Corpus Annotation and Data Analysis Equinox School: Introduction to Statistics

Maik Thalmann
maik.thalmann@gmail.com

September 2022

Contents

Chapter 1

Welcome

Hey everyone,

if you're looking at this, I assume that you signed up for the CAnDA equinox school in Göttingen in September 2022 and want to attend the Introduction to Statistics course. Here, I will cover some tools and concepts that are going to be instrumental in making sure that you get the most out of this class. I know that it is not ideal to have an introductory class that assumes basic familiarity with some material already, but unfortunately we do not have an entire semester together, but only one short week.

1.1 Prerequisites

Because our schedule is quite tight and the relevant material quite expansive, I will have to presuppose some familiarity with statistics. Below, I will briefly summarize what these requirements are (in the form of questions) and, if you do not yet feel comfortable with them, give some pointers on where to change that.

1.1.1 R

As you might have guessed, our technical analysis tool will be R. Because an introduction to R would be a class in and of itself, it would be beneficial for all attendees of the class to at least have some basic knowledge. This includes:

- What is an R script?

- How do I create and execute one?
- Which program should I use for my R environment, i.e., for editing R scripts, viewing data and plots, and for running statistical analyses. The popular choice here is undoubtedly RStudio, but you may also use Visual Studio Code and set it up for R if you're more comfortable with that.
- How do I load my data (csv, xlsx, txt file) into R?
- How do I use external packages to expand on the capabilities of base R?
- What are factors and integers and how do I switch between them in R?

As an aside, in case I do show code for data manipulation or plots, I will mostly rely on the packages in the so-called tidyverse, a collection of R packages. While I do not consider familiarity with all of these packages essential, they are important (and often a time saver) indepedently of this class if you want to use R for your own data analysis or data visualization projects. The packages I will most heavily rely on are dplyr for data wrangling and ggplot2 for visualization purposes.

If you know German (and prefer it over English resources), I have my own website to offer you as a way of (re)gaining familiarity with R. Sessions 1 through 6 should form a quite thorough background (with some skippable material).

Otherwise, I can recommend the relevant chapters in Gries (2013), which provide a gentle introduction to using R.

1.1.2 Statistics

As announced in the program for the class, I will, again for reasons of time, have to ask you to know your way around two widely used statistical tests and some basic notions of inferential statistics, detailed below. I am, of course, happy to answer questions during class and the practice sessions, but if you're not as confident with the topics below, I would advise to do some prepatory reading to get the most out of the class (and the one in the second week of the summer school).

- What do the following terms mean: Mean, median, variance, and standard deviation?
- What is the *t*-test? What does the output of the `t.test` command in R mean – see below?

```
# load the tidyverse
library(tidyverse)
# subset the data to only have two colors
diamond_sub <- diamonds %>%
  filter(color %in% c("E", "J"))
# show the first few rows of the data set
head(diamond_sub)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 5  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
## 6  0.22 Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
```

```
# perform a t-test
t.test(diamond_sub$price ~ diamond_sub$color, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  diamond_sub$price by diamond_sub$color
## t = -24.8811, df = 3766.32, p-value < 0.000000000000000222
## alternative hypothesis: true difference in means between group E and group J is not equal to 0
## 95 percent confidence interval:
##  -2424.1311 -2070.0000
## sample estimates:
## mean in group E mean in group J
##       3076.7525       5323.8180
```

- In which scenarios is the t-test applicable and when is it not (scale levels, assumptions of the t-test, etc.)? What is the effect of setting the `paired` argument in the R command of `t.test` to `TRUE`?
- What is the χ^2 test? Why does it find so much use in corpus linguistics compared to the t-test? What does the output below mean?

```
# get the frequencies for both diamond colors in the data set
(color_frequencies <- diamond_sub %>%
  group_by(color) %>%
  summarise(frequency = n())
)
```

```
## # A tibble: 2 x 2
##   color frequency
##   <ord>     <int>
## 1 E          9797
## 2 J          2808
```

```
# perform the chi^2 test
color_frequencies %>%
```

```
  select(frequency) %>%
  chisq.test()
```

```
##
##  Chi-squared test for given probabilities
##
## data:  .
## X-squared = 3875.14, df = 1, p-value < 0.000000000000000222
```

- What are proper and improper interpretations of a p-value? What is statistical significance?

To brush up on statistics, you can also read (the relevant chapters in) Gries (2013) Alternatively, I recommend Vasishth and Broe (2010) up to (and including) chapter 3 and Field et al. (2012) (chapters 1 through 3 as well).

1.2 Get in touch

If you have any questions about the information presented here or any other matters related to the class, please to do not hesitate to drop me a line via email.

Session Info

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Locale: en_US.UTF-8 / en_US.UTF-8 / en_US.UTF-8 / C / en_US.UTF-8 / en_US.UTF-8
##
## Package version:
##   forcats_0.5.1     stringr_1.4.0     dplyr_1.0.9       purrr_0.3.4       readr_2.1.2
##   tibble_3.1.7      ggplot2_3.3.6     tidyverse_1.3.1   styler_1.7.0      tidyselect_
##   rematch2_2.1.2    haven_2.5.0       colorspace_2.0-3  vctrs_0.4.1       generics_0.
##   yaml_2.3.5        utf8_1.2.2        rlang_1.0.3       R.oo_1.25.0       pillar_1.7.
##   glue_1.6.2        withr_2.5.0       DBI_1.1.3         dbplyr_2.2.1      modelr_0.1.
##   R.cache_0.15.0    lifecycle_1.0.1   munsell_0.5.0     gtable_0.3.0      cellranger_
##   rvest_1.0.2       codetools_0.2-18  evaluate_0.15     knitr_1.39        tzdb_0.3.0
##   fansi_1.0.3       broom_1.0.0       backports_1.4.1   scales_1.2.0      jsonlite_1.
##   hms_1.1.1         digest_0.6.29     stringi_1.7.6     bookdown_0.27     grid_4.2.1
##   tools_4.2.1       magrittr_2.0.3    crayon_1.5.1      pkgconfig_2.0.3   ellipsis_0.
##   reprex_2.0.1      lubridate_1.8.0   assertthat_0.2.1  rmarkdown_2.14    httr_1.4.3
##   R6_2.5.1          compiler_4.2.1
```

Chapter 2

First Session

(to be populated)

There will be a lot of things I cannot cover in class: - contrast coding - effect size - power analysis - some subtleties of when and when not to interpret null effects -

Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R*. London: Sage Publications.

Gries, Stefan T. 2013. *Statistics for linguistics with R: A practical introduction*. Berlin: de Gruyter.

Vasishth, Shravan, and Michael Broe. 2010. *The foundations of statistics: A simulation-based approach*. Berlin: Springer Science & Business Media.