Statistics Review

Anthony Tricarico

Table of contents

1 Intro

This document is a collection of useful statistical concepts that will help better understand the topics of data analysis and forecasting. Depending on whether you have already taken statistics class you might find some concepts that you already know well, if that is the case feel free to skip forward or just skim through this document. Otherwise, I am sure that you will find this document helpful. Although your textbook is very good, it assumes you have prior knowledge about R and statistics and if that is not the case you might find it hard at times to fully grasp what is in it. For the purpose of this short document, topics will be divided into:

1. Descriptive statistics
2. Inferential statistics

2 Descriptive statistics

Descriptive statistics is the field of statistics that deals with summarizing and making large quantities of data understandable, highlighting their main features to prepare for analysis. It does so through basic algebra and visuals, which might help grasp specific features of the data that would otherwise be very hard to spot. With R you can generate summaries and plots of large quantities of data, but it is important to understand what these results convey and how you should interpret them.

2.1 Measures of central tendency

Most of the times, you are interested in understanding what value your observations (i.e., rows in your dataset) cluster around. In other words, you might be interested in knowing what are the most frequent values in your dataset. For that, you can use various *measures of central tendency*. Among these, the *mean* (a.k.a., average) is probably the most popular followed by the *median* and the *mode*. Let's take a look at them one by one.

2.1.1 Mean

The mean is defined as the sum of all values that a specific variable (i.e., a column in a dataset) in the dataset takes divided by the number of observations for your variable (i.e., the number of rows). Compactly, you can write this with the following notation.

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

The summation notation above indicates that you should sum all your observations (x_i) and then divide by n (or multiplying by $\frac{1}{n}$ which is equivalent!) the sum you obtained. The mean is usually a good measure of central tendency, but it is very sensitive to extreme values (both low and high), so you should be cautious when interpreting the mean for highly skewed variables in your dataset. An example is in order.

```
var <- c(1,1,1,2,2,2,3,4,5,6,7,8,9)
var_skewed <- c(1,1,1,1,2,3,4,10000000)                    ①

mean(var)                                                  ②
```

① creating a vector containing an extreme value
② mean of the non-skewed variable

```
[1] 3.923077
```

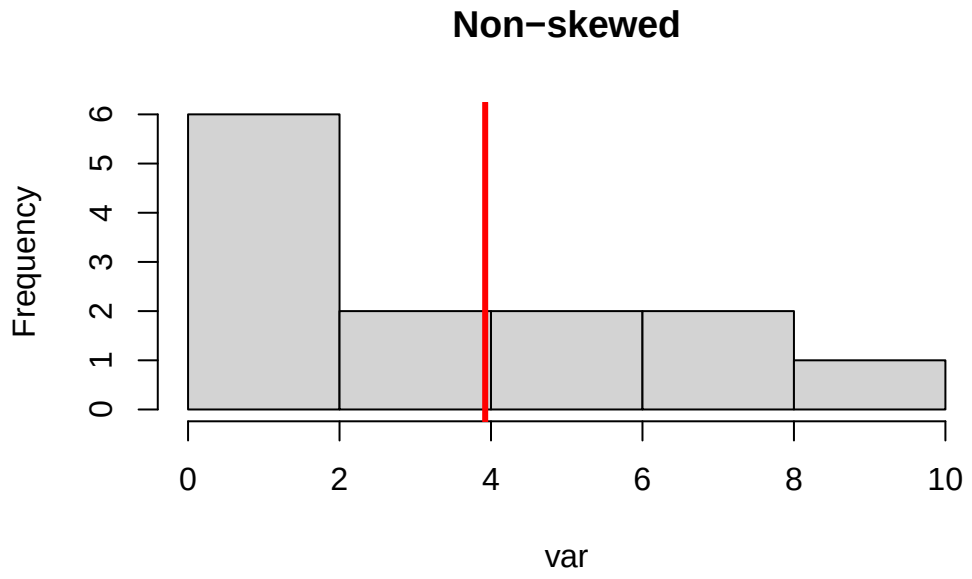
```
mean(var_skewed)
```

```
[1] 1250002
```

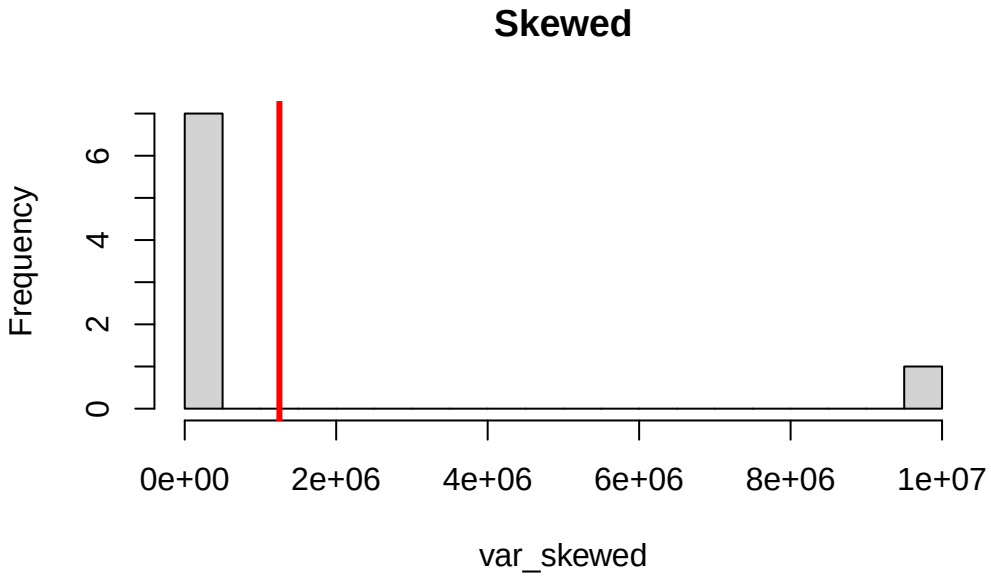
Notice how the vector containing the 10000000 has a much higher mean, illustrating how sensitive this measure is to extreme values. We can also plot this to show the difference in how those variables are *distributed*.

```
hist(var, main = "Non-skewed")                                    ①
abline(v = mean(var), col="red", lwd=3)                           ②
```

① plotting an histogram of the *var* variable
② adding a vertical (`v`) line where the mean of the variable is located




```
hist(var_skewed, breaks = 30, main = "Skewed")
abline(v = mean(var_skewed), col="red", lwd=3)
```



Notice how in the figure above the mean does not do a good job at identifying the value around which the data cluster. To improve on this, we introduce the *median*.

2.1.2 Median

The median is a measure of central tendency that is not sensitive to extreme values in the dataset (a.k.a., *outliers*). It corresponds to the middle point in an ordered set of data (when the total n is odd) or the mean of the two central middle observations (when the total n is even). To show how this measure works, let's take a look at how the median behaves when we plot it on the histogram of our data.

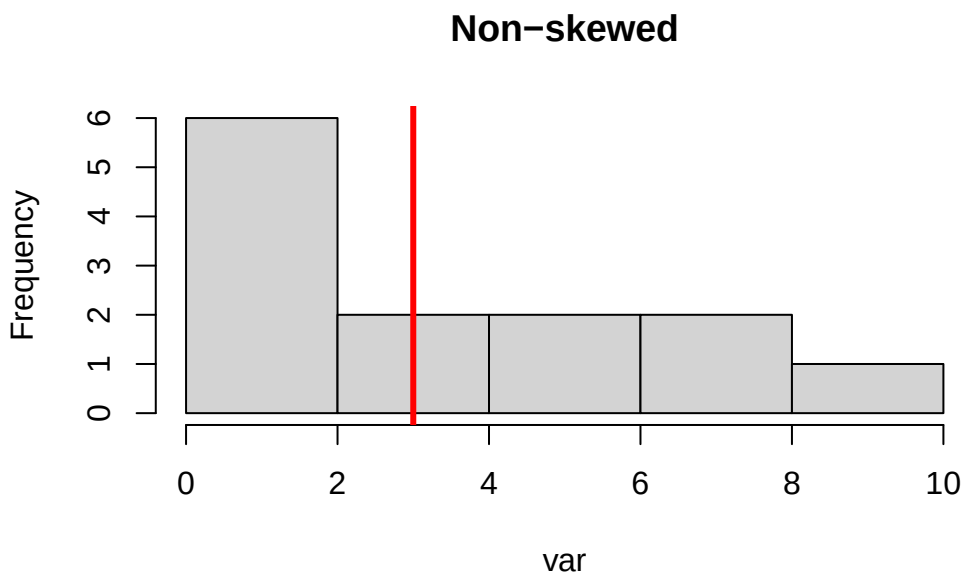
```
median(var)
```

```
[1] 3
```

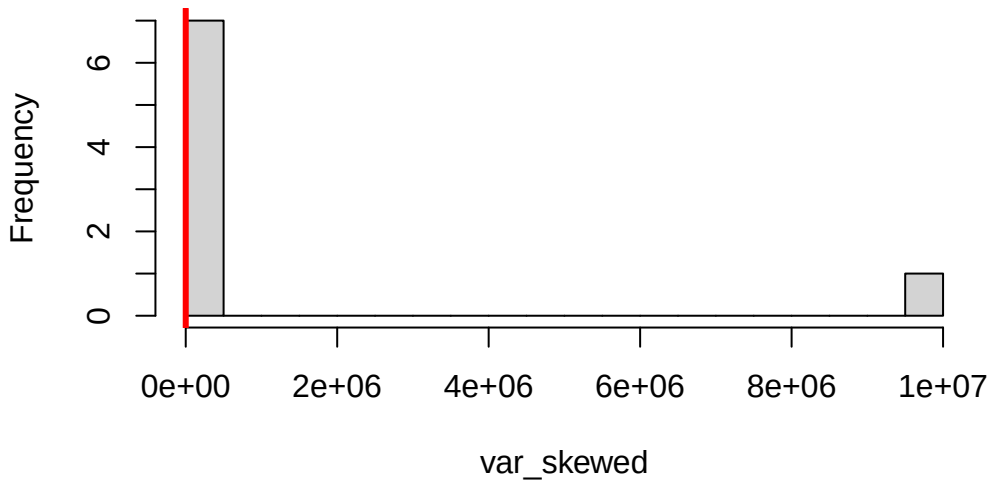
```
median(var_skewed)
```

```
[1] 1.5
```

```
hist(var, main = "Non-skewed")
abline(v = median(var), col="red", lwd=3)
```




```
hist(var_skewed, breaks = 30, main = "Skewed")
abline(v = median(var_skewed), col="red", lwd=3)
```



from here you can clearly see that when data is highly skewed, the median is able to better pinpoint where most of the data will be located.

2.1.3 Mode

Finally, there is the mode which is simply the value of the observation that occurs most frequently. Let's have a look at the sample data we have to compute it. For that we can use the `table()` function in R and pass in the variable `var` as an argument to it which will return a *frequency table* showing how often each value appears in our variable.

```
table(var)
```

```
var
1 2 3 4 5 6 7 8 9
3 3 1 1 1 1 1 1 1
```

The mode for the variable above is not unique (both 1 and 2 are modes). Indeed, the distribution of a variable might have one or more modes while others don't even have one.

```
table(var_skewed)
```

```
var_skewed
    1     2     3     4 1e+07
    4     1     1     1     1
```

Instead, for the case above, the mode is unique and is the value 1.

> **Note**
>
> As a side note, consider the following variable X which can take values 1,2,3,4,5. Since each value appears with the same frequency (e.g., $\frac{1}{5}$), X does not have a mode!

2.2 Measures of dispersion

It is also useful to quantify by how much variables vary on average. These measures are often referred to as measures of dispersion and there are a few which are fundamental for any statistical application.

2.2.1 Variance

The variance of a variable measures the spread between its values and its mean.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{1}$$

The term $(x_i - \mu)$ is often referred to as *deviation from the mean*. When we add all those up (notice the $_i$ to the bottom right of the x) and we square them we have a measure of how far values of a variable are from their mean. Since we want an average value, we divide the *squared sum of all the deviations* by the sample size N.

> **Note**
>
> μ is referred to as the population mean.
> Also, it is important to remember that the sum of all deviations from the mean is 0. This is why we square them so as to avoid negative terms in the sum and have a way to properly quantify the variance of a variable.

Variables with higher variance have values that are further away from their mean on average. The opposite is true. In R we use the `var()` function to compute the variance of a numeric variable passed as argument in the function.

```
var(var)                                                    ①
```

① here the outside `var` refers to the variance function while the *var* inside the parenthesis is the argument to the function, which in this case is our *var*iable!

```
[1] 7.910256
```

```
var(var_skewed) #result is 12500000000000
```

```
[1] 1.25e+13
```

Think about the two results above and make sure they make sense before moving on! (Don't get confused by the scientific notation employed here by R in the output)

Caution

When computing the variance of a variable you have to keep in mind that the resulting variance will be denominated by the square of the original unit of measure.
For example, if my variable R representing the return on a stock is denominated in USD its variance ($Var[R]$) will be denominated in USD^2. This makes the results less interpretable because in real life there is no such thing as squared dollars!
Next we learn how to address this problem.

Note

The formula for the variance introduced in Equation 1 is commonly referred to as the *population variance* of X. Note that there is a difference between this quantity and the sample variance which is computed as follows:

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

Notice the use of the s to represent a sample statistic. Sample statistics are usually denoted by Latin alphabet letters as opposed to population statistics denoted by letters from the Greek alphabet.
Also, we are dividing by $n-1$ due to the fact that we have one less degree of freedom which has been used to compute the sample mean $\bar{x}$ as we use it to estimate the unknown population mean μ.

2.2.2 Standard deviation

Standard deviation is defined as the square root of the variance. This solves the issue with the squared units of measures and brings everything back to unit scale so that the results become easier to interpret. In R you use `sd()` to compute the standard deviation of a variable.

```
sd(var)
```

```
[1] 2.812518
```

```
sd(var_skewed)
```

```
[1] 3535533
```

2.2.3 Additional resources

We've seen most of the basics by now, but in case you want to learn more about measures of dispersion here is a list of topics which you might find useful:

1. Quantiles and Quartiles
2. Interquartile range

2.2.4 The Two Variables Case

Sometimes, we are not interested only in describing the behavior of a single variable but we might want to study how it interacts with another variable. Quantifying this relationship is very easy thanks to some tools such as the **covariance** of two variables and the **correlation coefficient**.

2.2.4.1 Covariance

Covariance refers to how much two variables co-vary (i.e., vary together). It is useful to look at the formula of how it is computed to derive some intuition about how it works.

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

In other words, the covariance of two variables X and Y is determined by how many times they are both above (or below) their respective means, $\bar{x}$ and $\bar{y}$. Intuitively, if they are both above (or below) their mean we would be multiplying together two numbers with the same sign, and the result will be a positive number. Then, following simple reasoning, if we add together many positive numbers we get a number that is larger than if we were to add an equal number of some negative and some positive numbers together (assuming they are the same in absolute value). It follows that two variables that are more often than not above their means, at the same time, will have a higher covariance, meaning that these variables will be growing (or shrinking) together.

The opposite of this is the case when two variables have a covariance that is close to 0, which is the same as saying that the two variables are independent (i.e., knowing that one of the variables changes tells me nothing about how the other behaves).

> **Caution**
>
> Covariance has the same issue that variance has: squared units of measure. We learn how to deal with this next.

2.2.4.2 Correlation coefficient

In a similar way to how the standard deviation solves the issue of squared units of measures for the variance, so does the correlation coefficient for the covariance. Here is its formula:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

From here, we notice some important properties of r:

1. It is scaled (i.e., divided) by the product of the standard deviation of the two variables, and therefore r now always falls in the range [-1,1]. This allows us to have boundaries to interpret the strength of the relationship between X and Y.
2. The closer $|r|$ is to 1 (in absolute value) the stronger the relationship between the two variables.
3. The sign of r is determined solely by the sign of the $\text{Cov}(X, Y)$ at the numerator since the standard deviations at the bottom are always positive values.

> **!** Important
>
> Notice that r measures the strength of *linear correlation* between two variables. This means that if X and Y are linked by a non-linear relationship, r will not be able to detect it! We show this with an example.

```
lin_1 <- seq(1,10)                                                    ①
lin_2 <- seq(-1,-10)                                                  ②

cor(lin_1, lin_2)                                                     ③
```

① generate sequence of numbers from 1 to 10
② generate sequence of numbers from -1 to -10
③ compute the `cor`relation between these two linear sequences

```
[1] -1
```

```
sq_2 <- lin_1^2                                                       ①

cor(lin_1, sq_2)                                                      ②
```

① `sq_2` is the square of the sequence of numbers from 1 to 10 contained in `lin_1`.
② the correlation coefficient computed will now have more troubles recognizing this non-linear relationship and produce a lower correlation coefficient (in absolute value).

```
[1] 0.9745586
```

3 Inferential statistics

Inferential statistics refers to how we can make inferences about a population starting from a representative sample drawn from it.

3.1 Probability Distributions

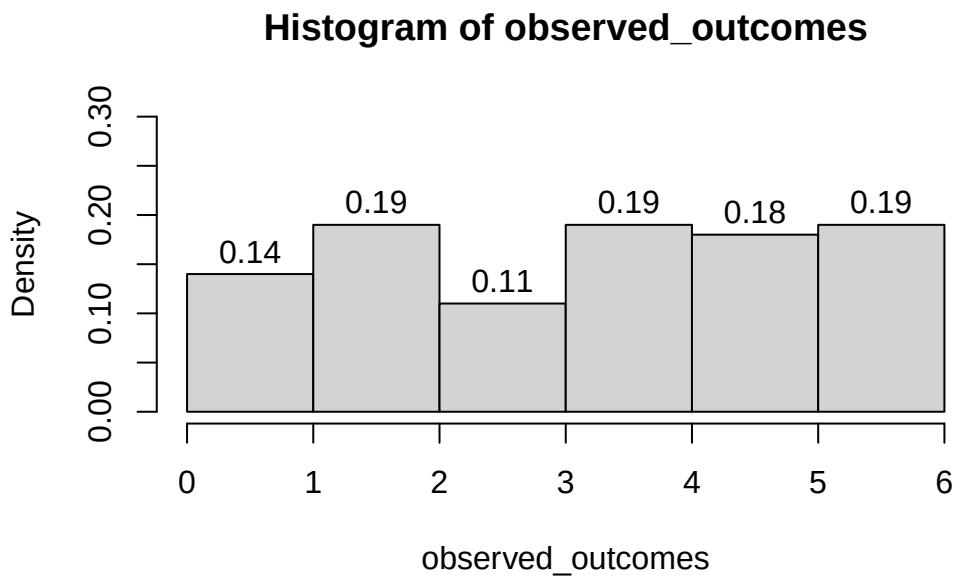
Probability theory refers to the set of tools mathematicians developed to study and model uncertainty. Here we are just going to introduce elementary topics to not make things too complicated.

When we try to understand what are the chances that something happen or does not happen, we are basically questioning whether an *event* will or will not take place. An event is an observable outcome which can either happen or not. For instance, if my event A is defined as observing a 2 on a single die roll it can either happen (I roll the die and I get a 2) or it does not happen (I get a number other than 2). Now, assuming that the die is fair (i.e., every number as an equal chance to show), we can assign a probability to the event that we get a 2, denoted by $P(A) = \frac{1}{6}$ (meaning there is just **a single 2**, over a **total of 6 possible numbers** that I can get from a single die roll). In general, we can denote by X a random variable that takes on each roll the value of the observed number. Then, X is random because its value cannot be predetermined exactly *a priori* but is the result of the following die roll.

3.1.1 Binomial distribution

Knowing this, we can use a probability distribution to compute the probability that we observe a specific number of 2's over a series of n die rolls. This is an example of a binomial experiment. If you are interested you can learn more about it but for now it suffices to say that we can model this using a binomial distribution. Plotting the distribution we observe that rolling the die 100 times yields the following.

```
set.seed(126)
possible_outcomes <- seq(1,6)
observed_outcomes <- sample(possible_outcomes, size = 100,
                            replace = T, prob = c(1/6,1/6,1/6,1/6,1/6,1/6))

hist(observed_outcomes, breaks = c(0,1,2,3,4,5,6), freq = F,
     labels = T, ylim = c(0,.3))
```



In this case 19% of all rolls (meaning 19 rolls) produced had as an outcome a 2. We could have computed this probability before conducting the experiment so as to form an expectation of what the outcome of the die roll would more likely be by multiplying the probability of observing a 2 on a die roll ($p = \frac{1}{6}$) by the number of times the die is rolled ($n = 100$) and the expected numbers of 2 would have been approximately 17, not that far from what we observed!

3.1.2 Normal distribution

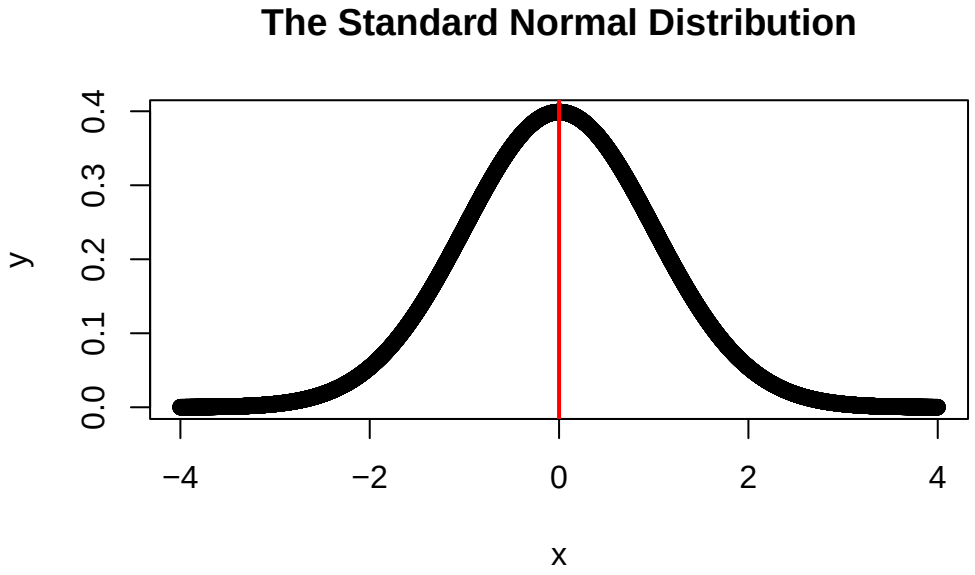
If you had to learn about only one distribution, let that be the Normal Distribution. There are so many important things there are to say about this distribution, but let's keep it simple:

1. it is used to model many real-life phenomena that take on a continuum of values (and not a discrete number of values, like the previous die roll example)
2. it depends on only two parameters: its mean ($ $), and its variance ($ ^2$)

3. given some assumptions, any distribution can be approximated by the Normal Distribution! This is a result of the Central Limit Theorem.
4. it has many nice properties (symmetry about μ, bell shape, etc.) and is used all over statistics and forecasting, along with its sample counterpart being the T distribution

Let's plot it out

```
x <- seq(-4,4,.001)
y <- dnorm(x)

plot(x,y, main = "The Standard Normal Distribution")
abline(v=mean(x), col = "red", lwd=2)
```



there is also a special case of the normal distribution which occur when $\mu = 0$ and $\sigma^2 = 1$. This is called the standard normal distribution (represented by Z). This means that one can *standardize* any normal random variable to bring it back to this standard scale and obtain a distribution like the one above. To do that, you simply need to subtract the *mean* of the variable you are standardizing and divide by its *standard deviation*.

$$Z = \frac{X - \mu}{\sigma}$$

3.1.3 Student's T distribution

Another useful distribution is the T distribution. Shape-wise the T distribution has the same nice properties of the standard normal distribution (of which it is a generalization). However, it is common to use this distribution when we do not know the true population value of σ^2 and we have to use its sample counterpart s^2 as its estimate. The T distribution comes with degrees of freedom which are used to assess how many data points are left free to vary after having estimated the population parameters that we need from the sample. In this case, note that since we are using s^2 to estimate σ^2 we lose one degree of freedom and our T distribution will be defined by $n - 1$ degrees of freedom.

3.2 Tests of hypotheses

Given two groups of people, you might be interested in understanding which group is taller on average. This can be translated into a test of hypothesis. Let's say you believe group 1 to be taller on average, how would you go about testing this quantitatively? First, let's define the two hypotheses:

$$H_0 : \mu_1 \leq \mu_2$$

$$H_a : \mu_1 > \mu_2$$

Let's break this down:

1. Here H_0 is known as the null hypothesis, and it represents the *status quo* or what we assume to be the truth.
2. H_a is the alternative (or research) hypothesis, and it is what we want to find evidence for. Note that finding supporting evidence for H_a makes H_0 less credible.
3. μ_i for $i = 1, 2$ represents the theoretical population mean of group 1 and 2, which we are trying to infer from our sample data.

Now, it would useful to have a metric that could help us determine whether there is enough supporting evidence for H_a to *reject* H_0. To do this, we introduce the concepts of significance level (denoted by α) and p-value (denoted by p). The p-value of a test is the probability of observing the result of our analysis if we establish that the null hypothesis is true. Of course, a small p-value tells us that what we observed is unlikely to happen if the null hypothesis was effectively true. How small the p-value needs to be to reject H_0 depends on the significance level we are using for our tests. Commonly, the .01, .05, and .1 significance levels are the most used ones, but there are scientific fields that require even smaller α (e.g., physics). Given the concepts of p-value and significance level just discussed, the following method is used to determine whether one should reject or do not reject the null hypothesis:

> **!** Important
>
> 1. If $p < \alpha \rightarrow$ reject H_0
> 2. if $p > \alpha \rightarrow$ do not reject H_0

To go back to our original height problem and to determine whether group 1 is truly taller on average compared to group 2 we need to introduce the t-statistic. The t-statistic comes from the T distribution and it is used to quantify by how much two sample means differ from each other.

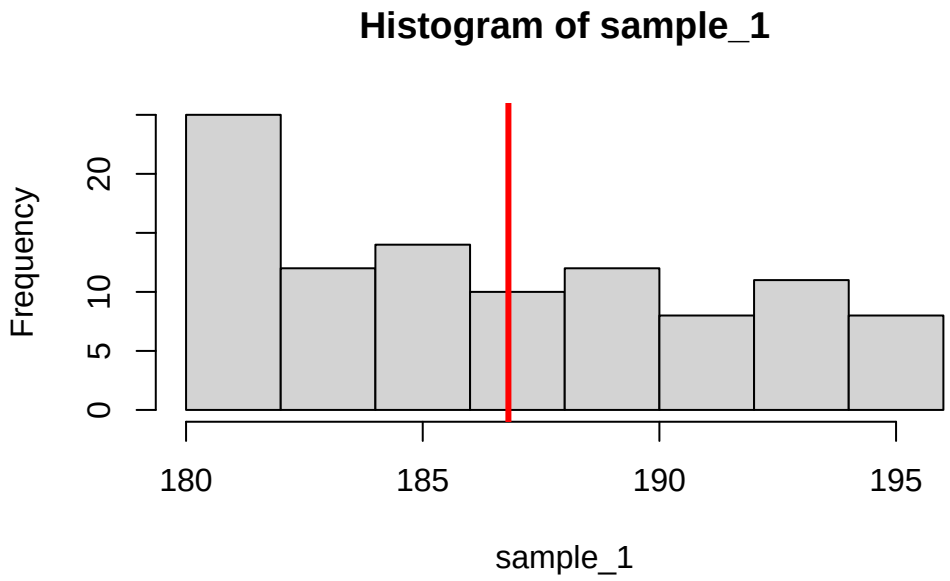
```
set.seed(128)                                                    ①
pop_1 <- seq(180,195)                                            ②
pop_2 <- seq(165,180)                                            ③

sample_1 <- sample(pop_1, 100, replace = T)                      ④
sample_2 <- sample(pop_2, 100, replace = T)                      ⑤
```

① Set the seed to get reproducible results across iterations
② specify that population 1 is made up of individuals with height between 180cm and 195cm

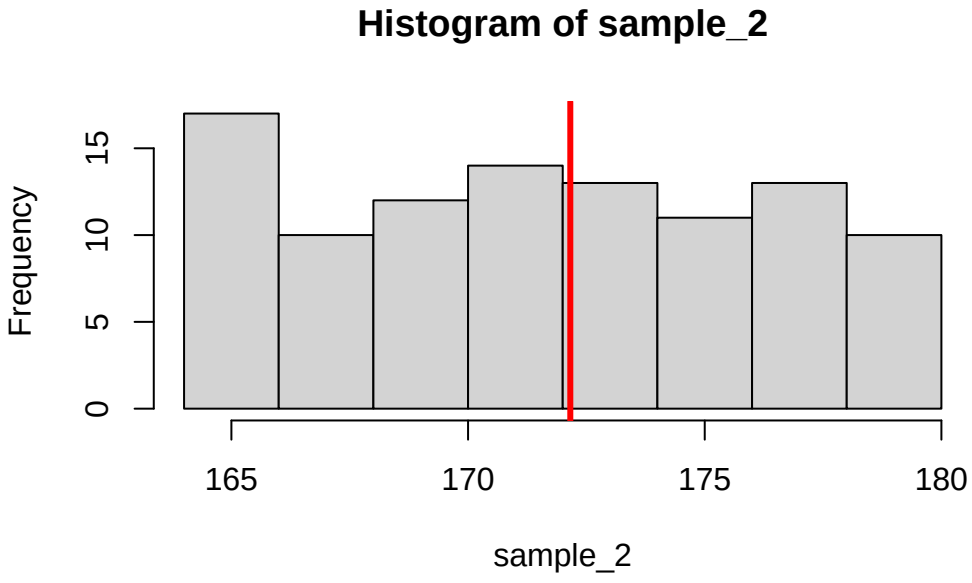
③ specify that population 2 is made up of individuals with height between 165cm and 180cm
④ sample 100 elements *with replacement* (i.e., allowing for the same value to be drawn multiple times) from population 1
⑤ sample 100 elements *with replacement* from population 2

Let's plot the distribution of the height of the two samples.

```
hist(sample_1)
abline(v=mean(sample_1), col="red", lwd=3)
```




```
hist(sample_2, xlim = c(164,180))
abline(v=mean(sample_2), col="red", lwd=3)
```



Finally, we conduct a t test to determine whether there is a significant difference in means between the two samples.

```
(t_test <- t.test(sample_1, sample_2, alternative = "greater"))
```

```
    Welch Two Sample t-test

data:  sample_1 and sample_2
t = 21.546, df = 197.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 13.52635      Inf
sample estimates:
mean of x mean of y
   186.81    172.16
```

```
t_test$p.value
```

```
[1] 4.361519e-54
```

Given a level of significance $\alpha = .01$ and a $p < .001$ we can reject H_0. This means that under the assumption that H_0 is true, we would be less than 0.1% likely to observe the height distributions that we ended up observing in the two groups. For this reason, H_0 does not seem to reflect well what we observed in our experiment and therefore is rejected.

> **!** Important
>
> Tests of hypotheses are everywhere in inferential statistics, so be sure to have familiarized with what discussed in this section.

3.3 (Linear) Regression

A linear regression is a model which estimates the (linear) relationship between one variable (dependent variable) and at least one other variable (independent variable). Linear regression is useful in that it quantifies how much the dependent variable changes as we let the independent variable(s) change. We now have a look at what should be clear about all this.

3.3.1 Simple regression

A simple regression is a regression model with a single independent variable. The equation that describes these models is usually of the form:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Where $\hat{Y}_i$ is the predicted value of Y when the independent variable X takes value X_i and it is multiplied by the estimated slope parameter $\hat{\beta}_1$ and added to the estimated intercept parameter $\hat{\beta}_0$. We now have a look at it in code.

```
library(MASS)                                                    ①
(animals <- MASS::Animals)                                       ②
```

① import the library `MASS`
② get the Animals dataset from inside the `MASS` library and assign it to the variable `animals`

```
                    body  brain
Mountain beaver     1.350    8.1
Cow               465.000  423.0
Grey wolf          36.330  119.5
Goat               27.660  115.0
Guinea pig          1.040    5.5
Dipliodocus     11700.000   50.0
Asian elephant   2547.000 4603.0
Donkey            187.100  419.0
Horse             521.000  655.0
Potar monkey       10.000  115.0
Cat                 3.300   25.6
Giraffe           529.000  680.0
Gorilla           207.000  406.0
Human              62.000 1320.0
African elephant 6654.000 5712.0
Triceratops      9400.000   70.0
Rhesus monkey       6.800  179.0
Kangaroo           35.000   56.0
Golden hamster      0.120    1.0
Mouse               0.023    0.4
Rabbit              2.500   12.1
Sheep              55.500  175.0
Jaguar            100.000  157.0
Chimpanzee         52.160  440.0
Rat                 0.280    1.9
Brachiosaurus   87000.000  154.5
Mole                0.122    3.0
Pig               192.000  180.0
```

```
regression <- lm(brain~body, animals)                            ①
summary(regression)                                              ②
```

① specify a linear model with the `lm()` function. Here we are specifying that `brain` is the dependent variable and `body` is the independent variable. After the comma, we tell R that these two variables should be retrieved from the `animals` dataset we saved before.

② we then call the `summary()` function on our regression to view some important statistics that are returned

```
Call:
lm(formula = brain ~ body, data = animals)

Residuals:
   Min     1Q Median     3Q    Max
-576.0 -554.1 -438.1 -156.3 5138.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.764e+02  2.659e+02   2.168   0.0395 *
body        -4.326e-04  1.589e-02  -0.027   0.9785
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1360 on 26 degrees of freedom
Multiple R-squared:  2.853e-05, Adjusted R-squared:  -0.03843
F-statistic: 0.0007417 on 1 and 26 DF,  p-value: 0.9785
```

From the output of the summary function we learn that the variable `body` is not a significant predictor of `brain` size! To see how we got to this conclusion, go to the column containing the p-value of the `body` variable (the last column). The p-value here is derived from the test of hypothesis that the t-statistic is different from 0. If we were to spell that out using the notation we developed in Section 3.2, we would get:

$$H_0 : t = 0$$

$$H_a : t \neq 0$$

If we compare the t-score of the `body` variable obtained from the `summary` of our regression we see that it is -0.027. Since after running the test the p-value is $.9785 > \alpha = .1$ it is concluded that body size is not a significant predictor of brain size in the samples of animals presented in this dataset. Therefore, we cannot interpret the estimated slope ($-4.326e^{-}4$). This is also the case because the significance for the overall model, reported as the p-value of the F-statistic at the bottom of the summary, is also greater than $\alpha = .1$.

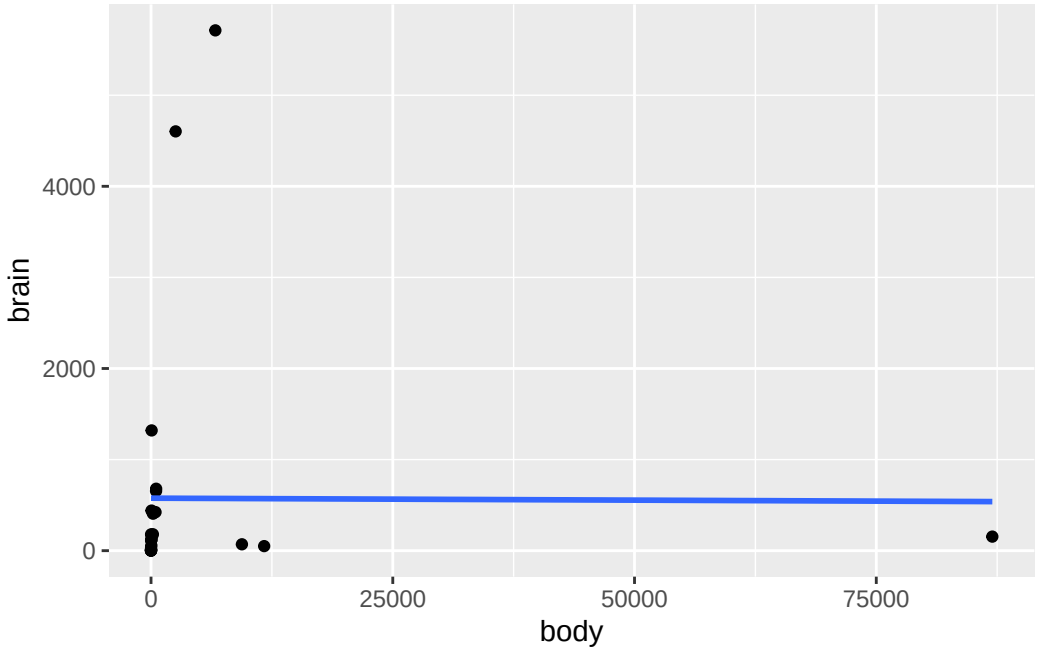
If you wanted to plot out the line obtained by the above linear model, you can use this code to visualize it

```
library(ggplot2)
ggplot(animals, aes(x=body, y=brain)) +                    ①
  geom_point() +                                           ②
  geom_smooth(method="lm", se=F)                           ③
```

① initializes a blank plot specifying what should be on the x and y axes

② adds a geom layer to plot points like in a scatterplot
③ adds the estimated regression line

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



Now that we plotted the data, however, we can clearly see that there is an observation that is far to the right and is skewing the distribution of `body`. Should we drop that observation to improve our regression model or should we leave it there? Feel free to continue the analysis with the tools you developed so far!

3.3.2 Additional resources

Since it is impossible to discuss everything about regressions in these few pages, here are some resources that you can find useful when trying to make sense of this topic:

1. YT short intro to regression Video
2. YT Playlist on regression