Computational Biology Group

D BSSE

ETH zürich

# Introduction to Bayesian Statistics with R

1: Exercise solutions

Jack Kuipers

28 November 2022

First we load the tidyverse and set a seed.

```
library(tidyverse); options(dplyr.summarise.inform = FALSE) # suppress summarise warnings
set.seed(42)
```

---

## Exercise 1.1 - a statistical report

*A small clinical trial on asthma patients has been run measuring the lung function of a control group on a placebo and a treatment group on a new drug.*

- *Read in the trial data (`lung_data.csv`),*
- *visualize the data for each group,*
- *test whether there is a difference in function between the two groups.*

We first read in the data

```
lung_data <- read.csv("./data/lung_data.csv")
```

Then we create a table of the descriptive statistics by *grouping* and *summarising*

```
lung_data %>% group_by(Trial.arm) %>%
  summarize(
    Mean =  signif(mean(Lung.function), 4),
    Sd = signif(sd(Lung.function), 2),
    Min = min(Lung.function),
    Median = median(Lung.function),
    Max = max(Lung.function),
    IQR = IQR(Lung.function),
    N = n()) %>%
  kable(caption = "Descriptive statistics of the lung data.",
        col.names = c("Trial arm", colnames(.)[-1]))
```

Table 1: Descriptive statistics of the lung data.

| Trial arm | Mean | Sd | Min | Median | Max | IQR | N |
|---|---|---|---|---|---|---|---|
| Control | 10.15 | 0.48 | 9.4 | 10.1 | 11.1 | 0.55 | 20 |
| Treatment | 10.48 | 0.47 | 9.4 | 10.5 | 11.4 | 0.60 | 20 |

We will also plot the data using explanatory plots. Here we store the plots as variables and use `cowplot` to

create labelled figures.

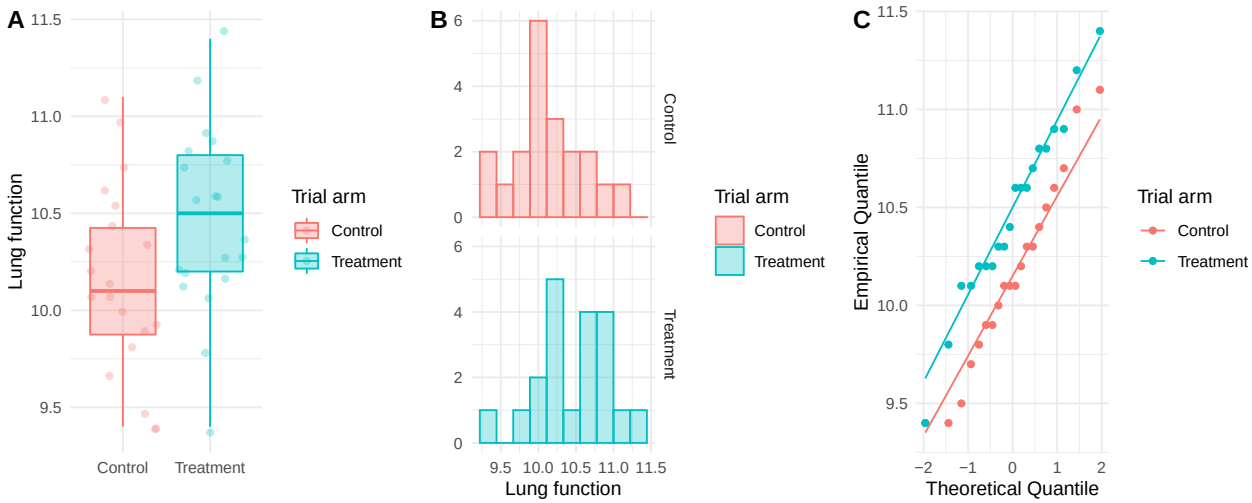
```
library(cowplot)
p <- ggplot(lung_data) + theme_minimal()

p1 <- p + geom_boxplot(aes(x = Trial.arm, y = Lung.function, color = Trial.arm,
                           fill=Trial.arm),
      outlier.shape = NA, alpha = 0.3) +
  geom_jitter(aes(x = Trial.arm, y = Lung.function, color = Trial.arm), alpha = 0.3) +
  scale_y_continuous("Lung function") + scale_x_discrete("") +
  scale_color_discrete("Trial arm") + scale_fill_discrete("Trial arm")

p2 <- p + geom_histogram(aes(x = Lung.function, color = Trial.arm, fill = Trial.arm),
      alpha = 0.3, bins = 10) +
  scale_y_continuous("") + scale_x_continuous("Lung function") +
  scale_color_discrete("Trial arm") + scale_fill_discrete("Trial arm") +
  facet_grid(Trial.arm ~ .)

p3 <- p + stat_qq(aes(sample = Lung.function, color = Trial.arm)) +
  stat_qq_line(aes(sample = Lung.function, color = Trial.arm)) +
  scale_color_discrete("Trial arm") +
  scale_x_continuous("Theoretical Quantile") + scale_y_continuous("Empirical Quantile")

plot_grid(p1, p2, p3,
  align = "vh", ncol = 3, labels = c("A", "B", "C"))
```




Finally we test for a significant difference in lung function of the two groups. Since we assume normally distributed data, we can use a two-sample (unpaired) t-test with non-equal variances between the two groups.

```
lung_t_test <- t.test(Lung.function ~ Trial.arm, lung_data)
lung_t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  Lung.function by Trial.arm
## t = -2.1569, df = 37.988, p-value = 0.0374
## alternative hypothesis: true difference in means between group Control and group Treatment is not equ
```

```
## 95 percent confidence interval:
##  -0.63003538 -0.01996462
## sample estimates:
##   mean in group Control mean in group Treatment
##                  10.150                  10.475
```

We can collate this into a statistical report:

**Report**

The following report summarizes the results obtained from a statistical analysis of the change in lung function of asthma patients when treated with a new drug which we assess by comparison to a control group. We are hypothesizing that treating the patients with the drug has an effect on lung function, and consequently formulate the null hypothesis

$$H_0 : \mu_T - \mu_C = 0 \quad \text{(the treatment has no effect)},$$

where $\mu_T$ and $\mu_C$ are the population means of treated and control patients, respectively. We shall reject the null at a significance level of $\alpha = 0.05$.

The data set we are analyzing consists of a total of $n = 40$ patients of two groups consisting of $n_T = 20$ patients that have been treated with the drug and $n_C = 20$ patients that have been treated with a placebo.

Table 2: Descriptive statistics of the lung data.

| Trial arm | Mean | Sd | Min | Median | Max | IQR | N |
|---|---|---|---|---|---|---|---|
| Control | 10.15 | 0.48 | 9.4 | 10.1 | 11.1 | 0.55 | 20 |
| Treatment | 10.48 | 0.47 | 9.4 | 10.5 | 11.4 | 0.60 | 20 |

The lung function of patients treated with the new drug overall has similar descriptive statistics as the control group. A total shift in means, however, can indeed be observed (Table 2, Figure 1).

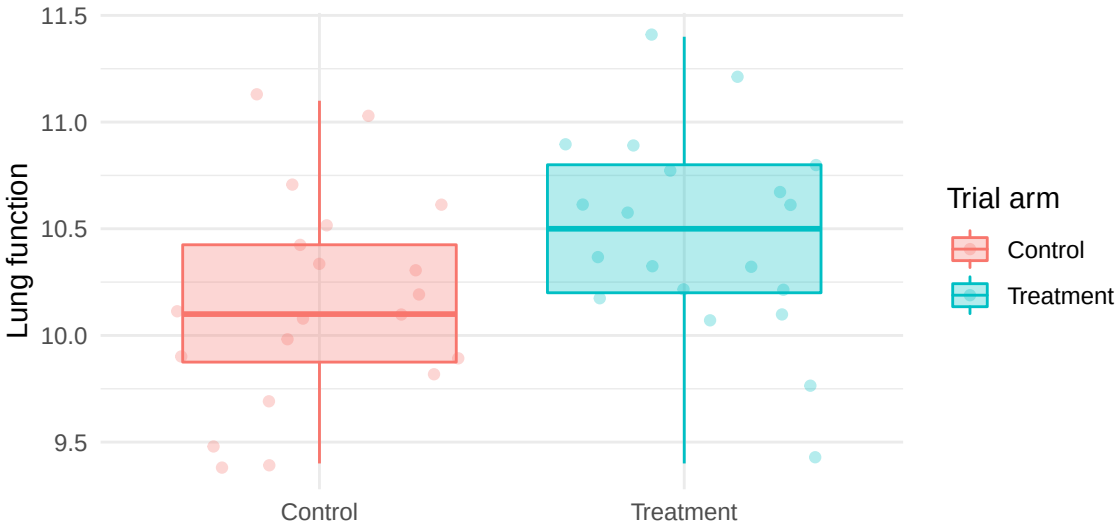



Figure 1: Boxplots of the two groups do not indicate any outliers.

Furthermore, the data do not reveal any outliers in either of the groups and both groups seem to follow a normal distribution (Figure 2). For normally distributed data with no outliers the most appropriate test is the $t$-test.

Thus, we conduct a two-sample $t$-test for independent means yielding a test statistic $t$ = -2.16 with $\nu = 38$ degrees of freedom and $p$-value $p = 0.037$. Since $p < \alpha$ we reject the null hypothesis that the two groups share the same lung function on average.

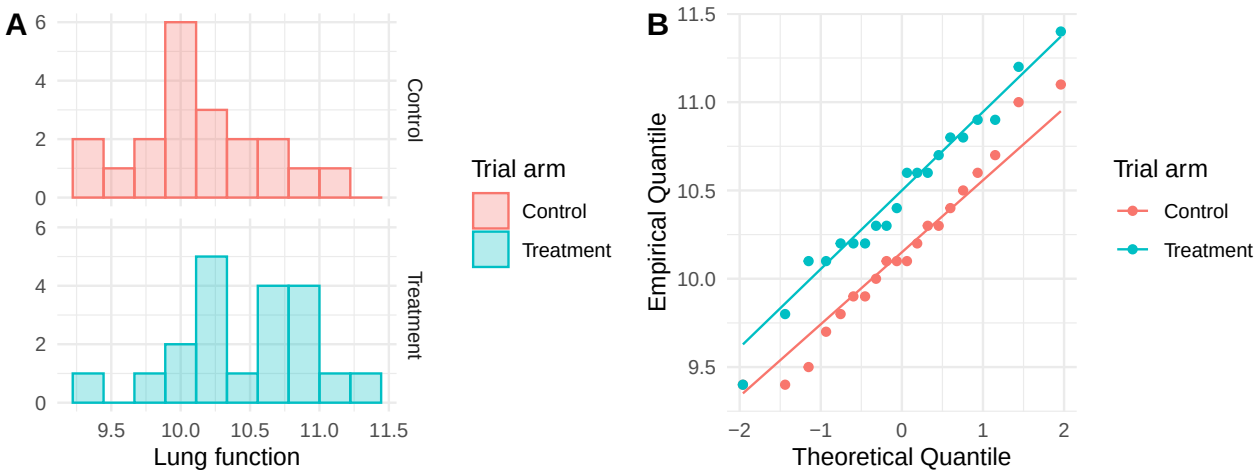



Figure 2: The two groups are approximately normally distributed.

---

## Bonus Exercise 1.2 - normality and outliers

*The t-test assumes normality and no outliers. To get a feel for how important those assumptions are, we can break them and check with simulated data.*

- *What happens to the power if we use a different distribution (with the same mean and sd) instead of a normal?*
- *What happens if we add an outlier (for example, shift one of the treatment group by a large negative value)?*

Let's first use the code from the Exercises and replace the Gaussian distribution with a uniform one. To have a variance of 1, the uniform distribution should span a range of $2\sqrt{3}$. We again shift the mean for the treatment group by $-0.25$.

```
n_reps <- 4e3 # how many repetitions
p_vals <- rep(NA, n_reps) # to store the p-values
for (ii in 1:n_reps) {
  test_samples <- runif(50, min = -sqrt(3), max = sqrt(3)) - 0.25 # treatment group
  control_samples <- runif(50, min = -sqrt(3), max = sqrt(3)) # control group
  p_vals[ii] <- t.test(test_samples, control_samples)$p.value # t-test
}
mean(p_vals < 0.05) # the power given by the fraction of significant tests
```

```
## [1] 0.233
```

The power is almost identical to the Gaussian case!

Let's look at the outlier instead

```
n_reps <- 4e3 # the number of repetitions
sample_shifts <- 0:20 # the possible shifts
p_vals_df <- data.frame() # start with an empty dataframe

for (s_shift in sample_shifts) { # loop over possible shifts
```

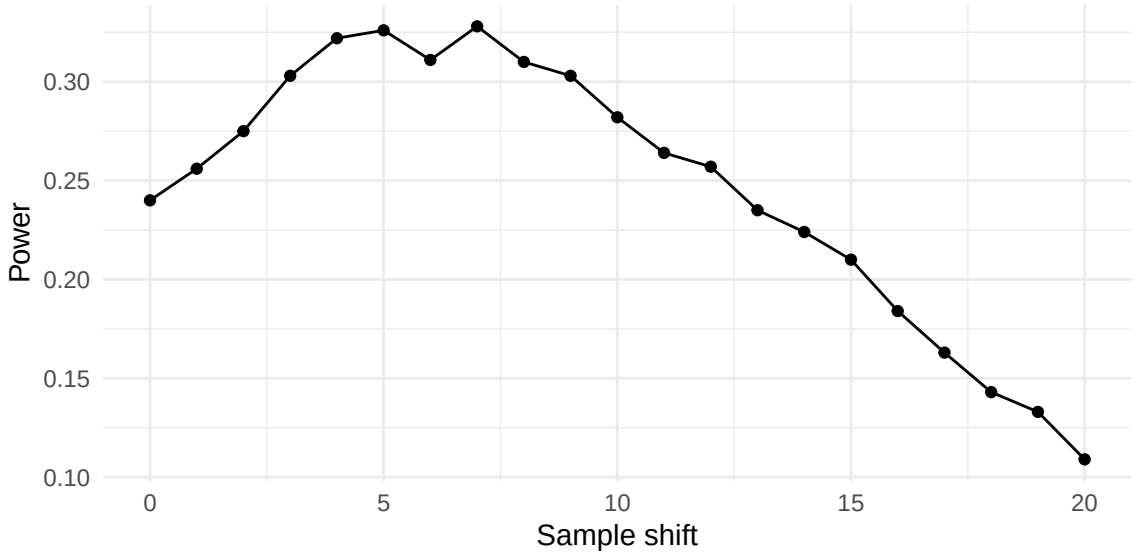
```
  p_vals <- rep(NA, n_reps) # to store the p-values
  for (ii in 1:n_reps) {
    test_samples <- rnorm(50, mean = -0.25, sd = 1) # treatment group
    test_samples[1] <- test_samples[1] - s_shift # shift one to make it an outlier
    control_samples <- rnorm(50, mean = 0, sd = 1) # control group
    p_vals[ii] <- t.test(test_samples, control_samples)$p.value # t-test
  }
  # build a local data frame for the repetitions with a given shift
  local_df <- data.frame(sample_shift = s_shift, p_vals = p_vals)
  p_vals_df <- rbind(p_vals_df, local_df) # append to the full data frame
} # end sample shift loop
```

We can then extract the empirical power

```
p_vals_df %>% group_by(sample_shift) %>%
  summarize(power = mean(p_vals < 0.05) %>% signif(3)) -> power_df
```

And plot it

```
power_df %>%
  ggplot() +
  geom_point(aes(x = sample_shift, y = power)) +
  geom_line(aes(x = sample_shift, y = power)) +
  scale_x_continuous("Sample shift") +
  scale_y_continuous("Power") +
  theme_minimal()
```




So after slightly increasing the power, having a large outlier actually ends up decreasing it!