Descriptive statistics

There are a myriad sources (books, websites, videos) explaining the concepts of descriptive statistics. We do not need to reiterate everything here. You can go through these sources to get started:

- R for non-programmers
- Science direct
- Descriptive statistics

The goal is to describe data in a meaningful and honest way. We **summarize** data to make them **more easily digestable** for us humans to answer questions like

- Where are the data points **located**? These questions are answered (at least attempted) by the location measures such as mean, median, and mode.
- How widely are they spread? How much do they **vary**? These questions are answered by the dispersion measures such as variance, standard deviation (root of the variance), interquantile range or just the range; or even Gini's mean difference.
- Are there any **outliers** (rare data points that are far away from the rest, Westfall, p. 405) and why? This is a bit more complicated.

Example: Descriptive statistics in health sciences

These are birds-eye views on the data. Let's look at a paper which was recently published in the Journal of Physiotherapy in order to get a running start:

Patients with worse disability respond best to cognitive functional therapy for chronic low back pain: a pre-planned secondary analysis of a randomised trial

(This should be open access.)

The research question was "Do five baseline moderators identify patients with chronic low back pain who respond best to cognitive functional therapy (CFT) when compared with usual care?".

In **Table 2** of the paper, the authors present the baseline characteristics of the patients stratified by the treatment group (ususal care vs CFT). We find absolute numbers, percentages, means, and standard deviations for the continuous variables, medians, and interquartile ranges for the ordinal variables. This should give an idea of the sample. In the population paradigm of statistics, we draw a sample from the population of interest and try to make inferences about the population. We want to learn more about the population respectively that data generating process (DGP) producing the data Westfall, p. 6-8. How did the data come about? Note that the **sample varies everytime we draw from the population**. We can either imagine an infinitely large population or a finite one (e.g. population of Switzerland).

Often, variables in a study are (approximately) normally distributed. We can then efficiently summarize the variable with its mean and standard deviation (location and scale parameter) as is done in the paper for *Age* in years for instance. We do not want to present p-values in such an overview table since we merely *describe* data instead of making inferences about the population or the DGP. It is by no means given that *age* has to be normally distributed in our sample. We could easily have a sample with many young people and few elderly ones. This would result in a (positively) *skewed* distribution. Having many elderly people and few young ones would result in a negatively skewed distribution of course.

To build a bridge to our concept of statistical independence from the previous chapter, the paper states: "This study found that baseline level of activity limitation was **associated** with different effects of CFT treatment". What they mean is that there was a "statistically significant effect" in the chosen model. We will get to that in chapter 5. If you want to be picky, you could argue, that they did not formally define an α-level for "significance".

Univarate vs. bivariate statisics

One can distinguish between univariate and bivariate statistics. In univariate statistics, we look at **one variable at a time**, for instance *Age* in the example above, where we could draw a boxplot or a histogram. Table 2 in the paper is a good example of univariate statistics. We are not so much interested in the relationship between variables.

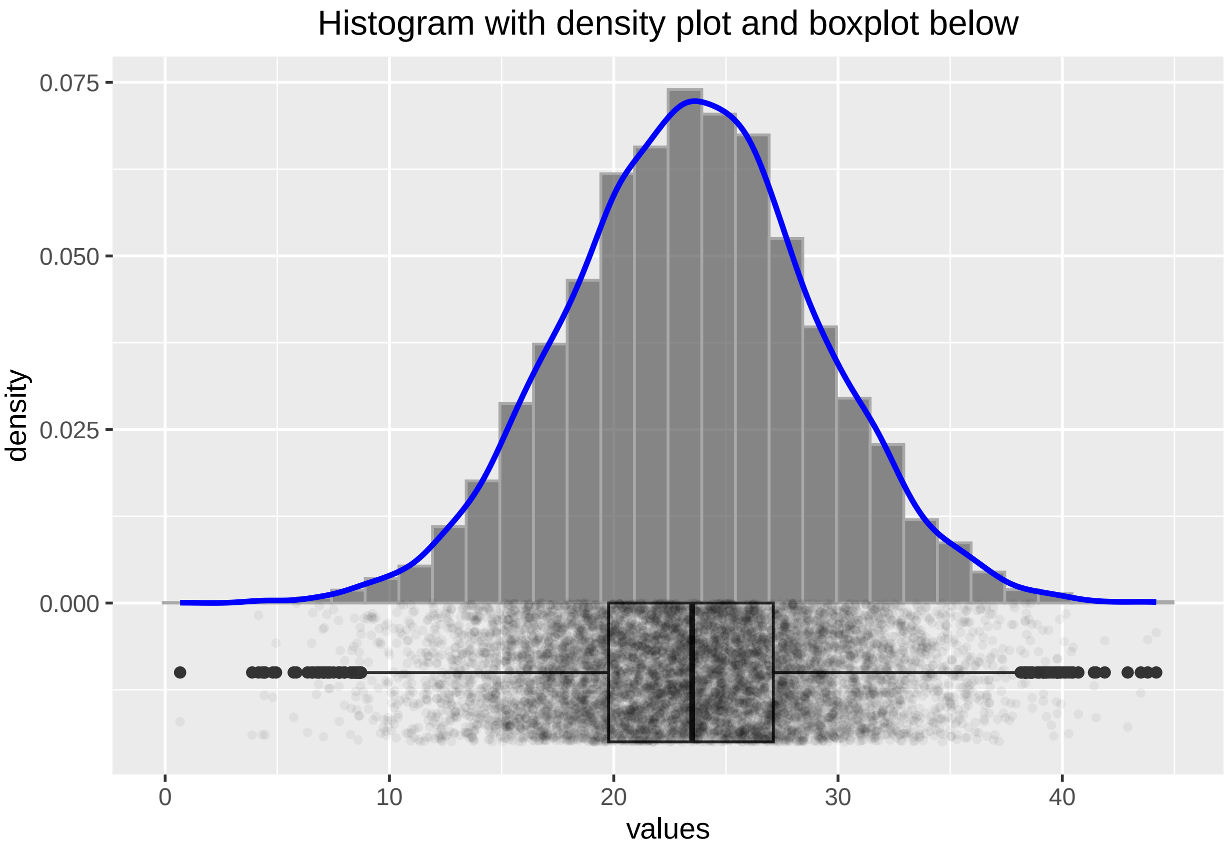
In bivariate statistics, we look at **two variables simultaneously**. An example could be a scatter plot of *Age* and *Cognitive flexibility*, where we would possible find a falling relationship (example in the wild, this would actually be a line plot, not a scatter plot). We are interested in the relationship between the two variables. How do they change (covary) together?

Multivariate statistics is the next step, where we look at more than two variables *at the same time*.

The histogram

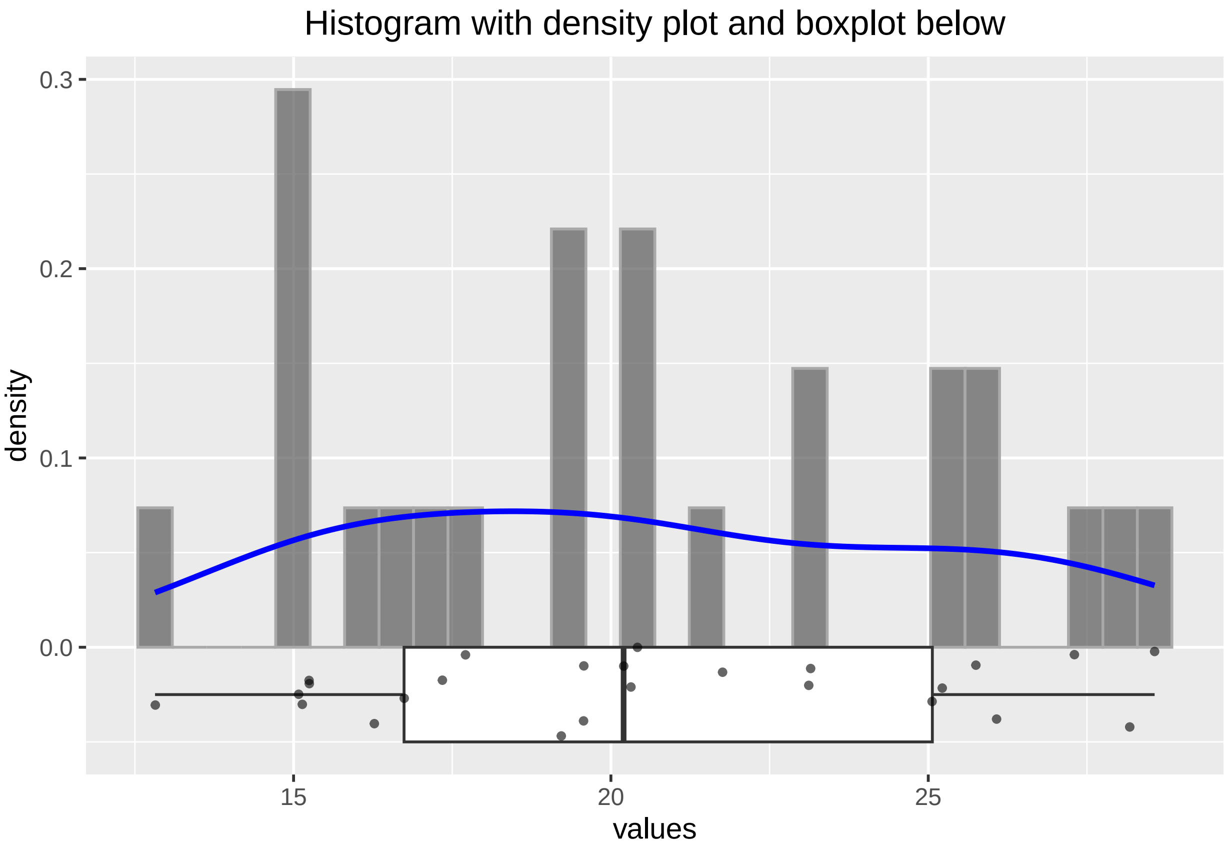
One way to visualize the distribution of a continuous variable is the histogram. This video might be helpful to cover the basics. I recommend plotting the histogram with a boxplot below; this helps to visualize the raw data points as well.

```r
library(pacman)
p_load(tidyverse)

set.seed(4433) # to get the same plot every time

# Generate normally distributed sample
x <- rnorm(10000, mean = 23.4, sd = 5.6)
df <- data.frame(values = x)

p2 <- ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 30, alpha = 0.7, color = "darkgrey") +
  geom_density(aes(y = after_stat(density)), color = "blue", linewidth = 1) +
  geom_boxplot(aes(y = -0.01, x = values),
               width = 0.02, position = position_nudge(y = -0.00)) +
  geom_point(aes(y = -0.01),
             position = position_jitter(width = 0.002, height = 0.01),
             size = 1, alpha = 0.05) +
  ggtitle("Histogram with density plot and boxplot below") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



Some researchers would discard the values below 10 and above 40 as outliers, but we know here that the data points are perfectly legitimate. One important thing we should be aware of in connection with small sample sizes is variability. Let's create not 10000 samples of a normally distributed variable, but only 25 samples:

```
# Generate normally distributed sample
set.seed(1245) # to get the same plot every time
x <- rnorm(25, mean = 23.4, sd = 5.6)
df <- data.frame(values = x)

p2 <- ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 30, alpha = 0.7, color = "darkgrey") +
  geom_density(aes(y = after_stat(density)), color = "blue", linewidth = 1) +
  geom_boxplot(aes(y = -0.025, x = values), width = 0.05,
               position = position_nudge(y = -0.00)) +
  geom_point(aes(y = -0.025),
             position = position_jitter(width = 0.005, height = 0.025),
             size = 1, alpha = 0.6) +
  ggtitle("Histogram with density plot and boxplot below") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



We *know* (in this case) that these values come from a normal distribution with a mean of 23.4 and a standard deviation of 5.6. Let's *estimate* the parameters (μ, σ) from the sample and use the often but not recommended Shapiro-wilk test for normality:

```
# Estimate mean and standard deviation
mean(x) # or
```

```
## [1] 20.6017
```

```
1 / length(x) * sum(x)
```

```
## [1] 20.6017
```

```
sd(x) # or
```

```
## [1] 4.622042
```

```
sqrt(1 / (length(x) - 1) * sum((x - mean(x))^2))
```

```
## [1] 4.622042
```

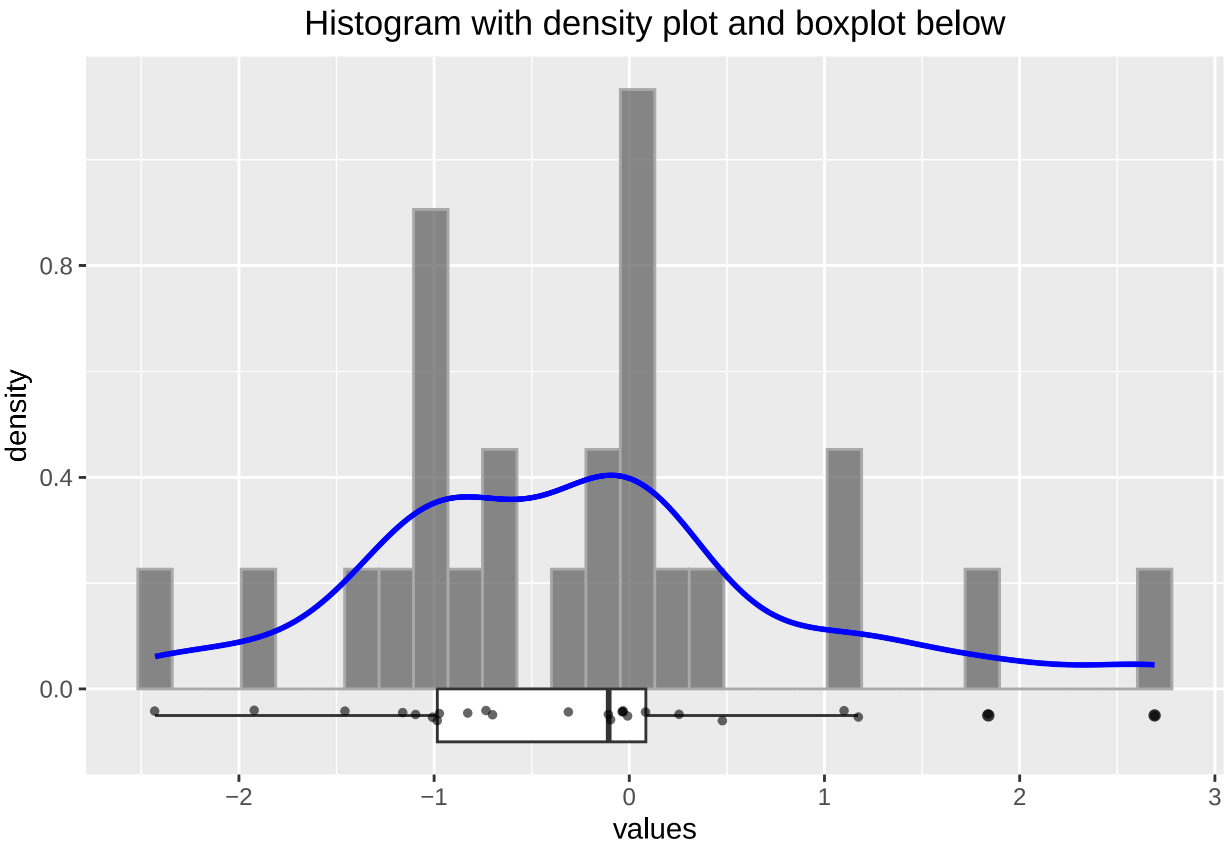
```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.95059, p-value = 0.2585
```

Firstly, the histogram looks rather differently when using a different seed. **Secondly**, we would probability not be able to tell if the data stems from a normal distribution. **Thirdly**, the sample mean and standard deviation are not all that bad estimators for the true (but unknown) mean and standard deviation. Nicely enough, the Shapiro test would not reject the null hypothesis of normality in this case.

Let's try the Shapiro test with a *t*-distribution with 3 degrees of freedom, which is not normal:

```r
set.seed(1245) # to get the same plot every time
x <- rt(25, df = 3) # random numbers from t-distribution
df <- data.frame(values = x)

p2 <- ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 30, alpha = 0.7, color = "darkgrey") +
  geom_density(aes(y = after_stat(density)), color = "blue", linewidth = 1) +
  geom_boxplot(aes(y = -0.05, x = values), width = 0.1,
               position = position_nudge(y = -0.00)) +
  geom_point(aes(y = -0.05),
             position = position_jitter(width = 0.002, height = 0.01),
             size = 1, alpha = 0.6) +
  ggtitle("Histogram with density plot and boxplot below") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



```r
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.95827, p-value = 0.3811
```

Some researchers would argue, that the distribution is normal (since the p-value is rather large), but we know it is not. We will be cautious with such ugly rule of thumbs. We will instead use Q-Q plots and histograms to decide if a variable is normally distributed.

Example in the wild

We will try to catch a histogram in the *wild* (= in a research paper): We find one here. Figure 1 shows a histogram of the the foot posture index (FPI-6) scores from the participants (3217 healthy children aged 3 to 15). "The FPI score may range from −12 (highly supinated) to +12 (highly pronated)". There are a few things to note here:

- Acccording to the Methods-section: "Testing for normality using a Kolmogorov-Smirnov test, found non-normal distribution of all data. . . ". Especially the BMI could be asymmetrically distributed, or also age. Unfortunately, we do not have histograms or boxplots for these variables. Nevertheless, the authors present the mean and standard deviation for these variables which implies (when not reading the methods section) that the data is (at least sufficiently) normally distributed.

- Strictly speaking, an FPI score cannot be normally distributed since it takes discrete values which are bounded between -12 and 12 (normal distribution can take values between $-\infty$ and ∞). But that should not be a problem with so many levels of an ordered categorical variable. It should be a sufficient approximation.

- "The FPI was analysed as continuous data, rather than as z-score data". z-scores are standardized scores:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

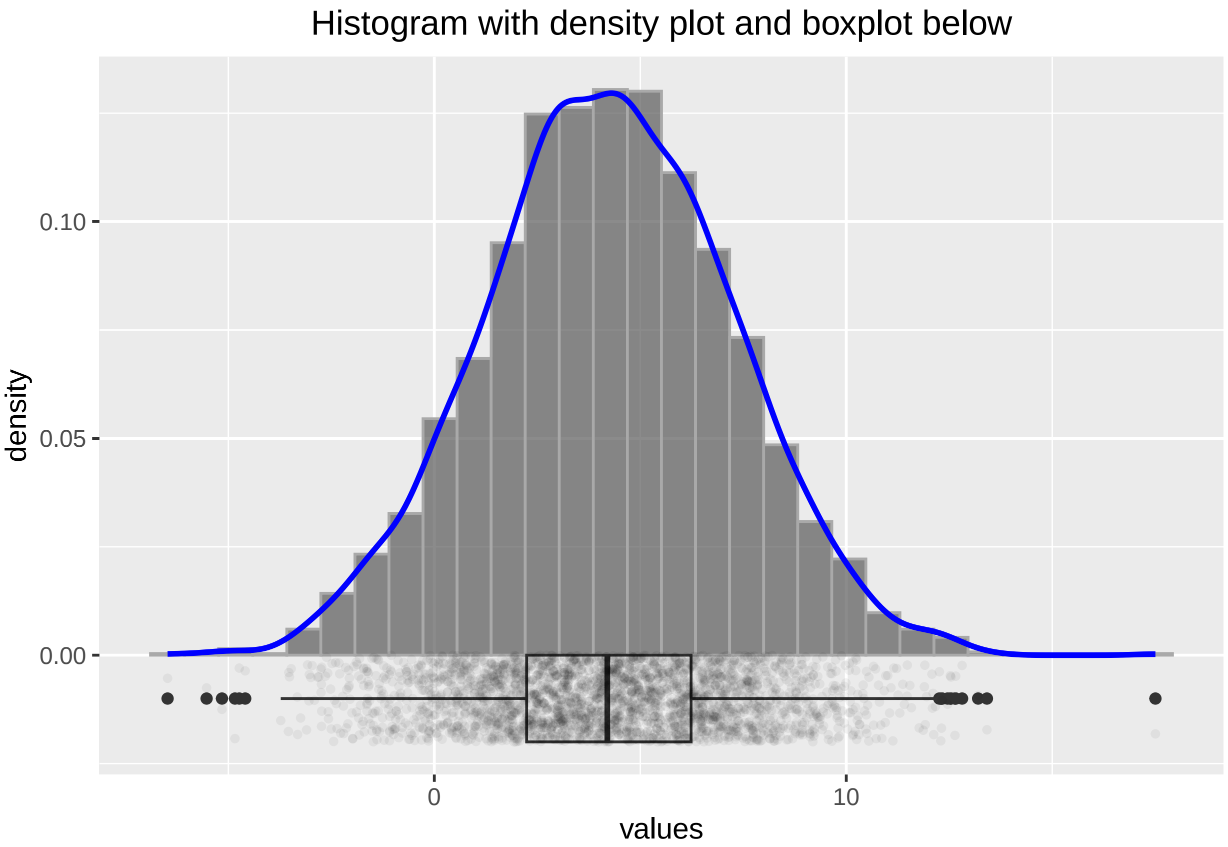
  Doing this for the FPI scores would give (FPI right, see Table 1):

$$z_i = \frac{FPI_i - \overline{FPI}}{SD(FPI)} = \frac{FPI_i - 4.20}{3.00}$$

  $\overline{FPI}$ ist the arithmetic mean of the FPI scores, $SD(FPI)$ the standard deviation of the FPI scores. μ_X is the expectation of X and σ_X the standard deviation of X. So we measure not in FPI units anymore, but in **standard deviations from the mean** which makes it easier to compare different variables with each other. If we would only center the data and define $FPI_{center} = FPI_i - \overline{FPI}$, we would interpret the values in FPI units, but the mean would be 0. Hence an FPI_{center}-value of 2 would mean that the FPI score is 2 units above the mean. A **nice property of the normal distribution** is: If X is normally distributed ($X \sim N(\mu, \sigma)$), then the z-scores are standard normally distributed ($Z \sim N(0, 1)$). See Exercise 4. Of course *both*, z-scores as well as FPI scores are considered "continuous" (in this context).

- With so many obsvervations ($n = 3217$), how would a truly normal distribution with the parameters $FPI \sim N(4.20, 3.00)$ look like? We can simulate this with the following code:

```r
set.seed(8345) # to get the same plot every time
x <- rnorm(3217, mean = 4.20, sd = 3.00)
df <- data.frame(values = x)

p2 <- ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 30, alpha = 0.7, color = "darkgrey") +
  geom_density(aes(y = after_stat(density)), color = "blue", linewidth = 1) +
  geom_boxplot(aes(y = -0.01, x = values), width = 0.02,
               position = position_nudge(y = -0.00)) +
  geom_point(aes(y = -0.01),
             position = position_jitter(width = 0.002, height = 0.01),
             size = 1, alpha = 0.05) +
  ggtitle("Histogram with density plot and boxplot below") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



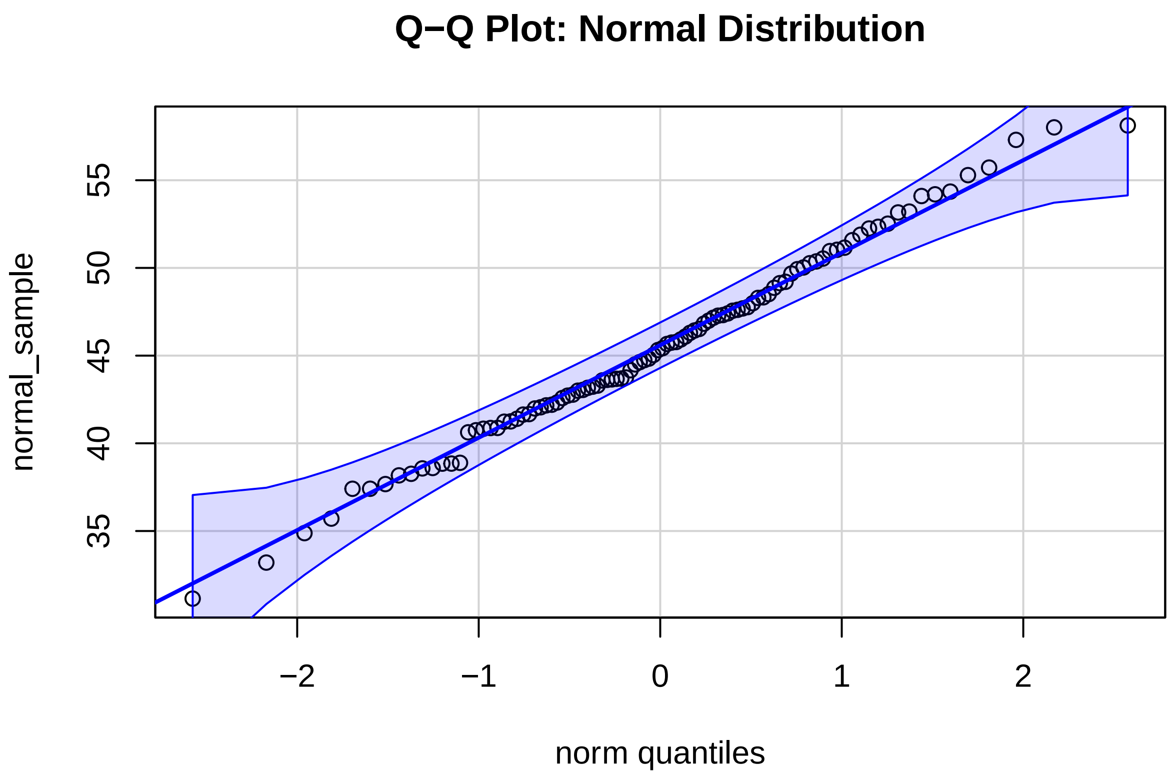
Comparing these histograms (using different seed-values), we could assume that for truly normally distributed FPI scores, the histogram might look smoother than the one in Figure 1 in the paper. Especially the values around 6 seem to deviate from the normal distribution.

Q-Q Plots

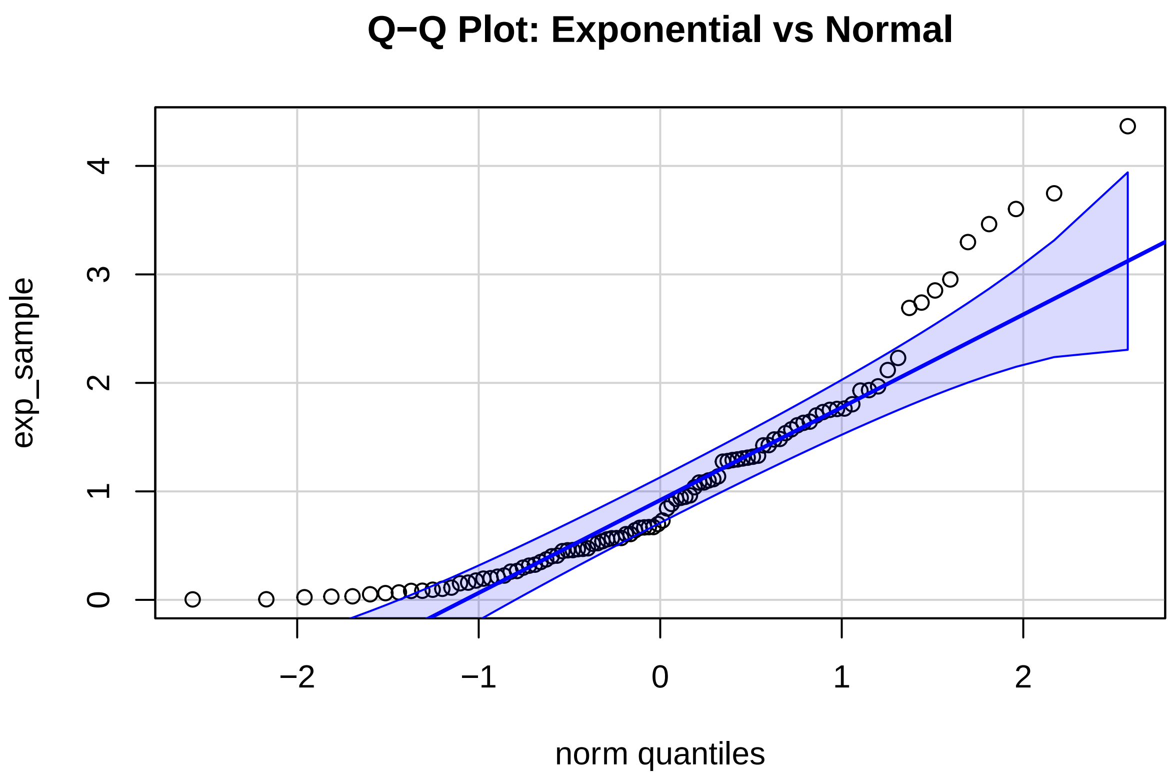
Another way to check for normality (or if the sample is distributed in another way) is the Q-Q plot. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. On the x-axis we plot the quantiles of the theoretical (for instance) normal distribution and on the y-axis the quantiles of the observed data. If the data is normally distributed, the points should lie on a straight line or at least sufficiently close to a confidence band. See exercise 9.

Let's look at multiple examples of normal Q-Q plots which demonstrate the concept:

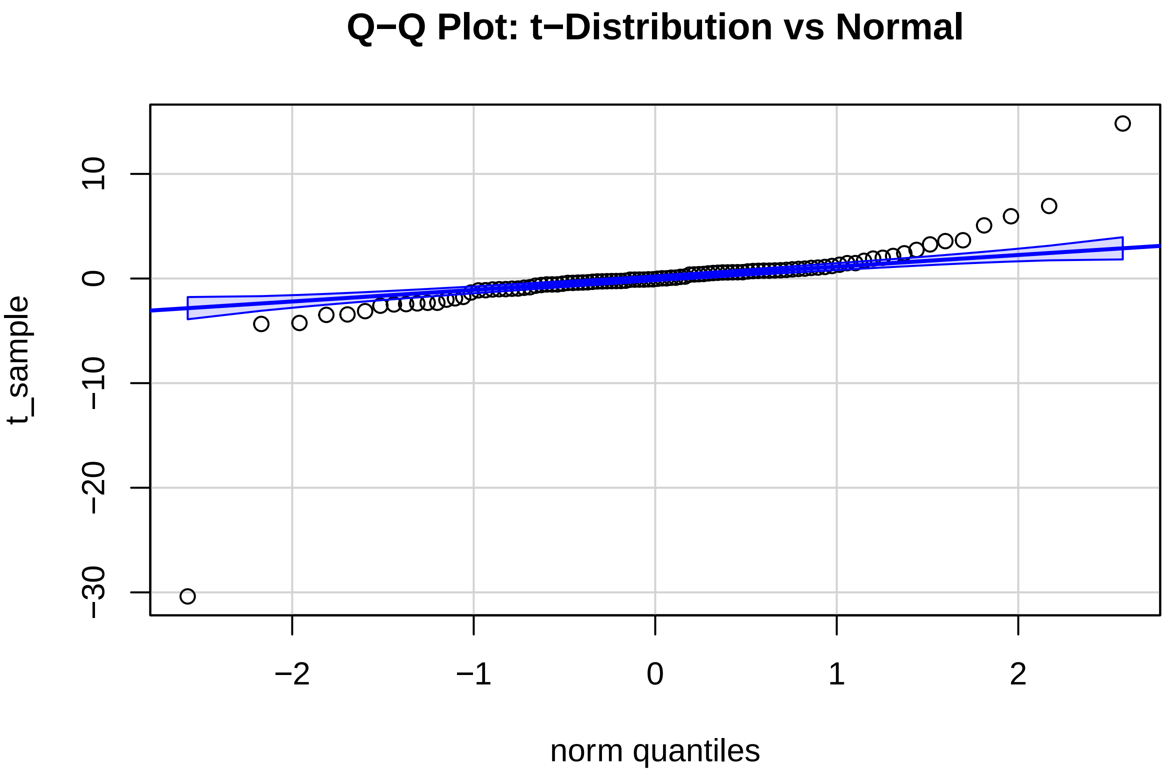
```r
library(pacman)
p_load(car)

# Set seed for reproducibility
set.seed(123)

# 1. Sample from a normal distribution
normal_sample <- rnorm(100, mean = 45, sd = 6)
qqPlot(normal_sample, main = "Q-Q Plot: Normal Distribution", id = FALSE)
```



```r
# 2. Sample from an exponential distribution
exp_sample <- rexp(100, rate = 1)
qqPlot(exp_sample, distribution = "norm",
       main = "Q-Q Plot: Exponential vs Normal", id = FALSE)
```



```r
# 3. Sample from a t-distribution with low degrees of freedom
t_sample <- rt(100, df = 2)
qqPlot(t_sample, distribution = "norm",
       main = "Q-Q Plot: t-Distribution vs Normal", id = FALSE)
```



The **first plot** shows a Q-Q plot of a sample from a normal distribution with $\mu = 45$ and $\sigma = 6$. The points lie on a straight line and are within a confidence band, indicating that the data is normally distributed.

The **second plot** shows a Q-Q plot of a sample from an exponential distribution, which is obviously not normal for a rate (λ) of 1. There are more observations to be expected in the lower range of x compared to normal for instance.

The **third plot** shows a Q-Q plot of a sample from a Student's *t*-distribution with 2 degrees of freedom, which has fater tails than the normal distribution.

Hint: There are many R packages to display Q-Q plots, for instance ggpubr.

Correlation

The correlation between two variables is a measure of the strength and direction of the **linear** relationship between them. It does **not** (directly) measure other kinds of relationships (for instance monotonic or polynomial). See also Anscombe's quartet.

You might play around with this online tool to get a better feeling what correlation is.

Correlation is often denoted by the Greek letter

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

for the population parameter (the true but unknown) where:

- $\text{Cov}(X, Y)$ is the covariance between variables X and Y,
- σ_X and σ_Y are the population standard deviations of X and Y, respectively.

and

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

for the sample estimate.

We can interpret the expression $\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$
Without loss of generality let's assume that $\text{Cov}(X, Y) > 0$. Assuming, one could change these quantities as one wishes:

- ρ is near zero, if the covariance is near zero compared to $\sigma_X \sigma_Y$.
- For fixed $\sigma_X \sigma_Y$, the correlation is larger, the larger the covariance $\text{Cov}(X, Y)$ is.
- For fixed covariance $\text{Cov}(X, Y)$, the correlation is larger, the smaller σ_X or σ_Y (or their product) are.

An interesting and rarely mentioned version of formula for the correlation coefficient is:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Here, we see that the correlation coefficient is the average of the products of the z-scores of the two variables. This preserves the covariance interpretation of adding a positive or negative product to the sum if the obeservation is above or below the respective mean.

Remember: We use r to estimate ρ using the sample.

The correlation coefficient can take values between -1 and 1.

- A value of 1 indicates a perfect positive linear relationship between the variables.
- A value of -1 indicates a perfect negative linear relationship between the variables.

See here and here for a nice illustration. Study this in detail. A correlation of ± 1 just means that the data points lie on a straight line. It does not say anything about the steepness of the line.

You can watch these videos to get a better understanding of the concept:

- StatQuest - correlation
- StatQuest - covariance
- Animated correlation

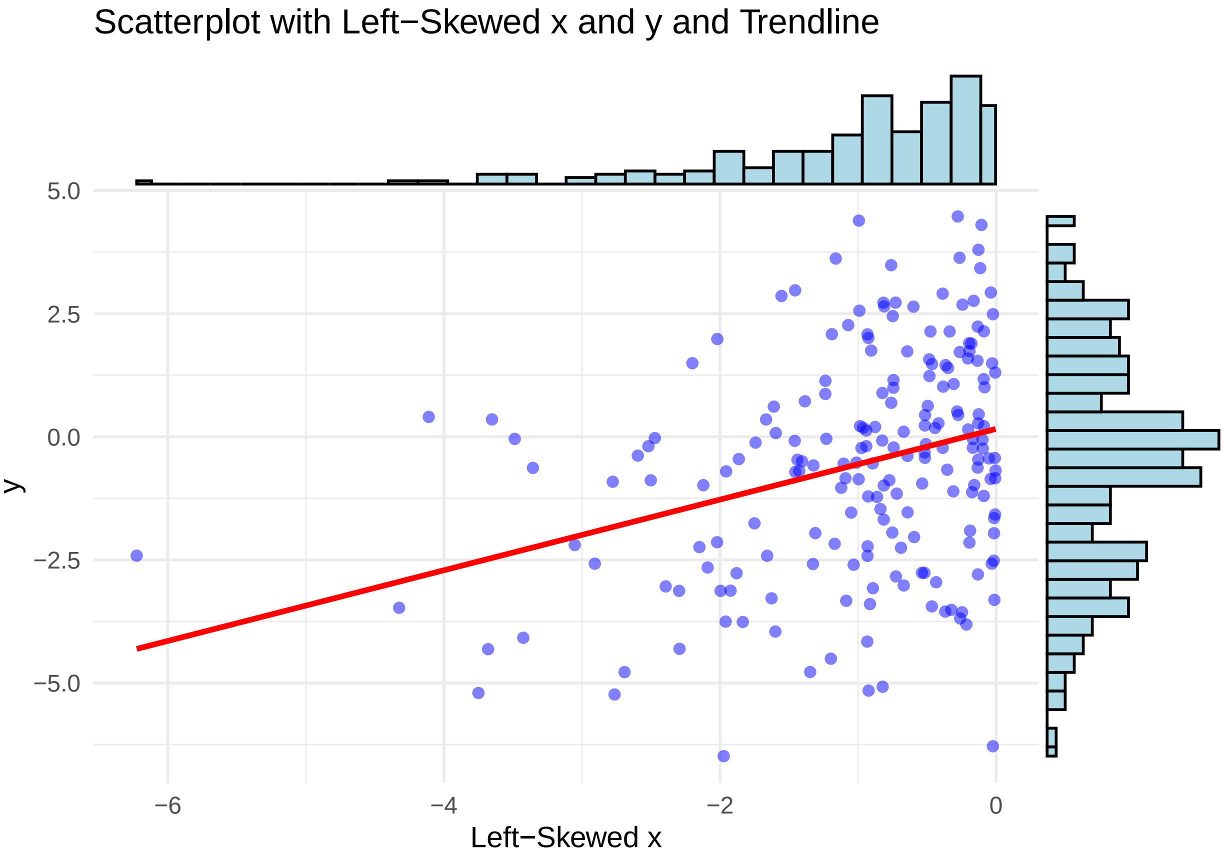
Correlation is an often (over-)used measure in research. It is important to note that correlation does not imply causation. For instance, chocolate consumption is positively correlated with the number of Nobel laureates in a country. This does - unfortunately - not mean that eating chocolate makes you smarter.

One thing many of us think when they see a high correlation is that if we know the value of one variable, we can predict the value of the other variable. This is not necessarily true. This section explains why. Study the visualization here. It depicts the reduction of the prediction interval for Y given X as the correlation increases (assuming that the variables are jointly normally distributed). **How much smaller is the interval where my next observation of Y will fall if I know the value of X?** As you can see, this curve is relatively flat in the beginning, meaning, on an individual level, the correlation does not tell us much about the value of Y given X. Of course with correlation 1, we can predict the value of Y perfectly, but even with a correlation of 0.5 (which is considered high in many areas), the prediction interval for Y is only 13% smaller.

One often reads that the two variables used for calculation of a Pearson correlation coefficient should be normally distributed. This is not necessary, at least not for descriptive purposes. The correlation coefficient is a measure of the linear relationship and we can think of an example with skewed data where the correlation coefficient is still meaningful:

```r
# Load required libraries
library(ggplot2)
library(ggExtra)

# Set seed for reproducibility
set.seed(1234)

# Parameters
n <- 200  # Sample size
beta <- 0.7  # Desired beta

# Generate left-skewed variable x (negative exponential distribution)
x <- -rexp(n, rate = 1)

# Generate correlated left-skewed variable y
error <- rnorm(n, mean = 0, sd = 2)
y <- beta * x + error

# Create a data frame
data <- data.frame(x = x, y = y)

# Create scatterplot with marginal histograms and trendline
p <- ggplot(data, aes(x = x, y = y)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Scatterplot with Left-Skewed x and y and Trendline",
       x = "Left-Skewed x", y = "y") +
  theme_minimal()

# Add marginal histograms using ggExtra
ggExtra::ggMarginal(p, type = "histogram", fill = "lightblue", color = "black")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



If X and Y are bivariate normally distributed, then one can use the variable $t = r\sqrt{\frac{n-2}{1-r^2}}$ as test statistic for the null hypothesis (next chapter) $H_0 : \rho = 0$. Note: It is not sufficient for X and Y to be individually normally distributed in order to be jointly normally distributed.

Example in the wild

Let's again visit the previous paper. In the results, correlations are presented in a dichotomous way with regards to p-values, which should be avoided. We will expand on this in the next chapter. For large sample sizes, even small (true) correlations will be "significant", especially, if one decides to use an arbitrary threshold for the p-values like 0.05. There is not much information in this fact. I would argue that all p-values for such large sample sizes and larger correlations do not carry much information. See also Exercise 5.

That was all nice, what could go wrong?

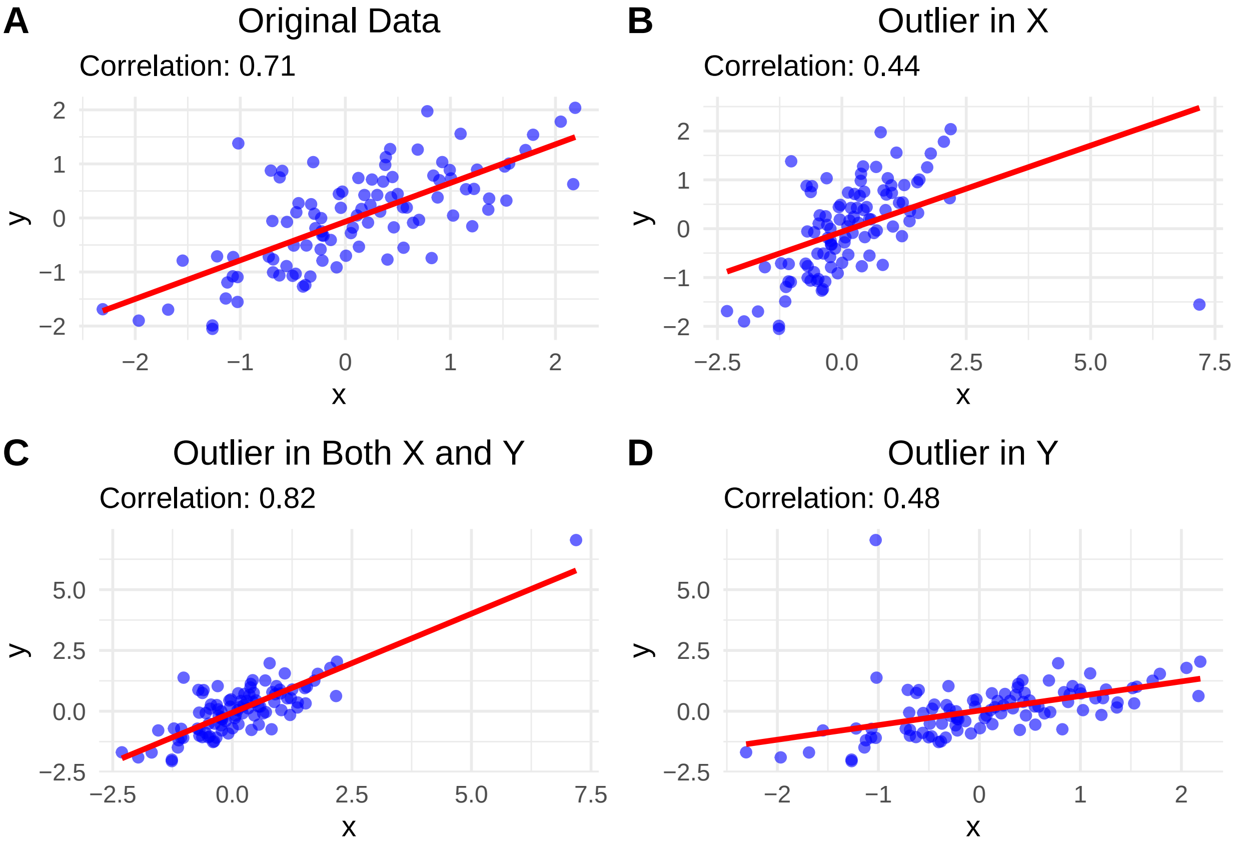
(The sample) Correlation (coefficient) is based (see the formula above) on the artihmetic mean $\bar{X}$ (capital X since we are talking about the random variable which is realised when the X_i have materialized into the x_i) and sample standard deviation s.

The **sample mean is not robust against outliers**. A single large value can distort the mean arbitrarily much. Let's see this in action:

```r
# Load necessary libraries
library(ggplot2)
library(ggpubr)

# Set seed for reproducibility
set.seed(123)
```

```r
# Parameters
n <- 100  # Sample size
rho <- 0.75  # Desired correlation

# Generate bivariate normal data with specified correlation
x <- rnorm(n)
y <- rho * x + sqrt(1 - rho^2) * rnorm(n)
# see e.g. https://www.probabilitycourse.com/chapter5/5_3_2_bivariate_normal_dist.php

# Store original data
data_original <- data.frame(x = x, y = y)

# Function to create scatterplot with trend line and correlation
create_plot <- function(data, title) {
  ggplot(data, aes(x = x, y = y)) +
    geom_point(alpha = 0.6, color = "blue") +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(title = title, subtitle = paste("Correlation:",
                                         round(cor(data$x, data$y), 2))) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))  # Center the title
}

# Original scatterplot
p_original <- create_plot(data_original, "Original Data")

# Add outlier in x
data_outlier_x <- data_original
data_outlier_x$x[n] <- max(x) + 5  # Extreme value in x
p_outlier_x <- create_plot(data_outlier_x, "Outlier in X")

# Add outlier in both x and y
data_outlier_xy <- data_original
data_outlier_xy$x[n] <- max(x) + 5  # Extreme value in x
data_outlier_xy$y[n] <- max(y) + 5  # Extreme value in y
p_outlier_xy <- create_plot(data_outlier_xy, "Outlier in Both X and Y")

# Add outlier in y
data_outlier_y <- data_original
data_outlier_y$y[n] <- max(y) + 5  # Extreme value in y
p_outlier_y <- create_plot(data_outlier_y, "Outlier in Y")

# Arrange plots using ggarrange
ggarrange(p_original, p_outlier_x, p_outlier_xy, p_outlier_y,
          labels = c("A", "B", "C", "D"),
          ncol = 2, nrow = 2)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



- A: $r = 0.71$ for the original data.
- B: When we add an outlier in x, r decreases to 0.44 and the trendline is notably shifted.

  The reason for the smaller correlation is that the outlier is above the mean of the x values and below the mean of the y values and therefore adds negatively to the covariance which is the numerator in the formula for r. The denominator is positive.
- C: When we add an outlier in both x and y, r stays relatively stable. (At least in this case: We could also add, e.g., an outlier above the y mean and below the x mean.) The reason is that the outlier is above the mean of the x values and above the mean of the y values, so it adds positively to the covariance. Also, the outlier is in line with the trend of the data. We could ask, what happend in this data point? Was another scale used for instance?
- D: When we add an outlier in y, the r decreases to 0.48. The reason for the smaller correlation is that the outlier is above the mean of the y values and below the mean of the x values and therefore adds negatively to the covariance.
- We could also ask why the trend lines (created with simple linear regression) change the way they do. We will come to this later in our courses. Please look at this really cool animation. This is the simulation of the regression line when an outlier with distance 8 to the center is introduced and wanders around the clock.

Spearman correlation

The Pearson correlation coefficient is a measure of the **linear relationship between two continuous variables**. Spearman's rank correlation coefficient, denoted by ρ, is a non-parametric measure of the **monotonic relationship** between two variables and can also be applied to ordered categorical variables (levels like "bad", "medium", "best").

It is defined as the Pearson correlation coefficient between the rank variables of the two variables. A rank is the position of an observation in an ordered list of observations. Since only the ranks are used, the Spearman correlation is less affected by outliers.

$$r_s = \rho\left[\mathrm{R}[X], \mathrm{R}[Y]\right] = \frac{\operatorname{cov}\left[\mathrm{R}[X], \mathrm{R}[Y]\right]}{\sigma_{\mathrm{R}[X]}\sigma_{\mathrm{R}[Y]}}$$

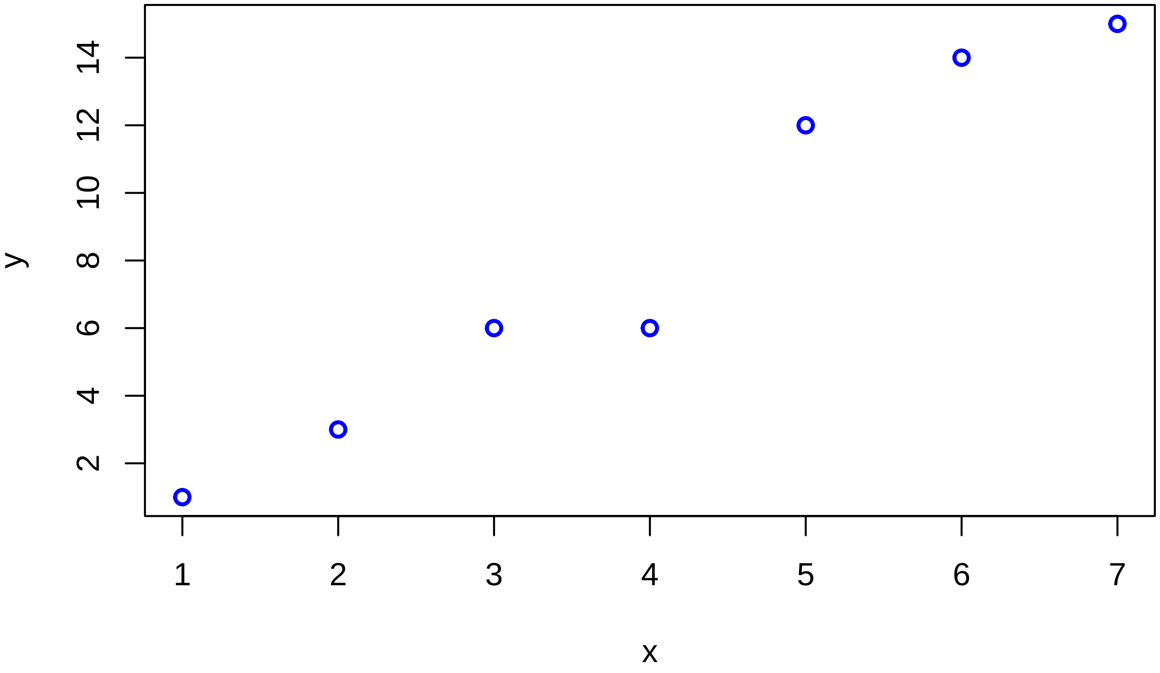
Interpretation: The sign of the Spearman correlation indicates the direction of association between X (the independent variable) and Y (the dependent variable). If Y tends to increase when X increases, the Spearman correlation coefficient is positive. If Y tends to decrease when X increases, the Spearman correlation coefficient is negative.

Here is an example of a scatterplot depicting a non-linear relationship between two variables X and Y. Spearman's rank correlation coefficient equals 1 in this case, since the function is perfectly monotonically increasing. Pearson's correlation coefficient would be < 1 in this case and argueably not appropriate here.

We can easily calculate both and compare them. If they differ noticably, we need to understand why.

Here is an example of a Spearman correlation:

```
set.seed(123)
x <- 1:7
y <- round(2 * x + rt(7, df = 3))
plot(x, y, type = "p", col = "blue", lwd = 2)
```



```
cor.test(x, y, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 10.553, df = 5, p-value = 0.000132
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8553851 0.9969116
## sample estimates:
##       cor
## 0.9782797
```

```
cor.test(x, y, method = "spearman")
```

```
## Warning in cor.test.default(x, y, method = "spearman"): Cannot compute exact
## p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  x and y
## S = 0.50225, p-value = 1.456e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9910312
```

```
# by hand:
rank(x) # already ordered
```

```
## [1] 1 2 3 4 5 6 7
```

```
rank(y) # order of y-values
```

```
## [1] 1.0 2.0 3.5 3.5 5.0 6.0 7.0
```

```
y
```

```
## [1]  1  3  6  6 12 14 15
```

```
cor(rank(x), rank(y), method = "pearson")
```

```
## [1] 0.9910312
```

```
cor(x, y, method = "spearman")
```

```
## [1] 0.9910312
```

The y-values contain two identical values (6), both get rank 3.5. These are so-called ties.

Exercises

Difficulty levels of exercises: **E**: easy, **M**: intermediate, **D**: difficult

[D] Exercise 1 - Recreate table with fake data

- Create fake data for the study mentioned above in R.
- Recreate Table 2 of the paper mentioned above with fake data in R (using GPT, the R package *gtsummary* and other useful packages). This is rather helpful later on in your master thesis.
- Try to export the table to Excel and Word.

[E] Exercise 2 - Outliers and estimates

Let's assume we know that the *Cognitive flexibility* is normally distributed with a mean of 60 and a standard deviation of 7.4 (Table 1): $CognFlex \sim N(60, 7.4)$.

- Draw a sample of 165 persons from this distribution and calculate the mean and standard deviation of the sample. How good is the estimate of the true (and in this case: known) mean and standard deviation?
- Let's replace some of the data points with outliers. Change the score of 5 persons with to impossible CognFlex score of 100. Calculate the mean and standard deviation of the sample. How do the estimates change?
- When we try to estimate the location of our *Cognitive flexibility* distribution with the median, how many outliers of what magnitude are necessary to disturb the estimate by 5 points?

[E] Exercise 3 - Recreating data in Table 2

We assume that *age* in both groups is normally distributed with a mean of 48 (47) years and a standard deviation of 16 (15) years:

$Age_{UsualCare} \sim Normal(\mu = 48, \sigma = 16)$ and

$Age_{CFT} \sim Normal(\mu = 47, \sigma = 15)$.

- Under these assumptions, what is the probability, that we would see a person of age 60 or older in a new sample (in either group)?
- What is the probability, that we would see a person of age 18 or younger in a new sample?
- Give a 99% interval for the age in CFT, where we would expect a new person drawn from the same population.

Let's assume *Sex* is binomially distributed with a probability of $p = 0.59$ for both groups.

- What is the probability, that we would see a woman as the next recruited person in the *Usual care* group?
- What is the probability, that we would see no man in a sample of 10 persons in the CFT group?
- Sometimes you want balanced samples. How many patients would we need to recruit to get at least 45 women with a probability of at least 90%. (We could solve this with simulation.)

[E] Exercise 4 - Z-scores

- Show with a simulation that the z-scores are standard normally distributed if the original variable is normally distributed.
- Try different parameter values for μ and σ and plot the histogram of the z-scores.
- Optional: Try to prove this mathematically.

[M] Exercise 5 - Correlation

The following R-code creates correlated samples with a (true) correlation ρ:

```r
set.seed(1234)
n <- 1000
rho <- 0.5
x <- rnorm(n)
y <- rho * x + sqrt(1 - rho^2) * rnorm(n)
cor(x, y)
```

```
## [1] 0.541743
```

- Try different values for ρ and n and plot the scatterplot of x and y.
- Execute the code above 1000 times and save the correlation coefficient in a vector.
- Calculate the sample quantiles and interquantile range of the simulated correlation coefficients. What do you observe with respect to variability?
- Plot the histogram of the simulated correlation coefficients.

[M] Exercise 6 - Bike parking locations in Switzerland

- Download the data from the bike parking locations json-file and execute the following code to see an interactive map of the bike parking locations in Switzerland (adapt file path if necessary):

```r
library(pacman)
p_load(jsonlite, tidyverse, leaflet)

bike <- fromJSON("./DATA/bike_parking.json")

# Extract the "features" data frame
features <- bike$features

# Extract and flatten geometry (coordinates)
coordinates <- do.call(rbind, lapply(features$geometry$coordinates, as.numeric))
colnames(coordinates) <- c("longitude", "latitude")

bike_data <- cbind(coordinates, features$properties)
head(bike_data)
```

```
##   longitude latitude                  name stopPlaceUic stopPlaceSloid
## 1  8.902417 47.48863 Veloparking überdacht      8506013             NA
## 2  8.902425 47.48814 Veloparking überdacht      8506013             NA
## 3  8.902277 47.48854 Veloparking überdacht      8506013             NA
## 4  8.050095 47.39113 Veloparking überdacht      8502113             NA
## 5  8.050337 47.39128           Velostation      8502113             NA
## 6  7.278020 47.04381 Veloparking überdacht      8504404             NA
##   source.name                            source.id category  subCategory
## 1       prail 2f767602-03fd-49e1-bca6-7fa60e9df655  parking bike_parking
## 2       prail 4d98b698-cd5b-4b40-abb4-004827e26e5b  parking bike_parking
## 3       prail 75dc4e71-5aec-4d58-b4f4-ef815f8f5f40  parking bike_parking
## 4       prail 51585d0e-ea2a-4763-a4e7-669b89f24f0b  parking bike_parking
## 5       prail 9caf1c56-5cac-423b-912a-5e551b4c69a5  parking bike_parking
## 6       prail 3e7aa3c8-4245-488b-9f00-b44ee0d8b79a  parking bike_parking
##   address_discriminatorType address_city address_postalCode capacity
## 1                   Address       Aadorf               8355       48
## 2                   Address       Aadorf               8355       48
## 3                   Address       Aadorf               8355       80
## 4                   Address        Aarau               5000      186
## 5                   Address        Aarau               5000      230
## 6                   Address      Aarberg               3270       72
##   properties_bookingSystem
## 1         {'type': 'NOVA'}
## 2         {'type': 'NOVA'}
## 3         {'type': 'NOVA'}
## 4         {'type': 'NOVA'}
## 5         {'type': 'NOVA'}
## 6         {'type': 'NOVA'}
##
## 1  [{'categoryType': 'STANDARD', 'total': 48}, {'categoryType': 'DISABLED_PARKING_SPACE', 'total': 0
## 2  [{'categoryType': 'STANDARD', 'total': 48}, {'categoryType': 'DISABLED_PARKING_SPACE', 'total': 0
## 3  [{'categoryType': 'STANDARD', 'total': 80}, {'categoryType': 'DISABLED_PARKING_SPACE', 'total': 0
## 4 [{'categoryType': 'STANDARD', 'total': 186}, {'categoryType': 'DISABLED_PARKING_SPACE', 'total': 0
## 5 [{'categoryType': 'STANDARD', 'total': 230}, {'categoryType': 'DISABLED_PARKING_SPACE', 'total': 0
## 6  [{'categoryType': 'STANDARD', 'total': 72}, {'categoryType': 'DISABLED_PARKING_SPACE', 'total': 0
##   properties_operator properties_parkingFacilityCategory
## 1                 SBB                               BIKE
## 2            Sonstige                               BIKE
## 3                 SBB                               BIKE
## 4            Sonstige                               BIKE
## 5            Sonstige                               BIKE
## 6                 SBB                               BIKE
##   properties_parkingFacilityType
## 1           BIKE_PARKING_COVERED
## 2           BIKE_PARKING_COVERED
## 3           BIKE_PARKING_COVERED
## 4           BIKE_PARKING_COVERED
## 5                   BIKE_STATION
## 6           BIKE_PARKING_COVERED
##
## 1
## 2
## 3
## 4
## 5 [{'bikeFacilityTraitId': '552fd1ef-bbc7-4eb7-8ef4-12901aff5250', 'bikeFacilityTraitNameI18n': {'de
## 6
##                                                                                        openingHours
## 1                                                                                              <NA>
## 2                                                                                              <NA>
## 3                                                                                              <NA>
## 4                                                                                              <NA>
## 5 [{'slots': [{'weekdayFrom': 'Mo', 'weekdayTo': 'Su', 'timeFrom': '00:00', 'timeTo': '23:59'}]}]
## 6                                                                                              <NA>
##   address_streetName address_houseNumber
## 1               <NA>                <NA>
## 2               <NA>                <NA>
## 3               <NA>                <NA>
## 4               <NA>                <NA>
## 5       Bahnhofplatz                  3a
## 6               <NA>                <NA>
##
## 1
## 2
## 3
## 4
## 5 {'operatingFrom': '00:00:00', 'operatingTo': '00:00:00', 'daysOfWeek': ['MONDAY', 'TUESDAY', 'WEDN
## 6
##   properties_pricingModel
## 1                    <NA>
## 2                    <NA>
## 3                    <NA>
## 4                    <NA>
## 5                    <NA>
## 6                    <NA>
```

```
bike_cleaned <- bike_data %>%
  dplyr::select(longitude, latitude, name,
              address_city, address_postalCode, capacity, properties_operator)
head(bike_cleaned)
```

```
##   longitude latitude                  name address_city address_postalCode
## 1  8.902417 47.48863 Veloparking überdacht       Aadorf               8355
## 2  8.902425 47.48814 Veloparking überdacht       Aadorf               8355
## 3  8.902277 47.48854 Veloparking überdacht       Aadorf               8355
## 4  8.050095 47.39113 Veloparking überdacht        Aarau               5000
## 5  8.050337 47.39128           Velostation        Aarau               5000
## 6  7.278020 47.04381 Veloparking überdacht      Aarberg               3270
##   capacity properties_operator
## 1       48                 SBB
## 2       48            Sonstige
## 3       80                 SBB
## 4      186            Sonstige
## 5      230            Sonstige
## 6       72                 SBB
```

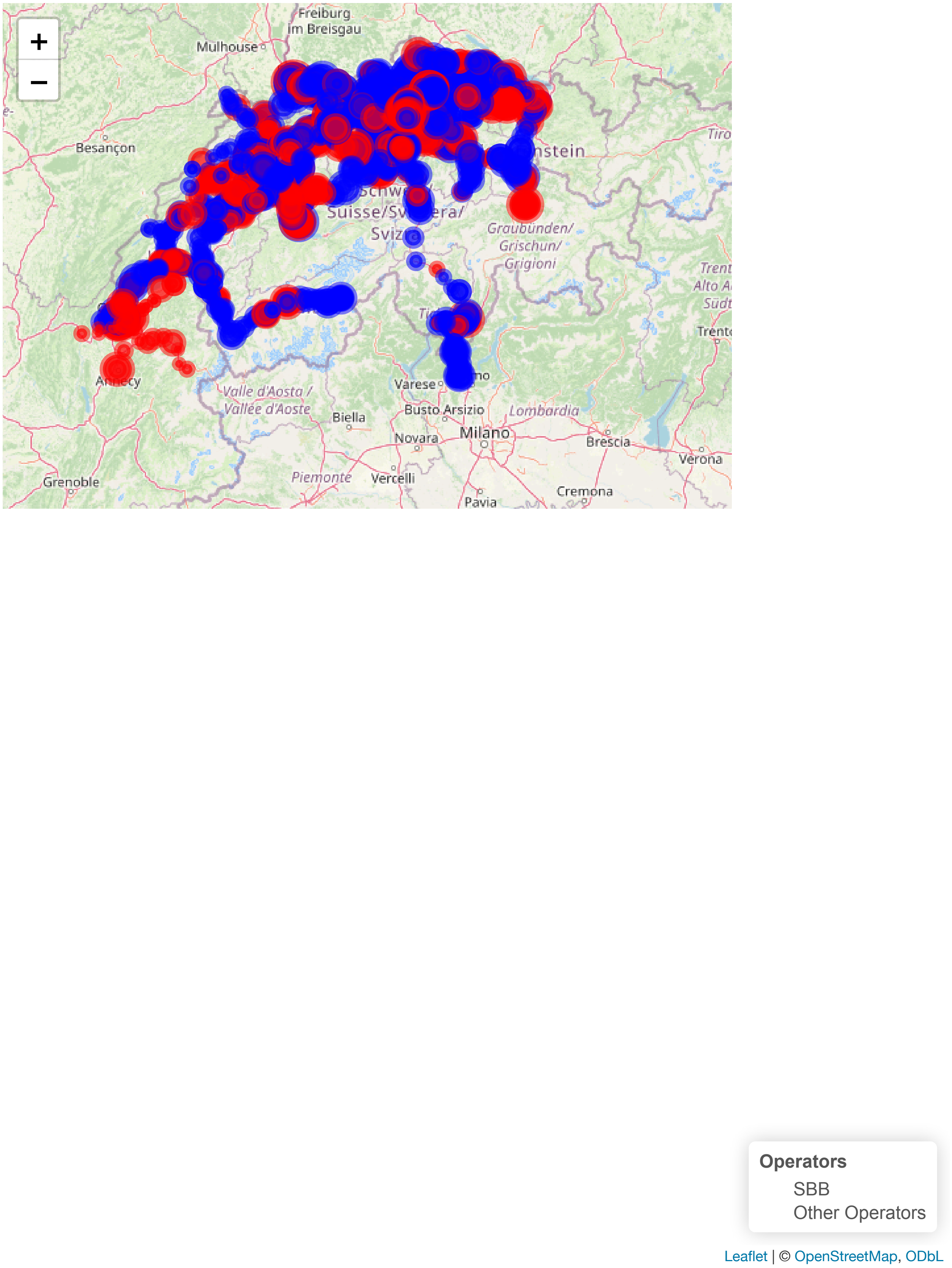
```
# Create an interactive map
# Check the structure of capacity
str(bike_cleaned$capacity)
```

```
##  chr [1:1394] "48" "48" "80" "186" "230" "72" "22" "45" "53" "44" "120" ...
```

```
# Convert capacity to numeric
bike_cleaned$capacity <- as.numeric(bike_cleaned$capacity)

# Check for warnings
if (any(is.na(bike_cleaned$capacity))) {
  warning("Some capacities could not be converted to numeric.
  Non-numeric values have been set to NA.")
}
```

```
# Retry the map creation
leaflet(bike_cleaned) %>%
  addTiles() %>%  # Add OpenStreetMap tiles
  addCircleMarkers(
    ~longitude, ~latitude,
    popup = ~paste(
      "<strong>Name:</strong>", name,
      "<br><strong>City:</strong>", address_city,
      "<br><strong>Capacity:</strong>", capacity,
      "<br><strong>Operator:</strong>", properties_operator
    ),
    radius = ~log(capacity + 1) * 2,  # Radius scaled by capacity
    color = ~ifelse(properties_operator == "SBB", "blue", "red"),
    fillOpacity = 0.7
  ) %>%
  addLegend("bottomright", colors = c("blue", "red"),
            labels = c("SBB", "Other Operators"),
            title = "Operators")
```

```
## file:////private/var/folders/pm/jd6n6gj10371_bml1gh8sc5w0000gn/T/RtmpRj7KUC/fileb7ea3195c1dc/widgetb
```



Let's look at the data frame *bike_cleaned* in R.

- In what city is the bike parking with the highest capacity located?
- Which cities are represented in the dataset?
- How many bike parkings are operated by the SBB (Swiss Federal Railways)? Create a bar plot to visualize this.
- What is the average capacity of bike parkings in the dataset? Draw a histogram and a boxplot below.
- What are the dimensions of the data set *bike_cleaned*?

[E] Exercise 7 - Median, Mean, and Mode

- Find a skewed continuous probability distribution and plot the theoretical density.
- Draw a sample from this distribution and calculate the mean, median, and mode. Plot a histogram with vertical lines for the mean, median, and mode. How do these values compare to each other and why do they look like this considering the definition of each?
- Find a skewed variable in a research paper.

[E] Exercise 8 - Correlation by hand

- Calculate the correlation coefficient by hand for the following data points and compare with the R-function *cor()*:

```r
library(pacman)
p_load(tidyverse)

set.seed(123) # Für reproduzierbare Ergebnisse
x <- c(1, 2, 3, 4, 5)
y <- c(1.5, 2.5, 3.5, 4.5, 5.5) + rnorm(5, 0, 2) # Rauschen hinzufügen
```

[E] Exercise 9 - Q-Q plot

- Create a Q-Q plot for a sample from a normal distribution and compare it with the theoretical quantiles.
- Repeat this multiple times again and again to get a feeling how truly normally distributed data looks like.

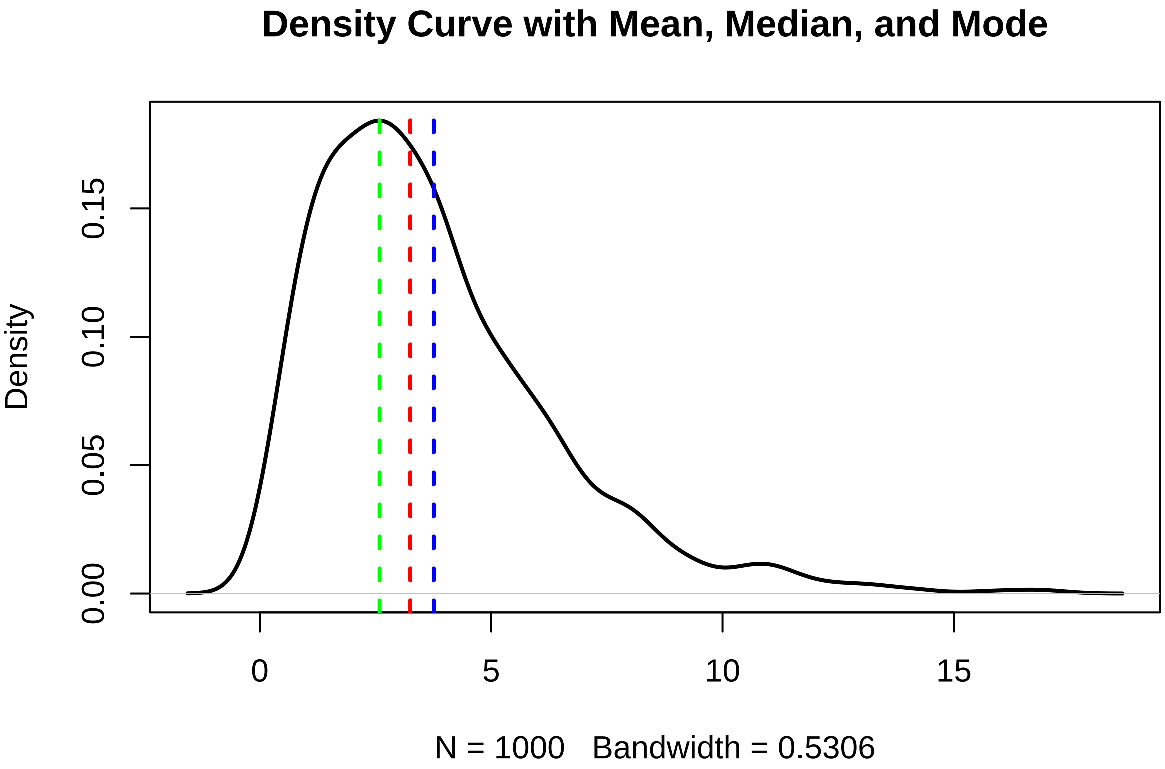
Solutions

Solutions for this chapter can be found here.

Sample exam questions for this chapter (in German since exam is in German)

Frage 1 - Mittelwert, Median, Modus

Gegeben sei ein Histogramm einer Stichprobe ($n = 1000$). Im Bild zu sehen ist die aus der Stichprobe geschätzte Wahrscheinlichkeitsdichte. Die 3 vertikalen strichlierten Linien zeigen den Mittelwert, Median und Modus der Stichprobe an (nicht notwendigerweise in der genannten Reihenfolge).

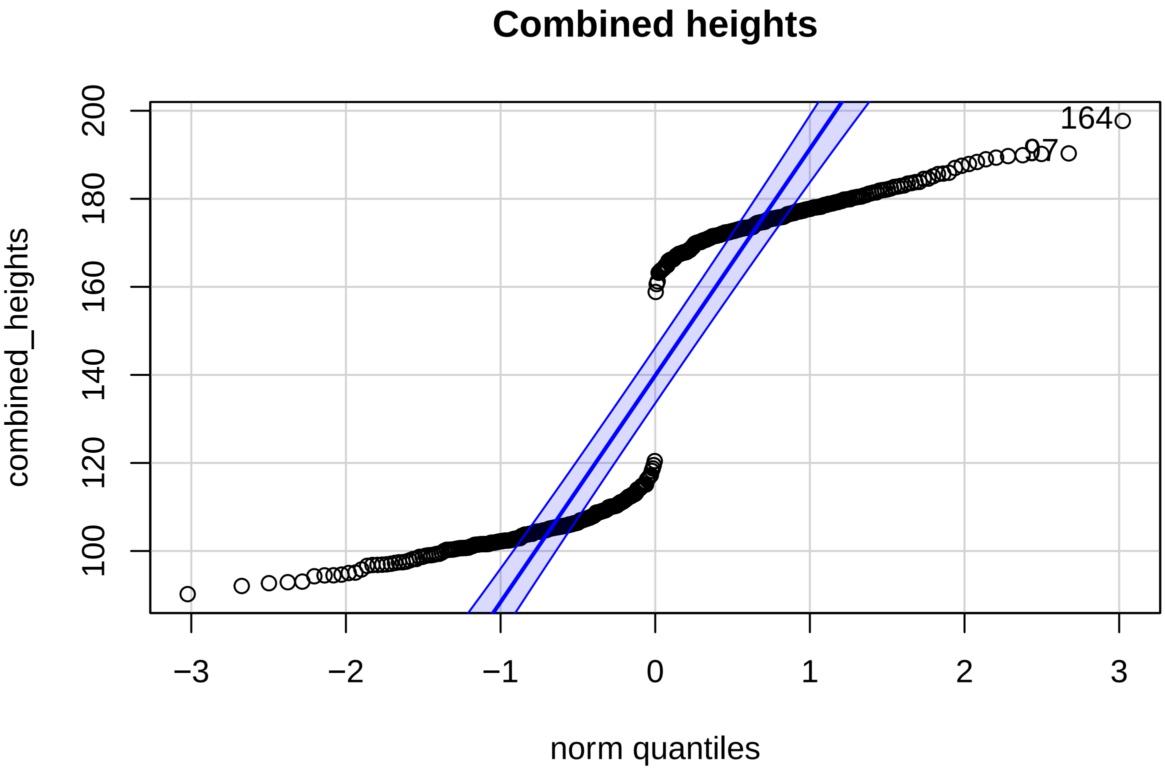


Welche der folgenden Aussage(n) ist/sind korrekt (0-4 korrekte Antworten)?

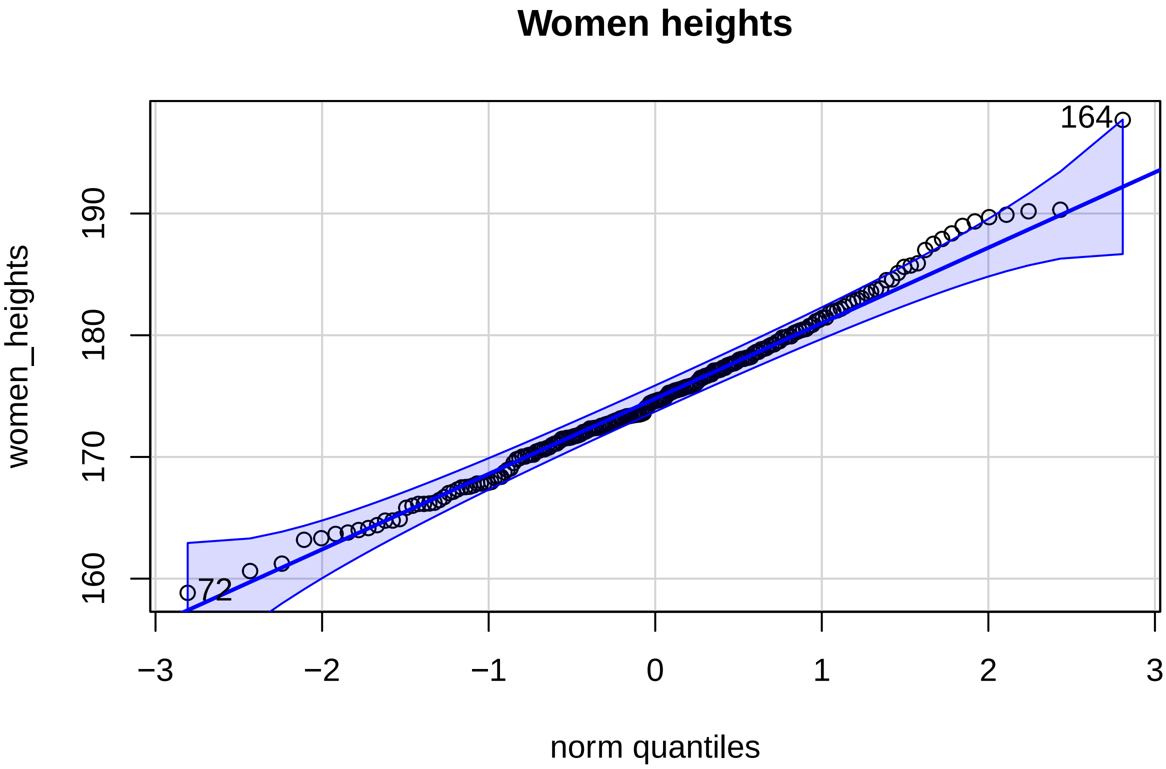
- Der Mittelwert ist größer als der Median.
- Der Modus ist größer als der Median.
- Die Daten wurden höchstwahrscheinlich aus einer symmetrischen Verteilung gezogen.
- Es handelt sich sehr wahrscheinlich um eine bimodale Verteilung.

Frage 2 - Normalverteilung

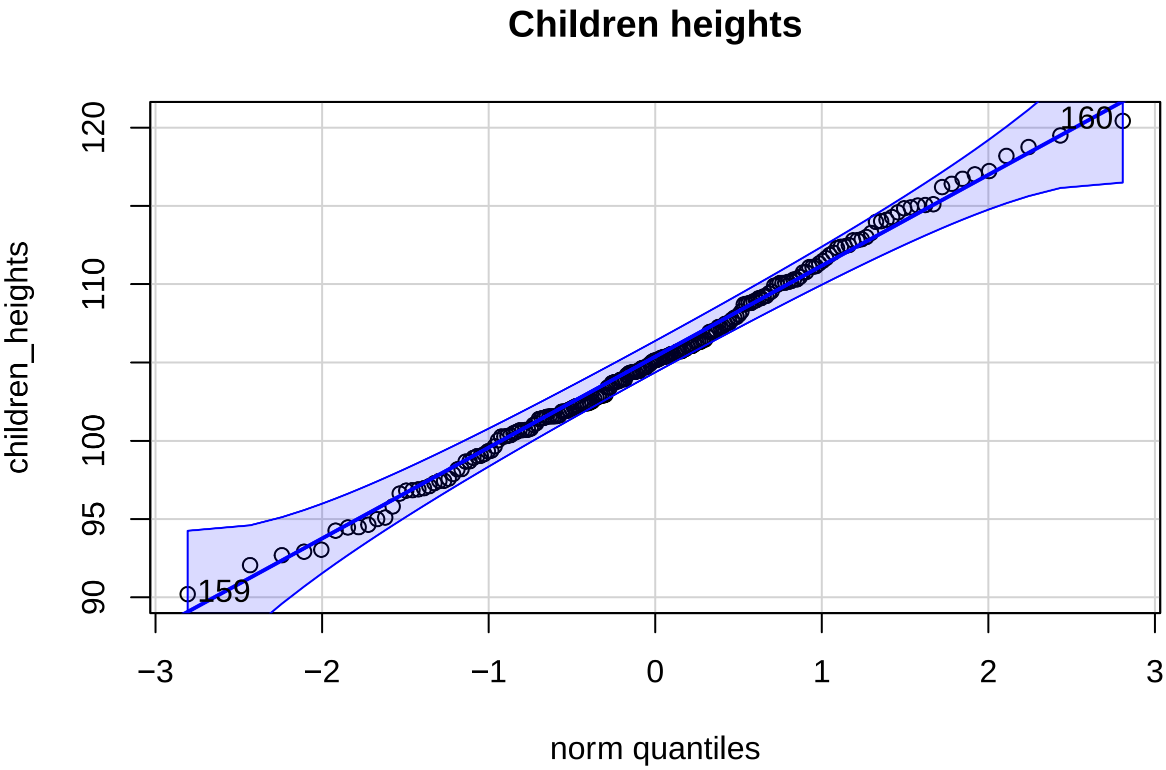
Es seien zwei Stichproben von Körpergrößen gegeben. 200 von Frauen, 200 von Kindern. Zu sehen sind 3 Q-Q-Plots.



```
## [1] 164  97
```



```
## [1] 164  72
```



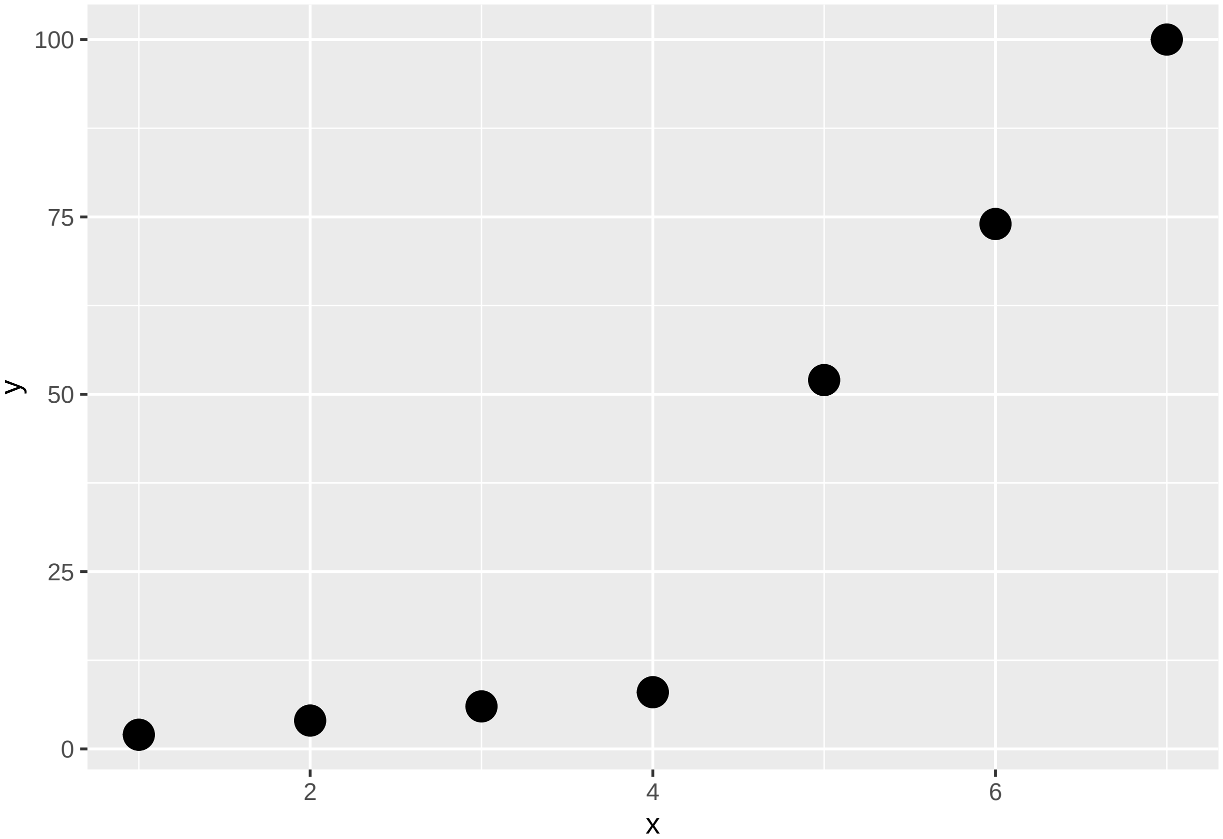
```
## [1] 160 159
```

Welche der folgenden Aussage(n) ist/sind korrekt (0-4 korrekte Antworten)?

- Der Q-Q-Plot der kombinierten Körpergrößen lässt auf eine Normalverteilung schließen.
- Der Q-Q-Plot der Frauen-Körpergrößen ist kompatibel mit einer Normalverteilung.
- Der Q-Q-Plot der Kinder-Körpergrößen ist kompatibel mit einer Normalverteilung.
- Das Histogramm der Kinder-Körpergrößen zeigt eine bimodale Verteilung.

Frage 3 - Korrelation

Gegeben sei folgende Stichprobe von Punktepaaren (x_i, y_i) der Variablen X und Y:



Welche der folgenden Aussage(n) ist/sind korrekt (0-4 korrekte Antworten)?

- Pearson’s Korrelationskoeffizient ist geeignet um den Zusammenhang zwischen X und Y zu messen.
- Pearson’s Korrelationskoeffizient ist gleich 1.
- Spearman’s Rangkorrelationskoeffizient ist < 1.
- Spearman’s Rangkorrelationskoeffizient misst die Stärke des linearen Zusammenhangs zwischen X und Y.

Frage 4

Gegeben sei folgende Stichprobe: $x = (9.83, 13.87, 5.90, 10.56, 6.95, 9.46)$.

Welche der folgenden Aussage(n) ist/sind korrekt (0-4 korrekte Antworten)?

- Median und Mittelwert sind identisch.
- Ein Modus ist 9.83.
- Die Stichprobenstandardabweichung ist 2.420024.
- Der Korrelationskoeffizient beträgt 0.78.