Exercise_05_Hypothesis_testing

Zoltan Kekecs

November 09, 2020

Hypothesis testing

Abstract

This exercise aims to introduce the basic concepts related to **hypothesis testing**. During this exercise we will learn about the most commonly used statistical test to test univariate and bivariate hypotheses (testing the effect/relationship between two variables).

Loading packages

```
library(tidyverse) # for dplyr and ggplot2
```

Null hypothesis significance testing, and its flipped logic

During hypothesis testing our aim often is to test establish the existence of an effect or relationship between variables. In other words, what is the probability that there is an effect. However, during **null-hypothesis significance testing (NHST)** we don't get an answer to this question directly. Rather, we apply a backward logic. We calculate what is the **probability of seeing the observed trend (or even more extreme trend) if the null hypothesis is true**.

A simple example: we suspect that **a coin is biased** in a way that when flipped it is more likely to show "heads" than "tails"). If I want to test whether the coin is biased with NHST, I first set up a **null-hypothesis**: in this case that the coin is fair (flipping "heads" and "tails" have the same probability).

- H1: the coin is biased (toward "heads")
- H0: the coin is fair

Let's say that we flip the coin and get 9 "heads" and 1 "tail". What is the probability that the coin is biased? We don't know. We cannot tell, partially because we did not specify "how much" bias are we looking for. However, we can easily tell what is the probability of getting this or even more extreme results (even more heads), if the coin is **fair**. It just requires a bit of probability theory.

Without going into the details, we can calculate with the code below that the probability of getting 9 or more heads out of 10 coin flips if the probability of heads is 0.5 (50%), is p = 0.0107 (**about 1% chance**).

```
1-pbinom(9-1, 10, 0.5)
```

```
## [1] 0.01074219
```

In other words if we kept repeating the same experiment many times (each experiment containing 10 coinflips) with fair coins, in only 1% of the experiments would we get 9 or more heads.

We can test this in R by running a simulation of a lot of similar "experiments". We can randomize a coinflip-like binomial outcome (0 or 1) using the **rbinom()** function. With the code below we randomize the outcome of 10 coinflips 10000 times, with the probability of getting 1 or 0 being equial: p = 0.5 (50%). In

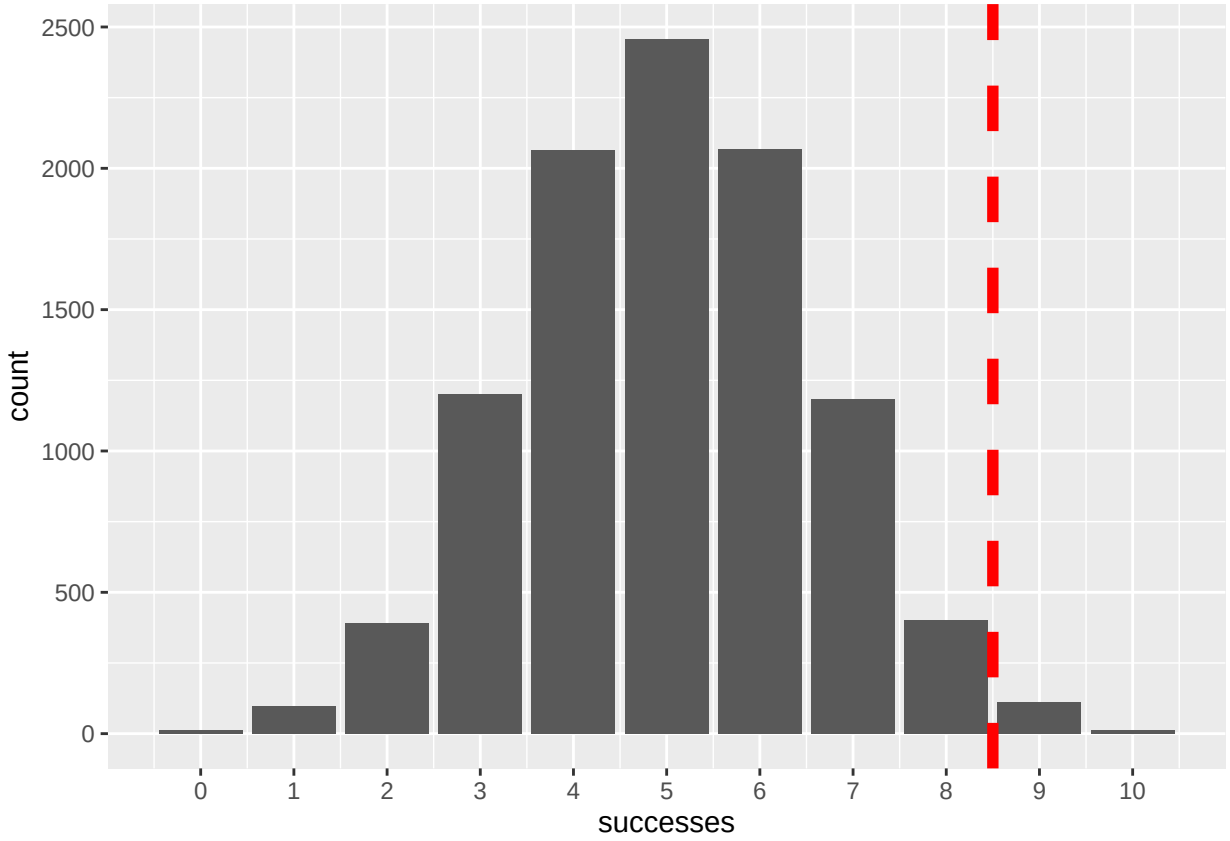
the plot we can see how many of the experiments ended up producing 9 or more ones ("heads"). It is close to the above calculated theoretically expected probability.

```
successes = rbinom(n = 10000, size = 10, prob = 0.5)
random_flips = data.frame(successes)

sum(successes > 9)/10000 # proportion of experiments with 9 or more "heads"
```

```
## [1] 0.0013
```

```
ggplot(data = random_flips) +
  aes(x = successes) +
  geom_bar()+
  scale_x_continuous(breaks = 0:10)+
  geom_vline(xintercept = 8.5, col = "red", linetype = "dashed", size = 2)
```



Furthermore, this value is equivalent to the p-value returned by a statistical test testing the "fairness" of the coin, in this case, the Binomial test (notice that the p-value returned by the test is the same as we calculated above using pbinom()).

```
binom.test(x = 9, n = 10, p = 0.5, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  9 and 10
## number of successes = 9, number of trials = 10, p-value = 0.01074
## alternative hypothesis: true probability of success is greater than 0.5
```

```
## 95 percent confidence interval:
##  0.6058367 1.0000000
## sample estimates:
## probability of success
##                    0.9
```

All in all, getting 9 or more heads is pretty **surprising** (rare) from 10 flips, if the coin is fair. But what does this mean regarding whether the coin is **actually fair or not**? We don't actually know. We only know **how rare it is to see the result that we got, if we assume that the coin is fair**. This is the backward logic of the NHST you need to understand.

Let's go further now and let's assume that we need to **make a decision** about whether we consider the coin to be fair (for example for a "coin flipping contest") or whether we should believe that the coin is biased. This is the **test** part in the NHST. We usually make this decision based on a pre-specified probability threshold. If the result we saw is **surpirising enough**, **rare enough** assuming the truth of the null hypothesis, we might consider to abonden the assumption that the null hypothesis is true (and with the process of elimination, the only thing we are left with is the alternative hypothesis).

In psychology, we usually set this decision threshold at **lower than 5% probability** ($p < 0.05$). If we see a probability (a p-value) lower than this, we say that we **reject the null**.

It is important to realize thought that we do not get information from the test about the true probability of the null or the alternative hypotheses being true in any single test. We can only know how probable or improbable the result we saw is "in a world where H0 is true". Nothing more and nothing less.

The benefit of using the NHST framework for statistical decision making is that **if we use our decision threshold consistently**, we can be **fairly certain** that we only reject the null falsely in 5% of the times **when the null was actually true**.

Please notice two things though: One is that I said **"fairly certain"** because the assumptions of our statistical tests must hold true, and we have to account for multiple testing. The other is that I said **"when the null was actually true"**. This is important to realize that the benefit that only 5% of our rejections are false does not mean that we are only wrong in 5% of our statistical decisions. This is because this also depends on the **base-rates of correct hypotheses**.

How many times are we actually right when making statistical decisions?

Consider for example of lie detectors. Lie detectors have a pretty high sensitivity, meaning that they tend to correctly identify liers in a high percentage of cases. Lets assume that we have a lie detector with a 95% sensitivity, and that the court uses decisions of lie detector operators in court rulings, so this means that we will catch 95% of the criminals, and only 5% of the criminals will get away unpunished (false negatives). This is all great and some might argue that we should use lie detectors more often in court cases. But notice that so far we only talked about criminals. What about the innocents? It turns out that while lie detectors have a great sensitivity, they tend to have a terrible specificity. That is, about 50% of non-liers are also detected as liers (false-positives) by the lie detector test. So how many times will our court rulings be correct if we rely on lie detectors? It depends on how many criminals and how many innocents we do the test with. If we mostly do the tests with criminals, we will be right most of the times. If we mostly do the test with innocents, we will be wrong most of the times. It is easy to see that if we blanket-test the population with lie detectors, and the actually proportion of criminals in the population is very low, we will end up with mostly innocents in the jails.

The same is true for NHST. NHST has a 5% false positive rate (false rejection of the null), but research studies tend to be powered to only have 80% of power (this means 20% false negative rate) in theory. So how many of our statistical decisions are correct then? As in the previous example, this depends on the base-rate. Of how often we do statistical tests of hypotheses which are actually true.

Lets say that the alternative hypothesis is true in 50% of the cases. So from 2000 hypotheses that we test, in 1000 H0 is true, and we will reject the null falsely in 5% of them, so in 50 of these studies we will (wrongly)

say that the hypothesis is true. In the rest of the studies, assuming 80% power, we will correctly reject the null in 800 studies. So in total we rejected the null in 850 studies, so in total, we our decisions to reject the null was incorrect in 50/850=0.0588 proportion of the null-rejection decisions (about 6%). Also, we retained the null hypothesis in 950 of the times when the null was actually correct, and we falesely retained the null in 200 studies where actually the alternative hypothesis was true. So our decision to retain the null was incorrect in 200/1150=0.174 proportion (17%) of the decisions in which we retained the null. In total, 250/2000 = 0.125, that is, 12.5% of our statistical decisions were incorrect.

But lets assume now that we (scientists) are not that good at coming up with good hypotheses. What if only 10% of the hypotheses that we test in research are actually true? If we crunch the numbers, we can see that in this case, from 2000 research studies we will end up with 90 false rejections of the null and 160 true rejections of the null, thus, we wrongly rejected the null in 36% of all the null-rejection decisions. This is pretty bad. You can say that OK, but in return only incorrectly retained the null in 2% of our null-retention decisions, which lands us at about only 6.5% of our statistical decisions being incorrect overall. That is true, but note that:

- the using a strict 5% decision threshold did not prevent us from making a lot of mistakes when we rejected nulls.
- the actual power in real research is much lower than the usually desired 80%. Some researchers estimate that depending on the field in question, the actual power is closer to 30-50%.
- there is a huge publication bias where positive results (rejecting the null) are overwhelmingly more easy to get published than negative results (retaining the null), so we mostly just see the null-rejection decisions, which means that out of all of the **published** statistical decisions, the actual rate of incorrect decisions is much higher overall.
- we did not factor in assumptions not holding true and multiple testing as I mentioned above, which might make things worse
- we don't actually know the true base-rate of "good ideas", so how many of our hypotheses baing tested in research are actually true. if most hypotheses we test are wrong, the published decisions can be wrong in even more times
- people's careers depend on publication, so researchers are incentivised to "get positive results", because that will increase their chances of being published, thus, there is a certain amount of fraud and questionable research practices in the field, that introduce bias.

In summary, unfortunaly it is a very real **possibility that most of the published statistical decisions are wrong** in the current literature. Recent multi-lab replication projects confirm this grim suspicion, where only about 40-60% of the findings published in top scientific journals could be replicated in high powered replication attempts.

How to do the most common statistical tests in R

As discussed above, statistical tests are designed to compute a p-value. In this class we will discuss the following statistical tests:

- binomial test
- Chi-squared test
- t-test
- one-way ANOVA
- correlation test

Binomial test

We can test the hypothesis that the probability of on of two binomial outcomes is different from a specified test value using the binomial test. In the coin-flip case we used this test to test the hypothesis that the probability of "heads" is higher than 0.5. In this test we need to specify the number of "successes" (x) and the number of total attempts (n), where "success" is defined as getting one of the two possible outcomes. We

also need to specify the propbability of success in case of the null hypothesis being true. This is denoted as "p", but don't let this confuse you, this is not the p-value that we use for statistical inference.

We specify that alternative = "greater" in the code because we have a specific directional hypothesis, the alternative hypothesis is not only that the probability of heads is "different" from 0.5, we suspect that the probability of heads is actually higher than 0.5.

```
binom.test(x = 9, n = 10, p = 0.5, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  9 and 10
## number of successes = 9, number of trials = 10, p-value = 0.01074
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.6058367 1.0000000
## sample estimates:
## probability of success
##                    0.9
```

Interpreting the results:

- p-value: This is the p-value used for statistical inference, giving the probability of seeing the observer results or even more extreme results if the null is true.
- alternative hypothesis: specifies the alternative hypothesis, in this case that "true probability of success is greater than 0.5", because we specified p = 0.5 when calling the statistical test, and because we specified that alternative = "greater".
- 95 percent confidence interval (95% CI for short): A 95% confidence interval, meaning that if the experiment was repeated many times and we calculated the the confidence intervals the same way, 95% of these intervals would contain the true population parameter (in this case, the actual probability of getting heads with this coin). It is important that just like the p-value, this 95% CI does not give any assurance about the current single study, it is designed to work over many studies. Note that the CI extend on one side to the highest possible value (in this case 1 is the highest possible value, since we are talking about probabilites). This is because we specified a "one-sided" statistical test with alternative = "greater".
- Sample estimates: the actual observed "empirical probability", the observed proportion of successes in our sample. This is the point-estimate.

The result can be written as:

"We observed 9 heads out of 10 coin flips. The probability of heads is statistically significantly larger (p = 0.011) than 50%: the probability of heads was 0.9 in the sample (95% CI = 0.61, 1)"

Generate data for the class

The code below will generate some data for our class today. It is not important to understand the data generation code at this point in the course, but if you are interested, you can learn from it.

```
n_per_group = 40

base_height_mean = 164
base_height_sd = 10
base_anxiety_mean = 18
base_anxiety_sd = 2
resilience_mean = 7
resilience_sd = 2
```

```r
treatment_effect = - 3
resilience_effect = - 0.8

gender_bias = 0.7
gender_effect = - 1
gender_effect_on_height = 12


treatment <- rep(c(1, 0), each = n_per_group)
set.seed(1)

gender_num <- rbinom(n = n_per_group * 2, size = 1, prob = 0.7)
gender <- NA
gender[gender_num == 0] = "female"
gender[gender_num == 1] = "male"

set.seed(2)
home_ownership <- sample(c("own", "rent", "friend"), n_per_group * 2, replace = T)

set.seed(3)
resilience <- rnorm(mean = resilience_mean, sd = resilience_sd, n = n_per_group*2)

set.seed(6)
anxiety_base <- rnorm(mean = base_anxiety_mean, sd = base_anxiety_sd, n = n_per_group*2)
anxiety_baseline <- anxiety_base + resilience * resilience_effect + gender_num * gender_effect + rnorm(
anxiety_post <- anxiety_base + treatment * treatment_effect + resilience * resilience_effect + gender_nu
participant_ID <- paste0("ID_", 1:(n_per_group*2))

set.seed(5)
height_base <- rnorm(mean = base_height_mean, sd = base_height_sd, n = n_per_group*2)
height <- height_base + gender_num * gender_effect_on_height


group <- rep(NA, n_per_group*2)
group[treatment == 0] = "control"
group[treatment == 1] = "treatment"

health_status <- rep(NA, n_per_group*2)
health_status[anxiety_post < 11] = "cured"
health_status[anxiety_post >= 11] = "anxious"

data <- data.frame(participant_ID)
data = cbind(data, gender, group, resilience, anxiety_baseline, anxiety_post, health_status, home_owners
data = as_tibble(data)

data = data %>%
  mutate(gender = factor(gender))

data = data %>%
  mutate(group = factor(group))

data = data %>%
  mutate(health_status = factor(health_status))
```

```
data = data %>%
  mutate(home_ownership = factor(home_ownership),
         anxiety_baseline = round(anxiety_baseline, 2),
         anxiety_post = round(anxiety_post, 2),
         resilience = round(resilience, 2),
         height = round(height, 2))
```

The data simlate data gathered in a randomized controlled clinical trial, where we assessed the effectiveness of psychotherapy to alleviate stress of hurricane survivors. People with high anxiety entered the study after they lost their homes in a hurricane. Participants were randomly allocated into a control or a treatment group. The treatment group got weekly CBT sessions for 6 weeks. The control group was put on a wait list, and did not get any intervention to decrease anxiety.

In the dataset we have the following variableS:

- participant_ID
- gender
- group - "treatment" or "control"
- resilience - a trait-like ability/capacity of a person to cope with challanges or stressful situations (a continuous variable)
- anxiety_baseline - anxiety levels measured at baseline
- anxiety_post - anxiety measured after 6 weeks
- health_status - whether the person can be considered to be "cured" or "anxious" according to clinical criteria.
- home_ownership - where did the person live before the hurricane: "friend" means that with family or friends, "own" means that in a property owned by the person, "rent" means that in property rented by the person
- height - height of the person

Checking data.

As always, we need the check the data structure and descriptives before doing any analysis.

```
data
```

```
## # A tibble: 80 x 9
##    participant_ID gender group resilience anxiety_baseline anxiety_post
##    <chr>          <fct>  <fct>      <dbl>            <dbl>        <dbl>
##  1 ID_1           male   trea~       5.08            18.7         10.5
##  2 ID_2           male   trea~       6.41            10.5          7.61
##  3 ID_3           male   trea~       7.52            17.2          9.72
##  4 ID_4           female trea~       4.7             17.7         14.7
##  5 ID_5           male   trea~       7.39             8.04         8.14
##  6 ID_6           female trea~       7.06            10.3         10.1
##  7 ID_7           female trea~       7.17            12.6          6.64
##  8 ID_8           male   trea~       9.23            10.6          8.09
##  9 ID_9           male   trea~       4.56            12.8         10.4
## 10 ID_10          male   trea~       9.53             5.82         4.28
## # ... with 70 more rows, and 3 more variables: health_status <fct>,
## #   home_ownership <fct>, height <dbl>
```

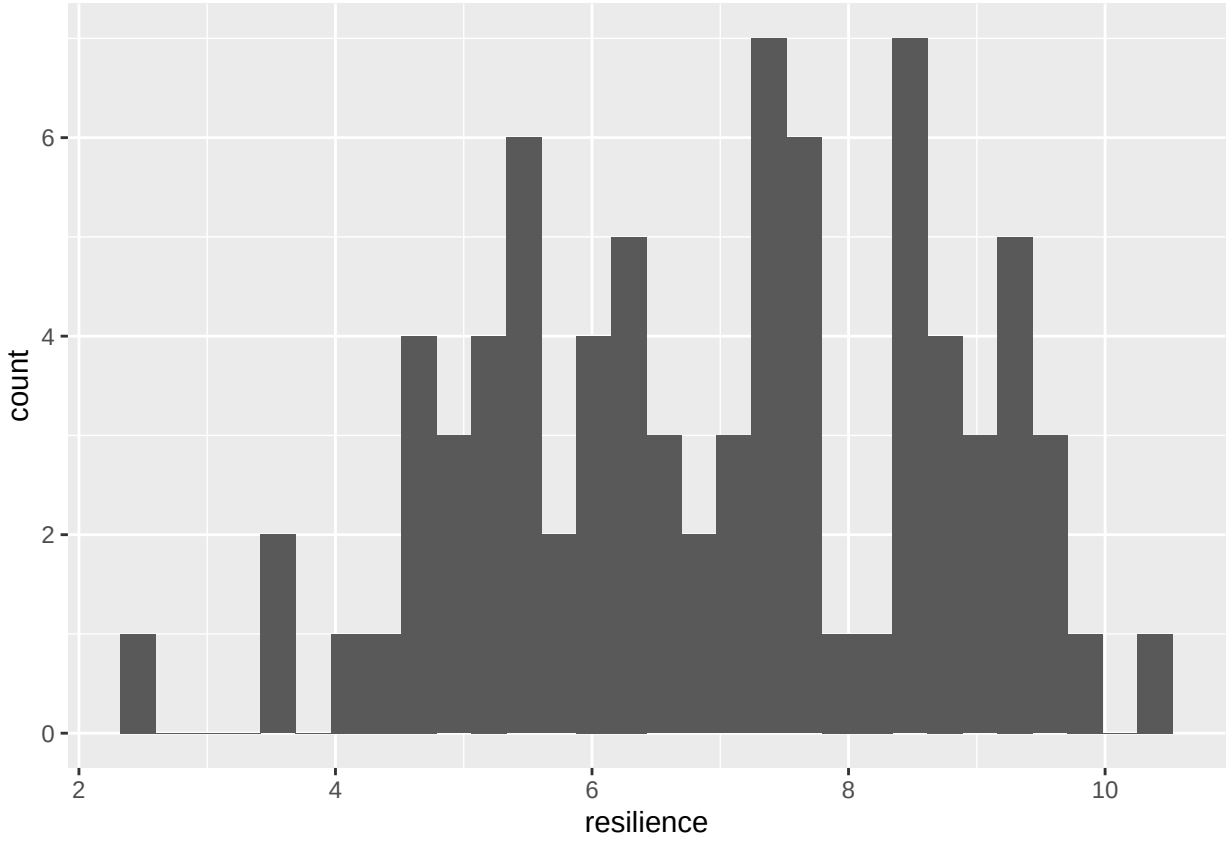
```
data %>%
  summary()
```

```
##  participant_ID        gender         group      resilience
##  Length:80          female:25   control  :40   Min.   : 2.470
```

```
##  Class :character   male  :55   treatment:40   1st Qu.: 5.518
##  Mode  :character                              Median : 7.125
##                                               Mean   : 6.981
##                                               3rd Qu.: 8.477
##                                               Max.   :10.400
##  anxiety_baseline  anxiety_post    health_status home_ownership     height
##  Min.   : 4.650   Min.   : 3.910   anxious:32    friend:22      Min.   :142.2
##  1st Qu.: 9.668   1st Qu.: 8.223   cured  :48    own   :31      1st Qu.:163.4
##  Median :11.155   Median :10.110                 rent  :27      Median :173.0
##  Mean   :11.393   Mean   :10.212                                Mean   :172.3
##  3rd Qu.:12.730   3rd Qu.:12.255                                3rd Qu.:179.7
##  Max.   :19.320   Max.   :16.710                                Max.   :198.2
```

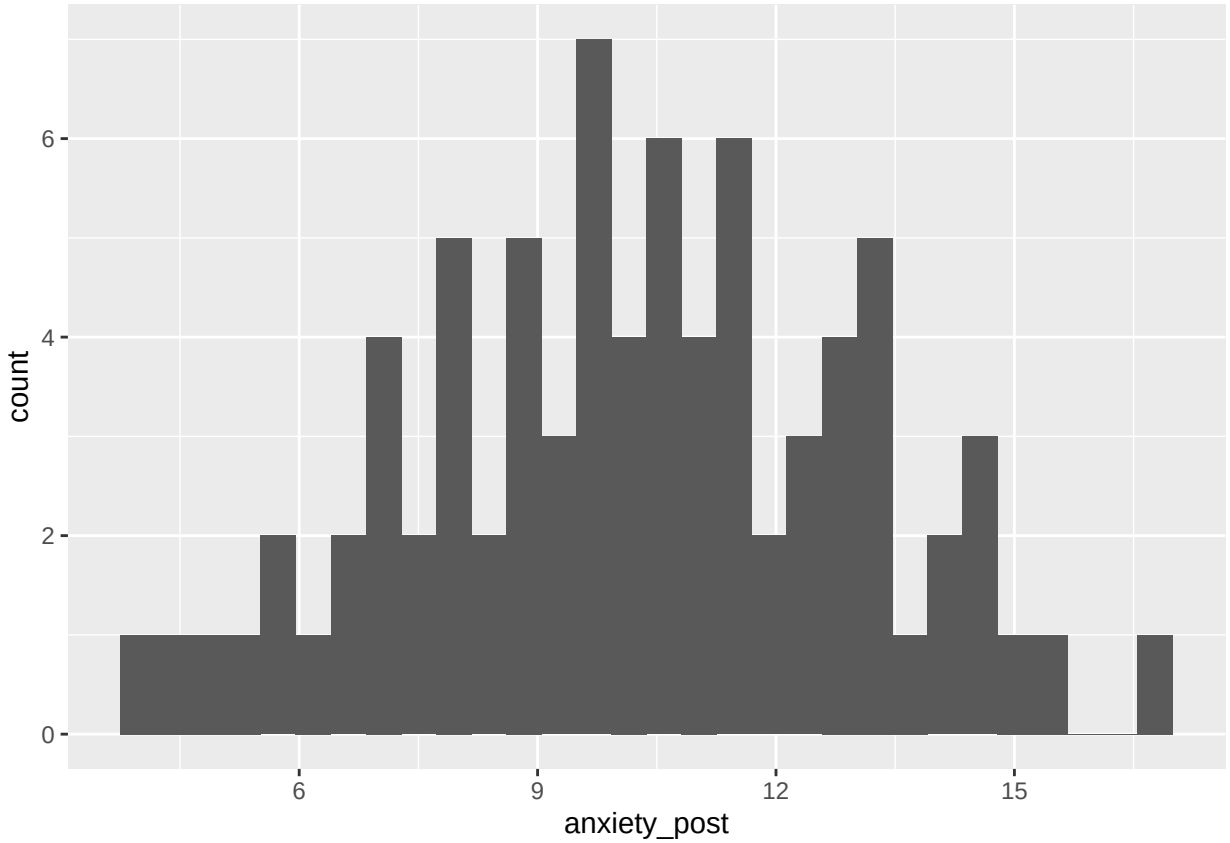
```
data %>%
  ggplot() +
    aes(x = resilience) +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

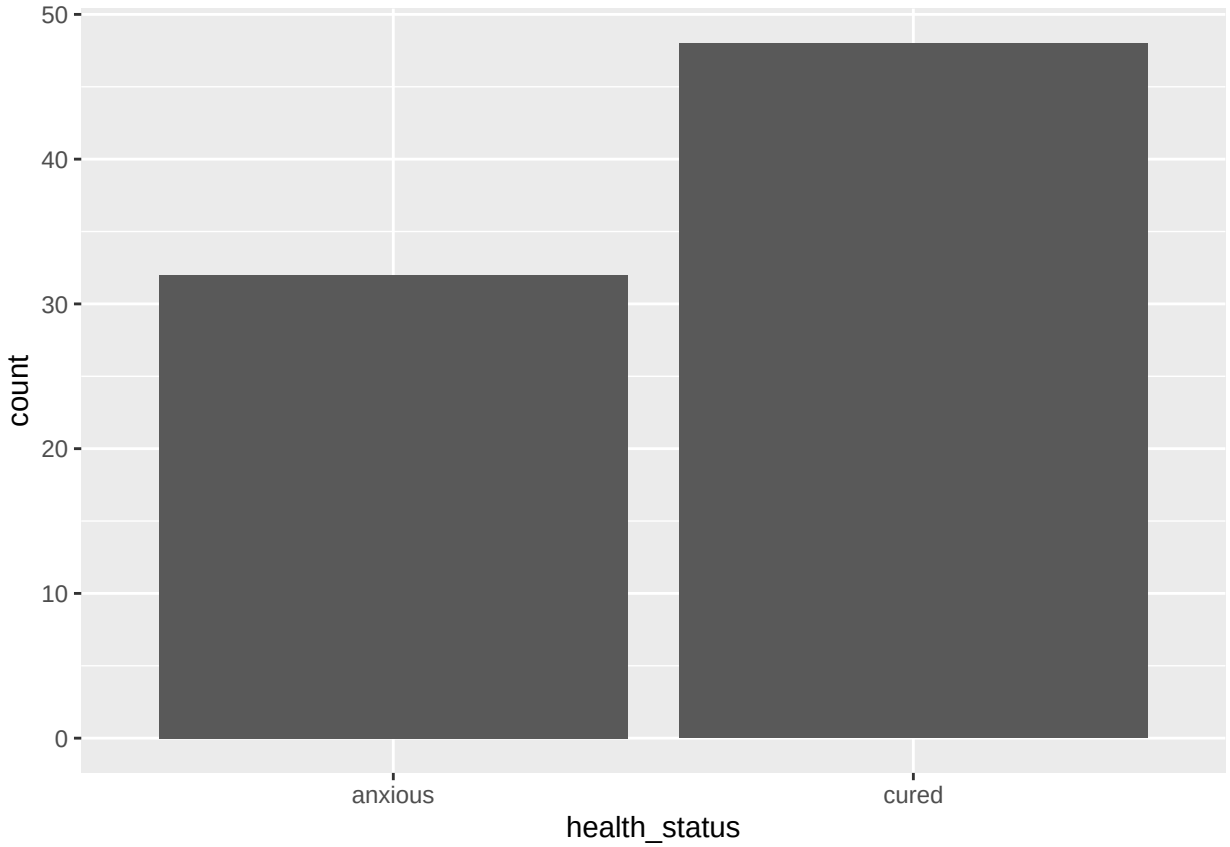



```
data %>%
  ggplot() +
    aes(x = anxiety_post) +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
data %>%
  ggplot() +
    aes(x = health_status) +
    geom_bar()
```




```
set.seed(Sys.time())
```

Hipotezisek

We will test the following hypotheses in this study:

1. More than 50% of the participants were male in the population (**gender** vs. 50%).
2. The research groups will differ from each other in the health_status of individuals. (**health_status** vs. **group**)
3. The average anxiety will be lower i the treatment group than in the control group at the 6 week (post-tretatment) measurement time (**anxiety_post** vs. **group**)
4. Resilience will be negatively correlated with post-test anxiety (those with higher resilience will have lower anxiety).(**anxiety_post** vs. **resilience**)

_____*Practice*_____

Test the hypothesis that there are more than 50% males (**gender** variable) in this clinical population.

Hint: We can test this hypothesis the same way as we tested the hypothesis above about the coin being biased.

Relationship between two categorical variables (Chi-squared test)

The Chi-squared test is used to assess whether there is a relationship between two categorical variables.

(For 2x2 tables the Fisher's exact test or likelihood ratio test is advised, but for tables larger than 2x2 the Chi-sqare test is advised.)

Assumptions for the Chi-sqare test:

- Each observation is independent of all the others (i.e., one observation per subject);
- No more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater

Lets now assess the relationship of **home_ownership** and **health_status**.

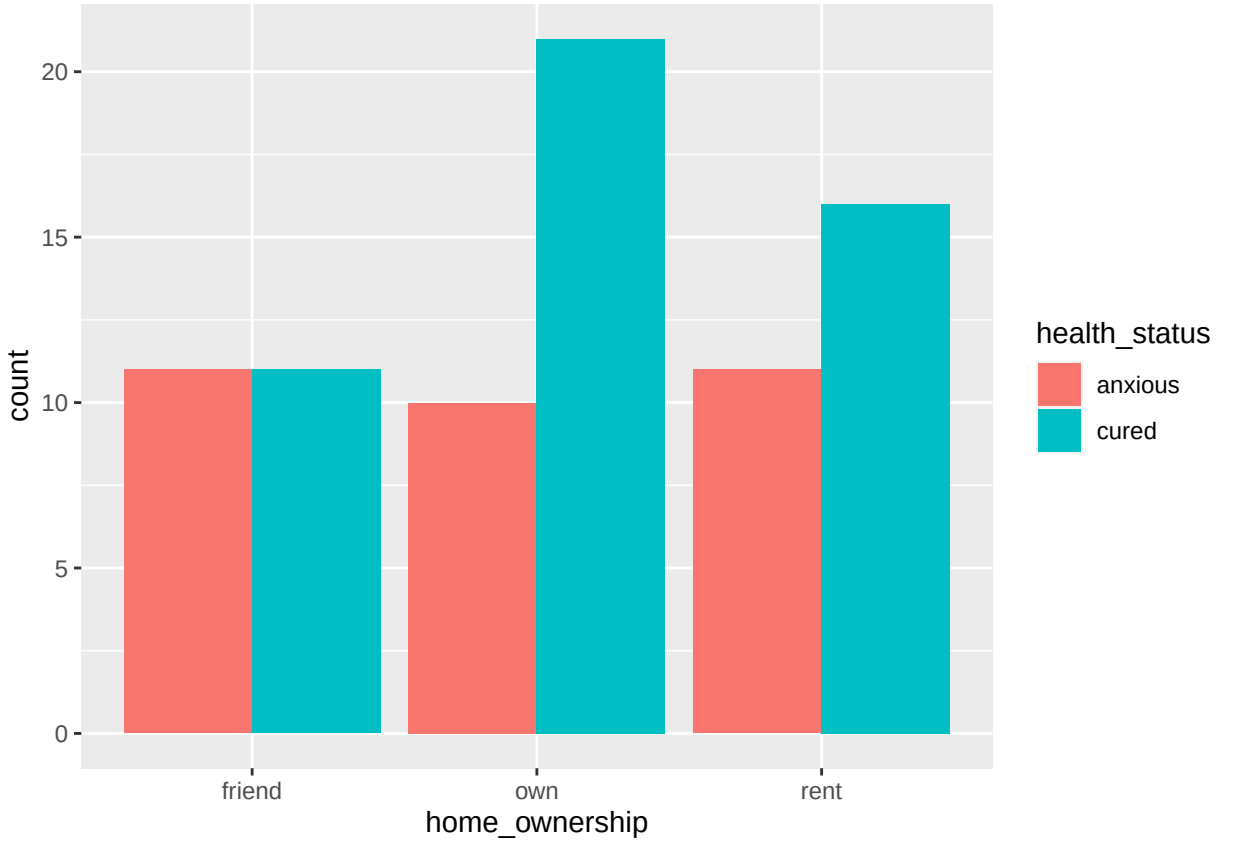
As always, we start with exploration:

- create a crosstab of the two variables
- visualize the relationship using a barchart (pl. geom_bar)

```
table(data$home_ownership, data$health_status)
```

```
##
##          anxious cured
##   friend      11    11
##   own         10    21
##   rent        11    16
```

```
data %>%
  ggplot() +
    aes(x = home_ownership, fill = health_status) +
    geom_bar(position = "dodge")
```



Then, we can perform the Chi-squared test. We do this on the crosstab we created for the exploration above.

In the Khi-squared test we are testing the null hypothesis that the probability distribution of one variable is the same in all categories of the other variable. (In our case this means that people in all home ownership

categoris display the same rate of being cured vs. being anxious)

```
ownership_health_status_table = table(data$home_ownership, data$health_status)
ownership_health_status_table
```

```
##
##          anxious cured
##   friend      11    11
##   own         10    21
##   rent        11    16
```

```
chisq.test(ownership_health_status_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ownership_health_status_table
## X-squared = 1.697, df = 2, p-value = 0.428
```

We can write this result down as:

"There is no statistically significant difference among the different home ownership groups in their health status. (X^2 = 1.7, df = 2, p = 0.428)."

_____*Practice*_____

Test hypothesis 2 mentioned above: "The research groups will differ from each other in the health_status of individuals." (**health_status** vs. **group**)

This can be tested with the same procedure as above.

- first, do explorative analysis
- create a crosstab of the two variables, save this table as a new object
- do the chisq.test() with this table object as the input
- write down the results according to the example in the class notes.

Comparing the average of a continuous variable across groups: t-test and ANOVA

t-test

The t-test is used to compare the mean of a continuous variable across two groups. For example in the code below we compare the mean of anxiety after 6 weeks of treatment (**anxiety_post**) of men and women (**gender**).

The assumptions of the independent sample t-test are:

- The dependent variable should be measured at the interval or ratio level
- The independent variable should consist of two categorical independent groups
- Independence of the observations. Each subject should belong to only one group. There is no relationship between the observations in each group.
- No significant outliers in the two groups
- Normality. the data for each group should be approximately normally distributed.
- Homogeneity of variances. the variance of the outcome variable should be equal in each group. The Welch t-test does not make this assumption, so this can be used in case of heteroscedasticity.

As usualy we start with exploratory statistics.

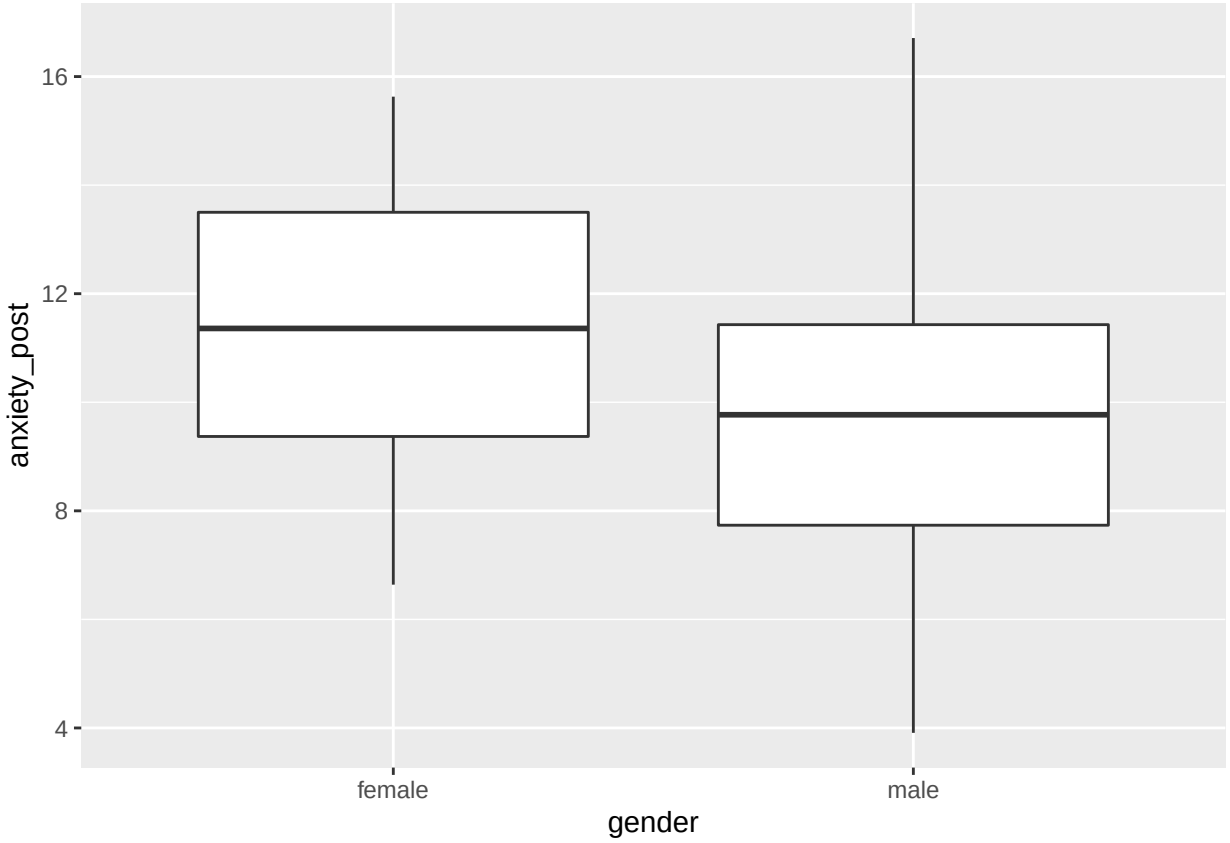
```
summary = data %>%
  group_by(gender) %>%
    summarize(mean = mean(anxiety_post), sd = sd(anxiety_post))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

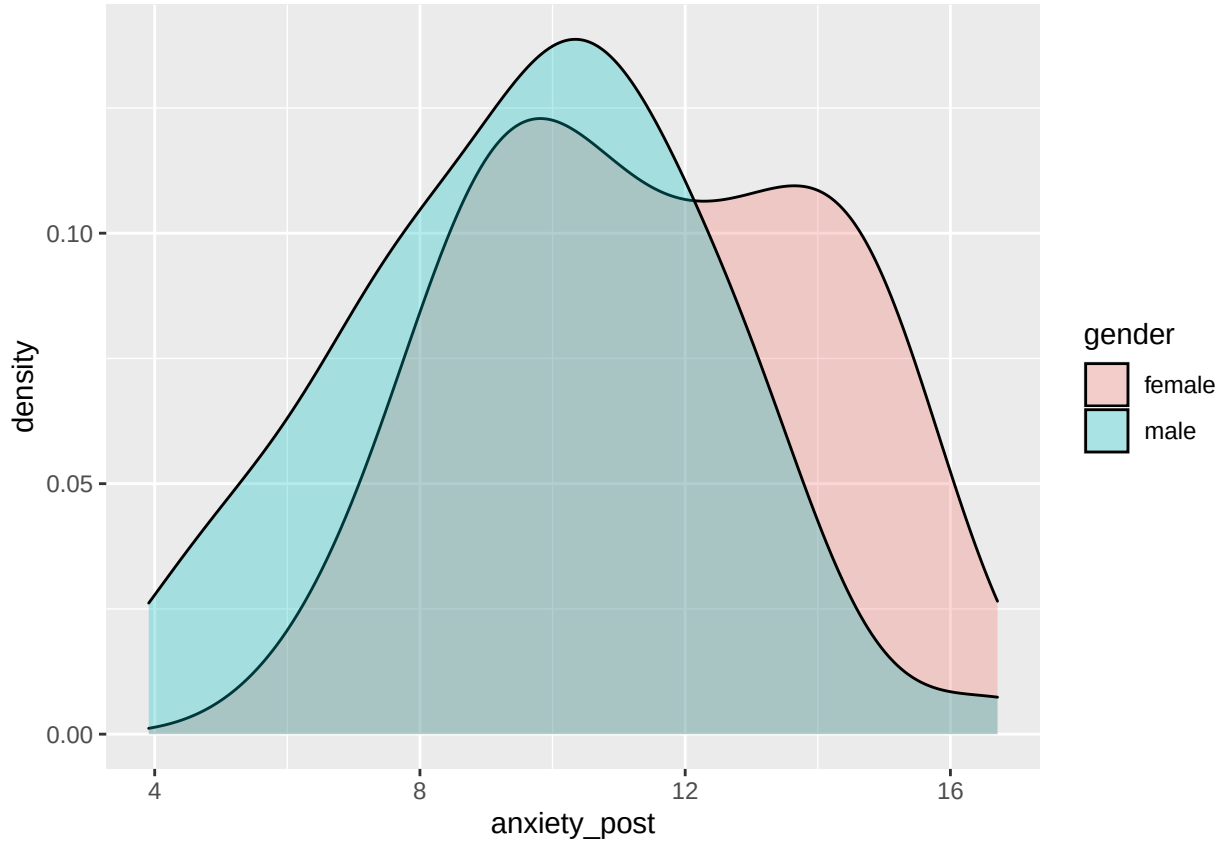
```
summary
```

```
## # A tibble: 2 x 3
##   gender  mean    sd
##   <fct>  <dbl> <dbl>
## 1 female 11.5   2.58
## 2 male    9.64  2.70
```

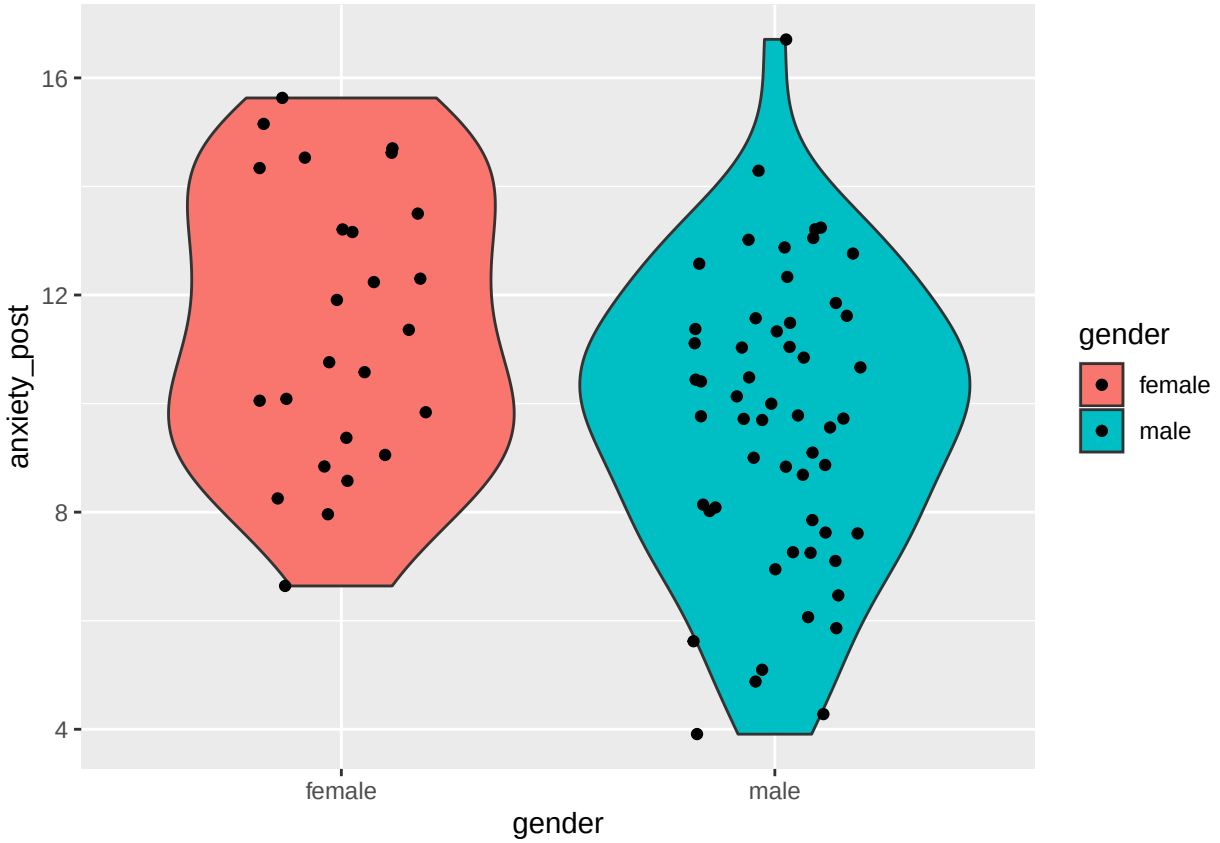
```
data %>%
  ggplot() +
    aes(x = gender, y = anxiety_post) +
    geom_boxplot()
```




```
data %>%
  ggplot() +
    aes(x = anxiety_post, fill = gender) +
    geom_density(alpha = 0.3)
```




```
data %>%
  ggplot() +
    aes(x = gender, y = anxiety_post, fill = gender) +
    geom_violin() +
    geom_jitter(width = 0.2)
```



The exploratory results that indicate that there is a difference between the gender groups, females having slightly higher anxiety than men.

To see whether there is a statistically significant difference, we need to perform a t-test using the t.test() function.

```
t_test_results = t.test(anxiety_post ~ gender, data = data)
t_test_results
```

```
##
##  Welch Two Sample t-test
##
## data:  anxiety_post by gender
## t = 2.8895, df = 48.518, p-value = 0.005754
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5553479 3.0941793
## sample estimates:
## mean in group female   mean in group male
##            11.466400             9.641636
```

```
mean_dif = summary %>%
    summarize(mean_dif = mean[1] - mean[2])
mean_dif
```

```
## # A tibble: 1 x 1
##   mean_dif
##      <dbl>
```

```
## 1     1.82
```

The result can be written down as follows:

"Man and women are significantly different in post-treatment anxiety. (t = 2.89, df = 48.52, p = 0.006. Mean anxiety in the groups were as follows: women: 11.47(2.58), men: 9.64(2.7). Women were on average more anxious by 1.82 points than men (95% CI = 0.56, 3.09)."

One-way ANOVA

If there are **more than 2 categories** in a categorical variable, we cannot use a t-test to compare all groups at the same time. Instead, we can use the ***One-way ANOVA*** with the **aov()** function. The formula is the same as for the t.test().

The assumptions of the one-way ANOVA:

- The dependent variable should be measured at the interval or ratio level
- The independent variable should consist of two or more categorical, independent groups
- Independence of observations
- No significant outliers in the groups
- Normality. The dependent variable should be approximately normally distributed for each category of the independent variable
- Homogeneity of variances.

Below we are testing the group differences in mean post-treatment anxiety (**anxiety_post**) between the home ownership groups (**home_ownership**).

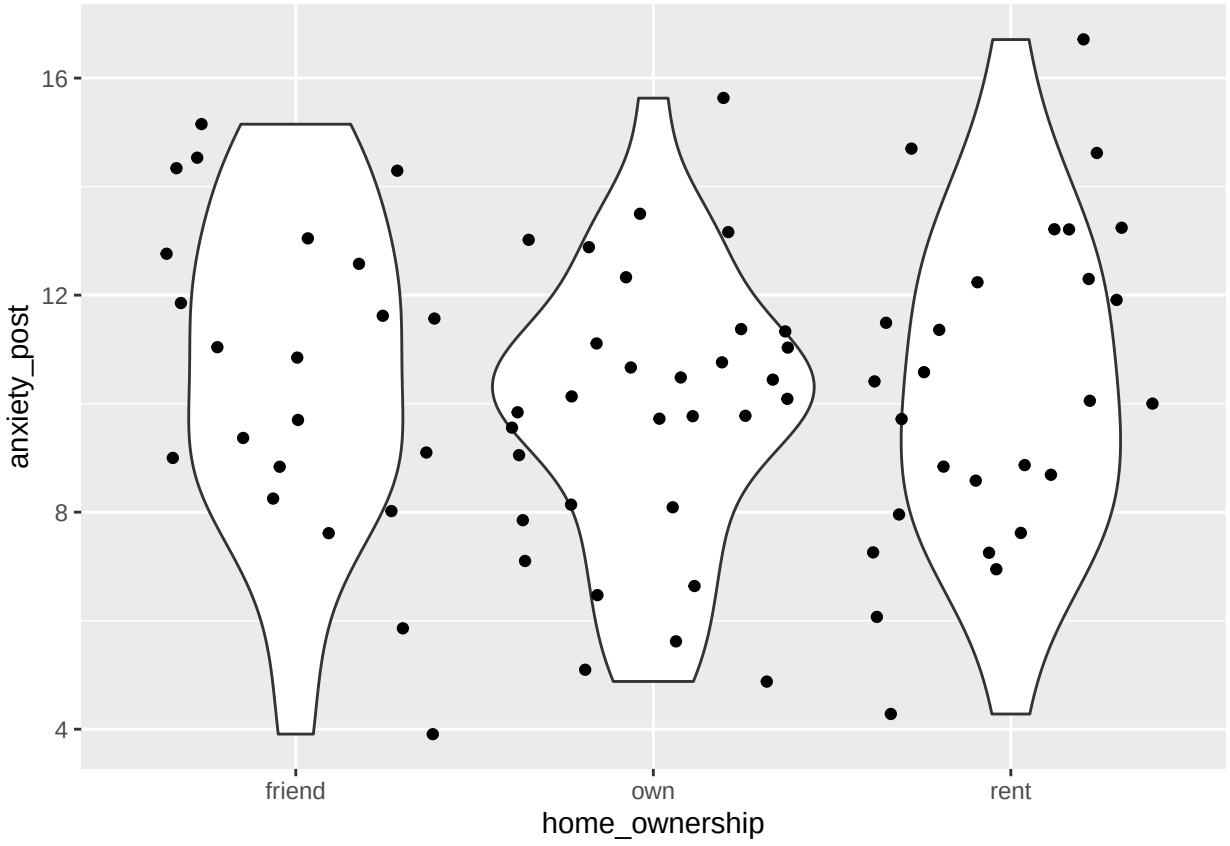
```
summary_home_ownership_vs_anxiety_post = data %>%
  group_by(home_ownership) %>%
    summarize(mean = mean(anxiety_post), sd = sd(anxiety_post))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
summary_home_ownership_vs_anxiety_post
```

```
## # A tibble: 3 x 3
##   home_ownership  mean    sd
##   <fct>          <dbl> <dbl>
## 1 friend         10.6   2.93
## 2 own             9.86  2.57
## 3 rent           10.3   2.94
```

```
data %>%
  ggplot() +
    aes(x = home_ownership, y = anxiety_post) +
    geom_violin() +
  geom_jitter()
```




```
ANOVA_result = aov(anxiety_post ~ home_ownership, data = data)
summary(ANOVA_result)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## home_ownership  2    7.5   3.760    0.48  0.621
## Residuals      77  603.3   7.835
```

The results can be written down as follows:

"There was no significant difference between the homw ownership groups in the average level of anxiety at post-treatment. (F (2 ,77) = 0.48, p = 0.621). Table 1 shows the mean and SD of anxiety for each group."

One-sided vs. two-sided tests

If we have an expectation about the direction of an effect (for example which group will show higher scores), that we use a **one-sided statistical test** instead of the degfault two-sided tests.

Let's say for example that we hypothesize that there will be a difference between the gender groups in post-treatment anxiety, and also that females will have higher anxiety. We can specify this by setting the "alternative" parameter to alternative = "greater".

If we compare this result to the previous result we got on the same t-test without setting the alternative parameter, see that all numerical results are the same, except for **the p-value and the confidence interval**. The p-value is exactly half of the size it was with the two-sided test, because the possible probability distribution was halved by using a one-sided hypothesis. This means that if we are right about the direction of the effect, specifying this in the hypothesis makes our test more powerful (have a higher probability of detecting the effect), so it is always worth it to specify a one-sided hypothesis if we have good reason to suspect it.

Also, as noted above, the confidence interval (CI) extends on one side to the highest possible value, in this case it is "Inf", that is, infinity.

One important thing to understand is that when we say **alternative = "greater"**, we mean that the **reference group** in the categorical variable will have a higher mean compared to the other group. If we thought that the reference category would have a lower mean than the other group, we would have to specify **alternative = "less"**.

As noted earlier, the reference group or reference category is determined by R automatically based on alphabetic order, unless specified otherwise. In our case, "female" is tha reference category, since it is earlier in the alphabet than "male". We can change this by using the **factor() function and setting the levels** = parameter as we have seen in the earlier exercise. The important thing is to always know what is the reference level when we specify a one-sided alternative hypothesis.

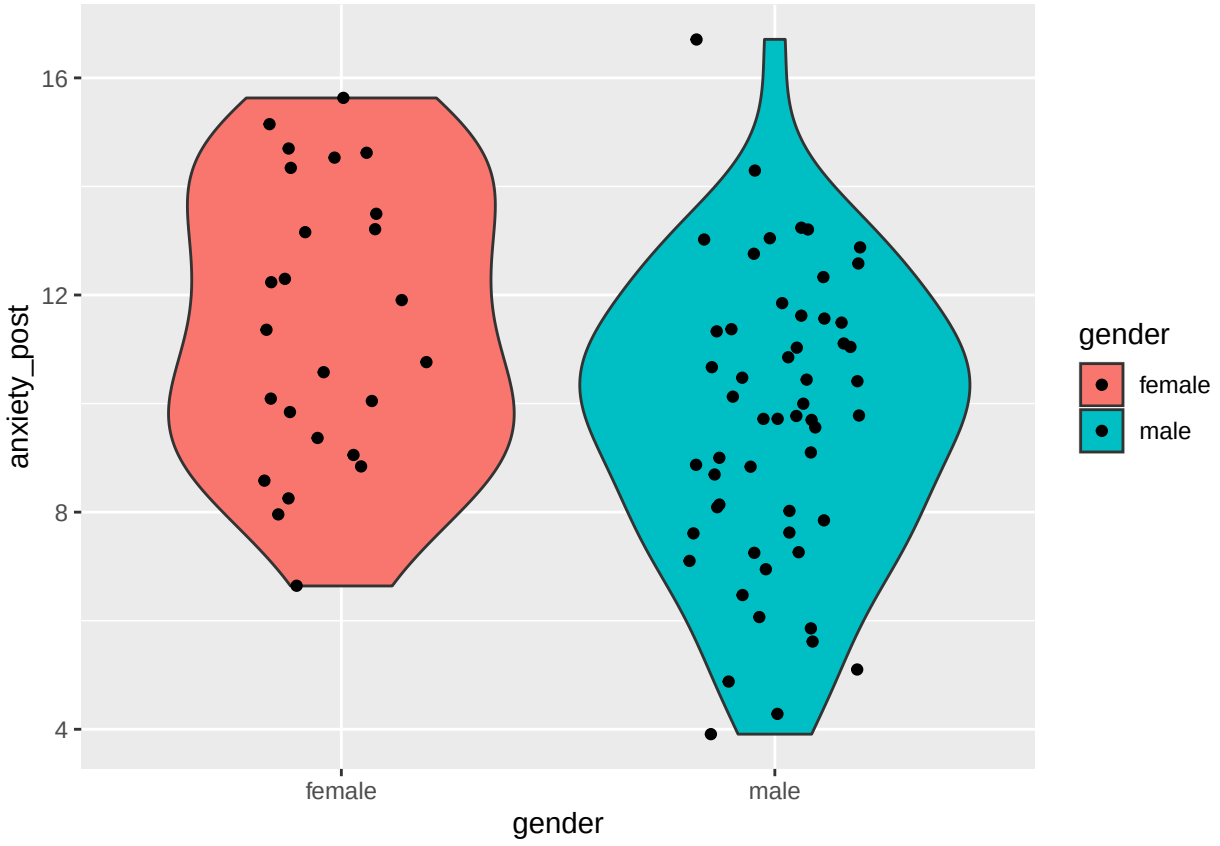
```
summary = data %>%
  group_by(gender) %>%
    summarize(mean = mean(anxiety_post), sd = sd(anxiety_post))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
summary
```

```
## # A tibble: 2 x 3
##   gender  mean    sd
##   <fct>  <dbl> <dbl>
## 1 female 11.5   2.58
## 2 male    9.64  2.70
```

```
data %>%
  ggplot() +
    aes(x = gender, y = anxiety_post, fill = gender) +
    geom_violin() +
    geom_jitter(width = 0.2)
```




```
t_test_results_one_sided = t.test(anxiety_post ~ gender, data = data, alternative = "greater")
t_test_results_one_sided
```

```
##
##  Welch Two Sample t-test
##
## data:  anxiety_post by gender
## t = 2.8895, df = 48.518, p-value = 0.002877
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7657774       Inf
## sample estimates:
## mean in group female   mean in group male
##            11.466400             9.641636
```

We can write down the result like this:

"The anxiety was significantly higher among women than men (t = 2.89, df = 48.52, p = 0.003. The mean and SD of anxiety in the groups were the following: women: 11.47(2.58), men: 9.64(2.7). women were on average 1.82 points more anxious (95% CI = 0.77, inf)."

Lets see what would have happened if we specified the alternative hypothesis in the other direction with **alternative = "less"**.

The p-values is almost 1 in this case, This is not surprising, since the trend in the data point into the opposite direction from what we specified in our alternative hypothesis, so this observation does not allow us at all to reject the null.

```
t_test_results_one_sided = t.test(anxiety_post ~ gender, data = data, alternative = "less")
t_test_results_one_sided
```

```
##
##  Welch Two Sample t-test
##
## data:  anxiety_post by gender
## t = 2.8895, df = 48.518, p-value = 0.9971
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf 2.88375
## sample estimates:
## mean in group female   mean in group male
##            11.466400             9.641636
```

Practice

Test the 3rd hypothesis mentioned above:

Teszteld a 3. hipotezist, hogy "The average anxiety will be lower i the treatment group than in the control group at the 6 week (post-tretatment) measurement time" (**anxiety_post** vs. **group**).

- Lets run exporatory analyses on the relationship of the two variables
- Decide whether to use one or two sided test, and if you decide to use one-sided, make sure to specify the alternative correctly based on what is the reference group
- Should we use the t.test or one-way ANOVA?
- Write down the results with words as shown in the example above.

Relationship between two continuous variables, correlation test using cor.test()

The **correlation test** determines whether there is a statistically significant relationship between **two conitnuous variables**

The assumptions of pearson's correlation are: - Continuous variables. If one or both of the variables are ordinal in measurement, then a Spearman correlation can be used - Each observation should have a pair of values - Absence of outliers - Linearity. A "straight line" relationship between the variable should be formed. - Normality. Both variables should be normally distributed. In case of non-normality a Spearman correlation can be conducted instead

Lets assess wehter there is a relationship between **resilience** and **height**.

As we learned before, the exploratory analysis can inlcude a scatterplot with a trend line and calculating the correlation coefficient. (Note that in geom_smooth we use method = "lm" to get a straight line instead of the curved line what is produced by the default method = "loess")

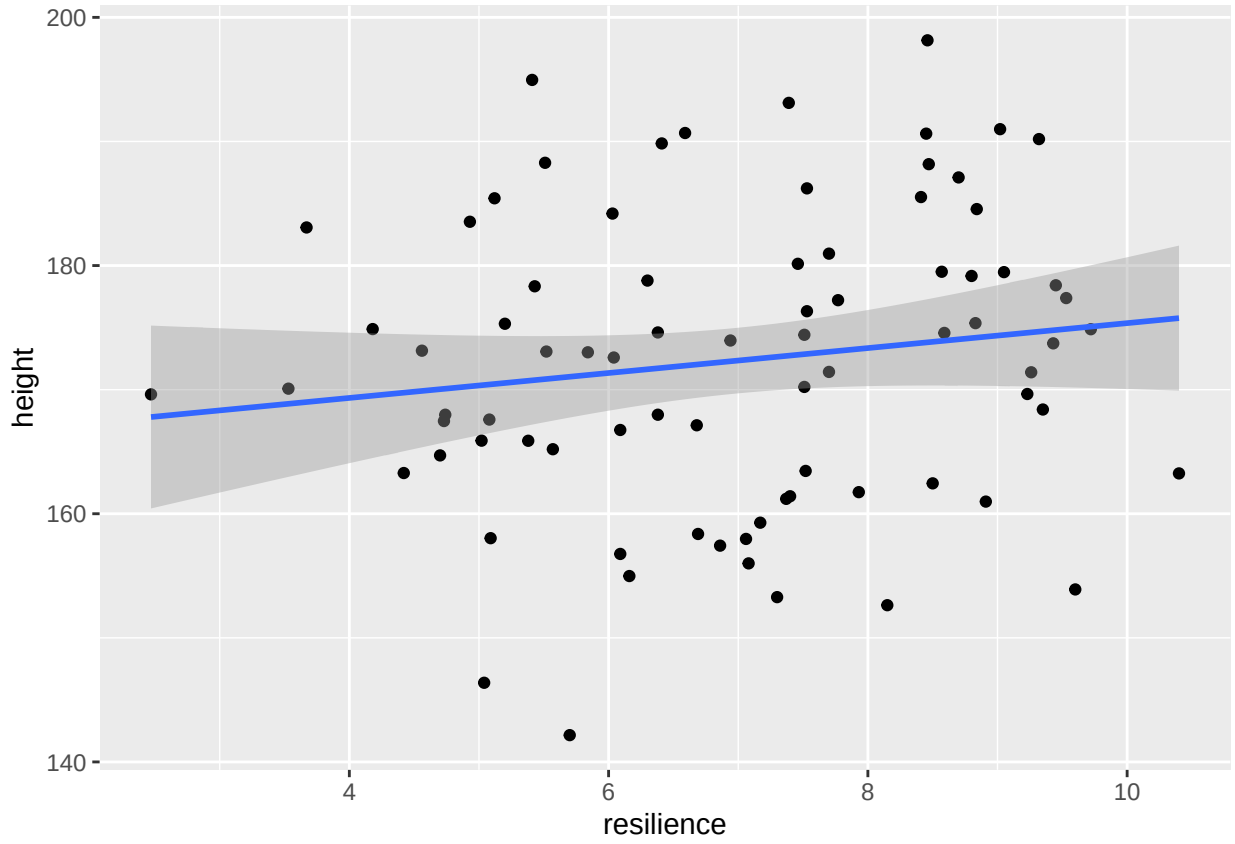
```
data %>%
  select(resilience, height) %>%
    cor()
```

```
##            resilience   height
## resilience   1.000000 0.146929
## height       0.146929 1.000000
```

```
data %>%
  ggplot() +
    aes(x = resilience, y = height) +
```

```
      geom_point() +
      geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The two variables seem to be unrelated based on the exploration.

We can test for whether there is a statistically significant relationship using Pearson correlation test via the **cor.test()** function:

```
correlation_result = cor.test(data$resilience, data$height)
correlation_result
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$resilience and data$height
## t = 1.3119, df = 78, p-value = 0.1934
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.07521612  0.35517967
## sample estimates:
##      cor
## 0.146929
```

We can write down the results as follows:

There is was no evidence for a significant relationship between resilience and height (r = 0.15, 95% CI = -0.08, 0.36, df = 78, p = 0.193)"

Similarly to the t.test, it is good to use a one-sided alternative hypothesis in the case of a correlation test as well, if we have a good reason to expect this. For example we can expect that there will be a positive correlation between resilience and height, so we could specify alternative = "greater", because we think that there will be a positive correlation. If we thought the correlation will be negative, we would specify alternative = "less".

```
correlation_result_greater = cor.test(data$resilience, data$height, alternative = "greater")
correlation_result_greater
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$resilience and data$height
## t = 1.3119, df = 78, p-value = 0.0967
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  -0.03942784  1.00000000
## sample estimates:
##      cor
## 0.146929
```

Practice

test the 4th hypothesis mentioned above:

Teszteld a 4. hipotezist, hogy "Resilience will be negatively correlated with post-test anxiety (those with higher resilience will have lower anxiety)." (**anxiety_post** vs. **resilience**)

- Lets run exporatory analyses on the relationship of the two variables
- Decide whether to use one or two sided test, and if you decide to use one-sided, make sure to specify the alternative correctly based on the expected direction of the relationship
- Write down the results of the test with words as shown in the example above.

About reporting statistical results in general:

When reporting statistical results, it can vary from test to test what kinds of data we get, but in general, we usually need to provide the following information in te description of the results: - text summary of the results - the test-statistic - degrees of freedon (in very simple tests with no df we usually report the sample size instead) - p-value - point estimate of the parameter or effect size - 95% confidence interval of the the parameter or effect size