

Übungszettel Harry Potter

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

Joshua Driesen (joshua.driesen@stud.uni-goettingen.de)

German

Links

[Übungszettel als PDF-Datei zum Drucken](#)

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 (Rechtsklick > Speichern unter...).
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](#)

Allgemeines

Wir möchten Ihnen in dieser Woche einen kleinen Vorgeschmack auf die Klausur geben. Dieser Übungszettel entspricht in seiner Form und Aufgabenstellung recht genau dem, was Sie in der Klausur erwartet, lediglich der Umfang ist hier geringer als er es in der sechsstündigen Klausur sein wird. Ihnen wird nach einer kurzen Einführung in den jeweiligen Themenbereich eine fertige Auswertung vorgelegt, die sie nachvollziehen sollen. Diese Auswertung ist manchen Bereichen nicht optimal, etwa wegen verletzter Annahmen oder falsch interpretierter Ergebnisse, aber auch inhaltlicher Ungereimtheiten. Ihre Aufgabe soll es sein, diese Probleme zu erkennen, und einzuschätzen, welche Aussagen der Auswertung Sie weiterhin für valide halten, welche Aussagen sich als unhaltbar herausstellen, und welche Verbesserungen die bestehenden Unklarheiten ausräumen könnten. Sie sollten dabei Ihre Verbesserungsvorschläge in R-Code umzusetzen können oder zumindest Ansätze dazu.

Ihre andere wichtige Aufgabe ist das Finden einer guten Visualisierung eines Aspekts der Daten. Auch diese Visualisierung sollte in R erstellt werden, um die volle Punktzahl zu erreichen. Ihre Darstellungsidee sollen Sie aber auch ganz analog mit Stift und Papier verdeutlichen, vor allem, wenn eine Umsetzung mit ggplot-Code nicht gelingen sollte. Die beste Übung für diese Art von Aufgabe ist das Lesen von Papern, bei denen Sie besonders auf die jeweilig verwendeten Visualisierungsformen achten. Ist es den Autoren geglückt, Ihre Befunde anschaulich darzustellen? Wenn ja, wie? Wenn nein, was fehlt, oder ist zu viel?

Aufgabe

Die Hogwarts Schule für Hexerei und Zauberei hat eine Studie durchgeführt, um Lernkultur und magische Fähigkeiten innerhalb Ihrer vier Häuser Gryffindor, Hufflepuff, Ravenclaw und Slytherin zu untersuchen. Hierzu wurde für alle Schüler_innen eines Jahrgangs erhoben, wieviele Stunden sie pro Woche durchschnittlich in der Bibliothek verbringen. Zudem führten alle Schüler_innen den Wingardium-Leviosa Zauber aus, um eine Feder zum Schweben zu bringen. Dabei wurde die erreichte Schwebehöhe in Zentimetern gemessen, um eine objektive Messung der Zauberkraft zu erhalten. Zusätzlich wurde die Hauszugehörigkeit erfasst, sodass insgesamt drei Variablen in die Auswertung eingehen.

```
library(tidyverse)
hogwarts_data <- read_csv("https://pzezula.pages.gwdg.de/data/HarryPotter.dat")
```

```
##
## -- Column specification -----
## cols(
##   Haus = col_character(),
##   Lernzeit = col_double(),
##   Schwebehoehe = col_double()
## )
```

```
hogwarts_data$Haus <- factor(hogwarts_data$Haus)
head(hogwarts_data)
```

```
## # A tibble: 6 x 3
##   Haus      Lernzeit Schwebehoehe
##   <fct>      <dbl>      <dbl>
## 1 Gryffindor    9.4         40.5
## 2 Gryffindor   10.2         46.1
## 3 Gryffindor    9.6         49.1
## 4 Gryffindor   10.2          0
## 5 Gryffindor    9.8         32.9
## 6 Gryffindor   10.6         40.4
```

```
library(psych)
describeBy(hogwarts_data, hogwarts_data$Haus)
```

```
##
## Descriptive statistics by group
## group: Gryffindor
##      vars  n  mean    sd median trimmed   mad min  max range skew kurtosis   se
## Haus*      1 30  1.00  0.00   1.00   1.00  0.00   1  1.0   0.0   NaN      NaN 0.00
## Lernzeit    2 30  9.24  1.16   9.55   9.27  1.19   7 11.7   4.7 -0.21   -0.70 0.21
## Schwebehoehe 3 30 29.75 16.57  33.90  31.00 11.93   0 51.8  51.8 -0.79   -0.73 3.02
## -----
## group: Hufflepuff
##      vars  n  mean    sd median trimmed   mad min  max range skew kurtosis   se
## Haus*      1 26  2.00  0.00   2.0   2.00  0.00   2  2.0   0.0   NaN      NaN 0.00
## Lernzeit    2 26 13.40  1.25  13.6  13.41  1.56  11 15.5   4.5 -0.09   -1.21 0.25
## Schwebehoehe 3 26 17.13 12.12  19.8  16.94 12.75   0 38.3  38.3 -0.10   -1.25 2.38
## -----
## group: Ravenclaw
```

```
##          vars  n  mean    sd median trimmed  mad min  max range  skew kurtosis  se
## Haus*      1 27  3.00  0.00   3.0   3.00 0.00 3.0  3.0   0.0   NaN      NaN 0.00
## Lernzeit   2 27 12.03  1.12  12.2  12.15 1.04 8.8 13.6   4.8 -1.01   0.79 0.21
## Schwebhoehe 3 27 60.78 17.45  64.5  63.40 8.45 0.0 80.3  80.3 -1.78   3.57 3.36
## -----
## group: Slytherin
##          vars  n  mean    sd median trimmed  mad min  max range  skew kurtosis  se
## Haus*      1 29  4.00  0.00   4.0   4.00 0.00 4.0  4.0   0.0   NaN      NaN 0.00
## Lernzeit   2 29  7.67  1.14   7.6   7.61 1.19 5.8 10.4   4.6  0.36   -0.33 0.21
## Schwebhoehe 3 29 23.03 15.80  24.8  23.08 20.61 0.0 47.4  47.4 -0.26   -1.44 2.93
```

Als Schule für Hexerei und Zauberei interessiert ist für uns vor allem die Schwebhöhe als Kriterium interessant, deshalb:

```
hogwarts_m1 <- lm(Schwebhoehe ~ Haus + Lernzeit, data = hogwarts_data)
summary(hogwarts_m1)
```

```
##
## Call:
## lm(formula = Schwebhoehe ~ Haus + Lernzeit, data = hogwarts_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -61.411  -9.118   2.057  10.871  26.685
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.5630    11.8327  -0.386  0.700541
## HausHufflepuff -28.0998     6.5866  -4.266 4.31e-05 ***
## HausRavenclaw  20.6523     5.3118   3.888 0.000176 ***
## HausSlytherin  -0.9114     4.3959  -0.207 0.836141
## Lernzeit       3.7149     1.2457   2.982 0.003544 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.13 on 107 degrees of freedom
## Multiple R-squared:  0.5709, Adjusted R-squared:  0.5549
## F-statistic: 35.59 on 4 and 107 DF,  p-value: < 2.2e-16
```

Die wichtigste Erkenntnis ist der signifikante Lernzeit-Prädiktor: Die Lehre an Hogwarts scheint zu funktionieren. Auffällig sind allerdings auch die teils signifikanten Dummyprädiktoren der Hauszugehörigkeit; anscheinend gibt es teils erhebliche Unterschiede zwischen den magischen Fähigkeiten der verschiedenen Häuser. Professor Snape betont vor allem die negative Sonderrolle Hufflepuffs: Der höchstsignifikante Dummyprädiktor ist für ihn ein klarer Beleg, dass es sich bei den Hufflepuffs um die mit Abstand schlechtesten Zauberer und Hexen Hogwarts handelt - und vielleicht sogar um hoffnungslose Fälle, bei denen jede weitere Lehranstrengung verschwendet sei.

Auf Wunsch von Professor Flitwick wurde eine zusätzliche Analyse durchgeführt. Er ist überzeugt davon, dass die in seinem Haus Ravenclaw verbreitete Kultur des Fleißes dazu führt, dass seine Schülerinnen und Schüler mehr lernen als in den anderen Häusern üblich. Hierzu liess er eine logistische Regression durchführen, bei der aus der Lernzeit die Zugehörigkeit zu Ravenclaw vorhergesagt wurde:

```

hogwarts_data <- mutate(hogwarts_data,
                        Ravenclaw_ja_nein = ifelse(
                          Haus == "Ravenclaw",
                          1,
                          0)) #Wenn Ravenclaw 1, sonst 0

hogwarts_m2 <- glm(Ravenclaw_ja_nein ~ Lernzeit,
                  data = hogwarts_data,
                  family = "binomial")

summary(hogwarts_m2)

##
## Call:
## glm(formula = Ravenclaw_ja_nein ~ Lernzeit, family = "binomial",
##      data = hogwarts_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3858  -0.7013  -0.4795  -0.3123   2.0452
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.1608     1.2351  -4.178 2.94e-05 ***
## Lernzeit      0.3638     0.1051   3.462 0.000537 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 123.72  on 111  degrees of freedom
## Residual deviance: 109.34  on 110  degrees of freedom
## AIC: 113.34
##
## Number of Fisher Scoring iterations: 4

```

Lernzeit erweist sich als signifikanter Prädiktor für Ravenclaw-Zugehörigkeit, wodurch sich Professor Flitwick darin bestätigt sieht, dass die Fleißkultur seines Hauses zu mehr Lernen führt.

In der ursprünglichen Auswertung befand sich zudem eine visuelle Darstellung der erhobenen Daten, diese Seite scheint aber von der Briefeule verspeist worden zu sein. Können Sie mit ggplot, dem Muggle-Ersatz für graphische Zauberei, eine angemessene Visualisierung erstellen?

English

Links

[Exercise sheet in PDF](#)

Some hints

1. Please give your answers in a .Rmd file. You may generate one from scratch using the file menu: ‘File > new file > R Markdown ...’ Delete the text below *Setup Chunk* (starting from line 11). Alternatively you may use this [sample Rmd](#) by downloading it.
2. You may find the informations useful that you can find on the [start page of this course](#).
3. Don’t hesitate to google for solutions. Effective web searches to find solutions for R-problems is a very useful ability, professionals to that too ... A really good starting point might be the R area of the programmers platform [Stackoverflow](#)
4. You can find very useful [cheat sheets](#) for various R-related topics. A good starting point is the [Base R Cheat Sheet](#).

General info

This week we want to give you a little taste of the final exam. This sheet’s form and exercise type matches those of the final exam pretty well, although the final exam will of course be longer. After all, you’ll be given six hours for the final exam, so you shouldn’t worry about that aspect too much. After a short introduction to the general topic, you’ll be given a finished data analysis which you’re supposed to understand and evaluate. This analysis is sub-optimal in multiple ways, in the sense of e.g. violated assumptions or wrongly interpreted results. There can also be inconsistencies contentwise. Your task is to identify these problems and to give an estimate of the severity of their impact on the overall integrity of the results presented. Additionally, you’ll be able to make proposals for possible improvements. You should be able to implement these improvements in R or at least give us an idea of what you intend with your code. Your other big task will be the design of a good visual representation for a key aspect of the data. If you can implement that using ggplot or other R graphical possibilities - great. But you should show us at least an idea of your idea using “old-school” pen and paper. For this sheet however, we’ll be focusing on the ggplot-part. The best way to prepare for this part of the exam is to read real papers, with a special focus on the way the authors chose to visualize their results. Did they succeed in showing their results clearly? If so, how? If not, what’s missing, what’s superfluous?

Main exercise

The Hogwarts School of Witchcraft and Wizardry conducted a study investigating learning culture and magical abilities within the four houses Gryffindor, Hufflepuff, Ravenclaw and Slytherin. For this, they asked all students from one age cohort how many hours per week they spent in the library on average. Additionally, all students had to perform the Wingardium Leviosa spell to levitate a feather. The distance that the feather levitate above the ground was used as an objective measure of magical ability. Finally, they noted which of the four houses the student belonged to, thus including three variables in total for this study.

```
library(tidyverse)
hogwarts_data <- read_csv("https://pzezula.pages.gwdg.de/data/HarryPotter.dat")
```

```
##
## -- Column specification -----
## cols(
##   Haus = col_character(),
##   Lernzeit = col_double(),
##   Schwebehoehe = col_double()
## )
```

```
hogwarts_data$Haus <- factor(hogwarts_data$Haus)
#Haus = house
head(hogwarts_data)
```

```
## # A tibble: 6 x 3
##   Haus      Lernzeit Schwebehoehe
##   <fct>      <dbl>      <dbl>
## 1 Gryffindor    9.4        40.5
## 2 Gryffindor   10.2        46.1
## 3 Gryffindor    9.6        49.1
## 4 Gryffindor   10.2         0
## 5 Gryffindor    9.8        32.9
## 6 Gryffindor   10.6        40.4
```

```
library(psych)
describeBy(hogwarts_data, hogwarts_data$Haus)
```

```
##
## Descriptive statistics by group
## group: Gryffindor
##      vars  n  mean    sd median trimmed  mad min  max range  skew kurtosis  se
## Haus*      1 30  1.00  0.00   1.00   1.00  0.00   1  1.0   0.0   NaN      NaN 0.00
## Lernzeit    2 30  9.24  1.16   9.55   9.27  1.19   7 11.7   4.7 -0.21   -0.70 0.21
## Schwebehoehe 3 30 29.75 16.57  33.90  31.00 11.93   0 51.8  51.8 -0.79   -0.73 3.02
## -----
## group: Hufflepuff
##      vars  n  mean    sd median trimmed  mad min  max range  skew kurtosis  se
## Haus*      1 26  2.00  0.00   2.0   2.00  0.00   2  2.0   0.0   NaN      NaN 0.00
## Lernzeit    2 26 13.40  1.25  13.6  13.41  1.56  11 15.5   4.5 -0.09   -1.21 0.25
## Schwebehoehe 3 26 17.13 12.12  19.8  16.94 12.75   0 38.3  38.3 -0.10   -1.25 2.38
## -----
## group: Ravenclaw
##      vars  n  mean    sd median trimmed  mad min  max range  skew kurtosis  se
## Haus*      1 27  3.00  0.00   3.0   3.00  0.00  3.0  3.0   0.0   NaN      NaN 0.00
## Lernzeit    2 27 12.03  1.12  12.2  12.15  1.04  8.8 13.6   4.8 -1.01    0.79 0.21
## Schwebehoehe 3 27 60.78 17.45  64.5  63.40  8.45  0.0 80.3  80.3 -1.78    3.57 3.36
## -----
## group: Slytherin
##      vars  n  mean    sd median trimmed  mad min  max range  skew kurtosis  se
## Haus*      1 29  4.00  0.00   4.0   4.00  0.00  4.0  4.0   0.0   NaN      NaN 0.00
## Lernzeit    2 29  7.67  1.14   7.6   7.61  1.19  5.8 10.4   4.6  0.36   -0.33 0.21
## Schwebehoehe 3 29 23.03 15.80  24.8  23.08 20.61  0.0 47.4  47.4 -0.26   -1.44 2.93
```

As a school for witchcraft and wizardry, the levitation height of the feather was especially interesting as a criterion, therefore:

```
hogwarts_m1 <- lm(Schwebehoehe ~ Haus + Lernzeit, data = hogwarts_data)
#Schwebehoehe = levitation height
#Lernzeit = study time
summary(hogwarts_m1)
```

```
##
```

```
## Call:
## lm(formula = Schwebhoehe ~ Haus + Lernzeit, data = hogwarts_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -61.411  -9.118   2.057  10.871  26.685
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.5630    11.8327  -0.386  0.700541
## HausHufflepuff -28.0998     6.5866  -4.266 4.31e-05 ***
## HausRavenclaw  20.6523     5.3118   3.888 0.000176 ***
## HausSlytherin  -0.9114     4.3959  -0.207 0.836141
## Lernzeit       3.7149     1.2457   2.982 0.003544 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.13 on 107 degrees of freedom
## Multiple R-squared:  0.5709, Adjusted R-squared:  0.5549
## F-statistic: 35.59 on 4 and 107 DF,  p-value: < 2.2e-16
```

The most important result is the significance of the learning time predictor: Hogwarts's teaching style seems to be working. However, some of the dummy variables of the different houses reached significance, too, indicating substantial differences in magical abilities between the houses. Professor Snape especially emphasizes Hufflepuff's negative results. To him, their highly significant dummy predictor is clear evidence for the Hufflepuffs being by far the worst witches and wizards in Hogwarts. They might even be completely hopeless cases, whom no further teaching capacities should be wasted on.

Professor Flitwick expressed a wish for an additional analysis. He's convinced that the culture of striving found in his house Ravenclaw leads to his students spending more time learning than the other houses. To show this, he conducted a logistic regression, using learning time to predict whether the student in question is a Ravenclaw or not:

```
hogwarts_data <- mutate(hogwarts_data,
                        Ravenclaw_ja_nein = ifelse( #Ravenclaw_ja_nein = Ravenclaw_yes_no
                        Haus == "Ravenclaw",
                        1,
                        0)) #If student's a Ravenclaw 1, otherwise 0

hogwarts_m2 <- glm(Ravenclaw_ja_nein ~ Lernzeit,
                  data = hogwarts_data,
                  family = "binomial")

summary(hogwarts_m2)
```

```
##
## Call:
## glm(formula = Ravenclaw_ja_nein ~ Lernzeit, family = "binomial",
##      data = hogwarts_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3858  -0.7013  -0.4795  -0.3123   2.0452
##
```

```
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.1608      1.2351  -4.178 2.94e-05 ***
## Lernzeit      0.3638      0.1051   3.462 0.000537 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 123.72  on 111  degrees of freedom
## Residual deviance: 109.34  on 110  degrees of freedom
## AIC: 113.34
##
## Number of Fisher Scoring iterations: 4
```

Learning time proves to be a significant predictor of ‘Ravenclaw-ness’, leading professor Flitwick to feel validated in his opinion of his house’s striving culture leading to more learning.

In the original analysis, there was also a visual representation of the collected data. However, the owl carrying these documents must’ve eaten that page. Are you able to use ggplot, the muggle alternative to visual wizardry, to create an adequate visualization?

Version: 12 August, 2021 12:03