Gruppenarbeit 7 zur sonstigen Beteiligung

Fiete Ostkamp, Kai Meder, Jannik Winther

10/01/2021

Contents

Einlesen des Datensatzes:

```
load(file = "datensonstbet2.RData")
head(daten2)
```

```
##        alter jgst   geschl    schultyp  iq note_m note_p note_d note_e sk_m
## 101 14.91667    8 männlich Hauptschule  63      5      3      3      5    8
## 102 14.66667    8 weiblich Hauptschule  71      4      4      4      4   19
## 103 14.91667    8 weiblich Hauptschule  83      5      5      3      3   14
## 104 14.66667    8 männlich Hauptschule  77      4      4      4      4   37
## 105 13.41667    7 weiblich  Realschule 102      3      3      3      3   30
## 106 12.58333    7 weiblich  Realschule  82      4      3      4      4   26
##     sk_p sk_d sk_e note_mp note_de note_mpde sk_mp sk_de sk_mpde
## 101   30   27   11     4.0       4      4.00    19  19.0   19.00
## 102   21   18   14     4.0       4      4.00    20  16.0   18.00
## 103   12   34   36     5.0       3      4.00    13  35.0   24.00
## 104   31   36   36     4.0       4      4.00    34  36.0   35.00
## 105   42   36   40     3.0       3      3.00    36  38.0   37.00
## 106   28   29   30     3.5       4      3.75    27  29.5   28.25
```

Aufgabe 1

Die folgenden Berechnungen sollen nun für die ersten fünf Werte der Variable `note_p`, d.h der Werte für die ersten 5 Personen, per Hand vorgenommen werden. Beschreiben Sie dabei auch den Rechenweg. Berechnen Sie für diese Variable

a) Den Mittelwert, die Varianz und die Standardabweichung

mit $n = 5, X = \{3, 4, 5, 4, 3\}$:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\overline{x} = \frac{(3 + 4 + 5 + 4 + 3)}{5}$$

$$= \frac{19}{5}$$

$$= 3.8$$

mit $n = 5, X = 3, 4, 5, 4, 3$ und $\overline{x} = 3.8$

$$
\begin{aligned}
s^2 &= \frac{1n-1}{\sum_{i=1}^{n}}(x_i - \overline{x})^2 \\
s^2 &= \frac{(3-3.8)^2 + (4-3.8)^2 + (5-3.8)^2 + (4-3.8)^2 + (3-3.8)^2}{4} \\
&= \frac{2.8}{4} \\
&= 0.7 \\
s &= \sqrt{s^2} \\
&= \sqrt{0.7} \\
&= 0.8366
\end{aligned}
$$

b) Die z-standartisierten Werte (für die ersten 5 Personen)

$z_i = \frac{y_i - \overline{y}}{s}$, mit $\overline{y} = 3.8$ und $s = 0.8366$
$z_1 = \frac{3-3.8}{0.8366} = -0.956, z_2 = \frac{4-3.8}{0.8366} = 0.239, z_3 = \frac{5-3.8}{0.8366} = 1.434, z_4 = \frac{4-3.8}{0.8366} = 0.239, z_5 = \frac{3-3.8}{0.8366} = -0.956$

c) Erklären Sie bitte kurz den wesentlichen Unterschied zwischen den z-standardisierten Wert und dem ursprünglichen Wert (Note) für die erste Person.

Der z-standardisierte Wert wurde in die Standardnormalverteilung überführt. Mit der Transformation in die Standardnormalverteilung werden Aussagen über die Population einer Stichprobe ermöglicht. Der z-Wert drückt aus, wie viele Standardabweichungen der Wert von dem Mittelwert der Population entfernt liegt. Über den ersten Wert $z = -0.956$ lässt sich sagen, dass dieser knapp eine Standardabweichung unter dem Mittelwert der Population liegt und somit zu den unteren 17% der Population gehört ($\phi(-0.956) = 0.1695$).

Aufgabe 2

a) Kovarianz zwischen den Variablen `note_p` und `sk_p`

mit $X_{note\_p} = \{3, 4, 5, 4, 3\}, Y_{sk\_p} = \{30, 21, 12, 31, 42\}$ und $n = 5$

$\overline{x} = 3.8$, siehe Aufgabe 1 a)

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{y} = \frac{(30 + 21 + 12 + 31 + 42)}{5}$$

$$= \frac{136}{5}$$

$$= 27.2$$

$$\begin{aligned} s_{note\_p,sk\_p} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \frac{1}{4}((3 - 3.8)(30 - 27.2) + (4 - 3.8)(21 - 27.2) + (5 - 3.8)(12 - 27.2) \\ &+ (4 - 3.8)(31 - 27.2) + (3 - 3.8)(42 - 27.2)) \\ &= \frac{(-0.8)(2.8) + (0.2)(-6.2) + (1.2)(-15.2) + (0.2)(3.8) + (-0.8)(14.8)}{4} \\ &= \frac{-2.24 + -1.24 + -18.24 + 0.76 + -11.84}{4} \\ &= \frac{-32.8}{4} \\ &= -8.2 \end{aligned}$$

b) Korrelation zwischen den Variablen `note_p` und `sk_p`

mit $s_{note\_p,sk\_p} = -8.2, s_{note\_p} = 0.8366$ und $\overline{y} = 27.2$

$$
\begin{aligned}
s_{sk\_p} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \sqrt{\frac{(30-27.2)^2 + (21-27.2)^2 + (12-27.2)^2 + (31-27.2)^2 + (42-27.2)^2}{4}} \\
&= \sqrt{\frac{2.8^2 + -6.2^2 + -15.2^2 + 3.8^2 + 14.8^2}{4}} \\
&= \sqrt{\frac{7.84 + 38.44 + 231.04 + 14.44 + 219.04}{4}} \\
&= \sqrt{\frac{510.8}{4}} \\
&= \sqrt{127.7} \\
&= 11.3
\end{aligned}
$$

$$
\begin{aligned}
r_{note\_p,sk\_p} &= \frac{s_{note\_p,sk\_p}}{s_{note\_p} s_{sk\_p}} \\
&= \frac{-8.2}{0.8366 * 11.3} \\
&= -0.8673963
\end{aligned}
$$

Aufgabe 3

Berechnen Sie bitte für die Variable `note_p`

a) die Häufigkeitstabelle mit den absoluten Häufigkeiten (`freq`), den relativen Häufigkeiten (`perc`) und der empirischen Verteilungsfunktion (`cumfreq`) anhand der absoluten und relativen Häufigkeiten (`cumperc`)

```
Freq(daten2$note_p)
```

```
##      level  freq   perc  cumfreq  cumperc
## 1  [1,1.5]     2   2.0%        2     2.0%
## 2  (1.5,2]    26  26.0%       28    28.0%
## 3  (2,2.5]     0   0.0%       28    28.0%
## 4  (2.5,3]    47  47.0%       75    75.0%
## 5  (3,3.5]     0   0.0%       75    75.0%
## 6  (3.5,4]    21  21.0%       96    96.0%
## 7  (4,4.5]     0   0.0%       96    96.0%
## 8  (4.5,5]     4   4.0%      100   100.0%
```

b) die Fünf-Punkte-Zusammenfassung

```
quantile(daten2$note_p)
```

```
##   0%  25%  50%  75% 100%
## 1.00 2.00 3.00 3.25 5.00
```

c) den Mittelwert, die Varianz und Standardabweichung

Der Mittelwert:

```
mean(daten2$note_p)
```

```
## [1] 2.99
```

Die Varianz:

```
var(daten2$note_p)
```

```
## [1] 0.7170707
```

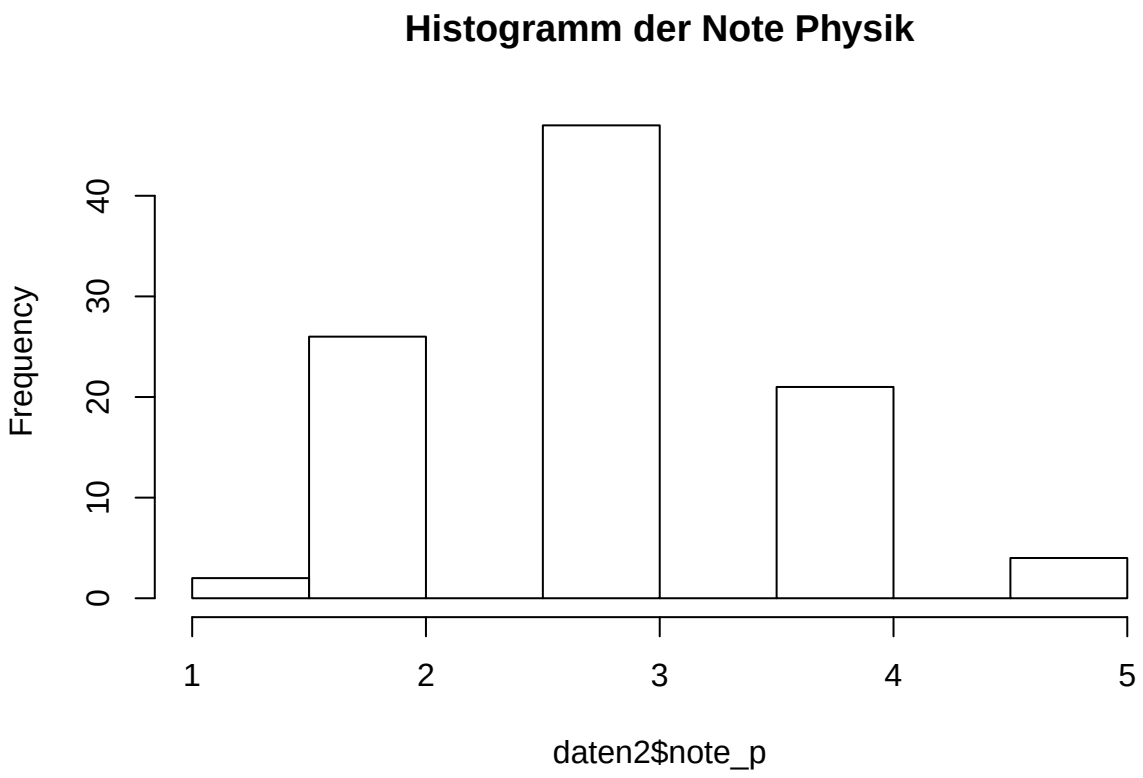
Die Standardabweichung:

```
sd(daten2$note_p)
```

```
## [1] 0.8468003
```

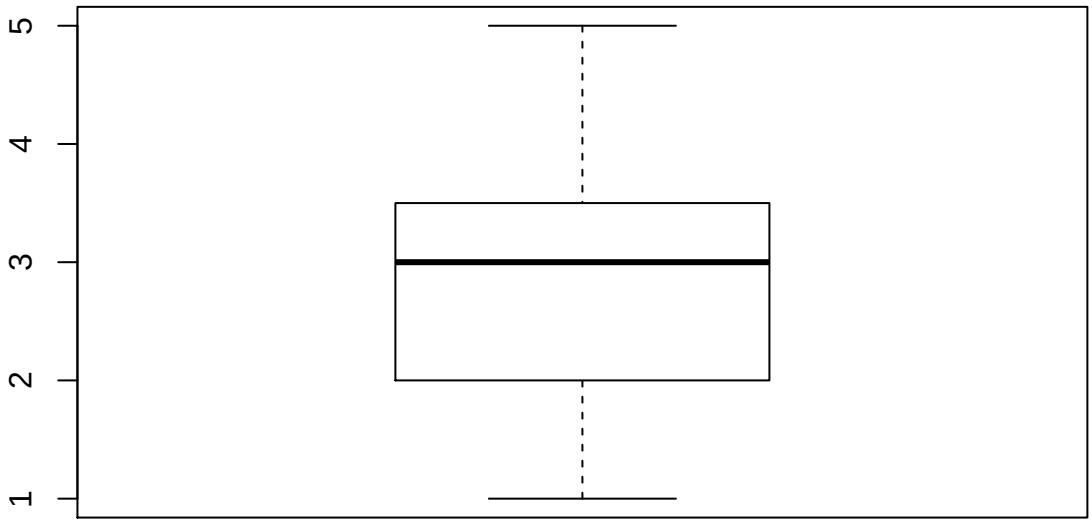
d) das Histogramm

```
hist(x = daten2$note_p, main = "Histogramm der Note Physik")
```



e) den Boxplot

```
boxplot(daten2$note_p)
```



i. Wieviele Ausreißer zeigt der Boxplot an?

Der Plot zeigt keine Ausreißer an.

ii. Wieviele Extremwerte zeigt der Boxplot an?

Der Plot zeigt keine Extremwerte an.

Aufgabe 4

a) den Korrelationskoeffizienten zwischen den Variablen `note_p` und `sk_p`

```
cor(daten2$note_p,daten2$sk_p)
```

```
## [1] -0.5446097
```

b) Wie ist das Ergebnis in inhaltlicher Hinsicht anhand von zwei, drei Sätzen zu interpretieren?

Es besteht eine moderate negative Korrelation zwischen den Variablen `note_p` und `sk_p`. Das heißt in der Regel korrelliert die eigene Einschätzung der Schüler zu ihren Leistungen auch mit der tatsächlichen Note im Fach Physik. Eine höhere Selbsteinschätzung führt zu einer numerisch niedrigeren Note.

c) Stellen Sie zu der inhaltlichen Hypothese, dass eine positive Korrelation zwischen diesen beiden Variablen besteht, die statistische Null- und Alternativhypothese auf.

$h_0 : \rho_{note\_p,sk\_p} = 0$
$h_1 : \rho_{note\_p,sk\_p} > 0$

d) Überprüfen Sie diese statistische Nullhypothese anhand eines Signifikanzniveaus von `=.05`.

```
cortest <- cor.test(x = daten2$note_p,y = daten2$sk_p, alternative = "greater", method =
cortest
```

```
##
##  Pearson's product-moment correlation
##
## data:  daten2$note_p and daten2$sk_p
## t = -6.4283, df = 98, p-value = 1
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  -0.6513819  1.0000000
## sample estimates:
##        cor
## -0.5446097
```

e) Wie groß ist der p-Wert?

Der p-Wert beträgt 1.

```
cortest$p.value
```

```
## [1] 1
```

f) Muss die Nullhypothese bei einem Signifikanzniveau von 5% verworfen werden? Begründen Sie bitte ihre Antwort.

Da der p-Wert $p = 1$ größer als das Signifikanzniveau $\alpha = .05$ ist, ist die Korrelation von -0,545 nicht signifikant. Daher kann die Nullhypothese nicht verworfen werden. Bei einem Korrelationstest auf eine negative Korrelation (`cor.test(daten2$note_p,daten2$sk_p, alternative = "less")`), würde mit einem p-Wert von 2.351e-09 ($< \alpha = .05$) eine signifikante Korrelation vorliegen.

g) Geben Sie bitte das 95%ige Konfidenzintervall an.

Das 95%ige Konfidenzintervall liegt zwischen -0.6513819 und 1.0000000.

```
cortest$conf.int
```

```
## [1] -0.6513819  1.0000000
## attr(,"conf.level")
## [1] 0.95
```

h) Kann man anhand dieses Konfidenzintervalls entscheiden, ob die Nullhypothese bei einem Signifikanzniveaus von `=.05` zu verwerfen ist? Begründen Sie bitte Ihre Meinung.

Das Konfidenzintervall gibt den Bereich an, in dem, unter Annahme von H_0, in 95% der Stichproben der Parameter θ (hier die Korrelation) enthalten ist. Entsprechend kann man einen Annahme- und Ablehnungsbereich für H_0 festlegen:
Annahmebereich: $-0.651 \leqslant \rho \leqslant 1.000$
Ablehnungsbereich: $\rho < -0.651, (\rho > 1)$, wobei $p > 1$ nicht möglich ist.

Aufgabe 5

Überprüfen Sie anhand eines geeigneten Tests, ob sich die Erwartungswerte der Variable `note_p` für Schülerinnen und Schüler signifikant unterscheiden.

a) Stellen Sie zu diesem Testproblem die statistische Null- und Alternativhypothese auf.

$H_0 : \mu_{note\_p_m} = \mu_{note\_p_w}$
$H_1 : \mu_{note\_p_m} \neq \mu_{note\_p_w}$

Kann von gleicher Varianz der `note_p` in beiden Samples ausgegangen werden?
$H_0 : Var(note\_p_m) = Var(note\_p_w)$
$H_1 : Var(note\_p_m) \neq Var(note\_p_w)$

```
LeveneTest(note_p ~ geschl, data = daten2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.1289 0.7203
##       98
```

Ja, da $Pr(> F) = 0.72 > \alpha$, mit $\alpha = 0.05$, wird H_0 nicht verworfen und es kann von einer gleichen Varianz der beiden Samples (m/w) ausgegangen werden.

b) Wie groß ist der p-Wert?

t-Test für homogene Varianzen:

```
t.test(note_p ~ geschl, alternative = "two.sided", conf.level = 0.95, var.equal = TRUE,
```

```
##
##  Two Sample t-test
##
## data:  note_p by geschl
## t = -1.3063, df = 98, p-value = 0.1945
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5554522  0.1144758
## sample estimates:
## mean in group männlich mean in group weiblich
##               2.877551               3.098039
```

Der p-Wert beträgt 0.1945.

c) Muss die Nullhypothese bei einem Signifikanzniveau von 5% verworfen werden? Begründen Sie bitte ihre Antwort.

Da der p-Wert 0.1945 größer ist als das Signifikanzniveau von 0.05, kann die Nullhypothese nicht verworfen werden.

d) Interpretieren Sie das Ergebnis dieses Tests aus inhaltlicher Sicht.

Man kann davon ausgehen, dass sich die Erwartungswerte der `note_p` nicht signifikant unterscheiden je Geschlecht. Im Fach Physik sind Jungen und Mädchen also im Schnitt gleich gut.