Übungszettel: Regression

M.Psy.205, Dozent: Dr. Peter Zezula

Kai Schneider (k.schneider01@stud.uni-goettingen.de)

20 Mai, 2021 09:06

Deutsche Version

Links

Übungszettel als PDF-Datei zum Drucken

Übungszettel mit Lösungen

Lösungszettel als PDF-Datei zum Drucken

Der gesamte Übugszettel als .Rmd-Datei (Zum Downloaden: Rechtsklick > Speichern unter...)

Ziel des Übungszettels

1. Datawrangling wiederholen
2. ggplot wiederholen
3. Regressionsanalyse wiederholen
4. Zusammenhang zwischen t.test für unabhängige Stichproben und der einfachen linearen Regression nominalen binären Prädiktor verstehen.

Hinweise zur Bearbeitung

1. Bitte beantworten Sie die Fragen in einer .Rmd Datei. Sie können Sie über `Datei > Neue Datei > R Markdown...` eine neue R Markdown Datei erstellen. Den Text unter dem *Setup Chunk* (ab Zeile 11) können Sie löschen. Unter diesem Link können Sie auch unsere Vorlage-Datei herunterladen.
2. Informationen, die Sie für die Bearbeitung benötigen, finden Sie auf der Website der Veranstaltung
3. Zögern Sie nicht, im Internet nach Lösungen zu suchen. Das effektive Suchen nach Lösungen für R-Probleme im Internet ist tatsächlich eine sehr nützliche Fähigkeit, auch Profis arbeiten auf diese Weise. Die beste Anlaufstelle dafür ist der R-Bereich der Programmiererplattform Stackoverflow
4. Auf der Website von R Studio finden Sie sehr hilfreiche Übersichtszettel zu vielen verschiedenen R-bezogenen Themen. Ein guter Anfang ist der Base R Cheat Sheet

Ressourcen

Da es sich um eine praktische Übung handelt, können wir Ihnen nicht alle neuen Befehle einzeln vorstellen. Stattdessen finden Sie hier Verweise auf sinnvolle Ressourcen, in denen Sie für die Bearbeitung unserer Aufgaben nachschlagen können.

Ressource	Beschreibung
Field, Kapitel 7 (7.1 - 7.5, 7.9)	Buchkapitel, das Schritt für Schritt erklärt, worum es geht, und wie man Regressionen in R durchführt. **Große Empfehlung!**
R for Data Science	Einsteiger-Buch von R-Gott Hadley Wickham. Hier wird topaktuell in die Arbeit mit R, insbesondere zur Datenaufbereitung und Visualisierung, eingeführt.
R Tutorial	Schritt-für-Schritt Einführung in das Arbeiten mit R von Christian Treffenstädt. Nützlich, falls Sie grundlegende Dinge noch einmal nachschlagen möchten

1) Daten einlesen

1. Setzen Sie ein sinnvolles Arbeitsverzeichnis für den Übungszettel (in der Regel der Ordner, in dem Ihre .Rmd liegt). Aber Vorsicht: Beim Rendern (Knit) geht RStudio davon aus, dass das Working-Directory das ist, in dem auch die .Rmd-Datei liegt. Dies ist besonders wichtig, wenn es um relative Links geht.
2. Laden Sie den Datensatz starwars.csv herunter (*rechtsklick > Ziel speichern unter* oder *rechtsklick > Verknüpfte Datei laden*) und speichern Sie ihn in Ihrem Arbeitsverzeichnis (idealerweise haben Sie noch den Ordner vom letzten Übungszettel - speichern Sie den Datensatz im Unterordner /data).
3. Laden Sie die Pakete des `tidyverse` und fügen Sie eine entsprechende Code-Zeile an den Beginn Ihres Dokuments ein.
4. Lesen Sie den Datensatz `starwars.csv` unter dem Namen `sw_data` in R ein.

Lösung

Unteraufgabe 1 Bitte folgen Sie den Anweisungen im Aufgabentext.

Unteraufgabe 2 Bitte folgen Sie den Anweisungen im Aufgabentext.

```
library(tidyverse)
```

Unteraufgabe 3 Anmerkung: `library()` und `require()` sind beides Befehle zum Laden von Paketen. `require()` ist prinzipiell für die Verwendung innerhalb von Funktionen gedacht. Siehe `?library` oder `?require` für Details.

```
# syntax for locally stored data file
# sw_data <- read_csv("data/starwars.csv")
sw_data <- read_csv("http://md.psych.bio.uni-goettingen.de/mv/data/div/starwars.csv")
```

Unteraufgabe 4

```
##
## -- Column specification --------------------------------------------------------
## cols(
```

```
##   name = col_character(),
##   height = col_double(),
##   mass = col_double(),
##   hair_color = col_character(),
##   skin_color = col_character(),
##   eye_color = col_character(),
##   birth_year = col_double(),
##   gender = col_character(),
##   homeworld = col_character(),
##   species = col_character()
## )
```

2) Data Wrangling

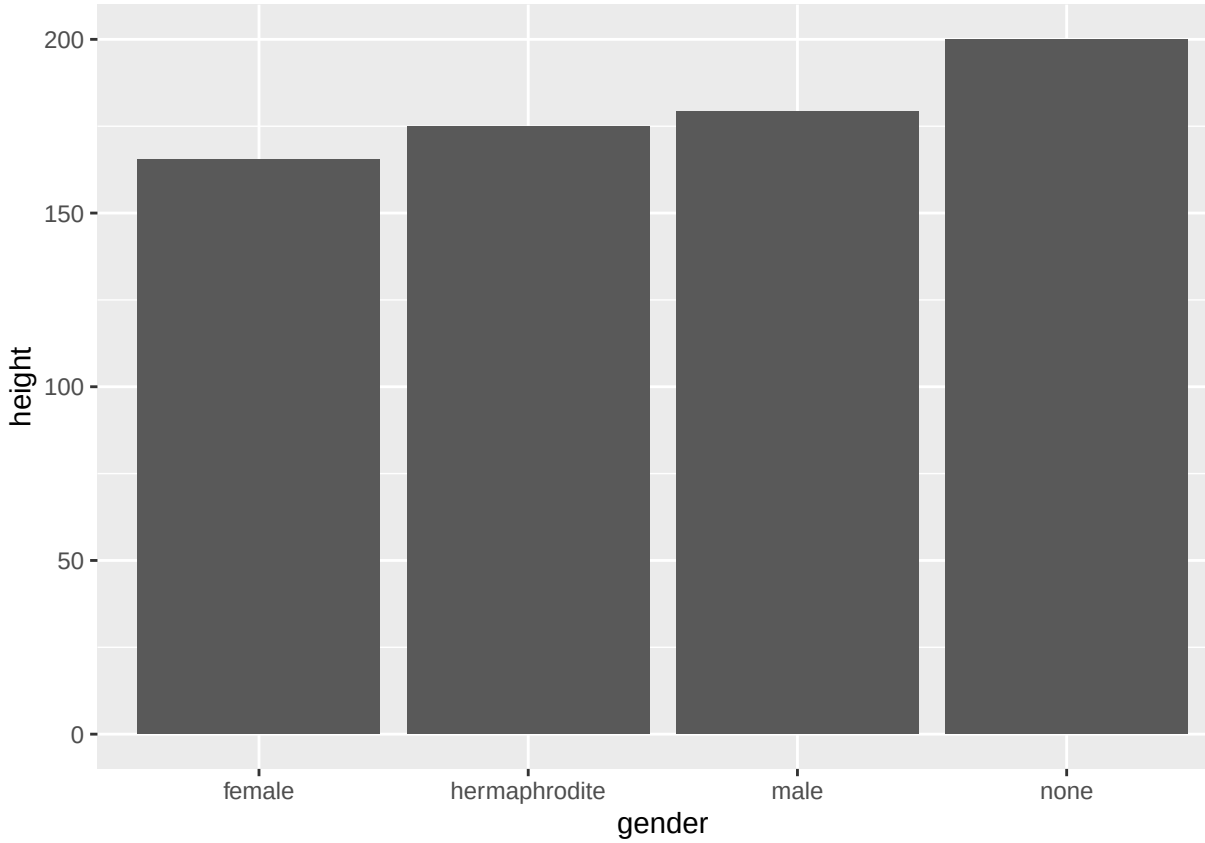
1. Erstellen Sie einen neuen Datensatz `sw_data2` welcher keine fehlenden Werte auf den Variablen `gender` und `height` enthält.
2. Erstellen Sie einen Barplot, welcher auf der x-Achse das Geschlecht und auf der y-Achse die mittlere Größe zeigt.
3. Ergänzen sie Fehlerbalken, färben Sie die Balken unterschiedlich ein und geben Sie ihrer Grafik einen passenden Titel. Was fällt ihnen bezüglich der Fehlerbalken auf? Warum werden für Hemaphroditen und Geschlechtlose keine Fehlerbalken gezeichnet?
4. Lassen Sie sich anzeigen wieviele Beobachtungseinheiten es je Geschlecht gibt. Lassen Sie sich auch den Mittelwert, Median und Varianz der Variable `height` für jedes Geschlecht angeben.
5. Erstellen Sie einen neuen Datensatz `sw_data3` welcher keine fehlenden Werte auf den Variablen `gender` und `height` enthält und in welchem nur noch die Geschlechter “male” und “female” vorkommen

Lösung

```
sw_data2 <- sw_data %>%
  drop_na(height, gender)
```

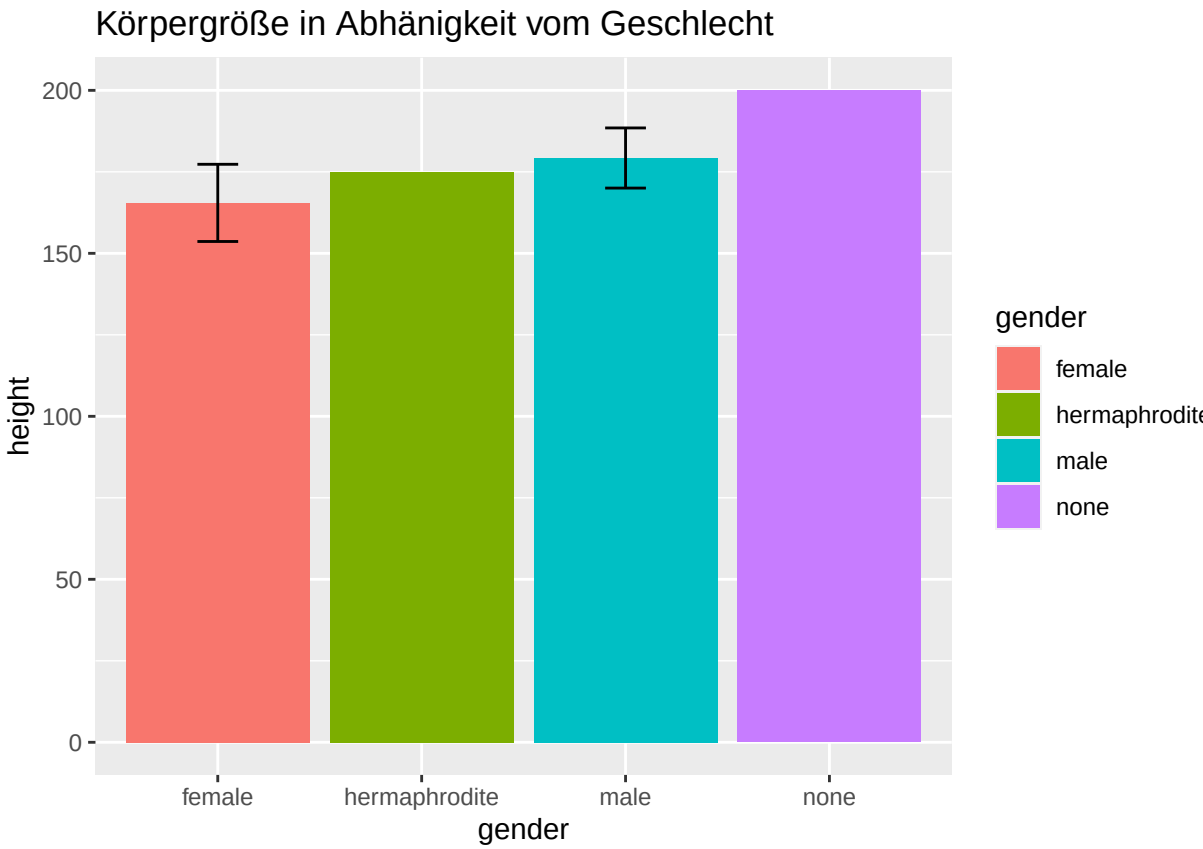
Unteraufgabe 1

```
mybarplot <- ggplot(sw_data2, aes(x = gender, y = height)) +
  stat_summary(fun = mean, geom = "bar")
mybarplot
```



Unteraufgabe 2

```
mybarplot <- ggplot(sw_data2, aes(x = gender, y = height, fill = gender)) +
  stat_summary(fun = mean, geom = "bar") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", width = 0.2) +
  labs(title = "Körpergröße in Abhänigkeit vom Geschlecht")
mybarplot
```



Unteraufgabe 3

```
table(sw_data2$gender)
```

Unteraufgabe 4

```
##
##        female hermaphrodite          male          none
##            17             1            59             1
```

```
sw_data2 %>%
  group_by(gender) %>%
  summarise(mean(height), median(height), var(height))
```

```
## # A tibble: 4 x 4
##   gender        `mean(height)` `median(height)` `var(height)`
##   <chr>                  <dbl>            <dbl>         <dbl>
## 1 female                 165.               166          530.
## 2 hermaphrodite          175                175           NA
## 3 male                   179.               183         1253.
## 4 none                   200                200           NA
```

```r
sw_data3 <- sw_data %>%
  drop_na(height, gender) %>%
  filter(gender == "male" | gender == "female")

table(sw_data3$gender)
```

Unteraufgabe 5

```
##
## female   male
##     17     59
```

```r
table(sw_data2$gender)
```

```
##
##        female hermaphrodite          male          none
##            17             1            59             1
```

3) t-test für unabhängige Stichproben

1. Nutzen Sie die Funktion `t.test()` um zu überprüfen, ob die Mittelwerte der Körpergröße zwischen Männern und Frauen signifikant voneinander abweichen. Nehmen Sie dafür (einfachhalber) Varianzhomogenität und Unabhängigkeit der Stichproben an. Testen Sie zweiseitig.
2. Treffen Sie eine Testentscheidung. Unterscheiden sich beide Gruppen signifikant voneinander? Genau genommen: Was testet der eigentlich?
3. Wie groß ist die Mittelwertsdifferenz?

Lösung

```r
t.test(sw_data3$height ~ sw_data3$gender,
       alternative = c("two.sided"),
       paired = FALSE,
       var.equal = TRUE)
```

Unteraufgabe 1

```
##
##  Two Sample t-test
##
## data:  sw_data3$height by sw_data3$gender
## t = -1.5104, df = 74, p-value = 0.1352
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -31.928219   4.394819
## sample estimates:
## mean in group female   mean in group male
##             165.4706             179.2373
```

```
# Die Annahme der gleichen Varianzen und einer ggf. normalverteilten abhänigen Variable
# ist übrigens diskutabel. Wir vertiefen das an dieser Stelle nicht. Hinweise geben aber:
# ggplot(sw_data3, aes(y = height, x = gender)) +
#        geom_boxplot()

# ggplot(sw_data3, aes(height)) +
#         geom_histogram(fill = "white", color = "grey30") +
#         facet_wrap(~ gender)
```

Unteraufgabe 2 Bei einem p-Wert > 0.05 lehnen wir die Nullhypothese nicht ab. Wir gehen von keinem signifikanten Mittelwertsunterschied aus.

Ganz genau genommen testet der Test, ob die Mittelwertsdifferenz signifikant von Null verschieden ist. Dies ist hier nicht der Fall.

```
mymeans <- sw_data3 %>%
  group_by(gender) %>%
  summarise(Means = mean(height))
mymeans
```

Unteraufgabe 3

```
## # A tibble: 2 x 2
##   gender Means
##   <chr>  <dbl>
## 1 female  165.
## 2 male    179.
```

```
abs(mymeans$Means[1] - mymeans$Means[2])
```

```
## [1] 13.7667
```

4) Einfache Regression: Zusammenhang Lineares Modell und t-Test

1. Rechnen Sie eine einfache lineare Regression mit abhäniger Variable `height` und Prädiktor `gender`.
2. Vergleichen Sie die Mittelwerte beider Gruppen mit dem Intercept und Slope der Regression. Erinnern Sie sich an die Mittelwertsdifferenz. Was fällt ihnen auf?
3. Schauen Sie den t-Wert und p-Wert für den Signifikanztest ob der Slope signifikant von Null verschieden ist an. Vergleichen Sie diesen mit Ergebnis des t-Tests aus Aufgabe 3.

Lösung

```
mymodel <- lm(formula = height ~ gender, data = sw_data3)
summary(mymodel)
```

Unteraufgabe 1

```
##
## Call:
## lm(formula = height ~ gender, data = sw_data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.237   -4.737    3.763   12.838   84.763
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  165.471      8.031   20.60   <2e-16 ***
## gendermale    13.767      9.115    1.51    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.11 on 74 degrees of freedom
## Multiple R-squared:  0.02991,    Adjusted R-squared:  0.0168
## F-statistic: 2.281 on 1 and 74 DF,  p-value: 0.1352
```

```
mymodel$coefficients
```

Unteraufgabe 2

```
## (Intercept)  gendermale
##    165.4706     13.7667
```

```
mymeans
```

```
## # A tibble: 2 x 2
##   gender Means
##   <chr>  <dbl>
## 1 female  165.
## 2 male    179.
```

```
mymeans$Means
```

```
## [1] 165.4706 179.2373
```

```
abs(mymeans$Means[1] - mymeans$Means[2])
```

```
## [1] 13.7667
```

```
sum(mymodel$coefficients)
```

```
## [1] 179.2373
```

Der Intercept entspricht dem Mittelwert der Körpergröße der Frauen. Der Slope entspricht der Mittelwertsdifferenz. Intercept + Slope entspricht dem Mittelwert der Männer.

```
t.test(sw_data3$height ~ sw_data3$gender,
       alternative = c("two.sided"),
       paired = FALSE,
       var.equal = TRUE)
```

Unteraufgabe 3

```
##
##  Two Sample t-test
##
## data:  sw_data3$height by sw_data3$gender
## t = -1.5104, df = 74, p-value = 0.1352
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -31.928219   4.394819
## sample estimates:
## mean in group female   mean in group male
##             165.4706             179.2373
```

```
summary(mymodel)
```

```
##
## Call:
## lm(formula = height ~ gender, data = sw_data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.237   -4.737    3.763   12.838   84.763
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  165.471      8.031   20.60   <2e-16 ***
## gendermale    13.767      9.115    1.51    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.11 on 74 degrees of freedom
## Multiple R-squared:  0.02991,    Adjusted R-squared:  0.0168
## F-statistic: 2.281 on 1 and 74 DF,  p-value: 0.1352
```

Der Absolutbetrag der t-Werte ist gleich. Auch die p-Werte sind identisch. In beiden Fällen testen wir ob die Mittelwertsdifferenz signifikant von Null verschieden ist.

Literature

Annotation: This exercise sheet bases in part on exercises, that you can find in the textbook *Dicovering Statistics Using R* (Field, Miles & Field, 2012). They were modified for the purpose of this sheet and the R-code was actualized.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: SAGE Publications Ltd.

Version: 20 Mai, 2021 09:06