**T-test and Hypothesis Testing (Explained Simply)**

Understand the concept and find how to avoid typical mistakes

Student’s t-tests are commonly used in inferential statistics for testing a hypothesis on the basis of a difference between sample means. However, people often misinterpret the results of t-tests, which leads to false research findings and a lack of reproducibility of studies. This problem exists not only among students. Even instructors and “serious” researchers fall into the same trap. To prove my words, I can link [this article](https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124), but there are others.

Another problem is that I’ve often seen and heard complaints from some students that their teachers don’t explain the concept of t-tests sufficiently. Instead, they focus on calculations and interpretation of the results. Nowadays, scientists use computers to calculate t-statistic automatically, so there is no reason to drill the usage of formulas and t-distribution tables, except for the purpose of understanding *how it works*. As for interpretation, there is nothing wrong with it, although without comprehension of the concept it may look like blindly following the rules. Actually, it is. Do you remember?

*“Absolute t-value is greater than t-critical, so the null hypothesis is rejected and the alternate hypothesis is accepted”.*

If you are familiar with this statement and still have problems with understanding it, most likely, you’ve been unfortunate to get the same training. These problems with intuition can lead to problems with decision-making while testing hypotheses. So, besides knowing what values to paste into the formula and how to use t-tests, it is necessary to know when to use it, why to use it, and the meaning of all that stuff.

This article is intended to explain two concepts: t-test and hypothesis testing. At first, I wanted to explain only t-tests. Later, I decided to include hypothesis testing because these ideas are so closely related that it would be difficult to tell about one thing while losing sight of another. Eventually, you will see that t-test is not only an abstract idea but has good common sense.

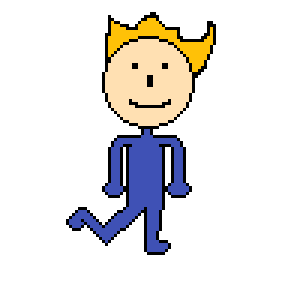
Be prepared, this article is pretty long. Take a look at the article outline below to not get lost.

**Article outline:**

1. Hypothesis testing
2. T-test definition and formula explanation
3. Choosing the level of significance
4. T-distribution and p-value
5. Conclusion

**Hypothesis testing**

Meet David! He is a high school student and he has started to study statistics recently.



Hi, I’m David created with Paint! | by Author

David wants to figure out whether his schoolmates from class A got better quarter grades in mathematics than those from class B. There is a 5-point grading system at school, where 5 is the best score. Students have no access to other students' grades because teachers keep their data confidential and there are approximately 30 students in both classes.

David cannot ask all the students about their grades because it is weird and not all the students are happy to tell about their grades. If he asks just his friends from both classes, the results will be biased. Why? Because we tend to make friends with people with similar interests. So, it is very likely that friends of David have more or less similar scores.

That is, David decided to take a sample of 6 random students from both classes and he asked them about math quarter grades. He got the following results:

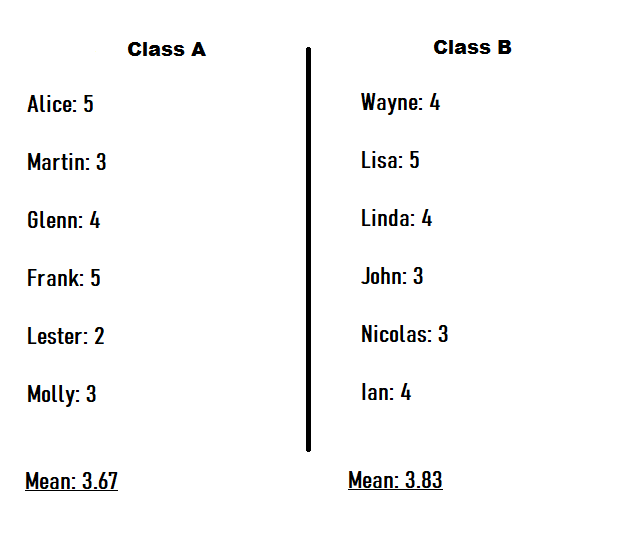


figure (1) — student’s quarter grades | by Author

It seems that students from class B outperform students from class A. But David did not ask other people! Maybe if he asked all the students, he could get the reverse result. Who knows? So, here is the problem and it needs to be solved scientifically.

To check whether the result was not likely to occur randomly or by chance, David can use the approach called **hypothesis testing**. **A hypothesis** is a claim or assumption that we want to check. The approach is very similar to a court trial process, where a judge should decide whether an accused person is guilty or not. There are two types of hypotheses:

* **Null hypothesis****(H₀)**— the hypothesis that we have by default, or the accepted fact. Usually, it means the absence of an effect. By analogy with the trial process, it is “presumption of innocence” — a legal principle that every person accused of any crime is considered innocent until proven guilty.
* **Alternative hypothesis** **(H₁)**— the hypothesis that we want to test. In other words, the alternative hypothesis will be accepted only if we gather enough evidence to claim that the effect exists.

The null hypothesis and alternative hypothesis are *always*mathematically opposite. The possible outcomes of hypothesis testing:

* **Reject the null hypothesis** —a person is found guilty.
* **Fail to reject the null hypothesis** — the accused is acquitted.

David decided to state hypotheses in the following way:

* **H₀** — There is no difference in the grade means of those students in class A and those from class B.
* **H₁**— There is a difference in the grade means of those students in class A and those from class B.

Now, David needs to gather enough evidence to show that students in two classes have different academic performances. But, what can he consider as “evidence”?

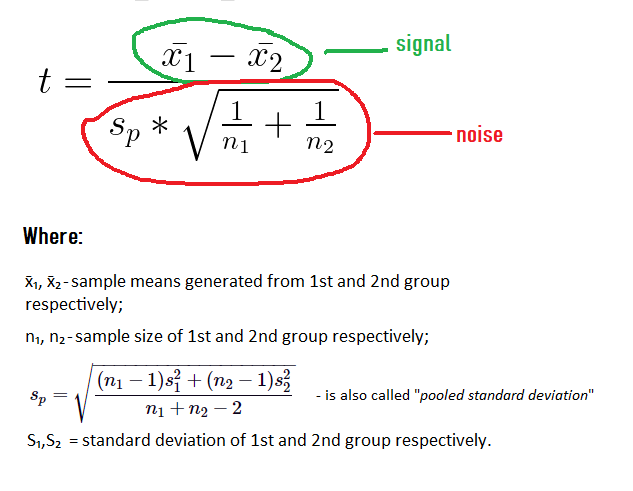
**T-test definition, formula explanation, and assumptions.**

The T-test is the test, which allows us to analyze one or two sample means, depending on the type of t-test. Yes, the t-test has several types:

* **One-sample t-test**— compare the mean of one group against the specified mean generated from a population. For example, a manufacturer of mobile phones promises that one of their models has a battery that supports about 25 hours of video playback on average. To find out if the manufacturer is right, a researcher can sample 15 phones, measure the battery life and get an average of 23 hours. Then, he can use a t-test to determine whether this difference is received not just by chance.
* **Paired sample t-test** — compares the means of two measurements taken from the same individuals, objects, or related units. For instance, students passed an additional course for math and it would be interesting to find whether their results became better after course completion. It is possible to take a sample from the same group and use the paired t-test.
* **An Independent two-sample t-test** —is used to analyze the mean comparison of two independent groups. Like two groups of students. Does it remind you of something?

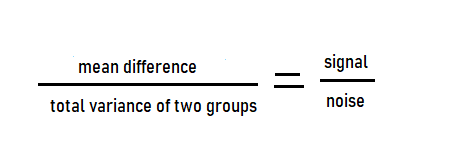
Exactly. David wants to use the independent two-sample t-test to check if there is a real difference between the grade means in A and B classes, or if he got such results by chance. Two groups are independent because students who study in class A cannot study in class B and reverse. And the question is how David can use such a test?

We have the following formula of t-statistic for our case, where the sample size of both groups is equal:



formula (1) — t-statistic, where equal variance is assumed | by Author

The formula looks pretty complicated. However, it can be presented in another way:



formula (2) — simplified formula of the t-statistic

Basically, **t-statistic is a signal-to-noise ratio**. When we assume that the difference between the two groups is real, we don’t expect that their means are exactly the same. Therefore, the greater the difference in the means, the more we are confident that the populations are not the same. However, if the data is too scattered (with high variance), then the means may have been a result of randomness and we got ones by chance. Especially, when we have a small sample size, like 3–5 observations.

Why is that? Take for example the salary of people living in two big Russian cities — Moscow and St. Petersburg.

There is a very high variance because the salary ranges from approximately $100 up to millions of dollars. So, if you decided to find whether the difference in means between the two cities exists, you may take a sample of 10 people and ask about their salaries. I know, it is very unlikely that you’ll face some millionaire on a street and I know, it is a bit strange to compare average salaries instead of median salaries. Nevertheless, if you took the sample correctly, you may find that the salary of people is highly scattered in both cities. For instance, in St. Petersburg, the mean is $7000 and the standard deviation is $990, in Moscow — $8000 is the mean and $1150 standard deviation. In such a situation, you can’t be confident whether the difference in means is statistically significant. That’s because you asked only 10 people and the variance of salary is high, hence you could get such results just by chance.

Thus, the concept of t-statistic is just a signal-to-noise ratio. With less variance, more sample data, and a bigger mean difference, we are more sure that this difference is real.

I could take an even closer look at the formula of t-statistic, but for the purpose of clarity, I won’t. If you want, you can read the proof [here](https://www.real-statistics.com/students-t-distribution/two-sample-t-test-equal-variances/two-sample-t-test-advanced/). Knowing the idea of the t-test would be enough for effective usage.

Let’s also cover some **assumptions regarding the t-test**. There are 5 main assumptions listed below:

1. The data is collected from a representative, randomly selected portion of the total population. This is necessary to generalize our findings to our target population (in the case of David — to all students in two classes).
2. Data should follow a continuous or discrete scale of measurement. We can consider grades as an example of discrete data.
3. Means should follow the normal distribution, as well as the population. Not sample data, as some people may think, but means *and*population. This needs a more detailed explanation, which I give in the section about t-distributions.
4. *(for independent t-test)* Independence of the observations. Each subject should belong to only one group. There is no relationship between the observations in each group. Otherwise, use the *paired t-test*.
5. *(for an independent t-test with equal variance)*Homogeneity of variances.Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal. It is possible to test for variance equality using F-test or Levene test. Otherwise, we should use Welch’s t-test.

So, t-statistic is the evidence that David needs to gather in order to claim that the difference in means of two groups of students is not taking place by chance. If there will be enough evidence, then David can reject the null hypothesis. The question is how much evidence is enough?

**Choosing the level of significance**

David needs to determine whether a result he has got is likely due to chance or to some factor of interest. He can find t-statistic as the evidence, but *how much risk David is willing to take for making a wrong decision*? This risk can be represented as the level of significance (α).

The **significance level is the desired probability of rejecting the null hypothesis when it is true**. For instance, if a researcher selects α=0.05, it means that he is willing to take a 5% risk of falsely rejecting the null hypothesis. Or, in other words, to take the 5% risk of conviction of an innocent. Statisticians often choose α=0.05, while α=0.01 and α=0.1 are also widely used. However, this choice is only a convention, based on R. Fisher’s argument that a 1/20 chance represents an unusual sampling occurrence. This arbitrary threshold was established in the 1920s when a sample size of more than 100 was rarely used.

We don’t want to set the level of significance mindlessly. But what approach we should use to choose this value? Well, describing such an approach in detail is a topic for another article because there are a lot of things to talk about. Still, I’m going to give a quick explanation of the factors to consider while choosing an optimal level of significance. According to J. Kim (2021), these factors include:

* losses from incorrect decisions;
* the researcher’s prior belief for the **H₀**and **H₁**;
* the power of the test;
* substantive importance of the relationship being tested.

By saying “the researcher should consider losses from incorrect decisions”, it is meant that the researcher has to figure out whether Type I error is more important than Type II error, or reverse.

**Type I error** means *rejecting the null hypothesis when it’s actually true*.

**Type II error**occurs when a statistician*fails to reject a null hypothesis that is actually false*.

Notice that Type I error has almost the same definition as the level of significance (α). The difference is that Type I error is the actual error, while the level of significance represents the desired *risk*of committing such error. The risk of committing Type II error is represented by the **β**sign and 1-β stands for the **power**of the test. In other words, the **power is the probability that the test correctly rejects the null hypothesis**. It is also called as “true positive rate”.

There may be cases when a Type I error is more important than a Type II error, and the reverse is also true. Take A/B testing as an example. A researcher wants to test two versions of a page on a website. After running the t-test one incorrectly concludes that version B is better than version A. As a consequence, the website starts to lose conversions. Another case is testing for pregnancy. Suppose, there are two tests available. Test 1 has a 5% chance of Type I error and a 20% chance of Type II error. Test 2 has a 20% chance of Type I error and 5% of Type II error. In this case, a doctor would prefer using Test 2 because misdiagnosing a pregnant patient (Type II error) can be dangerous for the patient and her baby.

The second thing that needs to be considered is the researcher’s prior belief in two hypotheses. The word “prior” means that a researcher has a personal assumption on the probability of H₀ relative to H₁ before looking at one’s data. However, the assumption should not be arbitrary or irrational just because it is “personal”. It needs to be based on good argumentation. For example, the judgment can preferably be informed by previous data and experiences. Let’s say that some researcher has invented a drug, which can cure cancer. There had been many researchers before him with similar “inventions”, whose attempts had failed. That is, the researcher believes that the probability of H₁ (i. e. the drug can cure cancer) is highly unlikely and is about 0.001. In another case, if a statistician a priori believes that H₀ and H₁ are equally likely, then the probability for both hypotheses will be 0.5.

The third factor is substantive importance or the effect size. It accounts for the question of how big the effect size is of the relationship being tested. When there is a big sample size, the t-test often shows the evidence in favor of the alternative hypothesis, although the difference between the means is negligible. While testing on small sample sizes, the t-test can suggest that H₀ should not be rejected, despite a large effect. That’s why it is recommended to set a higher level of significance for small sample sizes and a lower level for large sample sizes.

While reading all this, you may think: “OK, I understand that the level of significance is the desired risk of falsely rejecting the null hypothesis. Then, *why not set this value as small as possible in order to get the evidence as strongest as possible*? So, if I conduct a study, I can always set α around 0.00001 (or less) and get valid results”.

There is a reason why we shouldn’t set α as small as possible. Partially, we’ve already talked about it when presenting the concept of substantive importance — on small sample sizes we can miss a large effect if α is too small. But the answer is hidden in the fourth factor that we haven’t discussed yet. And it is the power.

There is a relationship between the level of significance and the power. These values depend on each other. Making decisions on them is like deciding where to spend money or how to spend free time. There are benefits in one area and there are losses in another area. The relationship between α and β is represented in a very simple diagram below. Note that β is the probability of Type II error, not power (power is 1-β).

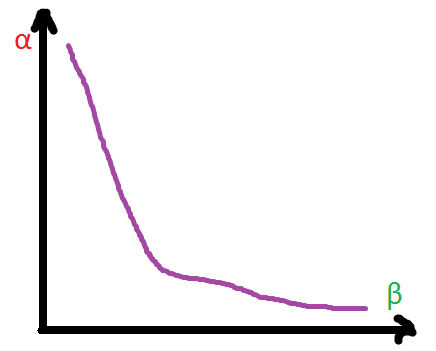


figure (2) — the relationship between α and β | by Author

As you see, there is a trade-off between α and β. The optimal value of α can be chosen after estimating the value of β. It can be done in one of the following two ways:

* using the assumption of normality
* using bootstrapping

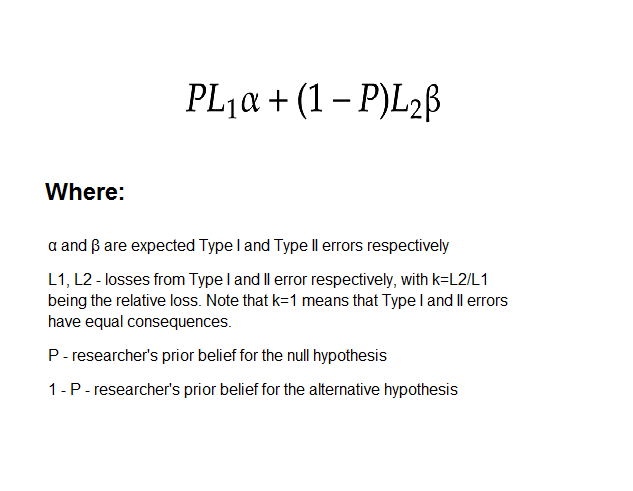
It is preferred to use the second method for calculating the power because there are many cases when the assumption of normality fails or is unjustifiable. The bootstrapping approach doesn’t rely on this assumption and takes full account of sampling variability. That’s why it is widely used in practice.

So, how to use bootstrapping to calculate the power?

In the case of David, there are 3 steps:

1. Generate independent samples from class A and class B;
2. Perform the test, comparing class A to class B, and record whether the null hypothesis was rejected;
3. Repeat steps 1–2 many times and find the rejection rate — this is the estimated power.

Calculating the power is only one step in the calculation of expected losses.



formula (3) — expected losses | image by Author, formula by J. Kim and I. Choi (2021)

**The optimal value of α can be chosen in 3 steps:**

1. Choose a grid of α ∈ (0,1)
2. For each value of α, calculate β (using the 3-step process described above) and expected loss by the formula above
3. Find the value of α that minimizes expected loss

Let’s get back to David. He wants to set the desired risk of falsely rejecting H₀. To do this correctly David considers 4 factors that we’ve already discussed. First, he thinks that Type I and Type II errors are equally important. Second, David believes that students in both classes do not have the same grades. That is, he gives more weight to his alternative hypothesis (P=0.4, 1-P=0.6). Third, because the sample size is small, David decides to raise α much higher than 0.05 to not to miss a possible substantial effect size. The last thing that he needs to do is to estimate the power. For estimating the power it is necessary to choose a grid of possible values of α and for each α carry out multiple t-tests to estimate the power. For now, David knows that the null hypothesis should be rejected if the p-value is greater than the level of significance. Otherwise, one fails to reject the null hypothesis. In the following section I explain the meaning of the p-value, but let’s leave this for now.

The whole process of calculating the optimal level of significance can be expressed in the R code below:

opt\_alpha = function(x, y, alpha\_list, P=0.5, k=1, sample\_size=6, is\_sampling\_with\_replacement=TRUE){  
   
 # This function estimates the power using simulation and returns a data frame with alpha, beta and losses that should be minimized  
   
 beta\_list = c()expected\_losses\_list = c()  
   
 P = 0.4  
 k = 1  
   
 for (alpha in alpha\_list){  
 set.seed(23)  
 rejection\_count = c()  
 for (i in 1:500){  
 sample\_x = sample(x, size = sample\_size, replace = is\_sampling\_with\_replacement)  
 sample\_y = sample(y, size = sample\_size, replace = is\_sampling\_with\_replacement)  
 pvalue = t.test(x = sample\_x, y=sample\_y)$p.value  
 if (pvalue < alpha){  
 rejection\_count = append(rejection\_count,1)  
 }else{  
 rejection\_count = append(rejection\_count,0)  
 }  
 }  
   
 rejected = sum(rejection\_count)  
 total = length(rejection\_count)  
   
 power = rejected/total  
 beta = 1 - power  
   
 beta\_list = append(beta\_list,beta)  
   
 expected\_losses = (P\*alpha)+((1-P)\*k\*beta)  
   
 expected\_losses\_list = append(expected\_losses\_list,expected\_losses)  
 }  
 solutions = data.frame(alpha\_list, beta\_list,expected\_losses\_list)   
 return(solutions)  
}alpha\_list = c(0.01,0.05,0.1,0.15,0.20,0.25,0.30,0.35,0.40,0.45,0.5,0.55,0.6,0.65,0.7,0.75,0.8,0.85,0.9,0.95)solutions = opt\_alpha(x = a\_score$Score, y = b\_score$Score,alpha\_list, P=0.4, k=1)optimal\_solution = solutions %>% filter(expected\_losses\_list==min(expected\_losses\_list))

https://miro.medium.com/max/310/1*dXm5LzAiSPf8uMHkd4bxZQ.png

figure (3) — optimal level of significance | by Author

David found that α = 0.8 is the optimal value. Notice how far it is from the conventional level of 0.05.

**T-distribution and p-value**

So, David set the level of significance equal to 0.8. Now, he can calculate the t-statistic.

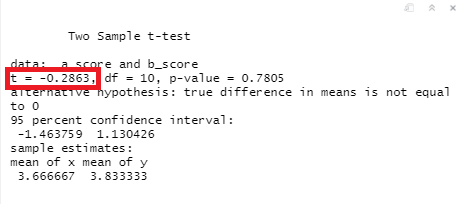


figure (4) — the results of the t-test | by Author

After calculation, he figured out that t-statistic = -0.2863. Why this value is negative? Because we observe a negative effect. In this sample, students from class B perform better in math, though David supposed that students from class A are better. The other thing that we found is that the signal is about 28.6% from the noise. It almost gets lost. Perhaps, the difference in the means is explained by variance. But how big t-statistic should be to reject the null hypothesis?

That’s where t-distribution comes in. It connects the level of significance and t-statistic so that we could compare the proof boundary and the proof itself. The idea of t-distribution is not as hard as one might think. Consider the example of comparing the mean SAT scores of two cities. We know that in both cities SAT scores follow the normal distribution and the means are equal, i.e. the null hypothesis is true. Note that SAT scores from both cities represent two populations, not samples.

From this point, we can start to develop our logic. We decided to emulate the actions of a person, who wants to compare the means of two cities but have no information about the population. Of course, one would take samples from each distribution. Let’s say, the sample size was 10. The following R code generates SAT distributions, takes samples from both, and calculates the t-statistic.

# 1. Generate two normal distributions with equal meansset.seed(123)  
city1 = rnorm(n = 10000, mean = 1150, sd = 150)  
city1 = as.data.frame(city1)  
city2 = rnorm(n = 10000, mean = 1150, sd = 150)  
city2 = as.data.frame(city2)ggplot(data = city1) + geom\_density(aes(x = city1), colour = 'red') + xlab("City1 SAT scores")ggplot(data = city2) + geom\_density(aes(x = city2), colour = 'green')+ xlab("City2 SAT scores")

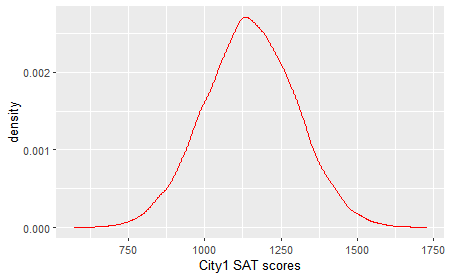


figure (5) — City 1 SAT scores distribution | by Author

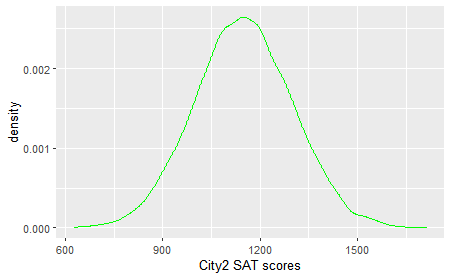


figure (6) — City 2 SAT scores distribution | by Author

# 2. Take samples from both distributionsset.seed(2356)  
sample\_city1 = sample(city1$city1, size = 10)  
sample\_city2 = sample(city2$city2, size = 10)# 3. Calculate t-value  
tvalue = t.test(x = sample\_city1, y=sample\_city2, var.equal = TRUE)$statistictvalue = as.numeric(tvalue)

We got value of t-statistic equal to 1.09. It shows some signal, which is strange because we know that H₀ is true and t-value should be equal to zero. Why is that? That’s because we got unlucky with our samples. It would be interesting to know how t-statistic would change if we take samples 70 thousand times. Let’s do it.

# 4. Do steps 2-3 70000 times and generate a list of t-valuestvalue\_list = c()for (i in 1:70000){  
 sample\_city1 = sample(city1$city1, size = 10)  
 sample\_city2 = sample(city2$city2, size = 10)  
   
 tvalue = t.test(x = sample\_city1, y=sample\_city2, var.equal = TRUE)$statistictvalue = as.numeric(tvalue)  
   
 tvalue\_list = append(tvalue\_list,tvalue)  
}

Well, we’ve got a huge list of t-values. Let’s plot ones.

# 5. Plot the list of t-valuesggplot(data = as.data.frame(tvalue\_list)) + geom\_density(aes(x = tvalue\_list)) + theme\_light()+xlab("t-value")

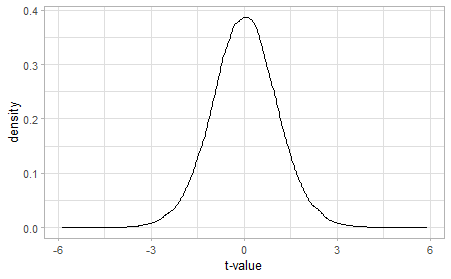


figure (7) — the distribution of the t-value | by Author

That’s it. Now we have a distribution of t-statistic that is very similar to Student’s t-distribution. T-distribution looks like the normal distribution but it has heavier tails. Also, it can look different depending on sample size, and with more observations, it approximates the normal distribution. T-distribution can be interpreted as follows. There is a high chance of getting a t-value equal to zero when taking samples. It makes sense — when the null hypothesis is true, the t-value should be equal to zero because there is no signal. But the further away the t-value is from zero, the less likely we are to get it. For instance, it is very unlikely to get t=6. But a question arises there. How much it is likely or unlikely to get a certain t-value?

*The probability of getting a t-value at least as extreme as the t-value actually observed under the assumption that the null hypothesis is correct* is called the **p-value**. In the figure below the probability of observing t>=1.5 corresponds to the red area under the curve.

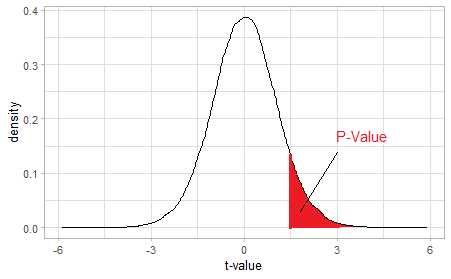


figure (8) — the distribution of the t-value with one-sided critical region | by Author

A very small p-value means that getting a such result is very unlikely to happen if the null hypothesis was true. The concept of p-value helps us to make decisions regarding H₀ and H₁. T-statistic shows the proportion between the signal and the noise, the p-value tells us how often we could observe such a proportion if H₀ would be true, and the level of significance acts as a decision boundary. By analogy to a court trial process, p-value=0.01 is somewhat similar to the next statement: “*If this man is innocent, there is a 1% probability that one would behave like this (change testimony, hide evidence) or even more weirdly*”. The jury can determine whether the evidence is sufficient by comparing the p-value with some standard of evidence (the level of significance). Thus, if α = 0.05 and p-value=0.01, the jury can deliver a “guilty” verdict.

Several notes need to be taken. First, **there is a common misinterpretation of the p-value, when people say that “the p-value is the probability that H₀ is true”**. Of course, the p-value doesn’t tell us anything about H₀ or H₁, it only assumes that the null hypothesis is true. Consider the example, when David took a sample of students in both classes, who get only 5’s. T-statistic would be obviously 0 because there is no observed difference in the means. In this case, a p-value would be equal to 1, but does it mean that the null hypothesis is true “for certain”? No, not at all! It rather means that David did sampling incorrectly, choosing only the “good” students in math, or that he was extremely unfortunate to get a sample like this. Second, t-distribution was not actually derived by bootstrapping (like I did for educational purposes). **In the times of Willam Gosset, there were no computers, so t-distribution was derived mathematically**. I decided not to dive deep into math, otherwise, it would be hard to agree that the t-test is “explained simply”. Third, because t-statistic have to follow t-distribution,**the t-test requires normality of the population**. However, the population should not necessarily have a “perfect” normal distribution, otherwise, the usage of the t-test would be too limited. There may be some skewness or other “imperfections” in the population distribution as long as these “imperfections” allow us to make valid conclusions.

Finally, the critical region (red area on the figure 8) doesn’t have to take only one side. If there is a possibility that the effect (the mean difference) can be positive or negative, it is better to use a **two-tailed t-test**. The two-tailed t-test can detect the effect from both directions. For David, it is appropriate to use a two-tailed t-test because there is a possibility that students from class A perform better in math (positive mean difference, positive t-value) as well as there is a possibility that students from class B can have better grades (negative mean difference, negative p-value). **The one-tailed t-test** can be appropriate in cases, when the consequences of missing an effect in the untested direction are negligible, or when the effect can exist in only one direction.

So…

David has calculated a p-value.

It equals 0.7805.

Because David set α = 0.8, he has to *reject the null hypothesis*.

That’s it. The t-test is done. David now can say with some degree of confidence that the difference in the means didn’t occur by chance. But David still has doubts about whether his results are valid. Perhaps, the problem is connected with the level of significance. David allowed himself to falsely reject the null hypothesis with the probability of 80%. On the other hand, if the level of significance would be set lower, there would be a higher chance of erroneously claiming that the null hypothesis should not be rejected.

Well, that’s the nature of statistics. We *never*know for certain. Maybe, David could get more confidence in results if he’d get more samples. Who knows what the result of the t-test would show?

**Conclusion**

Suppose, we are a head teacher, who has access to students’ grades, including grades from class A and class B. We can figure out whether David was right or wrong. Here are the actual results:

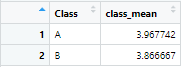


figure (9) — actual grade means of students from both classes | by Author

Indeed, students from class A did better in math than those from class B. There is a difference between the means, but it is pretty small. Therefore, the alternative hypothesis is true. Let’s calculate the true β (true α we cannot calculate because the null hypothesis is false, therefore, it is impossible to falsely reject the null hypothesis). For our α = 0.8, we found that β = 0.184. Comparing this value to the estimate of β = 0.14, we can say that our bootstrapping approach worked pretty well. Nevertheless, we underestimated the probability of Type II error.

What is the lesson to learn from this information?

Again, don’t be too confident, when you’re doing statistics. You shouldn’t rely on t-tests *exclusively*when there are other scientific methods available. Your logic and intuition matter. There is another thing to point out. David’s goal was to find out whether students from class A get better quarter grades than those from class B. Suppose that David conducted a rigorous study and figured out the right answer. But do the results have practical significance? Probably, not. What can he do with these results? Yes, students in class A got better quarter grades. But does it mean that students in class A are better in math than students from class B? It is impossible to answer this question, using the data only from one quarter. Perhaps, it would be useful to gather the information from other periods and conduct a time-series analysis. But still, using only observational data it is extremely difficult to find out some causal relationship, if not impossible. So here is another lesson. Do not try to make conclusions about the causality of the relationship observed while using statistical methods, such as t-test or regression.

If you want to take a look at David’s dataset and R code, you can download all of that using this [link](https://drive.google.com/file/d/1fyChFbcTXlBnJZy4L-Am7s6CxPBbGUCC/view?usp=sharing). A full dataset of students’ grades is also available in the archive. All the datasets were created by me.

Finally, if you have questions, comments, or criticism, feel free to write in the comments section. We all learn from each other.

Thank you for reading!