	4	5	6	7	8
a_i	2	4	5	6	7	8	11	16
h_i	2	3	3	4	1	2	3	2

Das Umfrageergebnis sei eine einfache Stichprobe der Grundgesamtheit aller Absolventen/-innen.

- Bestimmen Sie ein Konfidenzintervall für den Mittelwert der Anzahl der Bewerbungsgespräche zu einem Konfidenzniveau von 90 %.
- Wie müsste die Gegenhypothese H_1 lauten, wenn die Hochschulleitung statistisch bestätigen möchte, dass die durchschnittliche Anzahl an Bewerbungsgesprächen kleiner als 10 ist?
- Würden Sie eher ein hohes oder ein niedriges Signifikanzniveau wählen, wenn Sie diese Vermutung statistisch bestätigen wollen?
- Was bedeutet der Fehler 1. Art hier?
- Testen Sie zum Signifikanzniveau von 5 %, ob die durchschnittliche Anzahl an Bewerbungsgesprächen aller Absolventen in der Grundgesamtheit niedriger als 10 ist.

Lösungshinweis:

- $c = x_{0,95} = 1.729$, $\bar{x} = 7.15$, $s = 4.004$
 $\Rightarrow \left[\bar{x} \pm \frac{sc}{\sqrt{n}} \right] = [5.602, 8.698]$
- $H_0 : \mu = 10$ gegen $H_1 : \mu < 10$.
- Je größer α , desto eher wird H_0 abgelehnt, also sollte ein hohes Signifikanzniveau gewählt werden (dafür: höheres Risiko für Fehler 1. Art)
- Fehler 1. Art: Die Stichprobe führt zu einer Ablehnung der Nullhypothese ($\mu = 10$), obwohl H_0 stimmt.
- $\alpha = 0,05$. Verwerfungsbereich: $B = (-\infty; -1.729]$
 $v = \frac{\bar{x} - \mu_0}{s} \sqrt{n} = -3.184 \Rightarrow$ Ablehnen, also $\mu < 10$.