



## Lösungen zu den Aufgaben

### Empirischer Häufigkeitsverteilungen

1

Gib zur Stichprobe des Beispiels 3 den oder die Modalwert(e) an.

Am öftesten wurden im Beobachtungszeitraum 3 Tassen Kaffee pro Tag getrunken. Daher lautet der Modalwert: 3.

2

Gib für die Daten der Tabelle 3.4 das 0.65-Quantil an (berechnet nach dem Näherungsverfahren) und vergleiche dein Ergebnis mit dem exakten Wert.

$$k = 20 \cdot 0.65 = 13$$

Da dies eine ganze Zahl ist, müssen wir das Mittel aus dem 13. und 14. Element ausrechnen:

$$x_{0.65} = \frac{15.5 + 16}{2} = \frac{31.5}{2} = \underline{\underline{15.75}}$$

Der exakte, mit  $R$  berechnete Wert ist: 15.675. Unser genäherter Schätzwert ist also ein wenig größer als der exakte Wert.

3

Gegeben sei die Körpergröße der 7 Zwerge. Bestimme das Minimum, Maximum sowie das 1. - 3. Quartil sowohl näherungsweise als auch deren exakten Werte:

| Name    | Größe (in <i>cm</i> ) |
|---------|-----------------------|
| Doc     | 85                    |
| Grumpy  | 80                    |
| Happy   | 62                    |
| Sleepy  | 81                    |
| Bashful | 70                    |
| Sneezy  | 80                    |
| Dopey   | 88                    |

Wenn wir «von Hand» rechnen, ist es praktisch, die Daten der Größe nach zu ordnen:

62, 70, 80, 80, 81, 85, 88

Damit sind Minimum und Maximum schon klar:  $x_{\min} = 62$ ,  $x_{\max} = 88$ .

Für die Quartile von  $n = 7$  Werten bilden wir:

| $p$  | $n \cdot p$ | aufgerundet |
|------|-------------|-------------|
| 0.25 | 1.75        | 2           |
| 0.5  | 3.5         | 4           |
| 0.75 | 5.25        | 6           |

Und daraus ergeben sich die Quartile:

$x_{0.25} = x_2 = 70$ ,  $x_{0.5} = x_4 = 80$ ,  $x_{0.75} = x_6 = 85$ .

Für die exakte Berechnung verwenden wir die Excel-Funktion `QUARTILE.INKL` und erhalten:

$x_{0.25} = 75$ ,  $x_{0.5} = 80$  und  $x_{0.75} = 83$

#### 4

**Gegeben sei die Größe der 7 roten Zwerge, die innerhalb einer Entfernung von 10 Lichtjahren zur Erde liegen. Bestimme das Minimum, Maximum sowie das 1. - 3. Quartil sowohl «visuell» als auch deren exakten Werte:**

| Bezeichnung      | Durchmesser (in <i>Tausend km</i> ) |
|------------------|-------------------------------------|
| Luyten 726-8 A   | 195                                 |
| Luyten 726-8 B   | 195                                 |
| Proxima Centauri | 196.4                               |
| Wolf 359         | 222.8                               |
| Barnards Stern   | 273                                 |
| Ross 154         | 334.2                               |
| Lalande 21185    | 547.4                               |

Wir ordnen wieder der Größe nach und können gleich die Quartile einzeichnen:

195   195   196.4   222.8   273   334.2   547.4

Die exakten Werte sind:

$$x_{0.25} = 195.7, x_{0.5} = 222.8 \text{ und } x_{0.75} = 303.6$$

5

**Bestimme zu den Daten der Tabelle 3.4 die Spannweite und den Quartilsabstand. Gibt es auf Grund dieser Streuungswerte Anzeichen, dass die Daten Ausreißer enthalten?**

Das Minimum der Daten in Tab.3.4 ist 12, das Maximum 17. Die Spannweite beträgt daher  $17 - 12 = \underline{5 \text{ Jahre}}$ .

Die Quartile kann man exakt oder mit einem Näherungsverfahren berechnen. Bei der exakten Berechnung in R erhalten wir  $IQR = 1.75$ .

Wenden wir das Näherungsverfahren an, dann lauten die entsprechenden Quartile:  $x_{0.25} = 14.25$ ,  $x_{0.75} = 16.25$  und der Quartilsabstand ist 2.

Für den Ausreißertest nach Formel 3.20 und 3.20 gilt dann:  $A_u = 11.25$  und  $A_o = 19.25$ , d.h. dass keines der Elemente als Ausreißer zu betrachten ist.

6

**Wie groß ist die Varianz der in Beispiel 3 gegebenen Daten?**

In der Stichprobe des Bsp.3 sind  $n = 14$  Elemente enthalten. Die Summe ist 42, das Mittel daher  $\bar{x} = \frac{42}{14} = 3$ .

Die jeweiligen Differenzen zwischen gegebenen Elementen und dem Mittel sind:  $-2, 0, -2, 0, -1, -1, 2, 1, 0, -1, 0, 1, 3, 0$

bzw. deren Quadrate:

$4, 0, 4, 0, 1, 1, 4, 1, 0, 1, 0, 1, 9, 0$

Die Quadratsumme ist  $4 + 0 + 4 + 0 + 1 + 1 + 4 + 1 + 0 + 1 + 0 + 1 + 9 + 0 = 26$  und somit:

$$s^2 = \frac{26}{14 - 1} = \frac{26}{13} = \boxed{2}$$

7

Der folgende Datensatz besteht aus elf Elementen und hat einen arithmetischen Mittelwert von 25. Außerdem wissen wir, dass die Daten völlig symmetrisch um den Mittelwert gestreut sind.

$$x_1, 21, 22, 23, 24, 25, 26, 27, 28, 29, x_{11}$$

**Welche Werte musst du für  $x_1$  und  $x_{11}$  einsetzen, damit die Varianz 26 beträgt?**

Nachdem die Daten symmetrisch um den Mittelwert gestreut sind, müssen  $x_1$  und  $x_{11}$  den gleichen Abstand vom Mittelwert haben.

Für  $x_1$  schreiben wir dann  $(25 - a)$  und für  $x_{11} = (25 + a)$  und suchen nach einem passenden Wert  $a$ .

Die Varianz soll 26 betragen, d.h. die Quadratsumme der Abweichungen vom Mittelwert dividiert durch  $(n - 1) = 10$  muss 26 betragen.

Das ergibt folgende Gleichung:

$$26 = \frac{1}{11-1}[(25 - a - 25)^2 + (21 - 25)^2 + (22 - 25)^2 + (23 - 25)^2 + (24 - 25)^2 + (25 - 25)^2 + (26 - 25)^2 + (27 - 25)^2 + (28 - 25)^2 + (29 - 25)^2 + (25 + a - 25)^2]$$

$$260 = [a^2 + 60 + a^2]$$

$$200 = 2a^2$$

$$100 = a^2$$

$$10 = a$$

Somit ist  $x_1 = 15$  und  $x_{11} = 35$ .

8

**Kannst du – ohne konkret jede Kennzahl auszurechnen – «auf einen Blick» sagen und begründen, welche der drei folgenden Datensätze die größte Standardabweichung hat und welche die kleinste?**

A) 0, 20, 40, 50, 60, 80, 100

B) 0, 48, 49, 50, 51, 52, 100

C) 0, 1, 2, 50, 98, 99, 100

Die Standardabweichung ist abhängig von der Quadratsumme der Abstände der einzelnen Elemente vom Mittelwert.

A) B) und C) streuen alle um den Mittelwert 50, wobei C) wegen der jeweils größeren Abstände zwischen den einzelnen Werten und dem Mittelwert 50 auf die größte Quadratsumme kommen wird und B) auf die kleinste.

Daher hat C) die größte Standardabweichung und B) die kleinste.

9

**Gib zu den Daten aus Aufgabe 7 die standardisierten z-Werte an.**

Der Mittelwert der Daten beträgt laut Vorgaben der Angabe  $\bar{x} = 25$  und die Varianz  $s^2 = 26$ . Somit ist die Standardabweichung  $s = 5.099$  und daraus erhalten wir die standardisierten Werte:

|       |       |       |       |       |       |      |      |      |      |      |      |
|-------|-------|-------|-------|-------|-------|------|------|------|------|------|------|
| $x_i$ | 15    | 21    | 22    | 23    | 24    | 25   | 26   | 27   | 28   | 29   | 35   |
| $z_i$ | -1.96 | -0.78 | -0.59 | -0.39 | -0.20 | 0.00 | 0.20 | 0.39 | 0.59 | 0.78 | 1.96 |

## Merkmalszusammenhänge

10

In der folgenden Tabelle sind für sieben in Wien abgehaltenen Wahlen die Mittagstemperatur am jeweiligen Wahltag ( $x$ ) sowie das Verhältnis der abgegebenen Stimmen zur Anzahl der Wahlberechtigten, also die Wahlbeteiligung ( $y$ ) gegeben. Gib den Regressionskoeffizienten an.

|                       | 28.09.08 | 07.06.09 | 10.10.10 | 29.09.13 | 25.05.14 | 11.10.15 | 24.04.16 |
|-----------------------|----------|----------|----------|----------|----------|----------|----------|
| $x$ (Temperatur °C)   | 15       | 22       | 12       | 12       | 24       | 7        | 7        |
| $y$ (Wahlbeteiligung) | 0.74     | 0.43     | 0.68     | 0.70     | 0.35     | 0.75     | 0.64     |

Der Regressionskoeffizient ist der Anstieg der Regressionsgeraden. Den können wir z.B. mit der Excel-Funktion `STEIGUNG` berechnen und erhalten als Ergebnis:

$$\underline{\underline{k = -0.020}}$$

11

(Fortsetzung zu Aufgabe 10): In obiger Tabelle ist die Bundespräsidentenwahl 2010 nicht enthalten. Die Mittagstemperatur am Wahltag (25.4.2010) betrug 18°. Welche Wahlbeteiligung war bei dieser Temperatur zu erwarten?

Dafür benötigen wir die Parameter der Regressionsgeraden, also das  $k$  und das  $d$ , wobei wir den Anstieg  $k$  ja gerade im vorigen Beispiel (Aufgabe 10) ausgerechnet haben. Das  $d$  erhalten wir mit der Excel-Funktion `ACHSENABSCHNITT` und es ergibt sich insgesamt:

$$k = -0.020, d = 0.895$$

Der (statistische) Zusammenhang zwischen Temperatur ( $x$ ) und Wahlbeteiligung ( $y$ ) ist also gegeben durch

$$y = -0.020x + 0.895$$

d.h. bei 18°C wäre nach diesem Regressionsmodell eine Wahlbeteiligung von 54% zu erwarten gewesen. (Tatsächlich lag die Wahlbeteiligung bei der Bundespräsidentenwahl 2010 in Wien bei 52%).

12

In welchem mathematischen Zusammenhang stehen der Korrelationskoeffizient und der Regressionskoeffizient?

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad \text{und} \quad k = \frac{s_{xy}}{s_x^2}$$

und somit:

$$k \cdot s_x^2 = r_{xy} \cdot s_x s_y$$
$$k = \frac{s_y}{s_x} \cdot r_{xy}$$

13

**Besteht zwischen der in Aufgabe 10 gegebenen Wahlbeteiligung der Wienerinnen und Wiener und der am Wahltag vorherrschenden Temperatur eine hohe Korrelation? Wie groß ist der Korrelationskoeffizient?** Für den Korrelationskoeffizienten verwenden wir die Excel-Funktion `KORREL` und erhalten:

$$\underline{\underline{r = -0.85}}$$

Dieser Wert deutet auf eine hohe negative Korrelation hin, und somit auf einen starken negativen linearen Zusammenhang.

14

**Welche Werte kann ein Rangkorrelationskoeffizient annehmen?**

Der Rangkorrelationskoeffizient kann (so wie der «normale» Korrelationskoeffizient) einen beliebigen Wert von  $-1$  bis  $+1$  annehmen.

15

In einem bestimmten Jahrgang wurden bei einer Analyse der Test- und Prüfungsergebnisse aus MAT101 (Mathematik) und MAT102 (Statistik) u.a. folgende Zusammenhänge beobachtet:

Korrelation zwischen den Gesamtpunkten aus MT122 und den im Vorsemester erreichten Punkten aus MAT101:  $r = 0.37$  (Determinationskoeffizient:  $r^2 = 14\%$ ).

Korrelation zwischen den aus den Online-Tests in MT122 erreichten Punkten und der Zeit, die im Durchschnitt für die Bearbeitung der Online-Tests aufgewandt wurde:  $r = -0.08$  (Determinationskoeffizient:  $r^2 = 1\%$ ).

**Was lässt sich daraus über den Zusammenhang zwischen Mathematik- und Statistik-Kenntnissen bzw. den Zeitaufwand, den Studierende für die Online-Tests aufwenden, sagen?**

Antwort:

Ein Korrelationskoeffizient von  $r = 0.37$  bedeutet einen geringen bis mittleren positiven Zusammenhang zwischen den beiden Zufallsgrößen. D.h. tendenziell

werden zwar Studierende, die viele Punkte auf die Mathematikprüfung erhalten haben, auch in Statistik eher gut abschneiden, dieser Zusammenhang ist aber nicht besonders groß, sodass sich nicht von vornherein eine zuverlässige Prognose über das Abschneiden der einzelnen Studierenden in Statistik erstellen lässt, nur weil man ihre Mathematik-Punkte kennt.

Noch weniger ist ein solcher Zusammenhang aus der Zeit abzulesen, die die Studierenden für die Bearbeitung der Online-Tests aufbringen. Ein kurzer Zeitaufwand heißt nicht, dass die oder der Studierende so gut ist, dass sie oder er auch eine gute Note erhalten wird. Und umgekehrt: Länger am Test zu sitzen, garantiert auch keine gute Note.



## Wahrscheinlichkeitsverteilungen

16

In Oberösterreich gibt es insgesamt 6 630 Orte. 40 davon heißen «Au». Angenommen alle 6 630 Orte werden auf Kärtchen geschrieben und daraus eine beliebige Karte herausgezogen. Wie groß ist die Wahrscheinlichkeit, dass darauf *Au* steht?

$$P = \frac{40}{6\,630} = 0.006 = 0.6\%$$

17

Unter der Annahme, dass beim Würfeln jede Augenzahl gleich wahrscheinlich ist: Wie groß ist die Wahrscheinlichkeit, eine gerade Zahl zu würfeln?

Es gibt 6 verschiedene Ausgangsmöglichkeiten (nämlich die Augenzahlen 1 bis 6), und drei Eintrittsfälle (nämlich die drei geraden Zahlen 2, 4 und 6). Somit:

$$P(E) = \frac{\text{Eintrittsfälle}}{\text{Ausgangsmöglichkeiten}} = \frac{3}{6} = 0.5 = 50\%$$

18

Beim American Roulette gibt es je 18 rote und schwarze Nummernfelder sowie zwei grüne Felder mit einer Null. Gib an:

**Wie groß ist die Wahrscheinlichkeit auf eine «rote» Zahl?**

Es gibt insgesamt  $18 + 18 + 2 = 38$  Felder, davon sind 18 rot, daher ist die Wahrscheinlichkeit für eine rote Zahl  $P = \frac{18}{38} = \boxed{47.4\%}$ .

**Wie groß ist die Wahrscheinlichkeit auf eine Primzahl?**

Unter den Zahlen von 0 bis 36 sind 11 Primzahlen (nämlich 2, 3, 5, 7, 11, 13, 17, 19, 23, 29 und 31, daher ist  $P = \frac{11}{38} = \boxed{28.9\%}$ .

**Wie groß ist die Wahrscheinlichkeit auf eine rote Primzahl?**

Es gibt 4 rote Primzahlen (nämlich 3, 5, 7, 23) und wir erhalten:  $P = \frac{4}{38} = \boxed{10.5\%}$ .

19

Ein Statistiktest besteht aus 6 Single-Choice Fragen mit jeweils 4 Antwortmöglichkeiten. (Single-Choice = genau eine Antwort aus den 4 möglichen ist

richtig). Für jemanden, der sich nicht auf den Test vorbereitet hat und nach Belieben zufällige Antworten ankreuzt, beträgt die Erfolgswahrscheinlichkeit pro Frage  $p = 25\%$ . Wie groß ist die Wahrscheinlichkeit, dass diese Person den Test positiv besteht, d.h. mindestens 3 Fragen richtig beantwortet?

«Mindestens drei Fragen richtig» bedeutet: Es können drei, vier, fünf oder sechs Fragen richtig sein.

Mit Formel 5.13 können wir zunächst die Gegenwahrscheinlichkeit  $P(x \leq 2)$  ausrechnen, also die Wahrscheinlichkeit keine, eine oder zwei Fragen richtig zu haben. Sie beträgt 83% (aus EXCEL: `BINOM.VERT(2; 6; 0,25; WAHR)`) und somit:

$$1 - 0.83 = 0.17$$

d.h. die Wahrscheinlichkeit, den Test positiv zu bestehen, beträgt 17%

20

**Was ist der wahrscheinlichste Wert der in Abb. 5.5 dargestellten Binomialverteilung?**

Der Modalwert bzw. wahrscheinlichste Wert einer Verteilung ist derjenige, der am häufigsten vorkommt. In der in Abb. 5.5 dargestellten Binomialverteilung ist das der Wert 50. Das sieht man auch in Tab. 5.4, wo  $x = 50$  mit  $P(X = 50) \approx 8\%$  den höchsten Wahrscheinlichkeitswert hat.

21

**Hat eine Gleichverteilung auch einen Modalwert?**

Nein. Kennzeichen einer Gleichverteilung ist, dass allen Werten die gleiche Wahrscheinlichkeit zugeordnet wird. Es gibt also keinen, der «am wahrscheinlichsten» ist.

22

**Am Bahnhof Meidling wird folgende Information durchgesagt: «Der Zug Richtung Wiener Neustadt hat zwischen 5 und 15 Minuten Verspätung». Angenommen es gibt keinen Grund zur Annahme, dass die Verspätung eher bei 5 oder 15 Minuten liegt, sondern der Zug tatsächlich «irgendwann» in diesem Intervall eintreffen wird: Wie groß ist die Wahrscheinlichkeit, dass die Verspätung maximal 7 Minuten ausmacht?**

Ausgehend von einer Gleichverteilung mit  $a = 5$  und  $b = 15$  suchen wir also  $P(X \leq 7)$ , was nach Formel 5.20 den Wert

$$P(X \leq 7) = \frac{7 - 5}{15 - 5} = \frac{2}{10} = 0.2$$

ergibt. Die gesuchte Wahrscheinlichkeit beträgt also 20%.

## 23

Angenommen im Studiengang WIBA beträgt das (langjährig beobachtete) Durchschnittsalter der Studierenden  $\bar{x} = 37$  Jahre mit einer Standardabweichung von  $s = 5$  Jahren.

**Wie groß ist die Wahrscheinlichkeit, dass unter der Annahme einer Normalverteilung ein:e (beliebig ausgewählte:r) Student:in zwischen 32 und 42 Jahre alt ist?**

Für die Berechnung der gesuchten Wahrscheinlichkeit müssen wir entsprechend Formel 5.18) die Differenz bilden:

$$P(32 < x \leq 42) = F(42) - F(32)$$

und zwar für eine Normalverteilung mit dem Erwartungswert 37 und einer Standardabweichung 5.

Eingesetzt in EXCEL erhalten wir:

$$0.8413 - 0.1587 = 0.6827$$

Die Wahrscheinlichkeit, dass ein:e WIBA-Studierende:r zwischen 32 und 42 Jahre alt ist, beträgt also 68.3%.

Das stimmt im Übrigen auch mit der Aussage von Seite 121 überein, dass ca. 68% aller Realisierungen einer normalverteilten Zufallsgröße im Intervall  $[\mu \pm 1 \cdot \sigma]$  liegen.

## Konfidenzintervalle

24

Betrachte aus Tabelle 4.1 ausschließlich die Spalte «Gewicht» und nimm an, dass diese 20 zufällig ausgewählten Personen eine Stichprobe der Grundgesamtheit «Alle WIBA-Studierenden der FERNFH» sind.

**Gib ein Intervall an, das mit 95%-iger Wahrscheinlichkeit den Erwartungswert für das durchschnittliche Gewicht der WIBA-Studierenden beinhaltet.**

Wir können angeben:

$$\bar{x} = 77.75 \text{ kg} \quad s = 10.36 \text{ kg}$$

und für eine Irrtumswahrscheinlichkeit  $\alpha = 0.05$  das Quantil der  $t$ -Verteilung (für  $n = 20$ ) mit  $t = 2.093$ .

Die Grenzen des Konfidenzintervalls sind somit:

$$L = 77.75 - 2.093 \cdot \frac{10.36}{\sqrt{20}} = 72.9 \text{ kg}$$

$$U = 77.75 + 2.093 \cdot \frac{10.36}{\sqrt{20}} = 82.6 \text{ kg}$$

25

Gegeben sind für die Jahre 2005 bis 2014 die Anzahl polizeilich angezeigter Fälle in Österreich:

| <i>Jahr</i> | <i>Anzeigen</i> | <i>Jahr</i> | <i>Anzeigen</i> |
|-------------|-----------------|-------------|-----------------|
| 2005        | 604 229         | 2010        | 534 351         |
| 2006        | 588 229         | 2011        | 539 970         |
| 2007        | 592 636         | 2012        | 547 764         |
| 2008        | 570 952         | 2013        | 546 396         |
| 2009        | 589 961         | 2014        | 527 692         |

**In welchen Jahren lag die Anzahl der Anzeigen außerhalb (über oder unter) des 95%-Konfidenzintervalls (bezogen auf den Durchschnitt der Jahre 2005-2014)?**

Der Durchschnitt der Jahre 2005 bis 2014 beträgt 564 218, die Standardabweichung 28 078.36.

Für die Berechnung des Konfidenzintervalls können wir auch die Excel Funktion =KONFIDENZ.T(0,05;28078,36;10) verwenden. Wir erhalten den Wert 20 086.05, den wir nun einmal vom Mittelwert abziehen und einmal dazuzählen:

$$L = 564\,218 - 20\,086.05 = 544\,131.95$$

$$U = 564\,218 + 20\,086.05 = 584\,304.05$$

Somit lagen in den Jahren 2005, 2006, 2007 und 2009 die Anzahl der Anzeigen oberhalb der oberen und in den Jahren 2010, 2011 und 2014 unterhalb der unteren Grenze des Konfidenzintervalls.

## 26

Wir wollen mittels Umfrage herausfinden, mit wie viel Stunden Schlaf unsere Studierenden während des Semesters pro Nacht auskommen (müssen). Konkret wollen wir letztlich durch ein 90%-Konfidenzintervall den Erwartungswert der Schlafdauer auf eine Viertelstunde genau schätzen können. Methodisch bedienen wir uns dabei einer einfachen WhatsApp-Umfrage, d.h. wir ersuchen die Studierenden, uns ein WhatsApp mit den «Schlafstunden» der letzten Nacht zu schicken. Wir erwarten eine maximale Rücklaufquote von 20%, d.h. nur jeder fünfte Studierende, die wir einladen mitzutun, wird uns eine WhatsApp-Nachricht zurückschicken.

**Wie viele Personen sollten wir einladen, an der Untersuchung teilzunehmen?**

Zunächst müssen wir schätzen, was denn vermutlich die maximale und minimale Schlafdauer sein wird, die uns genannt wird. Anhaltspunkt könnte die Einschätzung unserer eigenen Zeiten sein. Gehen wir einmal davon aus, dass wir in der Regel mindestens 5 und höchstens 11 Stunden pro Nacht schlafen, was einer Spannweite von 6 Stunden entspricht. Eingesetzt in Formel 6.8 ergibt das eine geschätzte Standardabweichung von  $\hat{s} = 1\text{ h}$ .

Für ein Konfidenzniveau von 90% ist  $z = 1.64$ .

Eine Genauigkeit von einer Viertelstunde ergibt eine Intervalllänge von einer halben Stunde (= eine Viertelstunde auf oder ab).

Und somit:

$$n > \left( \frac{2 \cdot 1.64}{0.5} \cdot 1 \right)^2 = 43.3$$

Im Grunde benötigen wir daher eine Stichprobe von mindestens 44 Antworten. Allerdings beträgt die zu erwartende Rücklaufquote nur 20%, d.h. wir müssen 44 noch mit 100 multiplizieren und durch 20 dividieren und erhalten somit die geforderte Anzahl von 220 Personen, die wir befragen sollten.

## 27

**Welcher Anteil  $\hat{p}$  aus einer Stichprobe mit  $n = 183$  Personen müsste sich für**

einen Gesetzesvorschlag aussprechen, damit wir mit 95%iger Sicherheit darauf schließen können, dass eine (einfache) Mehrheit der Grundgesamtheit für den Beschluss dieses Gesetzes ist?

Die Fragestellung könnte auch so formuliert werden: Wie groß muss  $\hat{p}$  sein, damit die untere Grenze des Konfidenzintervalls für  $p$  mindestens 50 Prozent beträgt, also  $L > 50\%$ ? Dazu müssen wir die Ungleichung

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{183}} > 0.5$$

nach  $\hat{p}$  auflösen:

$$\begin{aligned}\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{183}} &> 0.5 \\ \hat{p} - 0.5 &> 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{183}} \\ (\hat{p} - 0.5)^2 &> 1.96^2 \left( \frac{\hat{p}(1-\hat{p})}{183} \right) \\ \hat{p}^2 - \hat{p} + 0.25 &> 3.84 \frac{\hat{p} - \hat{p}^2}{183} \\ 183\hat{p}^2 - 183\hat{p} + 45.75 &> 3.84\hat{p} - 3.84\hat{p}^2 \\ (183 + 3.84)\hat{p}^2 - (183 + 3.84)\hat{p} + 45.75 &> 0 \\ 186.84\hat{p}^2 - 186.84\hat{p} + 45.75 &> 0\end{aligned}$$

Um eine Ungleichung zu lösen, müssen wir zunächst die zugehörige Gleichung auflösen:

$$\begin{aligned}186.84\hat{p}^2 - 186.84\hat{p} + 45.75 &= 0 \\ 1\hat{p}_2 &= \frac{186.84 \pm \sqrt{186.84^2 - 4 \cdot 186.84 \cdot 45.75}}{2 \cdot 186.84} \\ \hat{p}_1 &= 0.572 \quad \hat{p}_2 = 0.428\end{aligned}$$

und dann in die Ungleichung einsetzen:

$$(\hat{p} - 0.572)(\hat{p} - 0.428) > 0$$

Ein Produkt ist bekanntlich dann positiv, wenn beide Faktoren positiv oder beide Faktoren negativ sind. Dass beide negativ sind, kann aber ausgeschlossen werden. (Warum? Weil das gesuchte  $\hat{p}$  sicher größer als 0.5 ist und somit  $(\hat{p} - 0.428)$  jedenfalls positiv ist). Es verbleibt also:

$$\begin{aligned}(\hat{p} - 0.572 > 0) \wedge (\hat{p} - 0.428 > 0) \\ (\hat{p} > 0.572) \wedge (\hat{p} > 0.428)\end{aligned}$$

und somit erhalten wir das Ergebnis:  $\hat{p} > 0.572$ , d.h. bei einer Stichprobe mit 183 Elementen bedarf es eines empirischen Anteilswerters von mindestens 57.2%, damit wir uns 95%ig sicher sein können, dass die dahinterliegende Grundgesamtheit einen theoretischen Anteilswert von mindestens 50% hat.

## Statistische Tests

28

Deine Ärztin empfiehlt dir, aus gesundheitlichen Gründen deinen Kaffeekonsum auf fünf Tassen pro Tag einzuschränken. Du bist dir eigentlich sicher, dass du das im Schnitt ohnehin nicht überschreitest ( $H_0 : \mu \leq 5$ ), dein besorgter Ehepartner schenkt dem aber keinen so rechten Glauben und behauptet, dass es mehr als fünf Tassen pro Tag sind ( $H_A : \mu > 5$ ). Ihr vereinbart, eine Zeitlang darüber Buch zu führen. Nach 40 Tagen hast du 210 Tassen Kaffee getrunken. Im Schnitt ergab das Experiment also 5,25 Tassen pro Tag, und das bei einer Standardabweichung von 0,25. **Wessen Hypothese kann bei diesem Ergebnis aufrecht erhalten werden?**

Die Testfunktion ergibt:

$$f = \frac{5.25 - 5}{0.25} \sqrt{40} = 6.32$$

und das Quantil der Normalverteilung:

$$z_{1-0.05} = 1.64$$

Nachdem  $f > z_{1-\alpha}$  müssen wir die Nullhypothese ablehnen und  $H_A$  akzeptieren – und unseren Kaffeekonsum in Zukunft wohl reduzieren. Er ist signifikant höher als die empfohlene Menge von 5 Tassen.

29

Eine Software-Entwicklerin überlegt, ihre Software als *Donationware* zur Verfügung stellen, d.h. sie kann grundsätzlich kostenlos verwendet werden, sie bittet aber um (freie) Spenden, damit auf Sicht wenigstens die bei ihr entstehenden Drittkosten abgedeckt sind. Sie schätzt: Nur wenn mindestens 25% der User bereit sind, 10 € zu spenden, werde ich kein Geld verlieren.

Sie startet einmal probeweise und stellt nach den ersten 500 Downloads fest: 145 User haben tatsächlich 10 € gespendet. Das sind sogar 29%. **Kann sie (aus statistischer Sicht) optimistisch sein, dass der Anteil der spendenfreudigen User tatsächlich größer als 25% ist?**

Wir wählen die Arbeitshypothese, dass der Anteil  $p$  größer als 25% ist, also:

$$H_A : p > 0.25 \rightarrow H_0 : p \leq 0.25$$

Aus den ersten 500 Downloads erhalten wir einen Schätzwert für  $p$ :

$$\hat{p} = \frac{145}{500} = 0.29$$

und damit die Prüfgröße:

$$f = \frac{0.29 - 0.25}{\sqrt{0.25(1 - 0.25)}} \sqrt{500} = 2.066$$

Der kritische Bereich beginnt bei 1.645 (Excel: =NORM.S.INV(1-0.05)). Mit 2.066 liegen wir da bereits drüber, d.h. wir lehnen  $H_0$  ab und akzeptieren  $H_A$ , also die Hypothese, dass auch in der Grundgesamtheit der Anteil der User, die 10 € spenden, über 25% ist.

### 30

Nach dem ersten Studienjahr wurde für alle Studierenden eines Jahrgangs ein nach ECTS gewichteter Notenschnitt (GPA) berechnet und daraus dann ein Durchschnittswert für jeweils alle männlichen ( $n_m = 84$ ) und alle weiblichen ( $n_w = 36$ ) Studierenden angegeben. Für die männlichen Studierenden beträgt er  $\bar{x}_m = 1.6$ , für die weiblichen  $\bar{x}_w = 1.4$ . Ermittelt wurde auch die zugehörigen empirischen Standardabweichungen:  $s_m = 0.6$ ,  $s_w = 0.4$ .

**Der empirisch ermittelte GPA der Frauen unterscheidet sich offenbar von dem der Männer. Ist dieser Unterschied signifikant bzw. inwiefern lässt sich aus den obigen Daten die Hypothese  $H_A : \mu_m - \mu_w \neq 0$  behaupten?**

Das zu untersuchende Hypothesenpaar lautet:  $H_A : \mu_m \neq \mu_w \rightarrow H_0 : \mu_m = \mu_w$ .

Einsetzen der Werte aus den beiden Stichproben ergibt die Testfunktion:

$$f = \frac{1.6 - 1.4}{\sqrt{0.6^2 \cdot 36 + 0.4^2 \cdot 84}} \sqrt{84 \cdot 36} = 2.14$$

Dieser Wert muss bei einem  $\alpha$  von 0.05 mit dem Wert  $z_{1-\alpha/2} = 1.96$  verglichen werden und ergibt, da  $|f| > z_{(1-\alpha/2)}$ , die Ablehnung der Nullhypothese und Akzeptanz der Arbeitshypothese. D.h. der Unterschied zwischen dem von den weiblichen und den von den männlichen erzielte GPA ist als signifikant einzustufen.