R Notebook

Jan Hohenheim

Setup

```
rm(list = ls())
library(ggplot2)
library(ggthemes)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(purrr)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggthemes':
##
##     theme_map
```

```
library(latex2exp)
library(glue)

theme <-
  theme_solarized() #+
  #theme(text = element_text(family = "Helvetica Neue"))
```

Problem 13

The following hemoglobin levels of blood samples from patients with HbSS and HbS/β sickle cell disease are given (Statistische Prinzipien für medizinische Projekte. Hüsler, J. and Zimmermann, H. (2010)):

```
HbSS <- c(7.2, 7.7, 8, 8.1, 8.3, 8.4, 8.4, 8.5, 8.6, 8.7, 9.1, 9.1, 9.1, 9.8, 10.1,
10.3)
HbSb <- c(8.1, 9.2, 10, 10.4, 10.6, 10.9, 11.1, 11.9, 12, 12.1)
Hb <- data.frame(level = c(HbSS, HbSb),
                 category = c(rep("HbSS", length(HbSS)), rep("HbSb", length(HbSb))))
print("HbSS")
```

```
## [1] "HbSS"
```

```
Hb |>
  filter(category == "HbSS") |>
  select(level) |>
  glimpse()
```

```
## Rows: 16
## Columns: 1
## $ level <dbl> 7.2, 7.7, 8.0, 8.1, 8.3, 8.4, 8.4, 8.5, 8.6, 8.7, 9.1, 9.1, 9.1,~
```

```
print("HbSb")
```

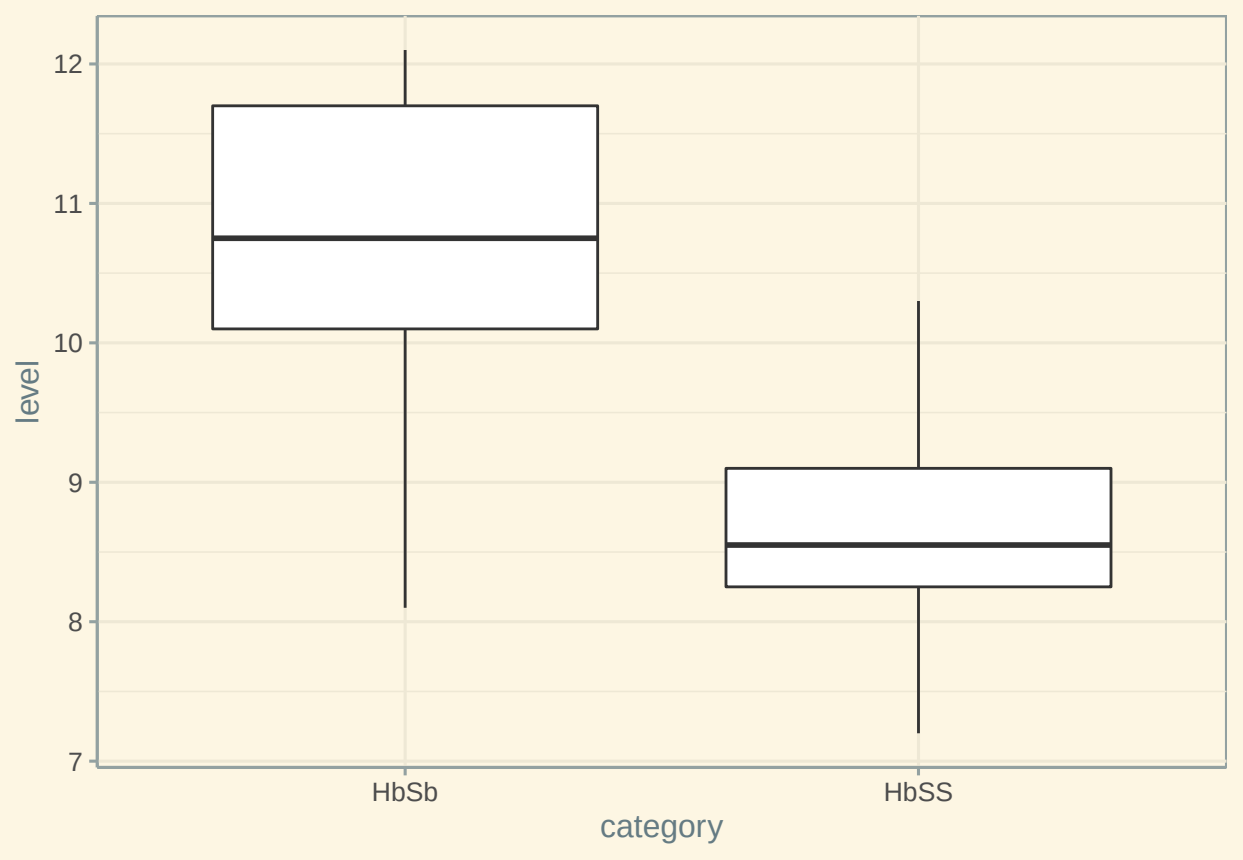
```
## [1] "HbSb"
```

```
Hb |>
  filter(category == "HbSb") |>
  select(level) |>
  glimpse()
```

```
## Rows: 10
## Columns: 1
## $ level <dbl> 8.1, 9.2, 10.0, 10.4, 10.6, 10.9, 11.1, 11.9, 12.0, 12.1
```

(a) Visualize the data with boxplots.

```
plot <- Hb |>
  group_by(category) |>
  ggplot(aes(x = category, y = level)) +
  geom_boxplot() +
  theme
plot
```

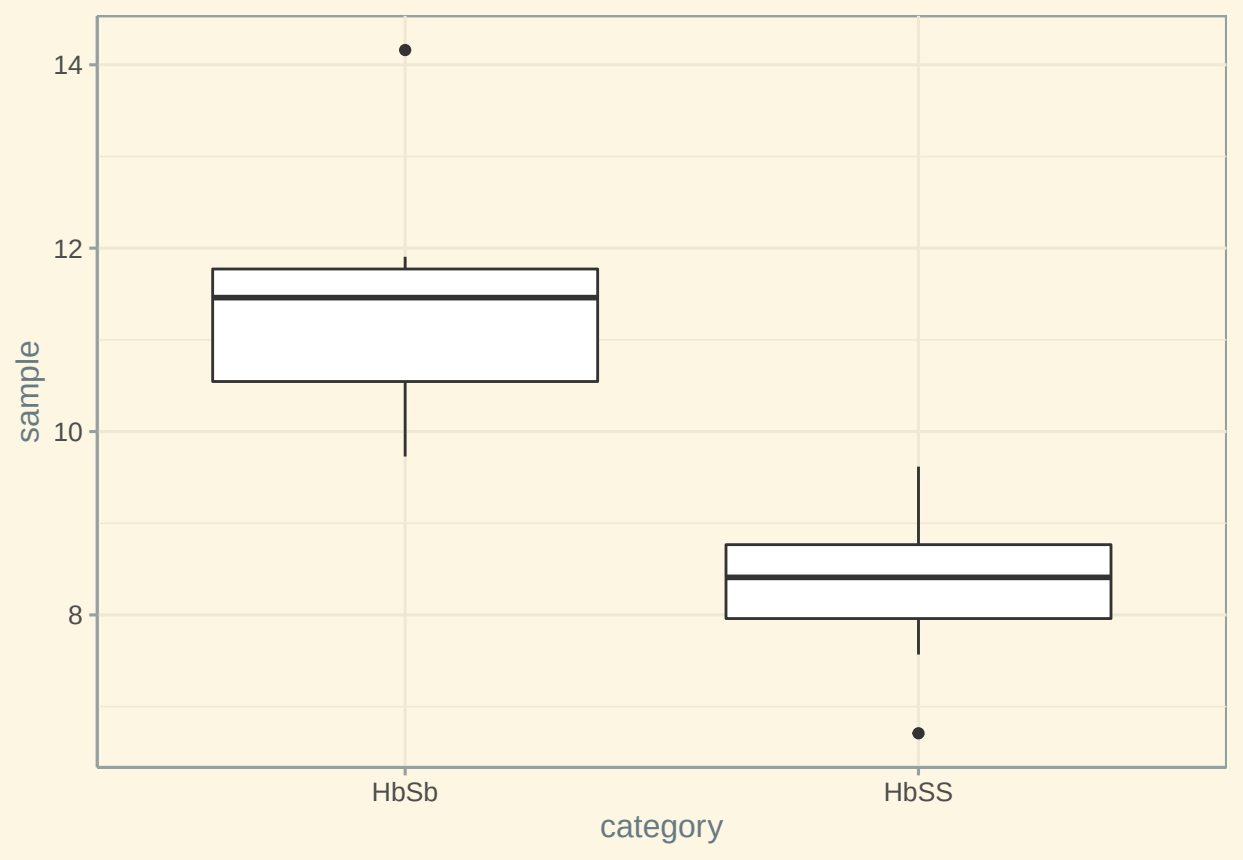


(b) Propose a statistical model for both diseases. What are the parameters? Estimate all parameters from your model based on intuitive estimators.

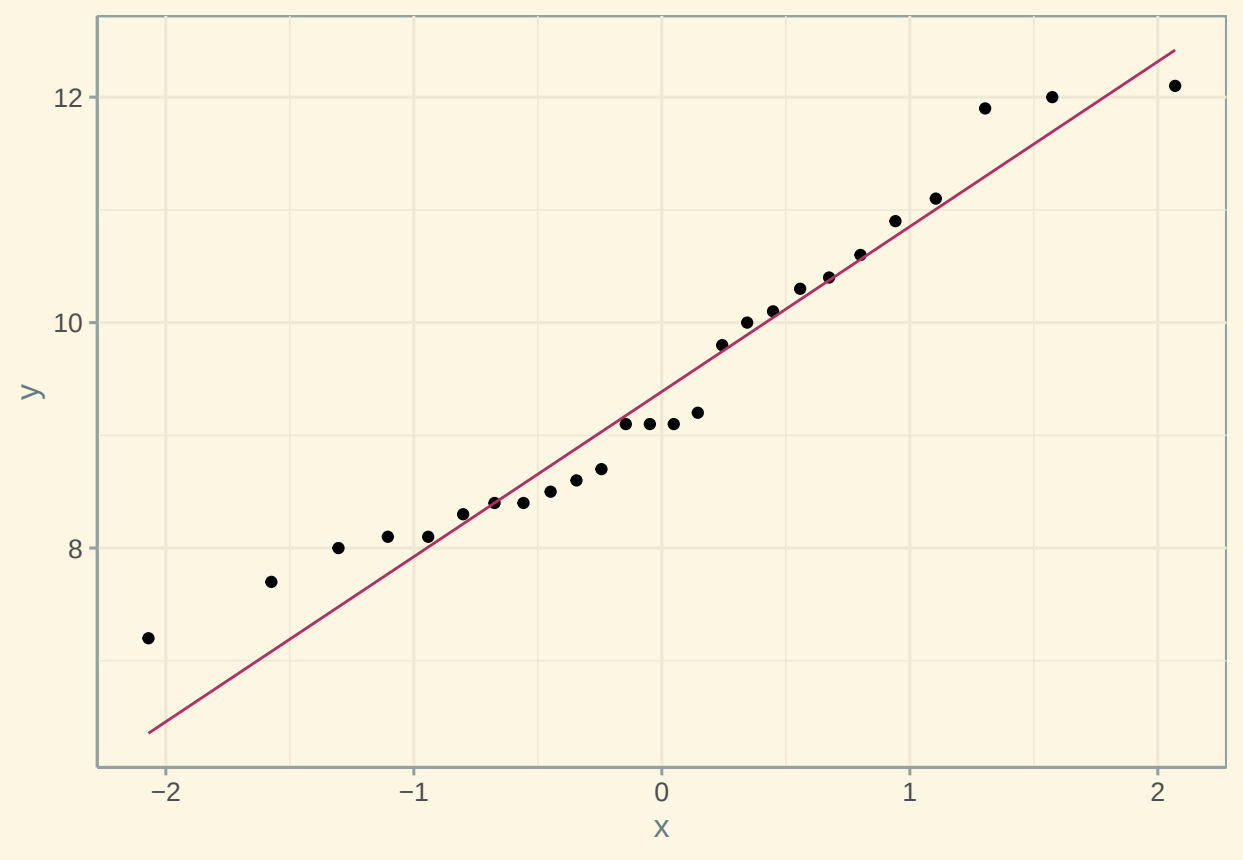
I propose: - HbSb ~ N(10.75, 1.5) and - HbSS ~ Bin(8.5, 1)

(c) Based on boxplots and QQ-plots, is there coherence between your model and the data?

```
Hb_sample <- Hb |>
  mutate(sample = c(rnorm(length(HbSS), mean = 8.5, sd = 1),
                    rnorm(length(HbSb), mean = 10.75, sd = 1.5)))
plot_box <- Hb_sample |>
  group_by(category) |>
  ggplot(aes(x = category, y = sample)) +
  geom_boxplot() +
  theme
plot_box
```




```
plot_qq <- Hb_sample |>
  ggplot(aes(sample = level)) +
  geom_qq() +
  geom_qq_line(colour = "maroon") +
  theme
plot_qq
```



It fits pretty well!

Problem 14

The dataset Oral is available in the R package spam and contains oral cavity cancer counts for 544 districts in Germany.

(a) Load the data and take a look at its help page using ?Oral

```
library(spam)
```

```
## Spam version 2.8-0 (2022-01-05) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.
```

```
##
## Attaching package: 'spam'
```

```
## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve
```

```
?Oral
```

(b) Compute summary statistics for all variables in the dataset. Which of the 544 regions has the highest number of expected counts E ?

```
Oral |>
  glimpse()
```

```
## Rows: 544
## Columns: 3
## $ Y   <int> 18, 62, 44, 12, 18, 27, 20, 29, 39, 21, 40, 23, 30, 20, 27, 373, 5~
## $ E   <dbl> 16.35051, 45.90600, 44.66248, 16.32308, 26.94854, 33.33713, 30.650~
## $ SMR <dbl> 1.1008834, 1.3505861, 0.9851669, 0.7351552, 0.6679397, 0.8099077, ~
```

```
Oral |>
  summary()
```

```
##        Y                E                SMR
##  Min.   :  1.00   Min.   :  3.011   Min.   :0.1460
##  1st Qu.:  9.00   1st Qu.: 10.883   1st Qu.:0.7219
##  Median : 19.00   Median : 19.503   Median :0.9279
##  Mean   : 28.43   Mean   : 28.430   Mean   :0.9753
##  3rd Qu.: 33.00   3rd Qu.: 33.217   3rd Qu.:1.1741
##  Max.   :501.00   Max.   :393.094   Max.   :2.3957
```

```
max_e_index <- Oral$E |>
  which.max()

print(glue("The region number {max_e_index} has the highest E at {max(Oral$E)}"))
```

```
## The region number 328 has the highest E at 393.09345467
```

(c)

Poisson distribution is common for modeling rare events such as death caused by cavity cancer (column Y in the data). However, the districts differ greatly in their populations. Define a subset from the data, which only considers districts with expected fatal casualties caused by cavity cancer between 35 and 45 (subset, column E). Perform a Q-Q Plot for a Poisson distribution. Hint: use qqplot() from the stats package and define the theoretical quantiles with qpois(ppoints(...), lambda=...)

```
Oral.subset <- Oral |>
  filter(35 >= E & E <= 45)

Oral.subset |> glimpse()
```

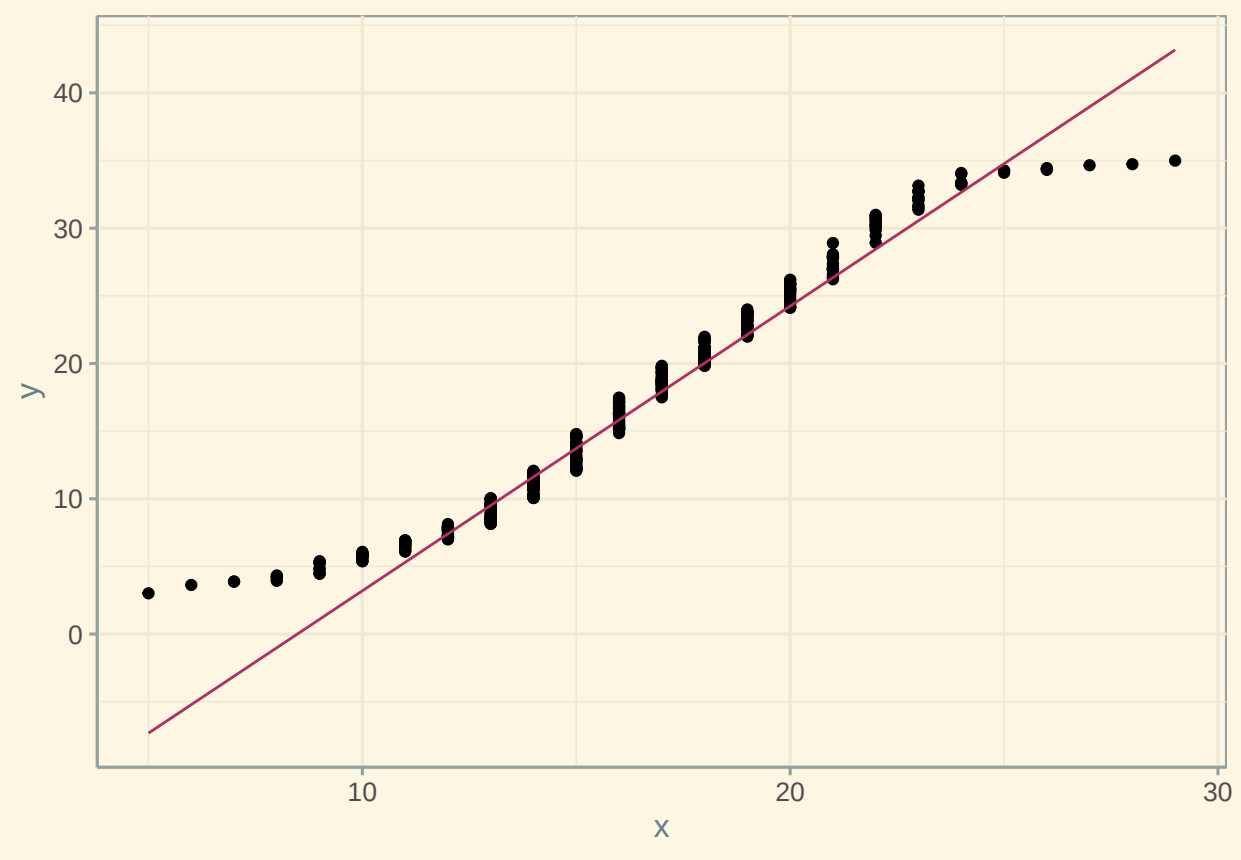
```
## Rows: 422
## Columns: 3
## $ Y   <int> 18, 12, 18, 27, 20, 21, 20, 22, 13, 18, 31, 22, 26, 27, 24, 21, 29~
## $ E   <dbl> 16.35051, 16.32308, 26.94854, 33.33713, 30.65087, 24.65196, 26.224~
## $ SMR <dbl> 1.1008834, 0.7351552, 0.6679397, 0.8099077, 0.6525099, 0.8518594, ~
```

```
lambda <- round(mean(Oral.subset$Y))
print(glue("Using lambda = {lambda}"))
```

```
## Using lambda = 16
```

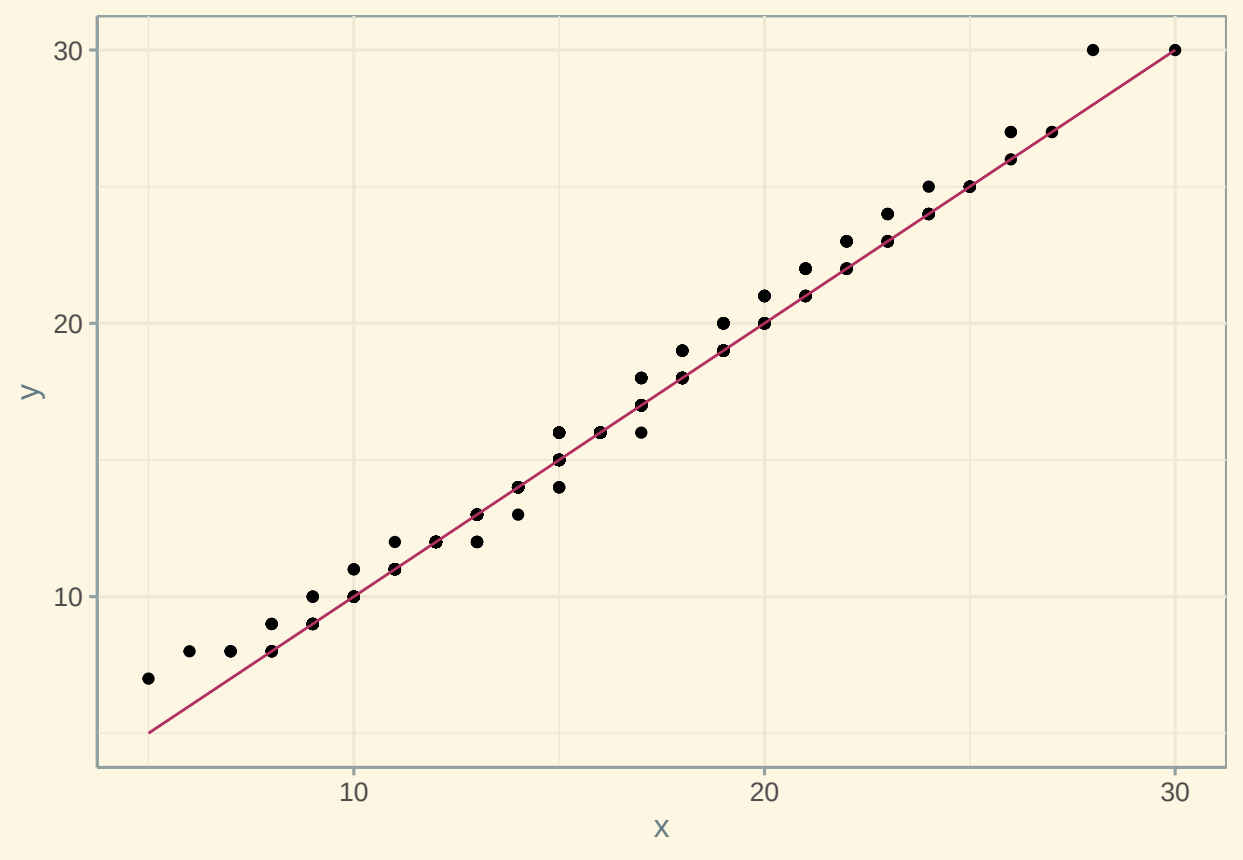
```
plot <- Oral.subset |>
  ggplot(aes(sample = E)) +
  geom_qq(distribution = qpois, dparams = lambda) +
  geom_qq_line(distribution = qpois, dparams = lambda, colour = "maroon") +
```

```
  theme
plot
```



(d) Simulate a Poisson distributed random variable with the same length and and the same lambda as your subset. Perform a QQ-plot of your simulated data. What can you say about the distribution of your subset of the cancer data?

```
sample <- rpois(nrow(Oral), lambda = lambda)
plot <- sample |>
  data.frame() |>
  ggplot(aes(sample = sample)) +
  geom_qq(distribution = qpois, dparams = lambda) +
  geom_qq_line(distribution = qpois, dparams = lambda, colour = "maroon") +
  theme
plot
```



> The assumed lambda does not fit very well

(e)

Assume that the standardized mortality ratio Zi = Yi/Ei is normally distributed, i.e., Z1,. . . , Z544 iid ~ N (μ, σ^2). Estimate μ and give a 95% (exact) confidence interval (CI). What is the precise meaning of the CI?

```
Z <- Oral$Y / Oral$E
mu <- mean(Z)
n <- length(Z)
sigma <- sd(Z)
alpha <- 0.05
ci <- mu + qt(p = c(alpha / 2, 1 - alpha / 2), df = n -  1) * sigma / sqrt(n)
print(glue("Estimating Z ~ N({mu}, {sigma^2})"))
```

```
## Estimating Z ~ N(0.975294559898491, 0.123965165620299)
```

```
print(glue("95% confidence interval: {ci[[1]]} - {ci[[2]]}"))
```

```
## 95% confidence interval: 0.94564163259642 - 1.00494748720056
```

The 95% CI means that when repeating the sampling many times, 95% of the time the mean will lie in the CI.

(f)

Simulate a 95% confidence interval based on the following bootstrap scheme (sampling with replacement): Repeat 10'000 times – Sample 544 observations Zi with replacement – Calculate and store the mean of these sampled observations Construct the confidence interval by taking the 2.5% and the 97.5% quantiles of the

stored means. Compare it to the CI from e).

```
n <- 544
means <- 1:10000 |>
  map(\(.) rnorm(n, mu, sigma) |> mean()) |>
  unlist()
quantile(means, probs = c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.9458035 1.0046173
```

The data fits extremely well