Simple guide to Student's t-test

Student's t-test is one of the most frequently used procedures in statistics. Why t-test is commonly used to study significance of differences is somewhat puzzling, considering it was originally used to guarantee beer quality in Guinness brewery in Ireland. Assumedly number of people would sample a barrel, give it a grade and then the statistics of that barrel would be compared to the statistics of other barrels. In short William Gosset (Student was his pen name) developed a method which allowed gave the brewers a guidelines to evaluate the results of the brewing process. How this method became the de-facto statistical test for all things scientific* is a mystery to me.

Unfortunately, people who frequently use t-tests often don't know exactly what happens when their data are wheeled away and operated upon behind the curtain using statistical software. Because if you know how a t-test works, you can understand what your results really mean. You can also better grasp why your study did (or didn't) achieve "statistical significance."

* *Where scientific refers to biology and medicine. Student's test is less popular among chemists and physicists for various reasons. One being that a 'statistically significant' difference does not imply actually significant difference. This will be discussed later.*

Anatomy of a t-test

A t-test is commonly used to determine whether the mean (average) of a population (set of measurements) significantly differs from a specific value (called the *hypothesized mean*) or from the mean of another population (set of measurements).

In a way, the t-value (not to be mixed with p-value) is a representation of trust we have the averages of the measurements being different from each other.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s/\sqrt{n}} = \frac{Differnce\ of\ averages}{Uncertainty\ of\ measurements},$$

The higher the difference of the averages is, the more we can trust our measurements arising from different populations. On the other hand, the more uncertain our measurements are, the less we can trust the measured difference. Here $\bar{X}_1$ and $\bar{X}_2$ are the averages of populations 1 and 2, *s* is the standard deviation (assumed to be constant, an assumption that doesn't apply most of the time) and *n* is the number of measurements, also assumed to be the same for both samples. We shall also note that increasing the sample size (*n*) decreases the uncertainty of measurements (duh).

As the above formula shows, the t-value simply compares the strength of the signal (the difference) to the amount of noise (the variation) in the data.

If the signal is weak *relative* to the noise, the (absolute) size of the t-value will be smaller. So the difference is not likely to be statistically significant. However, if the signal (difference of averages) is strong *relative* to the noise, the (absolute) size of the t-value will be larger.

When the standard deviations of the samples are not equal (i.e. always) and when there is unequal amount of measurements, the equation above changes into

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{Differnce\ of\ averages}{Uncertainty\ of\ measurements},$$

where $\bar{X}_1$ and $\bar{X}_2$ are the averages of populations 1 and 2, s_1 and s_2 are the standard deviations of the samples 1 and 2 while n_1 and n_2 are the number of measurements in samples 1 and 2. Let's note that if standard deviations and numbers of measurement are equal the equation above simplifies to the earlier form. This equation is also known as Welch's t-test.

For the sake of clarity, we'll remind ourselves of the fact that variance is dependent of standard deviation and they have the following relation:

$$variance = (stadrad\ deviation)^2$$

Hence the s^2 term is the variance of the data.

Statistical significance or p-values

In short, ***p*-value** is a function of the observed sample results (a test statistic such as t-value) relative to a statistical model, which measures how extreme the observation is. However, the desired significance level should be determined BEFORE the experiment is made, and then the measurements should be compared to the desired p-value. If the *p*-value is less than or equal to the chosen significance level (α) , the test suggests that the data sets differ from each other and we must reject the assumption that the two data sets (samples) would come from the same source.

However, that does not prove that the tested assumption (of difference) is true. When the *p*-value is calculated correctly, this test guarantees that the Type I error rate* is at most α. For typical analysis, using the standard $\alpha = 0.05$ cutoff, the assumption of the samples coming from the same source is rejected when $p < .05$ and not rejected when $p > .05$. The *p*-value does not in itself support reasoning about the probabilities of assumption but is only a tool for deciding whether to reject the assumption of the samples being the same.

American Statistical Association has stated that: *"The widespread use of "statistical significance" (generally interpreted as "p ≤ 0.05") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process."* Now consider the fact that a) to progress on your career you must publish and that b) most journals require some kind of statistical testing.

Richard Feynman noted that

In the Pacific there is a cargo cult of people. During the war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to imitate things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like

headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things cargo cult science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.

In short, just having the t-test on your paper won't make your paper any better, and if the only result you have are dependent of statistical interpretation, you might want to consider further experiments. Or in the words of Ernest Rutherford:

"If your experiment needs statistics, you ought to do a better experiment."

A good result is obvious without any kind of statistical analysis and will yield statistical significance regardless of the tools used. A bad result is only evident after statistical testing and such results be rightfully doubted*.

With that out of the way, let's look at how to choose the p-value for your experiment.

Statisticians** have estimated the error rates associated with different P values. While the precise error rate depends on various assumptions, the table below summarizes them for middle-of-the-road assumptions.

* There are cases where statistics ARE the result, e.g. genome wide and other high-throughput studies. Your experiment is not (really) a high-throughput experiment unless you have dozens of conditions.

***Thomas SELLKE, M. J. BAYARRI, and James O. BERGER, Calibration of p Values for Testing Precise Null Hypotheses, The American Statistician, February 2001, Vol. 55, No. 1*

p-value	Probability of incorrectly assuming samples to be different
0.05	At least 23% and typically close to 50%
0.01	At least 7% and typically close to 15%

Needless to say, the p = 0.05 required by the journals could be considered a relic of cargo cults in science. A magical number that has been agreed on sometime, but has actual little to do with data being valid or not. Now the reader as a scientist has two choices. Either conform to journals, looking for that elusive p = 0.05 and hope that they are on the right side of that 50% or aim for that p = 0.01 and risk missing a positive result. It should be perhaps noted that recent issues with unreproducible results is likely partially due to mishandled statistics. Notably, looking at equations above, high *n* will always lead to small 'noise' and thus potentially high 'statistical significance'.

So, now that we know that p-values have varying degrees of value for the scientists, we can have a look on how to determine whether our experiment has smaller p-value than we demanded. To reiterate, to use the t-test as you need to decide the p-value before you start the testing and then see if your data confers to the value. Basically, you need to look from a chart.

Table T Critical Values of the *t* Distribution

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Source: From *Biometrika Tables for Statisticians*, Vol. 1, Third Edition, edited by E. S. Pearson and H. O. Hartley, 1966, p. 146. Reprinted by permission of the Biometrika Trustees.

That's a damn large and complicated table. So let's have a look at this mess. First things first, the numbers if the table are t-values corresponding to the p-values on the upmost row of the chart. Now, there are two rows of p-values corresponding to one-tailed and two-tailed t-testing. Keep this in mind, we'll explore what

it means below. The leftmost column has been shortened as 'df'. This stands for 'degrees of freedom' which will be also discuss briefly below.

Degrees of freedom

Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter are called the degrees of freedom. The more *n* you have, the more degrees of freedom your data has.

The degrees of freedom for t-test presented above can be determined from Welch-Statterhwaite equation given below.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Well, that's one scary equation, I admit that. However, once inserted into excel getting numbers out requires you to only know the standard deviations (s_1 and s_2) and the number of measurements (n_1 and n_2) of your samples. Proof for this goes way (way) beyond the aims of this document. To maintain robustness fractional degrees of freedom should be rounded up.

Now assuming you know which column to look, you find the corresponding degree of freedoms from the column on the left and consult the table. The number you find shows you the lowest t-value you can have for your experiment to have the p-value you have determined.

Table T Critical Values of the *t* Distribution

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140

For two-tailed test to have p-value of 0.01 when degree of freedoms is 14, the t-value has to be at least 2.977. Any t-value greater than this means that your result agrees with the p-value you have determined.

One-tailed or two-tailed?

First let's start with the meaning of a two-tailed test. If you are using a significance level of 0.05, a two-tailed test allots half of your p-value to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction. This means that .025 is in each tail of the distribution of your test statistic. When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions.

For example, we may wish to compare the mean of a mutant to a wildtype (WT) using a t-test. Our null hypothesis (assumption) is that the mean is equal, and we want to disapprove our assumption. A two-tailed test will test if the mean of the mutant is either significantly higher or lower than the WT. In cases where you do not know, or cannot be sure about the direction of the change, you should use two-tailed test. This adds robustness to your test and reduces the risk of false positives.

Next, let's discuss the meaning of a one-tailed test. If you are using a significance level of .05, a one-tailed test allots all of your p-value to testing the statistical significance in the one direction of interest. When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction. A one-tailed test will test either if the mean of the mutant is significantly greater than the mean of the WT or if the mean is significantly less than the mean of the WT, but not both. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction.

When one-tailed test is appropriate?

Because the one-tailed test provides more power to detect an effect, you may be tempted to use a one-tailed test whenever you have a hypothesis about the direction of an effect. Before doing so, consider the consequences of missing an effect in the other direction. Imagine you are testing a potential new treatment that you believe is an improvement over an existing drug. You wish to maximize your ability to detect the improvement, so you opt for a one-tailed test. In doing so, you fail to test for the possibility that the new drug is less effective than the existing drug. The consequences in this example are extreme, but they illustrate a danger of inappropriate use of a one-tailed test.

So when is a one-tailed test appropriate? If you consider the consequences of missing an effect in the untested direction and conclude that they are negligible and in no way irresponsible or unethical, then you can proceed with a one-tailed test. For example, imagine again that you have developed a new drug. It is cheaper than the existing drug and, you believe, no less effective. In testing this drug, you are only interested in testing if it less effective than the existing drug. You do not care if it is significantly more effective. You only wish to show that it is not less effective. In this scenario, a one-tailed test would be appropriate.

When one-tailed test is NOT appropriate.

Choosing a one-tailed test for the sole purpose of attaining significance is not appropriate. Choosing a one-tailed test after running a two-tailed test that failed to reject the null hypothesis is not appropriate, no matter how "close" to significant the two-tailed test was. Using statistical tests inappropriately can lead to invalid results that are not replicable and highly questionable--a steep price to pay for a significance star in your results table!

Multiple Comparisons Problem

Multiple comparisons problem (also known as multiple testing problem) occurs when one considers a set of statistical inferences simultaneously. In short, if you compare more than ONE set of data with your control data, you are performing multiple comparisons.

Suppose we consider the efficacy of a drug in terms of the reduction of any one of a number of disease symptoms. As more symptoms are considered, it becomes more likely that the drug will appear to be an improvement over existing drugs in terms of at least one symptom.

Suppose we consider how much certain treatments effect expression of a protein in comparison to untreated control. As more treatments are compared with the control, it becomes more likely that the treatment and control groups will appear to differ *by random chance alone.*

For example, if one (statistical) test is performed at the 5% level, there is only a 5% chance of incorrectly rejecting the null hypothesis* if the null hypothesis is true. However, for 100 tests where all null hypotheses are true, the expected number of incorrect rejections is 5. If the tests are independent, the probability of at least one incorrect rejection is 99.4%.

*recall that null hypothesis was our assumption that our control has the same mean as our sample, a claim we wish to prove false.

To give another example, one might declare that a coin was biased if in 10 flips it landed heads at least 9 times. Indeed, if one assumes as a null hypothesis that the coin is fair, then the probability that a fair coin would come up heads at least 9 out of 10 times is $(10 + 1) \times (1/2)^{10} = 0.0107$. This is relatively unlikely, and under statistical criteria such as p-value < 0.05, one would declare that the null hypothesis should be rejected — i.e., the coin is unfair.

A multiple-comparisons problem arises if one wanted to use this test (which is appropriate for testing the fairness of a single coin), to test the fairness of many coins. Imagine if one were to test 100 fair coins by this method. Given that the probability of a fair coin coming up 9 or 10 heads in 10 flips is 0.0107, one would expect that in flipping 100 fair coins ten times each, to see *a particular* (i.e., pre-selected) coin comes up heads 9 or 10 times would still be very unlikely, **but seeing any coin behave that way, without concern for which one, would be more likely than not.** Precisely, the likelihood that all 100 fair coins are identified as fair by this criterion is $(1 - 0.0107)^{100} \approx 0.34$. Therefore, the application of our single-test coin-fairness criterion to multiple comparisons would be more likely to falsely identify at least one fair coin as unfair.

Correcting for multiple comparisons

Luckily, there are several simple methods to correct for multiple comparisons error. One of the easiest corrections is the **Bonferroni correction**. In this correction the desired p-value is adjusted in following manner.

Bonferroni Correction

For an experiment with n comparisons to have p-value of α, the p-values of individual comparisons must be set to α/n. Bonferroni correction is robust, and if your results are significant even after Bonferroni you have a good reason to trust your data. Unfortunately, the robustness of Bonferroni correction comes at a cost, often causing false negatives.