

Angewandte Datenanalyse mit R

Tag 3 - Datenimport und -modellierung

Andreas Mock

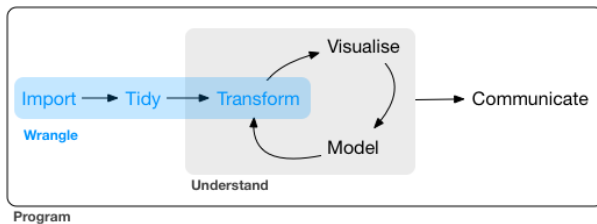
Abteilung für Medizinische Onkologie, Nationales Centrum für Tumorerkrankungen (NCT) Heidelberg

Sommersemester 2018



Ablauf - Tag 3

- ▶ Datenimport mit readr und readxl
- ▶ Datenmodellierung
- ▶ Häufige Missverständnisse über den p-Wert



Datenimport

Dateipfade

1. Wie finde ich den **absoluten** Dateipfad heraus, in dem sich meine R Umgebung befindet, das so genannte *working directory*?

```
getwd()
```

```
## [1] "/Users/Andy/Documents/Research/Github_page/presentations"
```

2. Welche Dateien befinden sich in dem aktuellen Dateipfad?

```
list.files()
```

```
## [1] "bell_curve.png"    "effect.tiff"       "Fisher.jpg"
## [4] "header.log"        "header.tex"        "header2.tex"
## [7] "info_pres.pdf"     "info_pres.Rmd"     "nct_logo_red.jpg"
## [10] "nct_logo.jpg"      "pres_day1.pdf"     "pres_day1.Rmd"
## [13] "pres_day2.pdf"     "pres_day2.Rmd"     "pres_day3.pdf"
## [16] "pres_day3.Rmd"     "RStudio.pdf"       "simulation.pdf"
```

Dateipfade

- Wie kann ich ausgehend vom *working directory* einen **relativen** Dateipfad angeben?

Mit dem Präfix `../` lässt sich eine Ordner Ebene, mit `../../` entsprechend 2 Ebenen nach oben gehen.

Der relative Pfad

```
list.files("../")
```

```
## [1] "_site"                "_site.yml"
## [3] "andreasmock.github.io" "archive"
## [5] "CVE.Rmd"              "data"
## [7] "day1.Rmd"              "day2.Rmd"
## [9] "day3.Rmd"              "Github_page.Rproj"
## [11] "index.Rmd"             "MetaboDiff.Rmd"
## [13] "p-value"               "pics"
## [15] "presentations"         "projects.Rmd"
## [17] "Readme.rtf"            "refs.Rmd"
## [19] "solution1.Rmd"         "solution2.Rmd"
## [21] "solution3.Rmd"         "teaching.Rmd"
```

entspricht damit dem **absoluten** Dateipfad

```
list.files("/Users/Andy/Documents/Research/Github_page")
```

Dateipfade

4. Wie kann ich Daten aus dem Internet einlesen?

Einfach den Webpfad benutzen:

```
library(readxl)
path <- paste0("https://tcga-data.nci.nih.gov/docs/publications/sarc_2017/",
               "SARC_264_Fusion_Gene_Profiles.txt")

read_tsv(file = path)
```

Datenimport

Benötigte Pakete: readr Paket im tidyverse, sowie das bisher noch nicht verwendete Paket readxl.

Funktionen zum Import nach Dateityp:

- ▶ read_tsv: tab-separated file
- ▶ read_csv: comma-separated file
- ▶ read_xlsx bzw. read_xls: Excel spread sheet
- ▶ read_delim: file mit beliebig anzugebendem delimiter, also Trennzeichen

Datenmodellierung

Datenmodellierung

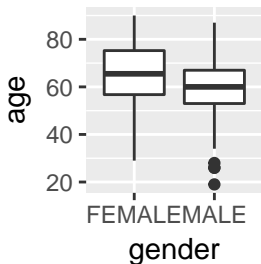
HNSCC Datensatz laden

```
load(url("http://andreasmock.github.io/data/hnsc.RData"))
```

T-Test

Der T-Test ist ein parametrischer (geht von normalverteilten Daten aus) Hypothesentest zum Vergleich von kontinuierlichen Daten zweier Gruppen.

```
ggplot(hnsc, aes(x=gender, y=age)) +  
  geom_boxplot()
```



T-Test

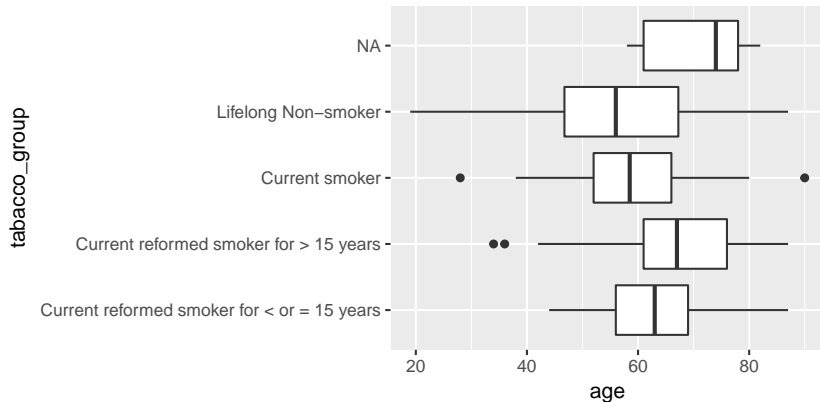
```
t.test(age ~ gender, data=hnscc)
```

```
##  
##  Welch Two Sample t-test  
##  
## data:  age by gender  
## t = 2.5518, df = 116.84, p-value = 0.01201  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
##  1.008323 7.998547  
## sample estimates:  
## mean in group FEMALE    mean in group MALE  
##           64.59211           60.08867
```

ANOVA (analysis of variance)

ANOVA bietet die Möglichkeit mehr als 2 Gruppen miteinander zu vergleichen.

```
ggplot(hnsc, aes(x=tabacco_group, y=age)) +  
  geom_boxplot() +  
  coord_flip()
```



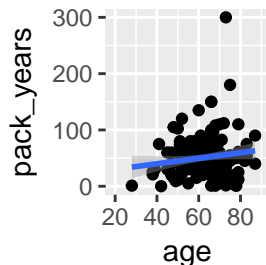
ANOVA (analysis of variance)

```
summary(aov(age ~ tobacco_group, data=hnscc))
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## tobacco_group   3   4385   1461.7    10.79 1.02e-06 ***
## Residuals      268  36294    135.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 7 observations deleted due to missingness
```

Korrelationstest

```
ggplot(hnsc, aes(x=age, y=pack_years)) +  
  geom_point() +  
  geom_smooth(method="lm")
```



```
cor.test(~ age + pack_years, data=hnsc)
```

```
##  
## Pearson's product-moment correlation  
##  
## data: age and pack_years  
## t = 1.7489, df = 152, p-value = 0.08233  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.01811495 0.29211981  
## sample estimates:  
## cor
```

Häufige Missverständnisse über den p-Wert

Der Erfinder des p-Wertes



Sir Ronald Fisher (1890-1962)
Gonville & Caius College, Cambridge

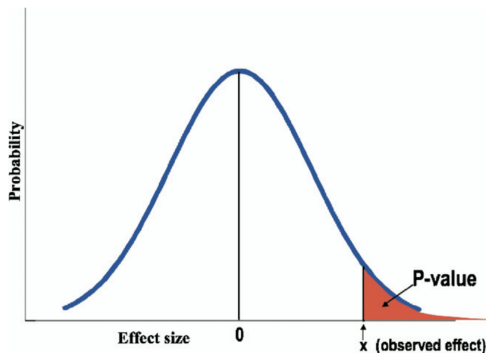
"Personally, the writer prefers to set a low standard of significance at the 5 percent point [. . .] A scientific fact should be regarded as experimentally established **only if** a properly designed **experiment rarely fails to give this level of significance.**"

Statistical Methods for Research Workers, 1926

Definition des p-Wertes

Wahrscheinlichkeit das gleiche Stichprobenergebnis oder ein noch extremeres zu erhalten, wenn die Nullhypothese wahr ist.

Algebraische Definition: $P(X \geq x \mid \sim H_0)$ wobei X eine Zufallsvariable und x der beobachtete Wert in den Daten ist



Goodman, 2008

#1 | Wenn $p < 0.05$, ist die Nullhypothese nur in 5% wahr

Dies ist die wohl **häufigste Fehlinterpretation** des p-Wertes.

Der p-Wert wird unter der Annahme berechnet, dass die Nullhypothese zutrifft ($P(\text{Daten} \mid \sim H_0)$), er kann daher nicht gleichzeitig die Wahrscheinlichkeit sein, dass die Nullhypothese zutrifft ($P(H_0 \mid \text{Daten})$).

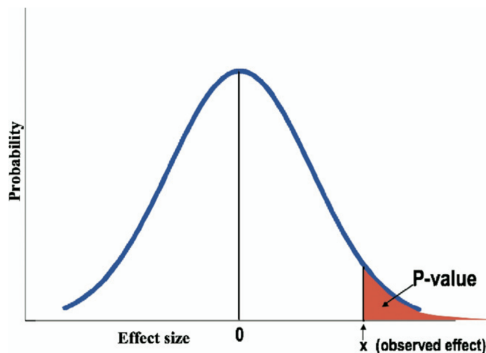
Beispiel: Die Wahrscheinlichkeit drei Mal hintereinander Kopf beim Münzwurf zu erhalten ist $p=0.125$. Dies bedeutet jedoch nicht, dass die Wahrscheinlichkeit, dass die Münze fair ist nur 12.5% beträgt.

#2 | $p > 0.05$ bedeutet, dass es keinen Unterschied zwischen den Gruppen gibt

Eine nicht signifikante Differenz bedeutet bloß, dass die beobachteten **Daten konsistent mit der Nullhypothese** sind und nicht, dass die Nullhypothese wahrscheinlicher ist.

#3 | $p=0.06$ ist substantiell schlechter als $p=0.04$

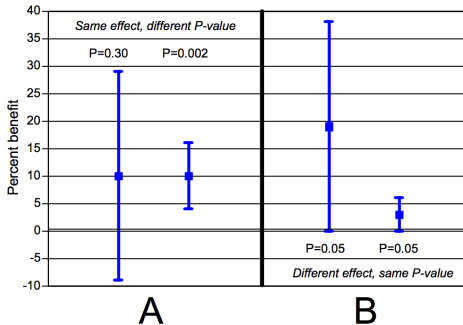
Fisher hat den p-Wert als **kontinuierliche Variable** eingeführt um abzuschätzen, ob ein Ergebnis es Wert ist weiter untersucht zu werden. Die magische p-Wert Grenze von 0.05 ist völlig arbiträr. p-Werte von 0.04 und 0.06 sind sehr ähnliche Wahrscheinlichkeiten!



Goodman, 2008

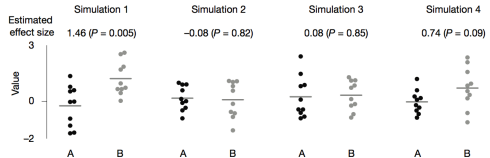
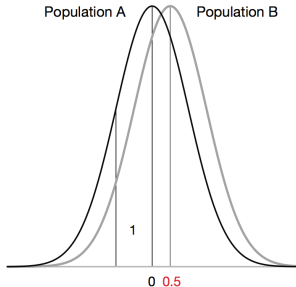
#4 | Studien mit gleichem p-Wert zeigen eine ähnlich starke Effektgröße

Der folgende Plot zeigt, dass dies nicht zutrifft. Der gleiche p-Wert kann einen völlig anderen Effekt indizieren (Fig. B). Umgekehrt, kann es einen identischen Effekt bei unterschiedlichem p-Wert geben (Fig. A):



Goodman, 2008

#5 | $p=0.05$ bedeutet, dass man bei Wiederholung des Experiments in 5% ein nicht signifikantes Ergebnis erhält



Halsey et al., 2015

Nur bei einer **großen Effektgröße** bzw. **Power** (i.e. Gruppengröße) sind p-Werte bei Wiederholung des Experiments mit einer anderen Stichprobe reproduzierbar!

A Dirty Dozen: Twelve P-Value Misconceptions

Goodman, S

Semin Hematol. 2008 Jul;45(3):135-40.

The fickle P value generates irreproducible results

Halsey LG, Curran-Everett D, Vowler SL & Drummond GB

Nat Methods. 2015 Mar;12(3):179-85.