Data analysis

The rationale of using quantitative measurements and statistics to test hypotheses and approximate truth

Before we do statistical tests on the collected data, it is important to first explain what the general purpose of doing statistics is. One branch of statistics is referred to as *descriptive statistics*, which, as indicated by the word "descriptive", is used to describe a set of observations that can be quantified. While the term *descriptive* may seem redundant since all branches of statistics do, in some way or another, describe data quantitatively (i.e., using numbers to describe real-life phenomena), this term is used deliberately to emphasise its *objectiveness* in that it is used to describe data *as it is*, i.e., it does not attempt to identify relevant factors hypothesised to explain *why* the observed data is as it is.

In contrast, there is another branch of statistics referred to as *inferential statistics* which is used in research to draw conclusions and make predictions about a larger population by analysing data from samples, assumed to be representative of the population of interest (i.e., the population the researcher wants to make generalisations and predictions about). While no sample is perfect and hence be presumed to be 100% representative of the population of interest, it is (in most cases) not feasible to collect data from the whole population and the use of samples is therefore a necessity. Nevertheless, despite the inherent limitations in using samples to draw conclusion about a wider population, there are several ways to reduce uncertainty associated with samples.

The fundamental role of error and uncertainty in statistics

In controlled experiments, one way to reduce uncertainty is the use of randomisation, e.g., a researcher may randomly assign participants to different treatment protocols in order to reduce systematic bias and hence irrelevant variables interfering with the research results, which may otherwise confound any findings and make interpretations ambiguous. However, randomisation is not always possible, especially in large observational studies like the one used in my thesis, but there are other ways to reduce bias and uncertainty in statistical tests.

The Law of Large Numbers and the importance of minimising error when modelling data

One way to reduce bias and increase the representativeness of a sample is to simply increase the sample size, and the reason why this is the case is due to the *Law of Large Numbers*, which states that as a sample size grows, its mean gets closer to the average of the whole population. The reason for this is that variation in a sample is not only associated with the nature of the statistical design and methods use to measure a certain variable. In fact, in addition to measurement error, variation is generally to large extent associated with the inherent nature of the variable being measured, since almost all quantifiable variables, i.e., variables we can measure numerically and subject to statistical tests, can be defined as *stochastic variables*, i.e., the thing you are trying to measure, for example the amount of hours a specific person sleeps each night, is influenced by random events.

Let's imagine that we knew how many hours each person in the world sleeps during the night, then it is conceivable that there would be a normal distribution, where the average would be the mean of all whole population, and that many would be a bit below or above the average, and that the frequency would lower as a function of the distance from the mean. More specifically, if the mean was 8 hours of sleep, many people would fall between 7 and 9 hours of sleep, but there would be fewer people that, for example, would sleep less than 5 or more than 11 hours each night, and so on. Therefore, if you measured a specific person and the hours of sleep one specific night, while the most likely prediction would be the mean of the population (i.e., 8 in this imaginary sample), due to random chance you could get a very different result. For example, if the person you measured was chosen from the whole population by chance, maybe you chose a person that sleeps extremely little and your measurement was 4 hours. Let's make the rather reasonable assumption that the amount of hours people sleep differs by some amount day by day, and let's imagine that we, in fact, by chance chose a person that generally sleeps above the average (for example 9 hours), but, since the person whose hours of sleep we measured, had an exam that was due in 2 days, he or she only slept 3 hours.

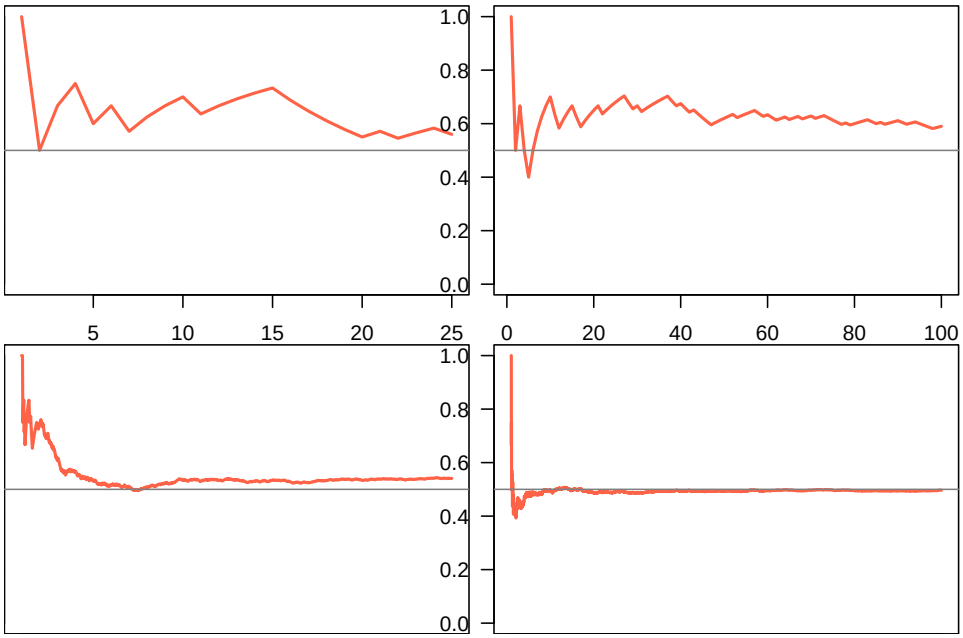
While this example is a imaged one, I think it is fair to assume that this scenario is not entirely realistic. More importantly, it exemplifies how variation in measurements not only can be influenced by scientific methodology and highlights the importance of the inherent randomness stochastic variables - a fundamental aspect of statistics, since all variation, whether it is due to measurement error or random chance, has an direct effect on statistical modelling, and hence what we can infer from our observations and therefore what conclusions we can derive from them.

demonstrating the law of large numbers by simulating a coin toss

A coin toss is a very

As Below, I will write a function that models the relative frequency of heads and tails of a coin-toss, assuming that it is a fair coin and the chance is 50-50.

We can then visualise the difference between the number of coin tosses, increasing from 25, 100, 1000 and 5000 tosses to demonstrate the Law of Large Numbers.



It is clear from the figures above that as the sample size increase, i.e., the amount of coin tosses increases, the relative frequency approaches 50% to get heads or tails, as we would expect. If a coin is tossed only a few times, one might get lucky and get 6 heads in a row and from the data erroneously conclude that the chance of tails is much lower than it really is. Nevertheless, if one keeps tossing the coin, it will inevitably approach the **true mean**. In summary, more data means less uncertainty and the same holds true for other types of data, not unlike the data that I have collected. For this reason, I have aimed to get a sample size as large as possible with the limited resources that I had, because it makes it more likely that my data is representative of the population of interest, namely students on their last year of school before University, and that any differences observed across factors, e.g., gender, can be subjected to statistical tests and detect those effects if they are present, or, indeed, conclude that the difference is too small to be statistically significant. I will return to the subject of statistical significance in a moment.

It is clear that statistics operates within the realm of uncertainty, but that is the nature of science. In fact, one of people who revolutionised science was Karl Popper who realised that the scientific method can never prove any statement completely true due to the *Problem of Induction*. Colloquially, it is often referred to as the *Black Swan Problem* since it illustrates a historical example of the induction problem, namely that for many centuries Europeans presumed that all swans were white, since all swans they had ever observed

had been white. This all changed in 1697 when the Dutch explorer Willem de Vlamingh discovered swans with black plumage in Australia - a continent that hitherto remained largely undiscovered by Europeans. The *black swan problem* illustrates the important scientific concept pioneered by Popper that, no matter how many observations we make and turn out to confirm our preconceived ideas, we cannot rule out that a future observation will contradict that them (Popper 1963). Thus, science is able to disprove claims, e.g., the observation of a single black swan effectively disproves the claim that all swans are white, but, in stark contrast, can never prove any claim with complete certainty.

Karl Popper revolutionised science by flipping the aim of the scientific method on its head: instead of trying to prove a claim, the only way for science to progress is to try to disprove hypotheses. If all attempts to disprove the hypothesis fail, which requires that the hypothesis is falsifiable or, in other words, can conceivably be wrong and hence be tested (if a hypothesis is true under any circumstance and no experiment can disprove it, then it is unfalsifiable and hence not a scientific hypothesis, because it cannot be tested. How can you test something if the result, namely that it is true, is the same regardless of the outcome of the experiment?), then it is rational to assume that it is more likely that the hypothesis is correct (Popper 1963), though, as stated previously, we can never be completely confident.

For this reason, statistics is used in science as a tool to both describe data *as it is* as well as making conclusions about underlying factors that can explain the observations. For example, if we observe that there is a difference between males and females in the average introspection score on a questionnaire such as self-perceived English reading ability, we can use statistical tests to answer the question: Is this observed difference real or did it occur just by random chance? As I already have explained, large sample sizes help decrease uncertainty, but there will always be uncertainty present, so whether a statistical test is significant or insignificant is also affected by the variation in the data and how large the difference is. Here, the term variation or *variance* (usually denoted as s^2) is a statistical term defined as how much a sample differs from its observed mean. In other words, it is a measure of how much the sample differs on average from the average and hence is an indication of the error (usually denoted as ε) in the model (which in this case is the mean, usually denoted as $\bar{x}$. A useful way to think about it is that if you had to guess the self-rated English speaking reading ability of a person in your sample, your best guess would be the observed mean of the sample. However, each person will have a different score with some rating themselves higher or lower than the average. Therefore, the variance in the data reflects the average error in our best guess. I will return to the concept of variance later and explain it in more depth, since it is a fundamental role in statistical tests designed to determine whether a given hypothesis is tenable or not.

In statistical testing, the tests used to determine whether there is a true effect or not take both the variance and size of an observed difference into account to determine whether the observed effect is, in fact, likely to be a true effect, i.e., it did not occur by chance. In

order to determine whether any observed effect is likely to be a true effect, p-values are calculated and used to determine whether the effect is statistically significant or not. As explained previously, Popper had a large influence on scientific thinking and argued that we can never prove anything with certainty, but we can, however, disprove hypotheses. Statistical testing is designed in this way and any statistical test always starts by testing the null-hypothesis, namely that there is no effect.

The p-value is a calculation of how likely it is to observe the result (e.g., a difference in the average score between males and females on an item in a questionnaire) or difference in your data set, given that the null-hypothesis is true, i.e., how likely is your data is given the assumption that there is no effect. The scientific consensus is that if the chance of observing your result, e.g., a difference between two means, or an even bigger effect is less than 5% (i.e., the p-value is $< .05$), then we have enough evidence to rule out the null-hypothesis and gain confidence in the alternative hypothesis, namely that there is a real difference. **Put simply, the p-value is a measure of the likelihood of observing your effect (or an even bigger one) by chance**. So, in accordance with the epistemology argued by Popper, the scientific method, and hence inferential statistics, cannot be designed to prove hypotheses. It can, however, be designed to disprove them. By failing to disprove them, we indirectly gain confidence in our alternative hypothesis (i.e., that there is a true effect), but we can, as Popper argued, never be 100% certain. Nevertheless, if a hypothesis and theory stands the test of time, i.e., all attempts to disprove it have failed, then, in this humble fashion, we can be pretty confident that we at least approximate something we can consider truth.

Mathematical example of the statistics used to calculate the observed results of quantitative variables in this thesis

In this next section, I will go through some of the mathematics of one of the statistical tests done in this thesis. Since it would be far to extensive to go through the mathematics of each test (and many statistical tests will be conducted in this thesis), hopefully this one example is comprehensive enough to give the reader a better understanding of how each test result and conclusion is mathematically derived.

Demonstration of the mathematics used to determine whether there is a difference between males and females in mean self-perceived reading ability

Using descriptive statistics to describe the data

Before doing a statistical null-hypothesis test, there has to be some indication that there is something to test. Therefore, it is often a good idea to look at descriptive statistics, because they might give a hint that there may be group effect on a given measure. First,

we can start by looking what the mean across all participant is on a given measure, e.g., self-perceived English reading ability. To calculate the mean, we can use the formula well known formula to calculate averages, i.e., $mean = \frac{\text{sum of all values}}{\text{number of values}}$. Formally, this can be expressed mathematically by the equation given below:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Here, x denotes the score or measurement for a specific subject i, so that $x_i ... x_n$ represents the score of each subject across all subjects in the sample, whose size is denoted by the letter n, i.e., the sample size.

To illustrate the calculation, we can use this script in R to remove any non-responses or *Na's* and give us the first 5 responses and sample size. It's important to note that all responses will be included in the calculations below, but only the first five will be shown for illustrative purposes (It's important to note that the sample size is over 200 and showing all the numbers in the calculations would be excessive and would not serve any benefit with respect to giving the reader a conceptual understanding of the mathematics, which is the purpose of this section).

```
# Removing Na's, i.e., people who did not submit an answer
d = subset(d, d$SP_EnglishReadingAbility != "Na")

# Calculating the length of the vector of answers, giving us the sample size
n = length(d$SP_EnglishReadingAbility)
cat("Sample =", d[1:5, "SP_EnglishReadingAbility"], "\n")
```

Sample = 5 5 4 4 5

```
cat("n =", n)
```

n = 226

From the numbers provided by the code above, we can represent our sample in the following way:

$$\begin{aligned} \underset{n_i\ =1\rightarrow 5}{Sample} &= (5,5,4,4,5\ ...i_{n=226}) \\ \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{226}\sum_{i=1}^{n} 5+5+4+4+5\ \ ...x_n \end{aligned}$$

To calculate the mean in R, we can simply write:

```r
SP_EngReadAbility_sum = sum(d$SP_EnglishReadingAbility)


SP_EngReadAbility_n = length(d$SP_EnglishReadingAbility)

SP_EngReadAbility_average = SP_EngReadAbility_sum * (1/SP_EngReadAbility_n)

SP_EngReadAbility_average = round(SP_EngReadAbility_average, digits = 2)

cat("The mean of self-perceived
  English reading ability
  across all subjects =",
    SP_EngReadAbility_average)
```

```
## The mean of self-perceived
##   English reading ability
##   across all subjects = 4.17
```

The code above to calculate the mean was used to illustrate the mathematics conceptually. We can do the same thing much easier by using the inbuilt function *mean()* in R, and obtain the same result.

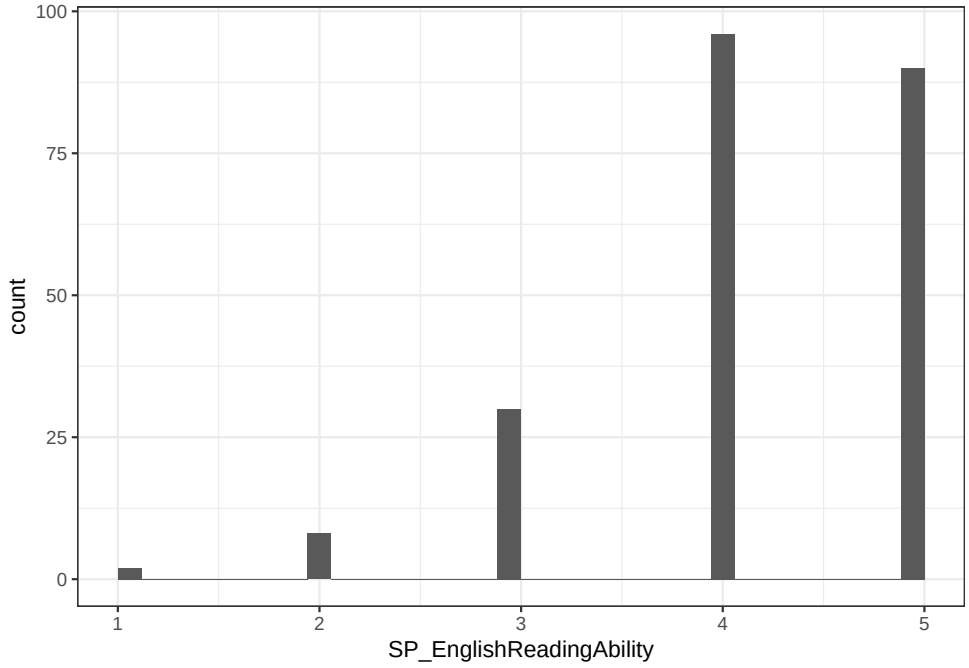
```r
SP_EngReadAbility_mean = round(mean(d$SP_EnglishReadingAbility), digits = 2)

cat("The mean of self-perceived
  English reading ability
  across all subjects =",
    SP_EngReadAbility_mean)
```

The mean of self-perceived English reading ability across all subjects = 4.17

Generally, I will apply inbuilt functions in R or from packages developed for R, since the same results can be obtained with less lines of code, and for some tests, they are more precise than doing the calculation manually.

Now that we have calculated the mean, one other thing we may wish to investigate is how much variation there is in the data. We can use R to plot a histogram to give us a clue about how the data is distributed.

```r
ggplot(d,
       aes(SP_EnglishReadingAbility)) +
  geom_histogram(
    aes(y = after_stat(count)),
    bins = 35) + theme_bw()
```



Next, we might ask the question, are there any variables that might affect respondents self-rated English reading ability? There may exist countless variables that may have an effect, but, more importantly, does my dataset contain any such variable?

One factor that could conceivably play a role with regards to how people perceive their own reading ability in English is gender, which happens to be one of the questions, and hence in my dataset. Here, it is important to emphasise that my data is questionnaire-based and hence based on introspection, and hence not necessarily indicative of the objective ability of a person to read English. Furthermore, it is reasonable to assume that introspection is confounded by people's ability to accurately describe and rate their own abilities, where some people may be more accurate than others. Nevertheless, measuring psychological variables is a complex endeavour - psychological variables are often known as latent variables, since they cannot be measures directly and hence need to be inferred by indirect measures assumed to some extent be valid measurements of the psychological variable of interest(Johnsen and Rytter 2021). In this regard, let's consider gender again. How can we determine whether it has an influence on a variable that I have measured?

As we did before when we calculated the mean across all subjects on the dependent variable *self-perceived English reading ability*, let's start by doing some descriptive statistics on *self-perceived English reading ability*, grouped by the factor *gender*.

We can start by sub-setting our data to create two vectors of data points for *self-perceived English reading ability*, and for illustrative purposes, print the first observations.

```
# creating a vector for females
d_fem_EngReadAbility =
  subset(d, d$gender == "Female")
# creating a vector for males
d_male_EngReadAbility =
  subset(d, d$gender == "Male")
# remove Na's
d_fem_EngReadAbility = subset(
  d_fem_EngReadAbility,
  d_fem_EngReadAbility$SP_EnglishReadingAbility != "Na")

d_male_EngReadAbility = subset(
  d_male_EngReadAbility,
  d_male_EngReadAbility$SP_EnglishReadingAbility!= "Na")
# calculate sample size for males and females
d_fem_EngReadAbility_n = length(
  d_fem_EngReadAbility[
    ,"SP_EnglishReadingAbility"])
d_males_EngReadAbility_n = length(
  d_male_EngReadAbility$SP_EnglishReadingAbility)
cat("The first five observations for
    males are:",
    d_male_EngReadAbility[1:5, "SP_EnglishReadingAbility"],
    "and the sample
    size for males is:",
    d_males_EngReadAbility_n, "The first
    five observations for females are:",
    d_fem_EngReadAbility[
      1:5, "SP_EnglishReadingAbility"],
    "and the sample size for females is:",
    d_fem_EngReadAbility_n)
```

The first five observations for males are: 5 4 4 5 4 and the sample size for males is: 79 The first five observations for females are: 5 4 3 5 5 and the sample size for females is: 147

```
d_male_EngReadAbility_mean = round(mean(
  d_male_EngReadAbility$SP_EnglishReadingAbility)
  , digits = 2)
d_fem_EngReadAbility_mean =
  round(mean(
    d_fem_EngReadAbility$SP_EnglishReadingAbility)
    , digits = 2)
cat("the mean for males and females is:",
```

```
    d_male_EngReadAbility_mean, "and",
    d_fem_EngReadAbility_mean, "respectively.")
```

the mean for males and females is: 4.44 and 4.02 respectively.

So, if we apply the formula for the mean for males and females we get, $\bar{x}_{male} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{5+4+4+5+4...x_i}{79} = 4.44$ and $\bar{x}_{female} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{5+4+3+5+5...x_i}{79} = 4.02$. In other words, the difference in mean score between males and females, $4.44 - 4.02 = .42$.

The question then becomes, is this difference enough to be significant, or may it just be due to chance that we observed this effect?

One way so find out is to model the data as a linear regression model or to use a two-sample t-test, but first we need to know a little about the variation in our samples. This is important, both in terms of gaining a better understanding of how much variation there is in our sample, but also because larger variation means more uncertainty, resulting in a worse fit when modelling the data. To understand this a bit better, lets think about our data as a linear model. Ultimately, we want to fit a model (i.e., estimating the parameters $\beta_0 + \beta_1 x_i$) and see how well it can predict scores on the dependent variable (i.e., the outcome variable being measured, which in this case is self-perceived English Reading Ability)

$$y_i = \beta_0 + \beta_1 x_i$$

However, inevitably there will be some error in our predictions, which, as previously explained, can arrive from different sources such as the measurement method, which in this case is a questionnaire, as well as randomness inherent to stochastic variables, i.e., random events and randomness associated with the sample itself. Therefore, we need to include an error term in the model, denoted as ε.

$$\begin{aligned} y_i &= \text{predicted self-perceived English reading ability} \\ y_i &= \beta_0 + \beta_{\text{gender}_{\text{male} \vee \text{female}}} x_i + \epsilon_i \end{aligned}$$

Here, y_i represents measurements on the dependent variable, namely self-perceived English reading ability, for subject x_i ... x_n, and β_0 and $\beta_1 x_i$ represents the fitted parameters, i.e., the intercept and slope, respectively. These parameters are fitted in a way that minimises the error in the model, denoted as ε. This procedure is often referred to as *The least squares method*, where the term *squares* refers to the squared error in the model, and can be calculated by using the following equation:

$$\sum_{i=1}^{n} \hat{\epsilon_i^2} = (y_i - \beta_0 + \beta_1 x)^2$$

To get a better grasp of what the equation means, it can be useful to break the equation down to its constituent parts.

By taking a closer look at the left-hand side of the equation inside the parentheses, one may notice that it contains formula for a linear regression model that has been re-arranged. As we can see below, the equation has been re-arranged so that the error term is isolated on one side, resulting in the following formula:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Instead of predicting values for the dependent or outcome variable, this re-arrangement results in the calculation of residuals, i.e., for $\text{subject}_{i=1}, \text{subject}_{i=2}$ ... subject_n, ϵ_i the equation can be expressed as a measure of the difference between the predicted value, $\hat{y}_i$, provided by the model, and the actual observed value, y_i, which is, for each observed value in the sample, is the error or ϵ_i.

for subject x_i ... x_n the deviance or offset, defined as offset or distance between the observed data points, $\hat{y}_{i=1}$, ...$\hat{y}_{i=n}$ across all subjects in the sample and the estimated parameters, denoted as $\hat{\beta}_0 + \hat{\beta}_1 x_i$, provided by the regression line fitted to the data.

More specifically, the prediction made by the model is subtracted from the actual observed value across all subjects, representing the offset or deviance for each observed value and the prediction made by the model. The sum of the deviance in the model squared, denoted as $\sum_{i=1}^{n} \hat{\epsilon_i^2}$, therefore represents how much error there is in the model, and hence is indicative of the accuracy of the model. The reason for squaring the deviance is due to the fact that the model will either *overshoot* or *undershoot* in each prediction, i.e., the deviance for each observation will either be a positive or negative number, and hence will cancel each other out. However, by squaring the deviance, the numbers will always be positive since squaring negative numbers always will, as is the case with squared positive numbers, result in positive numbers. Hence, by summing the difference between the predicted and observed values, the calculated error or sum of the deviance, will always grow as a function of how much error there is in the data.

Therefore, we can calculate the squared error in the model by using the re-arranged formula above by using the following formula $\sum_{i=1}^{n} \hat{\epsilon_i^2} = (y_i - \beta_0 + \beta_1 x)^2$ mentioned earlier. Since the value predicted by the model is a function of the intercept and slope and observed values plotted into the regression line, i.e., $\hat{y}_i = \beta_0 + \beta_1 x_i$, we can simplify the equation for the sum of squared errors (SSE) by substituting $\beta_0 + \beta_1 x$ with $\hat{y}_i$, resulting in the following formula:

$$\sum_{i=1}^{n} \hat{\epsilon_i^2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

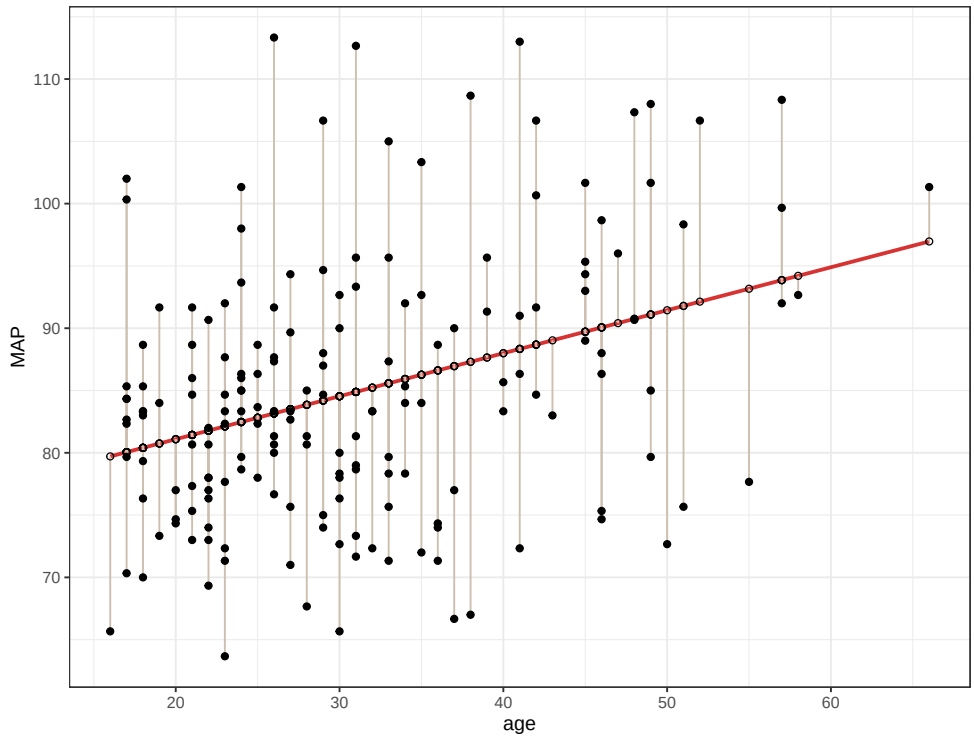
$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

While equations like this one may look intimidating, the explanation above should hopefully make understanding it a bit more intuitive. In essence, it says that the sum of the estimated squared error (the deviance from each datapoint) is equal to the the sum of each observed value, y_i minus the predicted value made by the model, $\hat{y}_i$.

The sum of the squared error is therefore a representation of relative accuracy or inaccuracy of the model, i.e., the relative inaccuracy of the model increases as the squared deviance gets larger, whereas when the squared deviance get smaller, the relative accuracy of the model increases. In statistical terms, accuracy represents how much of the variation can be explained by the model contra the amount of variance that remains unexplained.

This becomes more intuitive when illustrated, and I will use a fictional data set that turns out represent this concept graphically very well.

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



The concept of error is important to understand, since modelling data will always have

some degree of error, and when choosing between models, a general rule of thumb is to use the model that has the least amount of error.

The plot illustrates a regression line fitted to the data, i.e., the line connecting the dots that minimizes the error, also called the residual, in the model. In other words, for subject i the difference between the predicted and observed value, i.e. the residual or ϵ_i, is given by rewriting the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ to $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$. Therefore, the red line is fitted to the data so that average distance between each observed value (black dot) is as small as possible. Essentially, when a regression model is used to model data while minimising the error, the fitted equation can be defined as $y = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated parameters that provide the best fit to the data. In other words, for each observed value y_i and corresponding predictor value x_i the aim is to get the estimated or fitted value $\hat{y} = \beta_0 + \beta_1 x$ that minimises the sum of the squared error (SSE).

Results

Let's fit a linear model to our data where self-perceived English writing ability is plotted as a function of gender.

The results indicate that the model provides a significant fit, $R^2 = .06, 90\backslash\%$ CI $[0.02, 0.11]$, $F(1, 224) = 13.31$, $p < .001$ as well as a significant effect of gender, $b = 0.42$, 95% CI $[0.19, 0.65]$, $t(224) = 3.65$, $p < .001$.

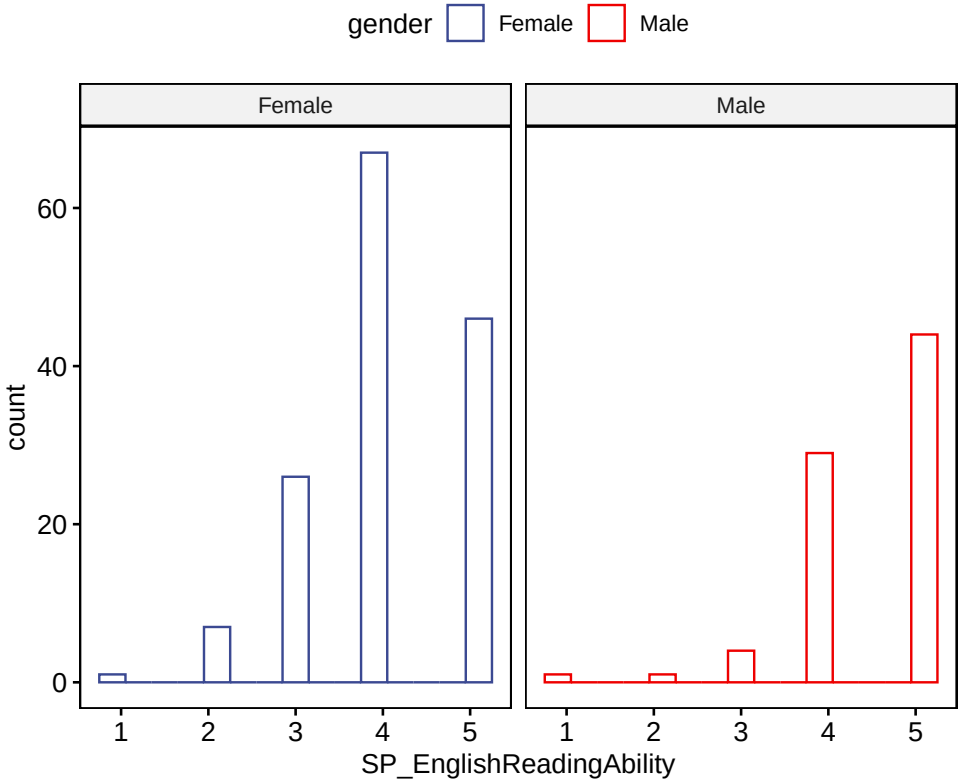
Fixed-Effects ANOVA Results for Sp_englishreadingability

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\,\eta^2$	$^{\text{partial}}\,\eta^2$ 90% CI [LL, UL]
(Intercept)	2376.06	1	2376.06	3446.41	<.001		
gender	9.18	1	9.18	13.31	<.001	.06	[.02, .11]
Error	154.43	224	0.69				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

This indicates that gender has a significant effect on self-perceived English Speaking Ability. By significant I mean, that the chance of observing a difference this large, assuming that in reality there is no difference as a function of gender, i.e., it happened by random chance, is less than $0.001 * 100$, i.e., 0.1%.

Before doing a t-test to confirm this, let's look at the distribution of responses as a function of gender.



First, we see that the sample size is much larger for females and that both groups generally rate themselves quite high. The real difference is that the males more consistently choose 4 or 5, whereas females also have many who choose 4 or 5, but a larger proportion of 3 and, to a lesser extent, 2's.

Let's do a t-test to confirm the result. Before doing that, let's test if the variance of the two groups are significantly different, which as the plot indicates, might be a possibility, which means we need to make a t-test that controls for unequal variances, namely the Welch's t-test. Let's first test if the variances are unequal and then, if they are, do a Welch's t-test. If they are equal, we can do a Students t-test.

Demonstration of a t-test

Testing null-hypothesis that the variances unequal

The variance is calculates using the following formula:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Essentially, it calculates, for subject x_i ... x_n how much each person differs from the mean $\bar{x}$ squared (as explained before, positive and negative cancel out unless the difference is

squared) divided by the sample size - 1.

```
## The variance for females and males is: 0.746156 and 0.5832522 respectively.
```

As we can see, there is more variance in the female group compared to males. Let's first the 95% confidence interval and see if it crosses 0 (which would indicate that there is no difference)

$$\begin{aligned}\text{CI95 for}\frac{\sigma_1^2}{\sigma_2^2} &= \left[\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha_{2,n_1-1,n_2-1}}} \mid \frac{s_1^2}{s_2^2}F_{\alpha_{2,n_1-1,n_2-1}}\right] \\ &= \left[1.279 \cdot (1/1.496); 1.279 \cdot 1.496\right] \\ &= [0.855; 1.869]\end{aligned}$$

As we can see, the confidence interval does not contain or cross 0. Therefore, to test the hypothesis, we would test the null-hypothesis H_0 (therer is no difference), and see whether it is true or should be rejected from the alternative hypothesis (there is a difference).

$H_0 :\ \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Test statistic :

$$F_{1-\alpha/2,n_1-1,n_2-1} \text{ and } F_{\alpha/2,n_1-1,n_2-1} = \frac{1}{F_{\alpha/2,n_1-1,n_2-1}}$$

$$F_{\alpha/2,n_1-1,n_2-1} = \frac{1}{F(0.025, 146, 78)}$$

$$F(0.025, 146, 78) = 0.684 \text{ and } 1.496$$

Conclusion: we do not reject the null hypothesis (H_0) $p = 0.229$. In other words, the variances are unequal, since the null-hypothesis could not be rejected. Our histogram did also suggest that the variation looked unequal.

Welch t-test: testing if there is a difference in self-perceived English speaking ability between males and females

We therefore need to use Welch's T-test.

$$\text{CI95} = \mu_1 - \mu_2 \ = \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{where } \ \nu \ = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = 177.339$$

$$\begin{aligned} \Rightarrow 95\text{ CI for } \mu_1 - \mu_2 &= 4.443 - 4.02 \pm (1.973 * 0.112) \\ &= [0.202; 0.643] \end{aligned}$$

As we can see, the confidence interval does not cross 0, so it indicates a significant difference. Let's state the hypotheses:

H_0 is $\mu_1 - \mu_2 = 0$ and H_1 is $\mu_1 - \mu_2 \neq 0$

Test statistic :

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

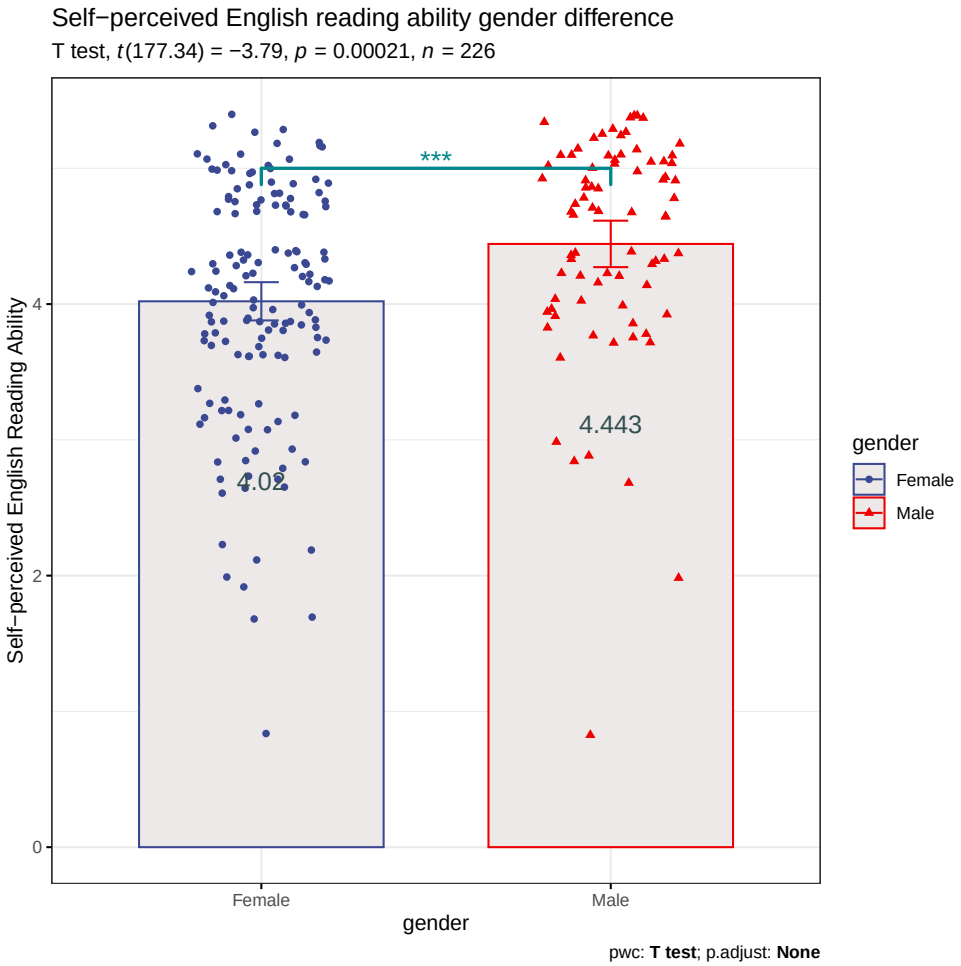
$$= \frac{(4.443 - 4.02 - 0)}{0.112} = 3.786$$

The Critical value: $\pm t_{\alpha/2,\nu} = \pm t(0.025, 177.339) = \pm 1.973$

The conclusion is therefore that we reject the null-hypothesis H_0 that the true difference in means is 0 $p < 0.001$.

Let's make a barplot to illustrate this:

```
## Warning in (function (mapping = NULL, data = NULL, stat = "count", position =
## "stack", : Ignoring unknown aesthetics: shape
```




```
## Warning in (function (mapping = NULL, data = NULL, stat = "count", position =
## "stack", : Ignoring unknown aesthetics: shape
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_summary()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

Tabelle 1

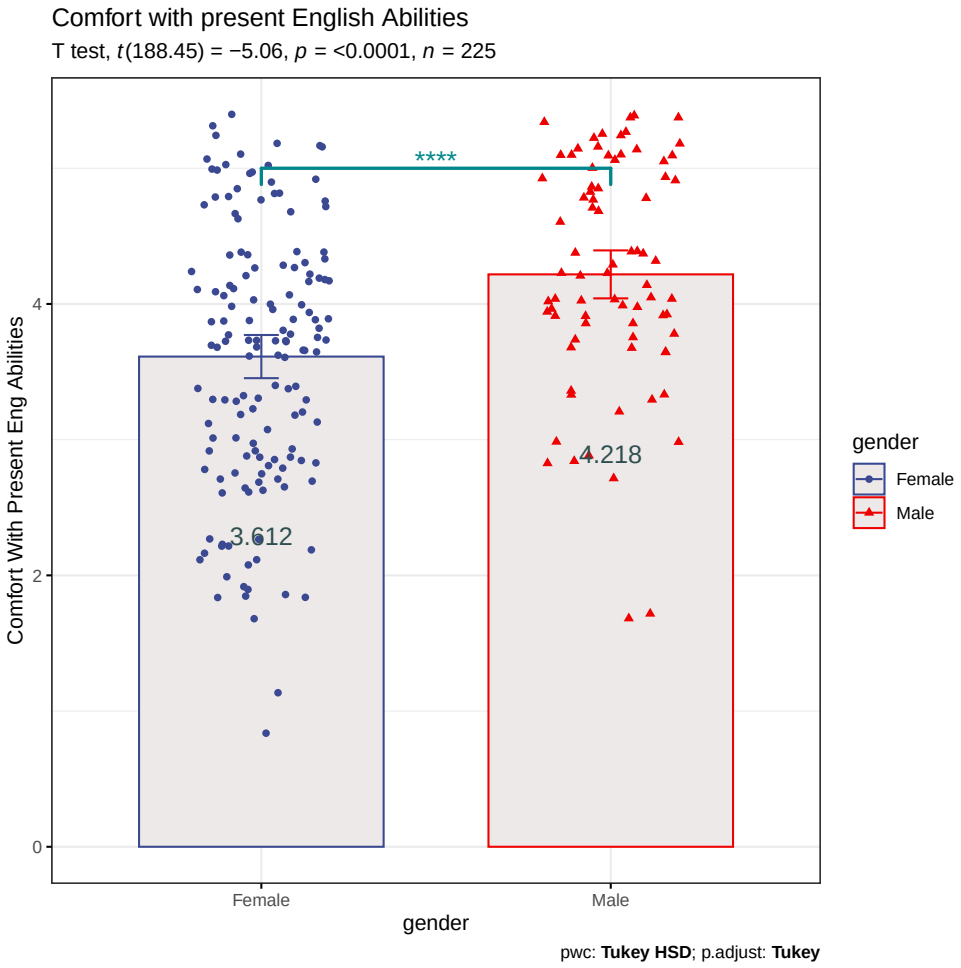
Females	n	mean	sd	se
SP_EnglishReadingAbility	147.00	4.02	0.86	0.07
SP_EnglishSpeakingAbility	147.00	3.68	0.86	0.07
SP_EnglishWritingAbility	147.00	3.63	0.91	0.08
SP_EnglishAbilityInGeneral	146.00	3.71	0.88	0.07
ComfortWithPresentEnglishAbilities	147.00	3.61	0.98	0.08
FreqReadWatchingEngInNonEngClasses	147.00	2.97	0.91	0.08
CollaborationBetweenEngVSnonEngClasses	145.00	2.27	1.00	0.08
ConsideredStudiesWithEngTeaching	147.00	3.11	1.28	0.11
ImportanceofEngToYouInFutureStudies	144.00	2.72	1.22	0.10
ExpectationsOfLikelihoodToReadEngTexts	145.00	3.21	1.06	0.09
ToWhichDegreeExpectCommunicationInEnglish	143.00	2.77	1.00	0.08
ComfortableWithEngTeachingMaterials	145.00	3.41	1.14	0.10
PreparednessForPotentialEngTeachingLater	145.00	3.14	1.27	0.10
WouldRequireExtraSupportForEng	145.00	2.03	1.02	0.08
DoesSchoolSufficientlyPrepareforEngTeachingLater	147.00	2.77	0.91	0.08
ImportanceOfEnglishToForFindingWork	143.00	2.32	1.14	0.10
RatedLikelihoodToReadEngTextsFutureWork	145.00	2.75	1.02	0.08
RatedLikelihoodOfEngCommunicationFutureWork	145.00	2.37	0.91	0.08

Tabelle 2

Males	n	mean	sd	se
SP_EnglishReadingAbility	79.00	4.44	0.76	0.09
SP_EnglishSpeakingAbility	79.00	4.05	0.75	0.08
SP_EnglishWritingAbility	78.00	4.01	0.80	0.09
SP_EnglishAbilityInGeneral	78.00	4.13	0.78	0.09
ComfortWithPresentEnglishAbilities	78.00	4.22	0.78	0.09
FreqReadWatchingEngInNonEngClasses	78.00	3.14	1.12	0.13
CollaborationBetweenEngVSnonEngClasses	76.00	2.17	1.00	0.12
ConsideredStudiesWithEngTeaching	77.00	3.26	0.98	0.11
ImportanceofEngToYouInFutureStudies	77.00	2.97	1.05	0.12
ExpectationsOfLikelihoodToReadEngTexts	77.00	3.48	0.88	0.10
ToWhichDegreeExpectCommunicationInEnglish	77.00	3.05	0.90	0.10
ComfortableWithEngTeachingMaterials	77.00	3.86	0.94	0.11
PreparednessForPotentialEngTeachingLater	77.00	3.66	1.11	0.13
WouldRequireExtraSupportForEng	77.00	1.65	0.86	0.10
DoesSchoolSufficientlyPrepareforEngTeachingLater	78.00	2.88	0.76	0.09
ImportanceOfEnglishToForFindingWork	78.00	2.59	1.14	0.13
RatedLikelihoodToReadEngTextsFutureWork	77.00	2.91	1.16	0.13
RatedLikelihoodOfEngCommunicationFutureWork	78.00	2.62	0.97	0.11

Tabelle 3

Outcome	n1	n2	p	p_adj	signif
ComfortWithPresentEnglishAbilities	147	78	0.00	0.00	****
ComfortableWithEngTeachingMaterials	145	77	0.00	0.00	**
ConsideredStudiesWithEngTeaching	147	77	0.33	0.35	ns
ToWhichDegreeExpectCommunicationInEnglish	145	78	0.07	0.12	ns
ExpectationsOfLikelihoodToReadEngTexts	145	77	0.32	0.35	ns
DoesSchoolSufficientlyPrepareforEngTeachingLater	147	78	0.31	0.35	ns
FreqReadWatchingEngInNonEngClasses	147	78	0.24	0.31	ns
ImportanceOfEnglishToForFindingWork	143	78	0.09	0.13	ns
CollaborationBetweenEngVSnonEngClasses	145	76	0.49	0.49	ns
SP_EnglishAbilityInGeneral	146	78	0.00	0.00	**
SP_EnglishReadingAbility	147	79	0.00	0.00	**
SP_EnglishSpeakingAbility	147	79	0.00	0.00	**
SP_EnglishReadingAbility	147	78	0.00	0.00	**
PreparednessForPotentialEngTeachingLater	145	77	0.00	0.00	**
ImportanceOfEnglishToForFindingWork	144	77	0.10	0.14	ns
RatedLikelihoodOfEngCommunicationFutureWork	143	77	0.03	0.07	ns
RatedLikelihoodToReadEngTextsFutureWork	145	77	0.05	0.09	ns
WouldRequireExtraSupportForEng	145	77	0.00	0.01	**



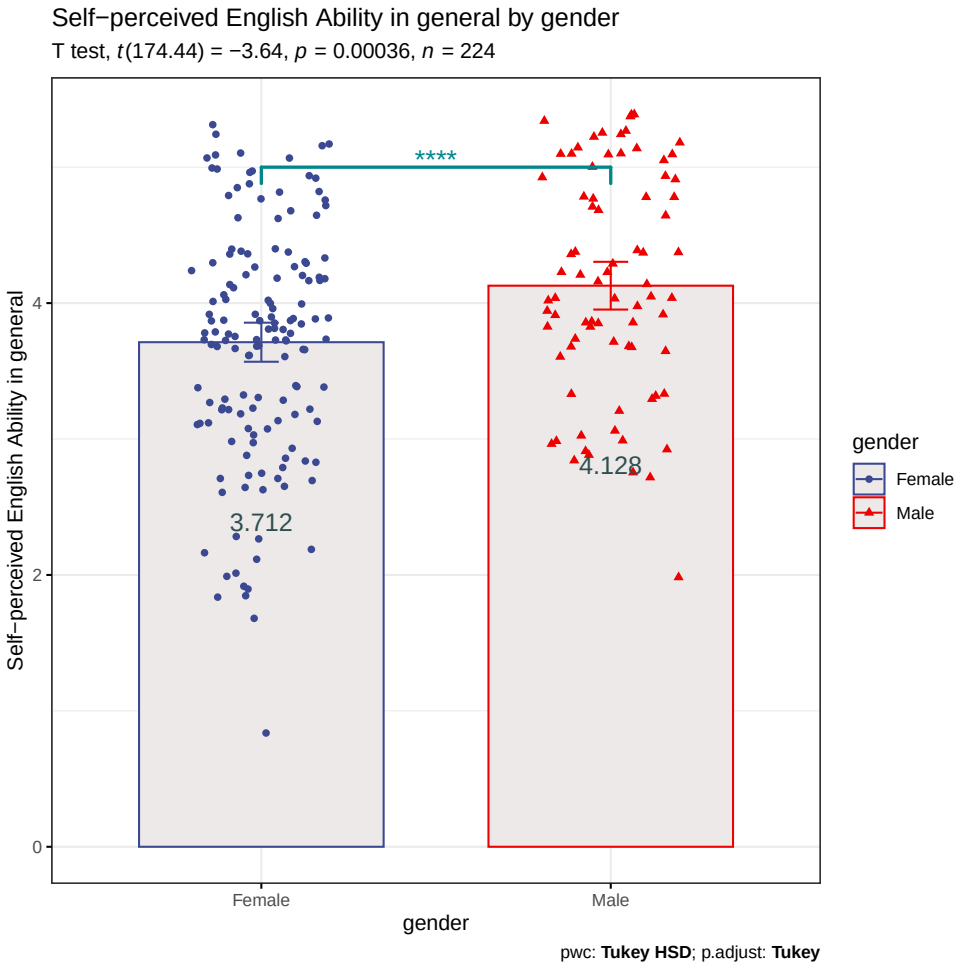
A tibble: 1 x 13

term group1 group2 null.value estimate conf.low conf.high p.adj 1 gender Female Male 0 0.606 0.353 0.858 0.00000395 # i 5 more variables: p.adj.signif , y.position , # groups , xmin , xmax

```
## Warning in (function (mapping = NULL, data = NULL, stat = "count", position =
## "stack", : Ignoring unknown aesthetics: shape
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_summary()`).
```

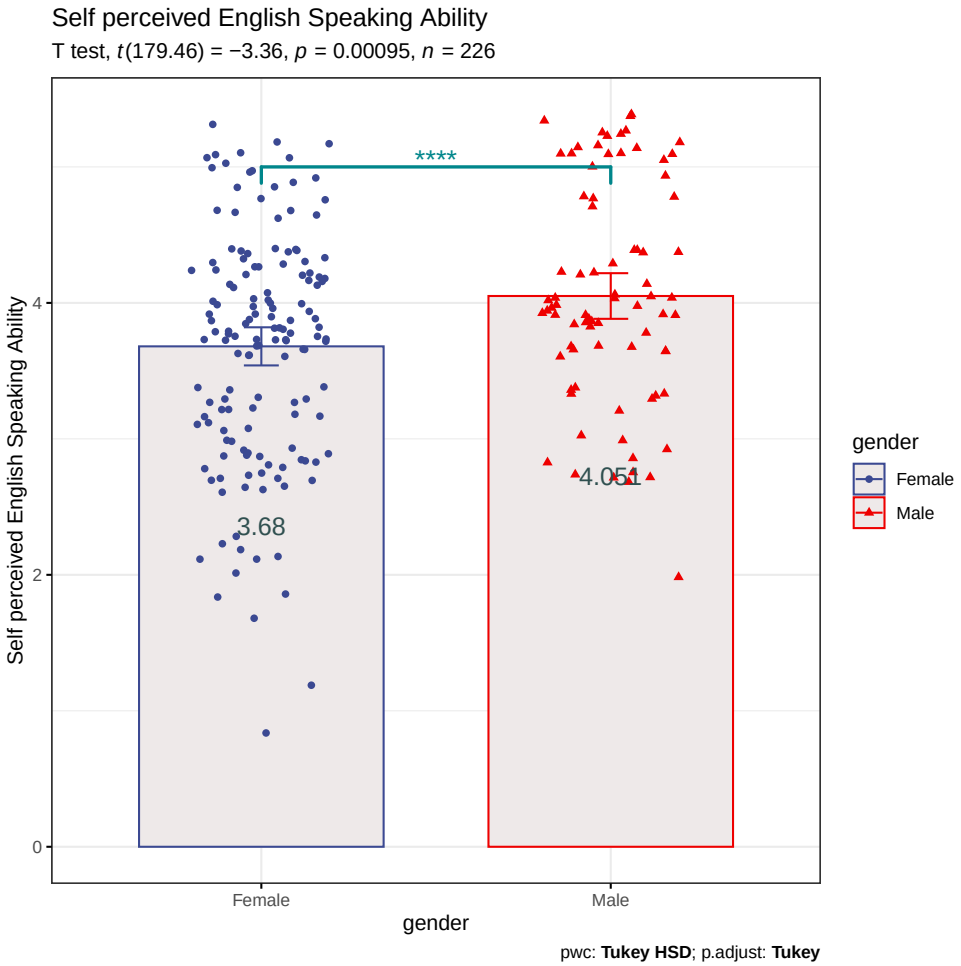
```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



A tibble: 1 x 13

term group1 group2 null.value estimate conf.low conf.high p.adj 1 gender Female Male 0 0.416 0.182 0.650 0.000547 # i 5 more variables: p.adj.signif , y.position , # groups , xmin , xmax

```
## Warning in (function (mapping = NULL, data = NULL, stat = "count", position =
## "stack", : Ignoring unknown aesthetics: shape
```



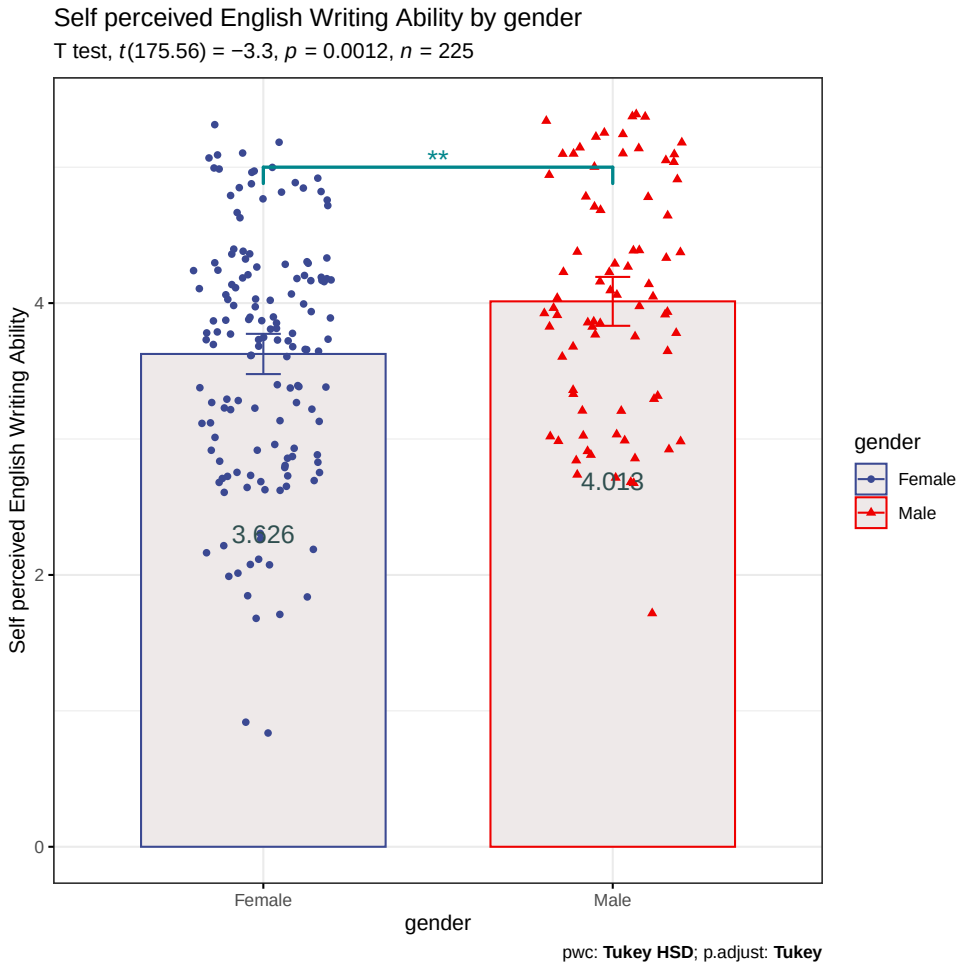
A tibble: 1 x 13

term group1 group2 null.value estimate conf.low conf.high p.adj 1 gender Female Male 0 0.370 0.144 0.597 0.00145 # i 5 more variables: p.adj.signif , y.position , # groups , xmin , xmax

```
## Warning in (function (mapping = NULL, data = NULL, stat = "count", position =
## "stack", : Ignoring unknown aesthetics: shape
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_summary()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



A tibble: 1 x 13

term group1 group2 null.value estimate conf.low conf.high p.adj 1 gender Female Male 0 0.387 0.146 0.628 0.00174 # i 5 more variables: p.adj.signif , y.position , # groups , xmin , xmax

References

Johnsen, Svend Heini W, and Hana Malá Rytter. 2021. "Dissociating Spatial Strategies in Animal Research: Critical Methodological Review with Focus on Egocentric Navigation and the Hippocampus." *Neuroscience & Biobehavioral Reviews* 126: 57–78.

Popper, Karl R. 1963. "Science as Falsification." *Conjectures and Refutations* 1 (1963): 33–39.