**Introduction**

One of the most important test within the branch of inferential statistics is the **Student’s t-test**. The Student’s t-test for two samples is used to **test whether two groups (two populations) are different** in terms of a quantitative variable, **based on the comparison of two samples** drawn from these two groups. In other words, a Student’s t-test for two samples allows to determine whether the two populations from which your two samples are drawn are different

The reasoning behind this statistical test is that if your two samples are markedly different from each other, it can be assumed that the two populations from which the samples are drawn are different. On the contrary, if the two samples are rather similar, we cannot reject the hypothesis that the two populations are similar, so there is no sufficient evidence in the data at hand to conclude that the two populations from which the samples are drawn are different. Note that this statistical tool belongs to the branch of inferential statistics because conclusions drawn from the study of the samples are generalized to the population, even though we do not have the data on the entire population.

To compare two samples, it is usual to compare a measure of central tendency computed for each sample. In the case of the Student’s t-test, the mean is used to compare the two samples. However, in some cases, the mean is not appropriate to compare two samples so the median is used to compare them via the Wilcoxon test. This article being already quite long and complete, the Wilcoxon test will be covered in a separate article, together with some illustrations on when to use one test or the other.

These two tests (Student’s t-test and Wilcoxon test) have the same final goal, that is, compare two samples in order to determine whether the two populations from which they were drawn are different or not. Note that the Student’s t-test is more powerful than the Wilcoxon test (it more often detects a significant difference if there is a true difference, so a smaller difference can be detected with the Student’s t-test) but the Student’s t-test is sensitive to outliers and data asymmetry. Furthermore, within each of these two tests, several versions exist, with each version using different formulas to arrive at the final result. It is thus necessary to understand the difference between the two tests and which version to use in order to carry out the appropriate analyses depending on the question and the data at hand.

In this article, I will first detail step by step how to perform all versions of the Student’s t-test for independent and paired samples by hand. The analyses will be done on a small set of observations for the sake of illustration and easiness. I will then show how to perform this test in R with the exact same data in order to verify the results found by hand. Reminders about the reasoning behind hypothesis testing, interpretations of the the *p*-value and the results, and assumptions of this test will also be presented.

Note that the aim of this article is to show how to compute the Student’s t-test by hand and in R, so we refrain from testing the assumptions and we assume all assumptions are met for this exercise. For completeness, we still mention the assumptions, how to test them and what other tests exist if an assumption is not met.

**Null and alternative hypothesis**

Before diving into the computations of the Student’s t-test by hand, let’s recap the null and alternative hypotheses of the this test:

* \(H\_0\): \(\mu\_1 = \mu\_2\)
* \(H\_1\): \(\mu\_1 \ne \mu\_2\)

where \(\mu\_1\) and \(\mu\_2\) are the means of the two populations from which the samples were drawn.

As mentioned in the introduction, although technically the Student’s t-test is based on the comparison of the means of the two samples, the final goal of this test is actually to test the following hypotheses:

* \(H\_0\): the two populations are similar
* \(H\_1\): the two populations are different

This is in the general case where we simply want to determine whether the two populations are **different** or not (in terms of the dependent variable). In this sense, we have no prior belief about a particular population being larger or smaller than the other. This type of test is referred as a **two-sided** or bilateral test.

If we have some prior beliefs about one population being larger or smaller than the other, the Student’s t-test also allows to test the following hypotheses:

* \(H\_0\): \(\mu\_1 = \mu\_2\)
* \(H\_1\): \(\mu\_1 > \mu\_2\)

or

* \(H\_0\): \(\mu\_1 = \mu\_2\)
* \(H\_1\): \(\mu\_1 < \mu\_2\)

In the first case, we want to test if the first population is significantly larger than the second, while in the latter case, we want to test if the first population is significantly smaller than the second. This type of test is referred as a **one-sided** or unilateral test.

Some authors argue that one-sided tests should not be used in practice for the simple reason that, if a researcher is so sure that one population is larger (smaller) than the other and would never be smaller (larger) than the other, why would she needs to test for significance at all? This a rather philosophical question and it is beyond the scope of this article. Interested readers are invited to see part of the discussion in Rowntree (2000).

**Hypothesis testing**

In statistics, many statistical tests is in the form of hypothesis tests. Hypothesis tests are used to determine whether a certain belief can be deemed as true (plausible) or not, based on the data at hand (i.e., the sample(s)). Most hypothesis tests boil down to the following 4 steps:[3](https://www.statsandr.com/blog/student-s-t-test-in-r-and-by-hand-how-to-compare-two-groups-under-different-scenarios/#fn3)

1. State the null and alternative hypothesis.
2. Compute the test statistic, denoted t-stat. Formulas to compute the test statistic differ among the different versions of the Student’s t-test but they have the same structure. See scenarios 1 to 5 below to see the different formulas.
3. Find the critical value given the theoretical statistical distribution of the test, the parameters of the distribution and the significance level \(\alpha\). For a Student’s t-test and its extended version, it is either the normal or the Student’s t distribution (*t* denoting the Student distribution and *z* denoting the normal distribution).
4. Conclude by comparing the t-stat (found in step 2.) with the critical value (found in step. 3). If the t-stat lies in the rejection region (determined thanks to the critical value and the direction of the test), we reject the null hypothesis, otherwise we do not reject the null hypothesis. These two alternatives (reject or do not reject the null hypothesis) are the only two possible solutions, we never “accept” an hypothesis. It is also a good practice to always interpret the decision in the terms of the initial question.

**Different versions of the Student’s t-test**

There are several versions of the Student’s t-test for two samples, depending on whether the samples are independent or paired and depending on the variances of the populations:

On the one hand, **independent** samples means that the two samples are collected on **different** experimental units or different individuals, for instance when we are working on women and men separately, or working on patients who have been randomly assigned to a control and a treatment group (and a patient belongs to only one group). On the other hand, we face **paired** samples when measurements are collected on the **same** experimental units, same individuals. This is often the case, for example in medical studies, when testing the efficiency of a treatment at two different times. The same patients are measured twice, before and after the treatment, and the dependency between the two samples must be taken into account in the computation of the test statistic by working on the **differences** of measurements for each subject. Paired samples are usually the result of measurements at two different times, but not exclusively. Suppose we want to test the difference in vision between the left and right eyes of 50 athletes. Although the measurements are not made at two different time (before-after), it is clear that both eyes are dependent within each subject. Therefore, the Student’s t-test for paired samples should be used to account for the dependency between the two samples instead of the standard Student’s t-test for independent samples.

Another criteria for choosing the appropriate version of the Student’s t-test is whether the variances of the populations (not the variances of the samples!) are known or unknown and equal or unequal. This criteria is rather straightforward, we either know the variances of the populations or we do not. The variances of the populations cannot be computed because if you can compute the variance of a population, it means you have the data for the whole population, then there is no need to do a hypothesis test anymore… So the variances of the populations are either given in the statement (use them in that case), or there is no information about these variances and in this case, it is assumed that the variances are unknown. In practice, the variances of the populations are most of the time unknown and the only thing to do in order to choose the appropriate version of the test is to check whether the variances are equal or not. However, we still illustrate how to do all versions of this test by hand and in R in the next sections following the 4 steps of hypothesis testing.

**How to compute Student’s t-test by hand?**

Note that the data are artificial and do not represent any real variable. Furthermore, remind that the assumptions may or may not be met. The point of the article is to detail how to compute the different versions of the test by hand and in R, so all assumptions are assumed to be met. Moreover, assume that the significance level \(\alpha = 5\)% for all tests.

**Scenario 1: Independent samples with 2 known variances**

For the first scenario, suppose the data below. Moreover, suppose that the two samples are independent, that the variances \(\sigma^2 = 1\) in both populations and that we would like to test whether the two populations are different.

| **value** | **sample** |
| --- | --- |
| 0.9 | 1 |
| -0.8 | 1 |
| 0.1 | 1 |
| -0.3 | 1 |
| 0.2 | 1 |
| 0.8 | 2 |
| -0.9 | 2 |
| -0.1 | 2 |
| 0.4 | 2 |
| 0.1 | 2 |

So we have:

* 5 observations in each sample: \(n\_1 = n\_2 = 5\)
* mean of sample 1: \(\bar{x}\_1 = 0.02\)
* mean of sample 2: \(\bar{x}\_2 = 0.06\)
* variances of both populations: \(\sigma^2\_1 = \sigma^2\_2 = 1\)

Following the 4 steps of hypothesis testing we have:

1. \(H\_0: \mu\_1 = \mu\_2\) and \(H\_1: \mu\_1 – \mu\_2 \ne 0\). (\(\ne\) because we want to test whether the two means are different, we do not impose a direction in the test.)
2. Test statistic: \[z\_{obs} = \frac{(\bar{x}\_1 – \bar{x}\_2) – (\mu\_1 – \mu\_2)}{\sqrt{\frac{\sigma^2\_1}{n\_1} + \frac{\sigma^2\_2}{n\_2}}} = \frac{0.02-0.06-0}{0.632} = -0.063\]
3. Critical value: \(\pm z\_{\alpha / 2} = \pm z\_{0.025} = \pm 1.96\)
4. Conclusion: The rejection regions are thus from \(-\infty\) to -1.96 and from 1.96 to \(+\infty\). The test statistic is outside the rejection regions so we do not reject the null hypothesis \(H\_0\). In terms of the initial question: At the 5% significance level, we do not reject the hypothesis that the two populations are the same, or there is no sufficient evidence in the data to conclude that the two populations considered are different.

**Scenario 2: Independent samples with 2 equal but unknown variances**

For the second scenario, suppose the data below. Moreover, suppose that the two samples are independent, that the variances in both populations are unknown but equal (\(\sigma^2\_1 = \sigma^2\_1\)) and that we would like to test whether population 1 is larger than population 2.

| **value** | **sample** |
| --- | --- |
| 1.78 | 1 |
| 1.5 | 1 |
| 0.9 | 1 |
| 0.6 | 1 |
| 0.8 | 1 |
| 1.9 | 1 |
| 0.8 | 2 |
| -0.7 | 2 |
| -0.1 | 2 |
| 0.4 | 2 |
| 0.1 | 2 |

So we have:

* 6 observations in sample 1: \(n\_1 = 6\)
* 5 observations in sample 2: \(n\_2 = 5\)
* mean of sample 1: \(\bar{x}\_1 = 1.247\)
* mean of sample 2: \(\bar{x}\_2 = 0.1\)
* variance of sample 1: \(s^2\_1 = 0.303\)
* variance of sample 2: \(s^2\_1 = 0.315\)

Following the 4 steps of hypothesis testing we have:

1. \(H\_0: \mu\_1 = \mu\_2\) and \(H\_1: \mu\_1 – \mu\_2 > 0\). (> because we want to test if the mean of the first population is larger than the mean of the second population.)
2. Test statistic: \[t\_{obs} = \frac{(\bar{x}\_1 – \bar{x}\_2) – (\mu\_1 – \mu\_2)}{s\_p\sqrt{\frac{1}{n\_1} + \frac{1}{n\_2}}}\] where \[s\_p = \sqrt{\frac{(n\_1-1)s^2\_1+ (n\_2 – 1)s^2\_2}{n\_1 + n\_2 – 2}} = 0.555\] so \[t\_{obs} = \frac{1.247-0.1-0}{0.555 \* 0.606} = 3.411\]  
   (Note that as it is assumed the variances of the two populations are equal, a pooled (common) variance, denoted \(s\_p\), is computed.)
3. Critical value: \(t\_{\alpha, n\_1 + n\_2 – 2} = t\_{0.05, 9} = 1.833\)
4. Conclusion: The rejection region is thus from 1.833 to \(+\infty\) (there is only one rejection region because it is a one-sided test). The test statistic lies within the rejection region so we reject the null hypothesis \(H\_0\). In terms of the initial question: At the 5% significance level, we conclude that the population 1 is larger than the population 2.

**Scenario 3: Independent samples with 2 unequal and unknown variances**

For the third scenario, suppose the data below. Moreover, suppose that the two samples are independent, that the variances in both populations are unknown and unequal (\(\sigma^2\_1 \ne \sigma^2\_1\)) and that we would like to test whether population 1 is smaller than population 2.

| **value** | **sample** |
| --- | --- |
| 0.8 | 1 |
| 0.7 | 1 |
| 0.1 | 1 |
| 0.4 | 1 |
| 0.1 | 1 |
| 1.78 | 2 |
| 1.5 | 2 |
| 0.9 | 2 |
| 0.6 | 2 |
| 0.8 | 2 |
| 1.9 | 2 |

So we have:

* 5 observations in sample 1: \(n\_1 = 5\)
* 6 observations in sample 2: \(n\_2 = 6\)
* mean of sample 1: \(\bar{x}\_1 = 0.42\)
* mean of sample 2: \(\bar{x}\_2 = 1.247\)
* variance of sample 1: \(s^2\_1 = 0.107\)
* variance of sample 2: \(s^2\_1 = 0.303\)

Following the 4 steps of hypothesis testing we have:

1. \(H\_0: \mu\_1 = \mu\_2\) and \(H\_1: \mu\_1 – \mu\_2 < 0\). (< because we want to test if the mean of the first population is smaller than the mean of the second population.)
2. Test statistic: \[t\_{obs} = \frac{(\bar{x}\_1 – \bar{x}\_2) – (\mu\_1 – \mu\_2)}{\sqrt{\frac{s^2\_1}{n\_1} + \frac{s^2\_2}{n\_2}}} = \frac{0.42-1.247-0}{0.268} = -3.084\]
3. Critical value: \(-t\_{\alpha, \upsilon}\) where \[\upsilon = \frac{\bigg(\frac{s^2\_1}{n\_1} + \frac{s^2\_2}{n\_2} \bigg)^2}{\frac{\bigg(\frac{s^2\_1}{n\_1}\bigg)^2}{n\_1 – 1} + \frac{\bigg(\frac{s^2\_2}{n\_2}\bigg)^2}{n\_2 – 1}} = 8.28\] so \[-t\_{0.05, 8.28} = -1.851\] The degrees of freedom 8.28 does not exist in the standard Student distribution table, so simply take 8, or compute it in R with qt(p = 0.05, df = 8.28).
4. Conclusion: The rejection region is thus from \(-\infty\) to -1.851. The test statistic lies within the rejection region so we reject the null hypothesis \(H\_0\). In terms of the initial question: At the 5% significance level, we conclude that the population 1 is smaller than the population 2.

**Scenario 4: Paired samples where the variance of the differences is known**

Student’s t-test with paired samples are a bit different than with independent samples, they are actually more similar to one sample Student’s t-test.  
  
Here is how it works. We actually compute the difference between the two samples for each pair of observations, and then we work on these differences as if we were doing a one sample Student’s t-test by computing the test statistic on these differences.

In case it is not clear, here is the fourth scenario as an illustration. Suppose the data below. Moreover, suppose that the two samples are dependent (matched), that the variance of the differences in the population is known and equal to 1 (\(\sigma^2\_D = 1\)) and that we would like to test whether the difference in the population is different than 0.

| **before** | **after** |
| --- | --- |
| 0.9 | 0.8 |
| -0.8 | -0.9 |
| 0.1 | -0.1 |
| -0.3 | 0.4 |
| 0.2 | 0.1 |

The first thing to do is to compute the differences for all pairs of observations:

| **before** | **after** | **difference** |
| --- | --- | --- |
| 0.9 | 0.8 | -0.1 |
| -0.8 | -0.9 | -0.1 |
| 0.1 | -0.1 | -0.2 |
| -0.3 | 0.4 | 0.7 |
| 0.2 | 0.1 | -0.1 |

So we have:

* number of pairs: \(n = 5\)
* mean of the difference: \(\bar{D} = 0.04\)
* variance of the difference in the population: \(\sigma^2\_D = 1\)
* standard deviation of the difference in the population: \(\sigma\_D = 1\)

Following the 4 steps of hypothesis testing we have:

1. \(H\_0: \mu\_D = 0\) and \(H\_1: \mu\_D \ne 0\)
2. Test statistic: \[z\_{obs} = \frac{\bar{D} – \mu\_0}{\frac{\sigma\_D}{\sqrt{n}}} = \frac{0.04-0}{0.447} = 0.089\]  
   (This formula is exactly the same than for one sample Student’s t-test with a known variance, except that we work on the mean of the differences.)
3. Critical value: \(\pm z\_{\alpha/2} = \pm z\_{0.025} = \pm 1.96\)
4. Conclusion: The rejection regions are thus from \(-\infty\) to -1.96 and from 1.96 to \(+\infty\). The test statistic is outside the rejection regions so we do not reject the null hypothesis \(H\_0\). In terms of the initial question: At the 5% significance level, we do not reject the hypothesis that the difference in the two populations is equal to 0.

**Scenario 5: Paired samples where the variance of the differences is unknown**

For the fifth and final scenario, suppose the data below. Moreover, suppose that the two samples are dependent (matched), that the variance of the differences in the population is unknown and that we would like to test whether a treatment is effective in increasing running capabilities (the higher the value, the better in terms of running capabilities).

| **before** | **after** |
| --- | --- |
| 9 | 16 |
| 8 | 11 |
| 1 | 15 |
| 3 | 12 |
| 2 | 9 |

The first thing to do is to compute the differences for all pairs of observations:

| **before** | **after** | **difference** |
| --- | --- | --- |
| 9 | 16 | 7 |
| 8 | 11 | 3 |
| 1 | 15 | 14 |
| 3 | 12 | 9 |
| 2 | 9 | 7 |

So we have:

* number of pairs: \(n = 5\)
* mean of the difference: \(\bar{D} = 8\)
* variance of the difference in the sample: \(s^2\_D = 16\)
* standard deviation of the difference in the sample: \(s\_D = 4\)

Following the 4 steps of hypothesis testing we have:

1. \(H\_0: \mu\_D = 0\) and \(H\_1: \mu\_D > 0\) (> because we would like to test whether the treatment is effective, so whether the treatment has a positive impact on the running capabilities.)
2. Test statistic: \[t\_{obs} = \frac{\bar{D} – \mu\_0}{\frac{s\_D}{\sqrt{n}}} = \frac{8-0}{1.789} = 4.472\]  
   (This formula is exactly the same than for one sample Student’s t-test with an unknown variance, except that we work on the mean of the differences.)
3. Critical value: \(t\_{\alpha, n-1} = t\_{0.05, 4} = 2.132\) (*n* is the number of pairs, not the number of observations!)
4. Conclusion: The rejection regions are thus from 2.132 to \(+\infty\). The test statistic lies within the rejection region so we reject the null hypothesis \(H\_0\). In terms of the initial question: At the 5% significance level, we conclude that the treatment has a positive impact on the running capabilities.

This concludes how to perform the different versions of the Student’s t-test for two samples by hand. In the next sections, we detail how to perform the exact same tests in R.

**How to compute Student’s t-test in R?**

A good practice before doing t-tests in R is to visualize the data by group thanks to a boxplot or a density plot or both. A boxplot with the two boxes overlapping each other gives a first indication that the two samples are similar, and thus, that the null hypothesis of equal means may not be rejected. On the contrary, if the two boxes are not overlapping, it indicates that the two samples are not similar, and thus, that the populations may be different. However, even if boxplots or density plots are great in showing a comparison between the two groups, only a sound statistical test will confirm our first impression.

After a visualization of the data by group, we replicate in R the results found by hand. We will see that for some versions of the t-test, there is no default function built in R (at least to my knowledge, do not hesitate to let me know if I’m mistaken). In these cases, a function is written to replicate the results by hand.

Note that we use the same data, the same assumptions and the same question for all 5 scenarios to facilitate the comparison between the tests performed by hand and in R.

**Scenario 1: Independent samples with 2 known variances**

For the first scenario, suppose the data below. Moreover, suppose that the two samples are independent, that the variances \(\sigma^2 = 1\) in both populations and that we would like to test whether the two populations are different.

dat1 <- data.frame(

sample1 = c(0.9, -0.8, 0.1, -0.3, 0.2),

sample2 = c(0.8, -0.9, -0.1, 0.4, 0.1)

)

dat1

## sample1 sample2

## 1 0.9 0.8

## 2 -0.8 -0.9

## 3 0.1 -0.1

## 4 -0.3 0.4

## 5 0.2 0.1

dat\_ggplot <- data.frame(

value = c(0.9, -0.8, 0.1, -0.3, 0.2, 0.8, -0.9, -0.1, 0.4, 0.1),

sample = c(rep("1", 5), rep("2", 5))

)

library(ggplot2)

ggplot(dat\_ggplot) +

aes(x = sample, y = value) +

geom\_boxplot() +

theme\_minimal()

Note that you can use the esquisse Rstudio Addin if you want to draw a boxplot with the package {ggplot2} without writing the code yourself. If you prefer the default graphics, use the boxplot() function:

boxplot(value ~ sample,

data = dat\_ggplot

)

The two boxes seem to overlap which illustrate that the two samples are quite similar, so we tend to believe that we will not be able to reject the null hypothesis that the two populations are similar. However, only a formal statistical test will confirm this belief.

Since there is no function in R to perform a t-test with known variances, here is one with arguments accepting the two samples (x and y), the two variances of the populations (V1 and V2), the difference in means under the null hypothesis (m0, default is 0), the significance level (alpha, default is 0.05) and the alternative (alternative, one of "two.sided" (default), "less" or "greater"):

t.test\_knownvar <- function(x, y, V1, V2, m0 = 0, alpha = 0.05, alternative = "two.sided") {

M1 <- mean(x)

M2 <- mean(y)

n1 <- length(x)

n2 <- length(y)

sigma1 <- sqrt(V1)

sigma2 <- sqrt(V2)

S <- sqrt((V1 / n1) + (V2 / n2))

statistic <- (M1 - M2 - m0) / S

p <- if (alternative == "two.sided") {

2 \* pnorm(abs(statistic), lower.tail = FALSE)

} else if (alternative == "less") {

pnorm(statistic, lower.tail = TRUE)

} else {

pnorm(statistic, lower.tail = FALSE)

}

LCL <- (M1 - M2 - S \* qnorm(1 - alpha / 2))

UCL <- (M1 - M2 + S \* qnorm(1 - alpha / 2))

value <- list(mean1 = M1, mean2 = M2, m0 = m0, sigma1 = sigma1, sigma2 = sigma2, S = S, statistic = statistic, p.value = p, LCL = LCL, UCL = UCL, alternative = alternative)

# print(sprintf("P-value = %g",p))

# print(sprintf("Lower %.2f%% Confidence Limit = %g",

# alpha, LCL))

# print(sprintf("Upper %.2f%% Confidence Limit = %g",

# alpha, UCL))

return(value)

}

test <- t.test\_knownvar(dat1$sample1, dat1$sample2,

V1 = 1, V2 = 1

)

test

## $mean1

## [1] 0.02

##

## $mean2

## [1] 0.06

##

## $m0

## [1] 0

##

## $sigma1

## [1] 1

##

## $sigma2

## [1] 1

##

## $S

## [1] 0.6324555

##

## $statistic

## [1] -0.06324555

##

## $p.value

## [1] 0.949571

##

## $LCL

## [1] -1.27959

##

## $UCL

## [1] 1.19959

##

## $alternative

## [1] "two.sided"

The output above recaps all the information needed to perform the test: the test statistic, the *p*-value, the alternative used, the two sample means and the two variances of the populations (compare these results found in R with the results found by hand).

The *p*-value can be extracted as usual:

test$p.value

## [1] 0.949571

The *p*-value is 0.95 so at the 5% significance level we do not reject the null hypothesis of equal means. There is no sufficient evidence in the data to reject the hypothesis that the two means in the populations are similar. This result confirms what we found by hand.

**A note on *p*-value and significance level \(\alpha\)**

For those unfamiliar with the concept of *p*-value, the *p*-value is a probability and as any probability it goes from 0 to 1. The ***p*-value is the probability of having observations as extreme as what we measured (via the samples) if the null hypothesis was true**. In other words, it is the probability of having a test statistic as extreme as what we computed, given that the null hypothesis is true. If the observations are not so extreme, i.e., not unlikely to occur if the null hypothesis was true, we do not reject this null hypothesis because it is deemed plausible to be true. And if the observations are considered too extreme, i.e., too unlikely to happen under the null hypothesis, we reject the null hypothesis because it is deemed too implausible to be true. Note that it does not mean that we are 100% sure that it is too unlikely, it happens sometimes that the null hypothesis is rejected although it is true (see the significance level \(\alpha\) later on).

In our example above, the observations are not really extreme and the difference between the two means is not extreme, so the test statistic is not extreme (since the test statistic is partially based on the difference of the means of the two samples). Having a test statistic which is not extreme is not unlikely and that is the reason why the *p*-value is quite high. The *p*-value of 0.95 actually tells us that the probability of having two samples with a difference in means of -0.04 (= 0.02 – 0.06), given that the difference in means in the populations is 0 (the null hypothesis), equals 95%. A probability of 95% is definitely considered as plausible, so we do not reject the null hypothesis of equal means in the populations.

One may then wonder, “What is too extreme for a test statistic?” Most of the time, we consider that a test statistic is too extreme to happen just by chance when the probability of having such an extreme test statistic given that the null hypothesis is true is below 5%. The threshold of 5% (\(\alpha = 0.05\)) that you very often see in statistic courses or textbooks is the threshold used in many fields. With a *p*-value under that threshold of 5%, we consider that the observations (and thus the test statistic) is **too unlikely** to happen just by chance if the null hypothesis was true, so the null hypothesis is rejected. With a *p*-value above that threshold of 5%, we consider that it is not really implausible to face the observations we have if the null hypothesis was true, and we therefore do not reject the null hypothesis.

Note that I wrote “we do not reject the null hypothesis”, and not “we accept the null hypothesis”. This is because it may be the case that the null hypothesis is in fact false, but we failed to prove it with the samples. Suppose the analogy of a suspect accused of murder and we do not know the truth. On the one hand, if we have collected enough evidence that the suspect committed the murder, he is considered guilty: we reject the null hypothesis that he is innocent. On the other hand, if we have *not* collected enough evidence against the suspect, he is presumed to be innocent although he may in fact have committed the crime: we failed to reject the null hypothesis of him being innocent. We are never sure that he did not committed the crime even if he is released, we just did not find sufficient evidence against the null hypothesis of the suspect being innocent. This is the reason why we do not reject the null hypothesis instead of accepting it, and why you will often read things like “there is no sufficient evidence in the data to reject the null hypothesis” or “based on the samples we fail to reject the null hypothesis”.

The **significance level \(\alpha\)**, derived from the threshold of 5% mentioned earlier, **is the probability of rejecting the null hypothesis when it is in fact true**. In this sense, it is an error (of 5%) that we accept to deal with, in order to be able to draw conclusions. If we would accept no error (an error of 0%), we would not be able to draw any conclusion about the population(s) since we only have access to a limited portion of the population(s) via the sample(s). As a consequence, we will never be 100% sure when interpreting the result of a hypothesis test unless we have access to the data for the entire population, but then there is no reason to do a hypothesis test anymore since we can simply compare the two populations. We usually allow this error (called Type I error) to be 5%, but in order to be a bit more certain when concluding that we reject the null hypothesis, the alpha level can also be set to 1% (or even to 0.1% in some rare cases).

**To sum up** what you need to remember about *p*-value and significance level \(\alpha\):

* If the *p*-value is smaller than the predetermined significance level \(\alpha\) (usually 5%) so if *p*-value < 0.05, we reject the null hypothesis
* If the *p*-value is greater than or equal to the predetermined significance level \(\alpha\) (usually 5%) so if *p*-value \(\ge\) 0.05, we do **not reject** the null hypothesis

This applies to all statistical tests without exception. Of course, the null and alternative hypotheses change depending on the test.

A rule of thumb is that, for most hypothesis tests, the alternative hypothesis is what you want to test and the null hypothesis is the status quo. Take this with extreme caution (!) because, even if it works for all versions of the Student’s t-test it does not apply to ALL statistical tests. For example, when testing for normality, you usually want to test whether your distribution follows a normal distribution. Following this piece of advice, you would write the alternative hypothesis \(H\_1:\) the distribution follows a normal distribution. Nonetheless, for normality tests such as the Shapiro-Wilk or Kolmogorov-Smirnov test, it is the opposite; the alternative hypothesis is \(H\_1:\) the distribution does not follow a normal distribution. So for every test, make sure to use the correct hypotheses, otherwise the conclusion and interpretation of your test will be wrong.

**Scenario 2: Independent samples with 2 equal but unknown variances**

For the second scenario, suppose the data below. Moreover, suppose that the two samples are independent, that the variances in both populations are unknown but equal (\(\sigma^2\_1 = \sigma^2\_1\)) and that we would like to test whether population 1 is larger than population 2.

dat2 <- data.frame(

sample1 = c(1.78, 1.5, 0.9, 0.6, 0.8, 1.9),

sample2 = c(0.8, -0.7, -0.1, 0.4, 0.1, NA)

)

dat2

## sample1 sample2

## 1 1.78 0.8

## 2 1.50 -0.7

## 3 0.90 -0.1

## 4 0.60 0.4

## 5 0.80 0.1

## 6 1.90 NA

dat\_ggplot <- data.frame(

value = c(1.78, 1.5, 0.9, 0.6, 0.8, 1.9, 0.8, -0.7, -0.1, 0.4, 0.1),

sample = c(rep("1", 6), rep("2", 5))

)

ggplot(dat\_ggplot) +

aes(x = sample, y = value) +

geom\_boxplot() +

theme\_minimal()

Unlike the previous scenario, the two boxes do not overlap which illustrates that the two samples are different from each other. From this boxplot, we can expect the test to reject the null hypothesis of equal means in the populations. Nonetheless, only a formal statistical test will confirm this expectation.

There is a function in R, and it is simply the t.test() function. This version of the test is actually the “standard” Student’s t-test for two samples. Note that it is assumed that the variances of the two populations are equal so we need to specify it in the function with the argument var.equal = TRUE (the default is FALSE) and the alternative hypothesis is \(H\_1: \mu\_1 – \mu\_2 > 0\) so we need to add the argument alternative = "greater" as well:

test <- t.test(dat2$sample1, dat2$sample2,

var.equal = TRUE, alternative = "greater"

)

test

##

## Two Sample t-test

##

## data: dat2$sample1 and dat2$sample2

## t = 3.4113, df = 9, p-value = 0.003867

## alternative hypothesis: true difference in means is greater than 0

## 95 percent confidence interval:

## 0.5304908 Inf

## sample estimates:

## mean of x mean of y

## 1.246667 0.100000

The output above recaps all the information needed to perform the test: the name of the test, the test statistic, the degrees of freedom, the *p*-value, the alternative used and the two sample means (compare these results found in R with the results found by hand).

The *p*-value can be extracted as usual:

test$p.value

## [1] 0.003866756

The *p*-value is 0.004 so at the 5% significance level we reject the null hypothesis of equal means. This result confirms what we found by hand.

Unlike the first scenario, the *p*-value in this scenario is below 5% so we reject the null hypothesis. At the 5% significance level, we can conclude that the population 1 is larger than the population 2.

If your data is formatted in the long format (which is even better), simply use the ~. For instance, imagine the exact same data presented like this:

dat2bis <- data.frame(

value = c(1.78, 1.5, 0.9, 0.6, 0.8, 1.9, 0.8, -0.7, -0.1, 0.4, 0.1),

sample = c(rep("1", 6), rep("2", 5))

)

dat2bis

## value sample

## 1 1.78 1

## 2 1.50 1

## 3 0.90 1

## 4 0.60 1

## 5 0.80 1

## 6 1.90 1

## 7 0.80 2

## 8 -0.70 2

## 9 -0.10 2

## 10 0.40 2

## 11 0.10 2

Here is how to perform the Student’s t-test in R with long data:

test <- t.test(value ~ sample,

data = dat2bis,

var.equal = TRUE,

alternative = "greater"

)

test

##

## Two Sample t-test

##

## data: value by sample

## t = 3.4113, df = 9, p-value = 0.003867

## alternative hypothesis: true difference in means is greater than 0

## 95 percent confidence interval:

## 0.5304908 Inf

## sample estimates:

## mean in group 1 mean in group 2

## 1.246667 0.100000

test$p.value

## [1] 0.003866756

The results are exactly the same.

**Scenario 3: Independent samples with 2 unequal and unknown variances**

For the third scenario, suppose the data below. Moreover, suppose that the two samples are independent, that the variances in both populations are unknown and unequal (\(\sigma^2\_1 \ne \sigma^2\_1\)) and that we would like to test whether population 1 is smaller than population 2.

dat3 <- data.frame(

value = c(0.8, 0.7, 0.1, 0.4, 0.1, 1.78, 1.5, 0.9, 0.6, 0.8, 1.9),

sample = c(rep("1", 5), rep("2", 6))

)

dat3

## value sample

## 1 0.80 1

## 2 0.70 1

## 3 0.10 1

## 4 0.40 1

## 5 0.10 1

## 6 1.78 2

## 7 1.50 2

## 8 0.90 2

## 9 0.60 2

## 10 0.80 2

## 11 1.90 2

ggplot(dat3) +

aes(x = sample, y = value) +

geom\_boxplot() +

theme\_minimal()

There is a function in R for this version of the test as well, and it is simply the t.test() function with the var.equal = FALSE argument. FALSE is the default option for the var.equal argument so you actually do not need to specify it. This version of the test is actually the Welch test, used when the variances of the populations are unknown and unequal. To test if two variances are equal, you can use the Levene’s test (leveneTest(dat3$value, dat3$sample) from the {car} package). Note that the alternative hypothesis is \(H\_1: \mu\_1 – \mu\_2 < 0\) so we need to add the argument alternative = "less" as well:

test <- t.test(value ~ sample,

data = dat3,

var.equal = FALSE,

alternative = "less"

)

test

##

## Welch Two Sample t-test

##

## data: value by sample

## t = -3.0841, df = 8.2796, p-value = 0.007206

## alternative hypothesis: true difference in means is less than 0

## 95 percent confidence interval:

## -Inf -0.3304098

## sample estimates:

## mean in group 1 mean in group 2

## 0.420000 1.246667

The output above recaps all the information needed to perform the test (compare these results found in R with the results found by hand).

The *p*-value can be extracted as usual:

test$p.value

## [1] 0.00720603

The *p*-value is 0.007 so at the 5% significance level we reject the null hypothesis of equal means, meaning that we can conclude that the population 1 is smaller than the population 2. This result confirms what we found by hand.

**Scenario 4: Paired samples where the variance of the differences is known**

For the fourth scenario, suppose the data below. Moreover, suppose that the two samples are dependent (matched), that the variance of the differences in the population is known and equal to 1 (\(\sigma^2\_D = 1\)) and that we would like to test whether the difference in the population is different than 0.

dat4 <- data.frame(

before = c(0.9, -0.8, 0.1, -0.3, 0.2),

after = c(0.8, -0.9, -0.1, 0.4, 0.1)

)

dat4

## before after

## 1 0.9 0.8

## 2 -0.8 -0.9

## 3 0.1 -0.1

## 4 -0.3 0.4

## 5 0.2 0.1

dat4$difference <- dat4$after - dat4$before

ggplot(dat4) +

aes(y = difference) +

geom\_boxplot() +

theme\_minimal()

Since there is no function in R to perform a t-test with paired samples where the variance of the differences is known, here is one with arguments accepting the differences between the two samples (x), the variance of the differences in the population (V), the mean of the differences under the null hypothesis (m0, default is 0), the significance level (alpha, default is 0.05) and the alternative (alternative, one of "two.sided" (default), "less" or "greater"):

t.test\_pairedknownvar <- function(x, V, m0 = 0, alpha = 0.05, alternative = "two.sided") {

M <- mean(x)

n <- length(x)

sigma <- sqrt(V)

S <- sqrt(V / n)

statistic <- (M - m0) / S

p <- if (alternative == "two.sided") {

2 \* pnorm(abs(statistic), lower.tail = FALSE)

} else if (alternative == "less") {

pnorm(statistic, lower.tail = TRUE)

} else {

pnorm(statistic, lower.tail = FALSE)

}

LCL <- (M - S \* qnorm(1 - alpha / 2))

UCL <- (M + S \* qnorm(1 - alpha / 2))

value <- list(mean = M, m0 = m0, sigma = sigma, statistic = statistic, p.value = p, LCL = LCL, UCL = UCL, alternative = alternative)

# print(sprintf("P-value = %g",p))

# print(sprintf("Lower %.2f%% Confidence Limit = %g",

# alpha, LCL))

# print(sprintf("Upper %.2f%% Confidence Limit = %g",

# alpha, UCL))

return(value)

}

test <- t.test\_pairedknownvar(dat4$after - dat4$before,

V = 1

)

test

## $mean

## [1] 0.04

##

## $m0

## [1] 0

##

## $sigma

## [1] 1

##

## $statistic

## [1] 0.08944272

##

## $p.value

## [1] 0.9287301

##

## $LCL

## [1] -0.8365225

##

## $UCL

## [1] 0.9165225

##

## $alternative

## [1] "two.sided"

The output above recaps all the information needed to perform the test (compare these results found in R with the results found by hand).

The *p*-value can be extracted as usual:

test$p.value

## [1] 0.9287301

The *p*-value is 0.929 so at the 5% significance level we do not reject the null hypothesis of the mean of the differences being equal to 0. There is no sufficient evidence in the data to reject the hypothesis that the difference in the two populations is equal to 0. This result confirms what we found by hand.

**Scenario 5: Paired samples where the variance of the differences is unknown**

For the fifth and final scenario, suppose the data below. Moreover, suppose that the two samples are dependent (matched), that the variance of the differences in the population is unknown and that we would like to test whether a treatment is effective in increasing running capabilities (the higher the value, the better in terms of running capabilities).

dat5 <- data.frame(

before = c(9, 8, 1, 3, 2),

after = c(16, 11, 15, 12, 9)

)

dat5

## before after

## 1 9 16

## 2 8 11

## 3 1 15

## 4 3 12

## 5 2 9

dat5$difference <- dat5$after - dat5$before

ggplot(dat5) +

aes(y = difference) +

geom\_boxplot() +

theme\_minimal()

There is a function in R for this version of the test, and it is simply the t.test() function with the paired = TRUE argument. This version of the test is actually the standard version of the Student’s t-test with paired samples. Note that the alternative hypothesis is \(H\_1: \mu\_D > 0\) so we need to add the argument alternative = "greater" as well:

test <- t.test(dat5$after, dat5$before,

alternative = "greater",

paired = TRUE

)

test

##

## Paired t-test

##

## data: dat5$after and dat5$before

## t = 4.4721, df = 4, p-value = 0.005528

## alternative hypothesis: true difference in means is greater than 0

## 95 percent confidence interval:

## 4.186437 Inf

## sample estimates:

## mean of the differences

## 8

Note that we wrote after and then before in this order. If you write before and then after, make sure to change the alternative to alternative = "less".

If your data is in the long format, use the ~:

dat5 <- data.frame(

value = c(9, 8, 1, 3, 2, 16, 11, 15, 12, 9),

time = c(rep("before", 5), rep("after", 5))

)

dat5

## value time

## 1 9 before

## 2 8 before

## 3 1 before

## 4 3 before

## 5 2 before

## 6 16 after

## 7 11 after

## 8 15 after

## 9 12 after

## 10 9 after

test <- t.test(value ~ time,

data = dat5,

alternative = "greater",

paired = TRUE

)

test

##

## Paired t-test

##

## data: value by time

## t = 4.4721, df = 4, p-value = 0.005528

## alternative hypothesis: true difference in means is greater than 0

## 95 percent confidence interval:

## 4.186437 Inf

## sample estimates:

## mean of the differences

## 8

The output above recaps all the information needed to perform the test (compare these results found in R with the results found by hand).

The *p*-value can be extracted as usual:

test$p.value

## [1] 0.005528247

The *p*-value is 0.006 so at the 5% significance level we reject the null hypothesis of the mean of the differences being equal to 0, meaning that we can conclude that the treatment is effective in increasing the running capabilities. This result confirms what we found by hand.

**Assumptions**

As for many statistical tests, there are some assumptions that need to be met in order to be able to interpret the results. When one or several assumptions are not met, although it is technically possible to perform these tests, it would be incorrect to interpret the results. Below are the assumptions of the Student’s t-test for two samples, how to test them and which other tests exist if an assumption is not met:

* The data, collected from a representative and randomly selected portion of the total population, should be independent. If observations between the two samples are dependent (for example if two measurements have been collected on the **same individual** as it is often the case in medical studies when measuring a value after and before a treatment), the paired version of the Student’s t-test, called the Student’s t-test for paired samples, should be preferred in order to take into account the dependency between the two groups to be compared.
* When the two samples are independent, observations in **both samples** should follow a normal distribution. When using the Student’s t-test for paired samples, it is the difference between the observations of the two samples that should follow a normal distribution. The normality assumption can be tested visually thanks to a histogram and a QQ-plot, and/or formally via a normality test such as the Shapiro-Wilk or Kolmogorov-Smirnov test (see more information about the normality assumption and how to test it). If, even after a transformation (logarithmic transformation, etc.), your data still do not follow a normal distribution, the Wilcoxon test (wilcox.test(variable1 ~ variable2, data = dat in R) can be applied. This test, robust to non normal distributions, compares the medians instead of the means in order to compare the two populations.
* When the two samples are independent, the variances of the two groups should be equal in the populations (an assumption called homogeneity of the variances, or even sometimes referred as homoscedasticity, as opposed to heteroscedasticity). This assumption can be tested thanks to the Levene’s test (leveneTest(variable ~ group) from the {car} package) or via a F test (var.test(variable ~ group)). If the hypothesis of equal variances is rejected, another version of the Student’s t-test can be used: the Welch test (t.test(variable ~ group, var.equal = FALSE)).