How to run a t test

Statistical hypothesis test Student's t-test is a statistical test used to test whether the difference between the response of two groups is statistically significant or not. It is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. It is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known (typically, the scaling term is unknown and is therefore a nuisance parameter). When the scaling term is estimated based on the data, the test statistic—under certain conditions—follows a Student's t distribution. The t-test's most common application is to test whether the means of two populations are significantly different. In many cases, a Z-test will yield very similar results to a t-test because the latter converges to the former as the size of the dataset increases. William Sealy Gosset, who developed the "t-statistic" and published it under the pseudonym of "Student" The term "t-statistic" is abbreviated from "hypothesis test statistic".[1] In statistics, the t-distribution was first derived as a posterior distribution in 1876 by Helmert[2][3][4] and Lüroth.[5][6][7] The t-distribution also appeared in a more general form as Pearson type IV distribution in Karl Pearson's 1895 paper.[8] However, the t-distribution, also known as Student's t-distribution, gets its name from William Sealy Gosset, who first published it in English in 1908 in the scientific journal Biometrika using the pseudonym "Student"[9][10] because his employer preferred staff to use pen names when publishing scientific papers.[11] Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples – for example, the chemical properties of barley with small sample sizes. Hence a second version of the etymology of the term Student is that Guinness did not want their competitors to know that they were using the t-test to determine the quality of raw material. Although it was William Gosset after whom the term "Student" is penned, it was actually through the work of Ronald Fisher that the distribution became well known as "Student's distribution"[12] and "Student's t-test". Gosset devised the t-test as an economical way to monitor the quality of stout. The t-test work was submitted to and accepted in the journal Biometrika and published in 1908.[9] Guinness had a policy of allowing technical staff leave for study (so-called "study leave"), which Gosset used during the first two terms of the 1906–1907 academic year in Professor Karl Pearson's Biometric Laboratory at University College London.[13] Gosset's identity was then known to fellow statisticians and to editor-in-chief Karl Pearson.[14] A one-sample Student's t-test is a location test of whether the mean of a population has a value specified in a null hypothesis. In testing the null hypothesis that the population mean is equal to a specified value μ0, one uses the statistic t = x ¯ − μ 0 s / n , {\displaystyle t={\frac {{\bar {x}}-\mu _{0}}{s/{\sqrt {n}}}},} where x ¯ {\displaystyle {\bar {x}}} is the sample mean, s is the sample standard deviation and n is the sample size. The degrees of freedom used in this test are n − 1. Although the parent population does not need to be normally distributed, the distribution of the population of sample means x ¯ {\displaystyle {\bar {x}}} is assumed to be normal. By the central limit theorem, if the observations are independent and the second moment exists, then t {\displaystyle t} will be approximately normal N (0 , 1) {\textstyle {\mathcal {N}}(0,1)} . Type I error of unpaired and paired two-sample t-tests as a function of the correlation. The simulated random numbers originate from a bivariate normal distribution with a variance of 1. The significance level is 5% and the number of cases is 60. Power of unpaired and paired two-sample t-tests as a function of the correlation. The simulated random numbers originate from a bivariate normal distribution with a variance of 1 and a deviation of the expected value of 0.4. The significance level is 5% and the number of cases is 60. A two-sample location test of the null hypothesis such that the means of two populations are equal. All such tests are usually called Student's t-tests, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called Welch's t-test. These tests are often referred to as unpaired or independent samples t-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.[15] Two-sample t-tests for a difference in means involve independent samples (unpaired samples) or paired samples. Paired t-tests are a form of blocking, and have greater power (probability of avoiding a type II error, also known as a false negative) than unpaired tests when the paired units are similar with respect to "noise factors" (see confounder) that are independent of membership in the two groups being compared.[16] In a different context, paired t-tests can be used to reduce the effects of confounding factors in an observational study. The independent samples t-test is used when two separate sets of independent and identically distributed samples are obtained, and one variable from each of the two populations is compared. For example, suppose we are evaluating the effect of a medical treatment, and we enroll 100 subjects into our study, then randomly assign 50 subjects to the treatment group and 50 subjects to the control group. In this case, we have two independent samples and would use the unpaired form of the t-test. Main article: Paired difference test Paired samples t-tests typically consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (a "repeated measures" t-test). A typical example of the repeated measures t-test would be where subjects are tested prior to a treatment, say for high blood pressure, and the same subjects are tested again after treatment with a blood-pressure-lowering medication. By comparing the same patient's numbers before and after treatment, we are effectively using each patient as their own control. That way the correct rejection of the null hypothesis (here: of no difference made by the treatment) can become much more likely, with statistical power increasing simply because the random interpatient variation has now been eliminated. However, an increase of statistical power comes at a price: more tests are required, each subject having to be tested twice. Because half of the sample now depends on the other half, the paired version of Student's t-test has only n/2 − 1 degrees of freedom (with n being the total number of observations). Pairs become individual test units, and the sample has to be doubled to achieve the same number of degrees of freedom. Normally, there are n − 1 degrees of freedom (with n being the total number of observations).[17] A paired samples t-test based on a "matched-pairs sample" results from an unpaired sample that is subsequently used to form a paired sample, by using additional variables that were measured along with the variable of interest.[18] The matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of other measured variables. This approach is sometimes used in observational studies to reduce or eliminate the effects of confounding factors. Paired samples t-tests are often referred to as "dependent samples t-tests". [dubious – discuss] Most test statistics have the form t = Z/s, where Z and s are functions of the data. Z may be sensitive to the alternative hypothesis (i.e., its magnitude tends to be larger when the alternative hypothesis is true), whereas s is a scaling parameter that allows the distribution of t to be determined. As an example, in the one-sample t-test t = Z s = X ¯ − μ σ ^ / n , {\displaystyle t={\frac {Z}{s}}={\frac {{\bar {X}}-\mu }{{\hat {\sigma }}/{\sqrt {n}}}},} where X ¯ {\displaystyle {\bar {X}}} is the sample mean from a sample X1, X2, …, Xn, of size n, s is the standard error of the mean, σ ^ = 1 n − 1 ∑ i (X i − X ¯) 2 {\displaystyle {\hat {\sigma }}={\sqrt {{\frac {1}{n-1}}\sum _{i}(X_{i}-{\bar {X}})^{2}}}} is the estimate of the standard deviation of the population, and μ is the population mean. The assumptions underlying a t-test in the simplest form above are that: X follows a normal distribution with mean μ and variance σ2/n. s2(n − 1)/σ2 follows a χ2 distribution with n − 1 degrees of freedom. This assumption is met when the observations used for estimating s2 come from a normal distribution (and i.i.d. for each group). Z and s are independent. In the t-test comparing the means of two independent samples, the following assumptions should be met: The means of the two populations being compared should follow normal distributions. Under weak assumptions, this follows in large samples from the central limit theorem, even when the distribution of observations in each group is non-normal.[19] If using Student's original definition of the t-test, the two populations being compared should have the same variance (testable using F-test, Levene's test, Bartlett's test, or the Brown–Forsythe test; or assessable graphically using a Q–Q plot). If the sample sizes in the two groups being compared are equal, Student's original t-test is highly robust to the presence of unequal variances.[20] Welch's t-test is insensitive to equality of the variances regardless of whether the sample sizes are similar. The data used to carry out the test should either be sampled independently from the two populations being compared or be fully paired. This is in general not testable from the data, but if the data are known to be dependent (e.g. paired by test design), a dependent test has to be applied. For partially paired data, the classical independent t-tests may give invalid results as the test statistic might not follow a t distribution, while the dependent t-test is sub-optimal as it discards the unpaired data.[21] Most two-sample t-tests are robust to all but large deviations from the assumptions.[22] For exactness, the t-test and Z-test require normality of the sample means, and the t-test additionally requires that the sample variance follows a scaled χ2 distribution, and that the sample mean and sample variance be statistically independent. Normality of the individual data values is not required if these conditions are met. By the central limit theorem, sample means of moderately large samples are often well-approximated by a normal distribution even if the data are not normally distributed. However, the sample size required for the sample means to converge to normality depends on the skewness of the distribution of the original data. The sample can vary from 30 to 100 or higher values depending on the skewness.[23][24] For non-normal data, the distribution of the sample variance may deviate substantially from a χ2 distribution. However, if the sample size is large, Slutsky's theorem implies that the distribution of the sample variance has little effect on the distribution of the test statistic. That is, as sample size n {\displaystyle n} increases: n (X ¯ − μ) → d N (0 , σ 2) {\displaystyle {\sqrt {n}}({\bar {X}}-\mu)\xrightarrow {d} N(0,\sigma ^{2})} as per the Central limit theorem, s 2 → p σ 2 {\displaystyle s^{2}\xrightarrow {p} \sigma ^{2}} as per the law of large numbers, ∴ n (X ¯ − μ) s → d N (0 , 1) {\displaystyle \therefore {\frac {{\sqrt {n}}({\bar {X}}-\mu)}{s}}\xrightarrow {d} N(0,1)} . Explicit expressions that can be used to carry out various t-tests are given below. In each case, the formula for a test statistic that either exactly follows or closely approximates a t-distribution under the null hypothesis is given. Also, the appropriate degrees of freedom are given in each case. Each of these statistics can be used to carry out either a one-tailed or two-tailed test. Once the t value and degrees of freedom are determined, a p-value can be found using a table of values from Student's t-distribution. If the calculated p-value is below the threshold chosen for statistical significance (usually the 0.10, the 0.05, or 0.01 level), then the null hypothesis is rejected in favor of the alternative hypothesis. Suppose one is fitting the model Y = α + β x + ε , {\displaystyle Y=\alpha +\beta x+\varepsilon ,} where x is known, α and β are unknown, ε is a normally distributed random variable with mean 0 and unknown variance σ2, and Y is the outcome of interest. We want to test the null hypothesis that the slope β is equal to some specified value β0 (often taken to be 0, in which case the null hypothesis is that x and y are uncorrelated). Let α ^ , β ^ = least-squares estimators , S E α ^ , S E β ^ = the standard errors of least-squares estimators . {\displaystyle {\begin{aligned}{\hat {\alpha }},{\hat {\beta }}&={\text{least-squares estimators}},\\SE_{\hat {\alpha }},SE_{\hat {\beta }}&={\text{the standard errors of least-squares estimators}}.\end{aligned}}} Then t score = β ^ − β 0 S E β ^ ∼ T n − 2 {\displaystyle t_{\text{score}}={\frac {{\hat {\beta }}-\beta _{0}}{SE_{\hat {\beta }}}}\sim {\mathcal {T}}_{n-2}} has a t-distribution with n − 2 degrees of freedom if the null hypothesis is true. The standard error of the slope coefficient: S E β ^ = 1 n − 2 ∑ i = 1 n (y i − y ^ i) 2 ∑ i = 1 n (x i − x ¯) 2 {\displaystyle SE_{\hat {\beta }}={\frac {\sqrt {\displaystyle {\frac {1}{n-2}}\sum _{i=1}^{n}(y_{i}-{\hat {y}}_{i})^{2}}}{\sqrt {\displaystyle \sum _{i=1}^{n}(x_{i}-{\bar {x}})^{2}}}}} can be written in terms of the residuals. Let ε ^ i = y i − y ^ i = y i − (α ^ + β ^ x i) = residuals = estimated errors , SSR = ∑ i = 1 n ε ^ i 2 = sum of squares of residuals . {\displaystyle {\begin{aligned}{\hat {\varepsilon }}_{i}&=y_{i}-{\hat {y}}_{i}=y_{i}-({\hat {\alpha }}+{\hat {\beta }}x_{i})={\text{residuals}}={\text{estimated errors}},\\{\text{SSR}}&=\sum _{i=1}^{n}{{\hat {\varepsilon }}_{i}}^{2}={\text{sum of squares of residuals}}.\end{aligned}}} Then tscore is given by t score = (β ^ − β 0) n − 2 S S R ∑ i = 1 n (x i − x ¯) 2 . {\displaystyle t_{\text{score}}={\frac {({\hat {\beta }}-\beta _{0}){\sqrt {n-2}}}{\sqrt {\frac {SSR}{\sum _{i=1}^{n}(x_{i}-{\bar {x}})^{2}}}}}.} Another way to determine the tscore is t score = r n − 2 1 − r 2 , {\displaystyle t_{\text{score}}={\frac {r{\sqrt {n-2}}}{\sqrt {1-r^{2}}}},} where r is the Pearson correlation coefficient. The tscore, intercept can be determined from the tscore, slope: t score,intercept = α β t score,slope s x 2 + x ¯ 2 , {\displaystyle t_{\text{score,intercept}}={\frac {\alpha }{\beta }}{\frac {t_{\text{score,slope}}}{\sqrt {s_{\text{x}}^{2}+{\bar {x}}^{2}}}},} where sx2 is the sample variance. Given two groups (1, 2), this test is only applicable when: the two sample sizes are equal, it can be assumed that the two distributions have the same variance. Violations of these assumptions are discussed below. The t statistic to test whether the means are different can be calculated as follows: t = X ¯ 1 − X ¯ 2 s p 2 n , {\displaystyle t={\frac {{\bar {X}}_{1}-{\bar {X}}_{2}}{s_{p}{\sqrt {\frac {2}{n}}}}},} where s p = s X 1 2 + s X 2 2 2 . {\displaystyle s_{p}={\sqrt {\frac {s_{X_{1}}^{2}+s_{X_{2}}^{2}}{2}}}.} Here sp is the pooled standard deviation for n = n1 = n2, and s 2X1 and s 2X2 are the unbiased estimators of the population variance. The denominator of t is the standard error of the difference between two means. For significance testing, the degrees of freedom for this test is 2n − 2, where n is sample size. This test is used only when it can be assumed that the two distributions have the same variance (when this assumption is violated, see below). The previous formulae are a special case of the formulae below, one recovers them when both samples are equal in size: n = n1 = n2. The t statistic to test whether the means are different can be calculated as follows: t = X ¯ 1 − X ¯ 2 s p ⋅ 1 n 1 + 1 n 2 , {\displaystyle t={\frac {{\bar {X}}_{1}-{\bar {X}}_{2}}{s_{p}\cdot {\sqrt {{\frac {1}{n_{1}}}+{\frac {1}{n_{2}}}}}}},} where s p = (n 1 − 1) s X 1 2 + (n 2 − 1) s X 2 2 n 1 + n 2 − 2 {\displaystyle s_{p}={\sqrt {\frac {(n_{1}-1)s_{X_{1}}^{2}+(n_{2}-1)s_{X_{2}}^{2}}{n_{1}+n_{2}-2}}}} is the pooled standard deviation of the two samples: it is defined in this way so that its square is an unbiased estimator of the common variance, whether or not the population means are the same. In these formulae, ni − 1 is the number of degrees of freedom for each group, and the total sample size minus two (that is, n1 + n2 − 2) is the total number of degrees of freedom, which is used in significance testing. The minimum detectable effect (MDE) is:[25] δ ≥ 2 S p 2 n (t 1 − α , ν + t 1 − β , ν) {\displaystyle \delta \geq {\sqrt {\frac {2S_{p}^{2}}{n}}}(t_{1-\alpha ,\nu }+t_{1-\beta ,\nu })} Main article: Welch's t-test This test, also known as Welch's t-test, is used only when the two population variances are not assumed to be equal (the two sample sizes may or may not be equal) and hence must be estimated separately. The t statistic to test whether the population means are different is calculated as t = X ¯ 1 − X ¯ 2 s Δ ¯ , {\displaystyle t={\frac {{\bar {X}}_{1}-{\bar {X}}_{2}}{s_{\bar {\Delta }}}},} where s Δ ¯ = s 1 2 n 1 + s 2 2 n 2 . {\displaystyle s_{\bar {\Delta }}={\sqrt {{\frac {s_{1}^{2}}{n_{1}}}+{\frac {s_{2}^{2}}{n_{2}}}}}.} Here si2 is the unbiased estimator of the variance of each of the two samples with ni = number of participants in group i (i = 1 or 2). In this case (s Δ ¯) 2 {\displaystyle (s_{\bar {\Delta }})^{2}} is not a pooled variance. For use in significance testing, the distribution of the test statistic is approximated as an ordinary Student's t-distribution with the degrees of freedom calculated using d.f. = (s 1 2 n 1 + s 2 2 n 2) 2 (s 1 2 / n 1) 2 n 1 − 1 + (s 2 2 / n 2) 2 n 2 − 1 . {\displaystyle {\text{d.f.}}={\frac {\left({\frac {s_{1}^{2}}{n_{1}}}+{\frac {s_{2}^{2}}{n_{2}}}\right)^{2}}{{\frac {(s_{1}^{2}/n_{1})^{2}}{n_{1}-1}}+{\frac {(s_{2}^{2}/n_{2})^{2}}{n_{2}-1}}}}.} This is known as the Welch–Satterthwaite equation. The true distribution of the test statistic actually depends (slightly) on the two unknown population variances (see Behrens–Fisher problem). The test[26] deals with the famous Behrens–Fisher problem, i.e., comparing the difference between the means of two normally distributed populations when the variances of the two populations are not assumed to be equal, based on two independent samples. The test is developed as an exact test that allows for unequal sample sizes and unequal variances of two populations. The exact property still holds even with extremely small and unbalanced sample sizes (e.g. m ≡ n X = 50 {\displaystyle \ m\equiv n_{\mathsf {X}}=50\ } vs. n ≡ n Y = 5 {\displaystyle \ n\equiv n_{\mathsf {Y}}=5\ }). The statistic to test whether the means are different can be calculated as follows: Let X = [X 1 , X 2 , … , X m] ⊤ {\displaystyle \ X=\left[\ X_{1},X_{2},\ldots ,X_{m}\ \right]^{\top }\ } and Y = [Y 1 , Y 2 , … , Y n] ⊤ {\displaystyle \ Y=\left[\ Y_{1},Y_{2},\ldots ,Y_{n}\ \right]^{\top }\ } be the i.i.d. sample vectors (for m ≥ n {\displaystyle \ m\geq n\ }) from N o r m (μ X , σ X 2) {\displaystyle \ {\mathsf {Norm}}\left(\ \mu _{\mathsf {X}},\ \sigma _{\mathsf {X}}^{2}\ \right)\ } and N o r m (μ Y , σ Y 2) {\displaystyle \ {\mathsf {Norm}}\left(\ \mu _{\mathsf {Y}},\ \sigma _{\mathsf {Y}}^{2}\ \right)\ } separately. Let (P ⊤) n × n {\displaystyle \ (P^{\top })_{n\times n}\ } be an n × n {\displaystyle n\times n} orthogonal matrix whose elements of the first row are all 1 n , {\displaystyle \ {\tfrac {1}{\sqrt {n\ }}}\ ,} similarly, let (Q ⊤) n × m {\displaystyle \ (Q^{\top })_{n\times m}\ } be the first n {\displaystyle \ n\ } rows of an m × m {\displaystyle \ m\times m\ } orthogonal matrix (whose elements of the first row are all 1 m {\displaystyle \ {\tfrac {1}{\sqrt {m\ }}}\ }). Then Z ≡ (Q ⊤) n × m X m − (P ⊤) n × n Y n {\displaystyle \ Z\equiv {\frac {\ \left(Q^{\top }\right)_{n\times m}\ X\ }{\sqrt {m\ }}}\ -{\frac {\ \left(P^{\top }\right)_{n\times n}\ Y\ }{\sqrt {n\ }}}\ } is an n-dimensional normal random vector: Z ∼ N o r m ([μ X − μ Y , 0 , 0 , … , 0] ⊤ , (σ X 2 m + σ Y 2 n) I n) . {\displaystyle Z~\sim ~{\mathsf {Norm}}\left(\ \left[\ \mu _{\mathsf {X}}-\mu _{\mathsf {Y}},\ 0,\ 0,\ \ldots ,\ 0\ \right]^{\top }\ ,\ \left({\frac {\ \sigma _{\mathsf {X}}^{2}\ }{m}}+{\frac {\ \sigma _{\mathsf {Y}}^{2}\ }{n}}\right)\ I_{n}\ \right)~.} From the above distribution we see that the first element of the vector Z is Z 1 = X ¯ − Y ¯ = 1 m ∑ i = 1 m X i − 1 n ∑ j = 1 n Y j , {\displaystyle Z_{1}={\bar {X}}-{\bar {Y}}={\frac {1}{\ m\ }}\sum _{i=1}^{m}\ X_{i}-{\frac {1}{\ n\ }}\sum _{j=1}^{n}\ Y_{j}\ ,} hence the first element is distributed as Z 1 − (μ X − μ Y) ∼ N o r m (0 , σ X 2 m + σ Y 2 n) , {\displaystyle Z_{1}-\left(\mu _{\mathsf {X}}-\mu _{\mathsf {Y}}\right)~\sim ~{\mathsf {Norm}}\left(\ 0,\ {\frac {\ \sigma _{\mathsf {X}}^{2}\ }{m}}+{\frac {\ \sigma _{\mathsf {Y}}^{2}\ }{n}}\ \right)\ ,} and the squares of the remaining elements of Z are chi-squared distributed ∑ i = 2 n Z i 2 n − 1 ∼ χ n − 1 2 n − 1 × (σ X 2 m + σ Y 2 n) {\displaystyle {\frac {\ \sum _{i=2}^{n}Z_{i}^{2}\ }{\ n-1\ }}~\sim ~{\frac {\ \chi _{n-1}^{2}\ }{\ n-1\ }}\times \left({\frac {\ \sigma _{\mathsf {X}}^{2}\ }{m}}+{\frac {\ \sigma _{\mathsf {Y}}^{2}\ }{n}}\right)} and by construction of the orthogonal matricies P and Q we have Z 1 − (μ X − μ Y) ⊥ ∑ i = 2 n Z i 2 , {\displaystyle Z_{1}-\left(\mu _{\mathsf {X}}-\mu _{\mathsf {Y}}\right)\quad \perp \quad \sum _{i=2}^{n}Z_{i}^{2}\ ,} so Z1, the first element of Z, is statistically independent of the remaining elements by orthogonality. Finally, take for the test statistic T e ≡ Z 1 − (μ X − μ Y) (∑ i = 2 n Z i 2) / (n − 1) ∼ t n − 1 . {\displaystyle T_{\mathsf {e}}~\equiv ~{\frac {\ Z_{1}-\left(\mu _{\mathsf {X}}-\mu _{\mathsf {Y}}\right)\ }{\ {\sqrt {\left(\sum _{i=2}^{n}Z_{i}^{2}\right)/\left(n-1\right)\ }}\ }}~\sim ~t_{n-1}~.} This test is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired". This is an example of a paired difference test. The t statistic is calculated as t = X ¯ D − μ 0 s D / n , {\displaystyle t={\frac {{\bar {X}}_{D}-\mu _{0}}{s_{D}/{\sqrt {n}}}},} where X ¯ D {\displaystyle {\bar {X}}_{D}} and s D {\displaystyle s_{D}} are the average and standard deviation of the differences between all pairs. The pairs are e.g. either one person's pre-test and post-test scores or between-pairs of persons matched into meaningful groups (for instance, drawn from the same family or age group: see table). The constant μ0 is zero if we want to test whether the average of the difference is significantly different. The degree of freedom used is n − 1, where n represents the number of pairs. Example of matched pairs Pair Name Age Test 1 John 35 250 1 Jane 36 340 2 Jimmy 22 460 2 Jessy 21 200 Example of repeated measures Number Name Test 1 Test 2 1 Mike 35% 67% 2 Melanie 50% 46% 3 Melissa 90% 86% 4 Mitchell 78% 91% This article may need to summarize its corresponding main article in better quality. Please help out to edit this article and make improvements to the summary. (Learn how and when to remove this message) Let A1 denote a set obtained by drawing a random sample of six measurements: A 1 = { 30.02 , 29.99 , 30.11 , 29.97 , 30.01 , 29.99 } {\displaystyle A_{1}=\{30.02,\ 29.99,\ 30.11,\ 29.97,\ 30.01,\ 29.99\}} and let A2 denote a second set obtained similarly: A 2 = { 29.89 , 29.93 , 29.72 , 29.98 , 30.02 , 29.98 } {\displaystyle A_{2}=\{29.89,\ 29.93,\ 29.72,\ 29.98,\ 30.02,\ 29.98\}} These could be, for example, the weights of screws that were manufactured by two different machines. We will carry out tests of the null hypothesis that the means of the populations from which the two samples were taken are equal. The difference between the two sample means, each denoted by Xi, which appears in the numerator for all the two-sample testing approaches discussed above, is X ¯ 1 − X ¯ 2 = 0.095. {\displaystyle {\bar {X}}_{1}-{\bar {X}}_{2}=0.095.} The sample standard deviations for the two samples are approximately 0.05 and 0.11, respectively. For such small samples, a test of equality between the two population variances would not be very powerful. Since the sample sizes are equal, the two forms of the two-sample t-test will perform similarly in this example. If the approach for unequal variances (discussed above) is followed, the results are s 1 2 n 1 + s 2 2 n 2 ≈ 0.04849 {\displaystyle {\sqrt {{\frac {s_{1}^{2}}{n_{1}}}+{\frac {s_{2}^{2}}{n_{2}}}}}\approx 0.04849} and the degrees of freedom d.f. ≈ 7.031. {\displaystyle {\text{d.f.}}\approx 7.031.} The test statistic is approximately 1.959, which gives a two-tailed test p-value of 0.09077. If the approach for equal variances (discussed above) is followed, the results are s p ≈ 0.08399 {\displaystyle s_{p}\approx 0.08399} and the degrees of freedom d.f. = 10. {\displaystyle {\text{d.f.}}=10.} The test statistic is approximately equal to 1.959, which gives a two-tailed p-value of 0.07857. The t-test provides an exact test for the equality of the means of two i.i.d. normal populations with unknown, but equal, variances. (Welch's t-test is a nearly exact test for the case where the data are normal but the variances may differ.) For moderately large samples and a one tailed test, the t-test is relatively robust to moderate violations of the normality assumption.[27] In large enough samples, the t-test asymptotically approaches the z-test, and becomes robust even to large deviations from normality.[19] If the data are substantially non-normal and the sample size is small, the t-test can give misleading results. See Location test for Gaussian scale mixture distributions for some theory related to one particular family of non-normal distributions. When the normality assumption does not hold, a non-parametric alternative to the t-test may have better statistical power. However, when data are non-normal with differing variances between groups, a t-test may have better type-1 error control than some non-parametric alternatives.[28] Furthermore, non-parametric methods, such as the Mann-Whitney U test discussed below, typically do not test for a difference of means, so should be used carefully if a difference of means is of primary scientific interest.[19] For example, Mann-Whitney U test will keep the type 1 error at the desired level alpha if both groups have the same distribution. It will also have power in detecting an alternative by which group B has the same distribution as A but after some shift by a constant (in which case there would indeed be a difference in the means of the two groups). However, there could be cases where group A and B will have different distributions but with the same means (such as two distributions, one with positive skewness and the other with a negative one, but shifted so to have the same means). In such cases, MW could have more than alpha level power in rejecting the Null hypothesis but attributing the interpretation of difference in means to such a result would be incorrect. In the presence of an outlier, the t-test is not robust. For example, for two independent samples when the data distributions are asymmetric (that is, the distributions are skewed) or the distributions have large tails, then the Wilcoxon rank-sum test (also known as the Mann–Whitney U test) can have three to four times higher power than the t-test.[27][29][30] The nonparametric counterpart to the paired samples t-test is the Wilcoxon signed-rank test for paired samples. For a discussion on choosing between the t-test and nonparametric alternatives, see Lumley, et al. (2002).[19] One-way analysis of variance (ANOVA) generalizes the two-sample t-test when the data belong to more than two groups. When both paired observations and independent observations are present in the two sample design, assuming data are missing completely at random (MCAR), the paired observations or independent observations may be discarded in order to proceed with the standard tests above. Alternatively making use of all of the available data, assuming normality and MCAR, the generalized partially overlapping samples t-test could be used.[31] Main article: Hotelling's T-squared distribution A generalization of Student's t statistic, called Hotelling's t-squared statistic, allows for the testing of hypotheses on multiple (often correlated) measures within the same sample. For instance, a researcher might submit a number of subjects to a personality test consisting of multiple personality scales (e.g. the Minnesota Multiphasic Personality Inventory). Because measures of this type are usually positively correlated, it is not advisable to conduct separate univariate t-tests to test hypotheses, as these would neglect the covariance among measures and inflate the chance of falsely rejecting at least one hypothesis (Type I error). In this case a single multivariate test is preferable for hypothesis testing. Fisher's Method for combining multiple tests with alpha reduced for positive correlation among tests is one. Another is Hotelling's T2 statistic follows a T2 distribution. However, in practice the distribution is rarely used, since tabulated values for T2 are hard to find. Usually, T2 is converted instead to an F statistic. For a one-sample multivariate test, the hypothesis is that the mean vector (μ) is equal to a given vector (μ0). The test statistic is Hotelling's t2: t 2 = n (x ¯ − μ 0) ′ S − 1 (x ¯ − μ 0) {\displaystyle t^{2}=n({\bar {\mathbf {x} }}-{{\boldsymbol {\mu }}_{0}})'{\mathbf {S} }^{-1}({\bar {\mathbf {x} }}-{{\boldsymbol {\mu }}_{0}})} where n is the sample size, x is the vector of column means and S is an m × m sample covariance matrix. For a two-sample multivariate test, the hypothesis is that the mean vectors (μ1, μ2) of two samples are equal. The test statistic is Hotelling's two-sample t2: t 2 = n 1 n 2 n 1 + n 2 (x ¯ 1 − x ¯ 2) ′ S pooled − 1 (x ¯ 1 − x ¯ 2) . {\displaystyle t^{2}={\frac {n_{1}n_{2}}{n_{1}+n_{2}}}\left({\bar {\mathbf {x} }}_{1}-{\bar {\mathbf {x} }}_{2}\right)'{\mathbf {S} _{\text{pooled}}}^{-1}\left({\bar {\mathbf {x} }}_{1}-{\bar {\mathbf {x} }}_{2}\right).} This section does not cite any sources. Please help improve this section by adding citations to reliable sources. Unsourced material may be challenged and removed. (March 2025) (Learn how and when to remove this message) The two-sample t-test is a special case of simple linear regression as illustrated by the following example. A clinical trial examines 6 patients given drug or placebo. Three (3) patients get 0 units of drug (the placebo group). Three (3) patients get 1 unit of drug (the active treatment group). At the end of treatment, the researchers measure the change from baseline in the number of words that each patient can recall in a memory test. A table of the patients' word recall and drug dose values are shown below. Patient drug.dose word.recall 1 0 1 2 0 2 3 0 3 4 1 5 5 1 6 6 1 7 Data and code are given for the analysis using the R programming language with the t.test and lmfunctions for the t-test and linear regression. Here are the same (fictitious) data above generated in R. > word.recall.data=data.frame(drug.dose=c(0,0,0,1,1,1), word.recall=c(1,2,3,5,6,7)) Perform the t-test. Notice that the assumption of equal variance, var.equal=T, is required to make the analysis exactly equivalent to simple linear regression. > with(word.recall.data, t.test(word.recall~drug.dose, var.equal=T)) Running the R code gives the following results. The mean word.recall in the 0 drug.dose group is 2. The mean word.recall in the 1 drug.dose group is 6. The difference between treatment groups in the mean word.recall is 6 – 2 = 4. The difference in word.recall between drug doses is significant (p=0.00805). Perform a linear regression of the same data. Calculations may be performed using the R function lm() for a linear model. > word.recall.data.lm = lm(word.recall~drug.dose, data=word.recall.data) > summary(word.recall.data.lm) The linear regression provides a table of coefficients and p-values. Coefficient Estimate Std. Error t value P-value Intercept 2 0.5774 3.464 0.02572 drug.dose 4 0.8165 4.899 0.000805 The table of coefficients gives the following results. The estimate value of 2 for the intercept is the mean value of the word recall when the drug dose is 0. The estimate value of 4 for the drug dose indicates that for a 1-unit change in drug dose (from 0 to 1) there is a 4-unit change in mean word recall (from 2 to 6). This is the slope of the line joining the two group means. The p-value that the slope of 4 is different from 0 is p = 0.00805. The coefficients for the linear regression specify the slope and intercept of the line that joins the two group means, as illustrated in the graph. The intercept is 2 and the slope is 4. Compare the result from the linear regression to the result from the t-test. From the t-test, the difference between the group means is 6-2=4. From the regression, the slope is also 4 indicating that a 1-unit change in drug dose (from 0 to 1) gives a 4-unit change in mean word recall (from 2 to 6). The t-test p-value for the difference in means, and the regression p-value for the slope, are both 0.00805. The methods give identical results. This example shows that, for the special case of a simple linear regression where there is a single x-variable that has values 0 and 1, the t-test gives the same results as the linear regression. The relationship can also be shown algebraically. Recognizing this relationship between the t-test and linear regression facilitates the use of multiple linear regression and multi-way analysis of variance. These alternatives to t-tests allow for the inclusion of additional explanatory variables that are associated with the response. Including such additional explanatory variables using regression or anova reduces the otherwise unexplained variance, and commonly yields greater power to detect differences than do two-sample t-tests. Many spreadsheet programs and statistics packages, such as QtiPlot, LibreOffice Calc, Microsoft Excel, SAS, SPSS, Stata, DAP, gretl, R, Python, PSPP, Wolfram Mathematica, MATLAB and Minitab, include implementations of Student's t-test. Language/Program Function Notes Microsoft Excel pre 2010 TTEST(array1, array2, tails, type) See [1] Microsoft Excel 2010 and later T.TEST(array1, array2, tails, type) See [2] Apple Numbers TTEST(sample-1-values, sample-2-values, tails, test-type) See [3] LibreOffice Calc TTEST(Data1; Data2; Mode; Type) See [4] Google Sheets TTEST(range1, range2, tails, type) See [5] Python scipy.stats.ttest_ind(a, b, equal_var=True) See [6] MATLAB ttest(data1, data2) See [7] Mathematica TTest[{data1,data2}] See [8] R t.test(data1, data2, var.equal=TRUE) See [9] SAS PROC TTEST See [10] Java tTest(sample1, sample2) See [11] Julia EqualVarianceTTest(sample1, sample2) See [12] Stata ttest data1 == data2 See [13] Mathematics portal Conditional change model Equivalence test – Tool used to draw statistical inferences from observed data F-test – Statistical hypothesis test, mostly using multiple restrictions Noncentral t-distribution in power analysis – Probability distribution Student's t-statistic – Ratio in statisticsPages displaying short descriptions of redirect targets Z-test – Statistical test Mann–Whitney U test – Nonparametric test of the null hypothesis Šidák correction for t-test – Statistical method Welch's t-test – Statistical test of whether two populations have equal means Analysis of variance – Collection of statistical models (ANOVA) t-distribution – Probability distribution Confidence intervals for the mean of a normal distribution (also here)

^ The Microbiome in Health and Disease. Academic Press. 2020-05-29. p. 397. ISBN 978-0-12-820001-8. ^ Szabó, István (2003). "Systeme aus einer endlichen Anzahl starrer Körper". Einführung in die Technische Mechanik (in German). Springer Berlin Heidelberg. pp. 196–199. doi:10.1007/978-3-642-61925-0_16 (inactive 1 November 2024). ISBN 978-3-540-13293-6.{{cite book}}: CS1 maint: DOI inactive as of November 2024 (link) ^ Schlyvitch, B. (October 1937). "Untersuchungen über den anastomotischen Kanal zwischen der Arteria coeliaca und mesenterica superior und damit in Zusammenhang stehende Fragen". Zeitschrift für Anatomie und Entwicklungsgeschichte (in German). 107 (6): 709–737. doi:10.1007/bf02118337. ISSN 0340-2061. S2CID 27311567. ^ Helmert (1876). "Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehlers directer Beobachtungen gleicher Genauigkeit". Astronomische Nachrichten (in German). 88 (8–9): 113–131. Bibcode:1876AN.....88..113H. doi:10.1002/asna.18760880802. ^ Lüroth, J. (1876). "Vergleichung von zwei Werthen des wahrscheinlichen Fehlers". Astronomische Nachrichten (in German). 87 (14): 209–220. Bibcode:1876AN.....87..209L. doi:10.1002/asna.18760871402. ^ Pfanzagl, J. (1996). "Studies in the history of probability and statistics XLIV. A forerunner of the t-distribution". Biometrika. 83 (4): 891–898. doi:10.1093/biomet/83.4.891. MR 1766040. ^ Sheynin, Oscar (1995). "Helmert's work in the theory of errors". Archive for History of Exact Sciences. 49 (1): 73–104. doi:10.1007/BF00374700. ISSN 0003-9519. S2CID 121241599. ^ Pearson, Karl (1895). "X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material". Philosophical Transactions of the Royal Society of London A. 186: 343–414. Bibcode:1895RSPTA.186..343P. doi:10.1098/rsta.1895.0010. ^ a b Student (1908). "The Probable Error of a Mean" (PDF). Biometrika. 6 (1): 1–25. doi:10.1093/biomet/6.1.1. hdl:10338.dmlcz/143545. Retrieved 24 July 2016. ^ "T Table". ^ Wendl, Michael C. (2016). "Pseudonymous fame". Science. 351 (6280): 1406. doi:10.1126/science.351.6280.1406. PMID 27013722. ^ Walpole, Ronald E. (2006). Probability & statistics for engineers & scientists. Myers, H. Raymond (7th ed.). New Delhi: Pearson. ISBN 81-7758-404-9. OCLC 818811849. ^ Raju, T. N. (2005). "William Sealy Gosset and William A. Silverman: Two 'Students' of Science". Pediatrics. 116 (3): 732–735. doi:10.1542/peds.2005-1134. PMID 16140715. S2CID 32745754. ^ Dodge, Yadolah (2008). The Concise Encyclopedia of Statistics. Springer Science & Business Media. pp. 234–235. ISBN 978-0-387-31742-7. ^ Fadem, Barbara (2008). High-Yield Behavioral Science. High-Yield Series. Hagerstown, MD: Lippincott Williams & Wilkins. ISBN 9781451130300. ^ Rice, John A. (2006). Mathematical Statistics and Data Analysis (3rd ed.). Duxbury Advanced.[ISBN missing] ^ Weisstein, Eric. "Student's t-Distribution". mathworld.wolfram.com. ^ David, H. A.; Gunnink, Jason L. (1997). "The Paired t Test Under Artificial Pairing". The American Statistician. 51 (1): 9–12. doi:10.2307/2684684. JSTOR 2684684. ^ a b c d Lumley, Thomas; Diehr, Paula; Emerson, Scott; Chen, Lu (May 2002). "The Importance of the Normality Assumption in Large Public Health Data Sets". Annual Review of Public Health. 23 (1): 151–169. doi:10.1146/annurev.publhealth.23.100901.140546. ISSN 0163-7525. PMID 11910059. ^ Markowski, Carol A.; Markowski, Edward P. (1990). "Conditions for the Effectiveness of a Preliminary Test of Variance". The American Statistician. 44 (4): 322–326. doi:10.2307/2684360. JSTOR 2684360. ^ Guo, Beibei; Yuan, Ying (2017). "A comparative review of methods for comparing means using partially paired data". Statistical Methods in Medical Research. 26 (3): 1323–1340. doi:10.1177/0962280215577111. PMID 25834090. S2CID 46598415. ^ Bland, Martin (1995). An Introduction to Medical Statistics. Oxford University Press. p. 168. ISBN 978-0-19-262428-4. ^ "Central limit theorem and the normality assumption > Normality > Continuous distributions > Distribution > Statistical Reference Guide | Analyse-it® 6.15 documentation". analyse-it.com. Retrieved 2024-05-17. ^ DEMİR, Süleyman (2022-06-26). "Comparison of Normality Tests in Terms of Sample Sizes under Different Skewness and Kurtosis Coefficients". International Journal of Assessment Tools in Education. 9 (2): 397–409. doi:10.21449/ijate.1101295. ISSN 2148-7456. ^ Minimum Detectable Difference for Two-Sample t-Test for Means. Equation and example adapted from Zar, 1984 ^ Wang, Chang; Jia, Jinzhu (2022). "Te Test: A New Non-asymptotic T-test for Behrens-Fisher Problems". arXiv:2210.16473 [math.ST]. ^ a b Sawilowsky, Shlomo S.; Blair, R. Clifford (1992). "A More Realistic Look at the Robustness and Type II Error Properties of the t Test to Departures From Population Normality". Psychological Bulletin. 111 (2): 352–360. doi:10.1037/0033-2909.111.2.352. ^ Zimmerman, Donald W. (January 1998). "Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions". The Journal of Experimental Education. 67 (1): 55–68. doi:10.1080/00220979809598344. ISSN 0022-0973. ^ Blair, R. Clifford; Higgins, James J. (1980). "A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to That of Student's t Statistic Under Various Nonnormal Distributions". Journal of Educational Statistics. 5 (4): 309–335. doi:10.2307/1164905. JSTOR 1164905. ^ Fay, Michael P.; Proschan, Michael A. (2010). "Wilcoxon–Mann–Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules". Statistics Surveys. 4: 1–39. doi:10.1214/09-SS051. PMC 2857732. PMID 20414472. ^ Derrick, B; Toher, D; White, P (2017). "How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)" (PDF). The Quantitative Methods for Psychology. 13 (2): 120–126. doi:10.20982/tqmp.13.2.p120. O'Mahony, Michael (1986). Sensory Evaluation of Food: Statistical Methods and Procedures. CRC Press. p. 487. ISBN 0-82477337-3. Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, Brian P. (1992). Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press. p. 616. ISBN 0-521-43108-5. Boneau, C. Alan (1960). "The effects of violations of assumptions underlying the t test". Psychological Bulletin. 57 (1): 49–64. doi:10.1037/h0041412. PMID 13802482. Edgell, Stephen E.; Noon, Sheila M. (1984). "Effect of violation of normality on the t test of the correlation coefficient". Psychological Bulletin. 95 (3): 576–583. doi:10.1037/0033-2909.95.3.576. Wikiversity has learning resources about t-test Wikisource has original text related to this article: The Probable Error of a Mean "Student test". Encyclopedia of Mathematics. EMS Press. 2001 [1994]. Trochim, William M.K. "The T-Test", Research Methods Knowledge Base, conjoint.ly Econometrics lecture (topic: hypothesis testing) on YouTube by Mark Thoma Retrieved from "

The t test is one of the simplest statistical techniques that is used to evaluate whether there is a statistical difference between the means from up to two different samples. The t test is especially useful when you have a small number of sample observations (under 30 or so), and you want to make conclusions about the larger population.The characteristics of the data dictate the appropriate type of t test to run. All t tests are used as standalone analyses for very simple experiments and research questions as well as to perform individual tests within more complicated statistical models such as linear regression.In this guide, we'll lay out everything you need to know about t tests, including providing a simple workflow to determine what t test is appropriate for your particular data or if you'd be better suited using a different model.What is a t test?A t test is a statistical technique used to quantify the difference between the mean (average value) of a variable from up to two samples (datasets). The variable must be numeric. Some examples are height, gross income, and amount of weight lost on a particular diet.A t test tells you if the difference you observe is "surprising" based on the expected difference. They use t-distributions to evaluate the expected variability. When you have a reasonable-sized sample (over 30 or so observations), the t test can still be used, but other tests that use the normal distribution (the z test) can be used in its place.Sometimes t tests are called "Student's" t tests, which is simply a reference to their unusual history. It got its name because a brewer from the Guinness Brewery, William Gosset, published about the method under the pseudonym "Student". He wanted to get information out of very small sample sizes (often 3-5) because it took so much effort to brew each keg for his samples.When should I use a t test?A t test is appropriate to use when you've collected a small, random sample from some statistical "population" and want to compare the mean from your sample to another value. The value for comparison could be a fixed value (e.g., 10) or the mean of a second sample. For example, if your variable of interest is the average height of sixth graders in your region, then you might measure the height of 25 or 30 randomly-selected sixth graders. A t test could be used to answer questions such as, "Is the average height greater than four feet?"How does a t test work?Based on your experiment, t tests make enough assumptions about your experiment to calculate an expected variability, and then they use that to determine if the observed data is statistically significant. To do this, t tests rely on an assumed "null hypothesis." With the above example, the null hypothesis is that the average height is less than or equal to four feet.Say that we measure the height of 5 randomly selected sixth graders and the average height is five feet. Does that mean that the "true" average height of all sixth graders is greater than four feet or did we randomly happen to measure taller than average students?To evaluate this, we need a distribution that shows every possible average value resulting from a sample of five individuals in a population where the true mean is four. That may seem impossible to do, which is why there are particular assumptions that need to be made to perform a t test.With those assumptions, then all that's needed to determine the "sampling distribution of the mean" is the sample size (5 students in this case) and standard deviation of the data (let's say it's 1 foot).That's enough to create a graphic of the distribution of the mean, which is:Notice the vertical line at x = 5, which was our sample mean. We (use software to) calculate the area to the right of the vertical line, which gives us the P value (0.09 in this case). Note that because our research question was asking if the average student is greater than four feet, the distribution is centered at four. Since we're only interested in knowing if the average is greater than four feet, we use a one-tailed test in this case.Using the standard confidence level of 0.05 with this example, we don't have evidence that the true average height of sixth graders is taller than 4 feet.What are the assumptions for t tests?One variable of interest: This is not correlation or regression, where you are interested in the relationship between multiple variables. With a t test, you can have different samples, but they are all measuring the same variable (e.g., height).Numeric data: You are dealing with a list of measurements that can be averaged. This means you aren't just counting occurrences in various categories (e.g., eye color or political affiliation).Two groups or less: If you have more than two samples of data, a t test is the wrong technique. You most likely need to try ANOVA.Random sample: You need a random sample from your statistical "population of interest" in order to draw valid conclusions about the larger population. If your population is so small that you can measure everything, then you have a "census" and don't need statistics. This is because you don't need to estimate the truth, since you have measured the truth without variability.Normally Distributed: The smaller your sample size, the more important it is that your data come from a normal, Gaussian distribution bell curve. If you have reason to believe that your data are not normally distributed, consider nonparametric t test alternatives. This isn't necessary for larger samples (usually 25 or 30 unless the data is heavily skewed). The reason is that the Central Limit Theorem applies in this case, which says that even if the distribution of your data is not normal, the distribution of the mean of your data is, so you can use a z-test rather than a t test.How do I know which t test to use?There are many types of t tests to choose from, but you don't necessarily have to understand every detail behind each option. You just need to be able to answer a few questions, which will lead you to pick the right t test. To that end, we put together this workflow for you to figure out which test is appropriate for your data.Do you have one or two samples?Are you comparing the means of two different samples, or comparing the mean from one sample to a fixed value? An example research question is, "Is the average height of my sample of sixth grade students greater than four feet?" If you only have one sample of data, you can click here to skip to a one-sample t test example, otherwise your next step is to ask: Are observations in the two samples matched up or related in some way?This could be as before-and-after measurements of the same exact subjects, or perhaps your study split up "pairs" of subjects (who are technically different but share certain characteristics of interest) into the two samples. The same variable is measured in both cases.If so, you are looking at some kind of paired samples t test. The linked section will help you dial in exactly which one in that family is best for you, either difference (most common) or ratio. If you aren't sure paired is right, ask yourself another question:Are you comparing different observations in each of the two samples?If the answer is yes, then you have an unpaired or independent samples t test. The two samples should measure the same variable (e.g., height), but are samples from two distinct groups (e.g., team A and team B). The goal is to compare the means to see if the groups are significantly different. For example, "Is the average height of team A greater than team B?" Unlike paired, the only relationship between the groups in this case is that we measured the same variable for both. There are two versions of unpaired samples t tests (pooled and unpooled) depending on whether you assume the same variance for each sample.Have you run the same experiment multiple times on the same subject/observational unit?If so, then you have a nested t test (unless you have more than two sample groups). This is a trickier concept to understand. One example is if you are measuring how well Fertilizer A works against Fertilizer B. Let's say you have 12 pots to grow plants in (6 pots for each fertilizer), and you grow 3 plants in each pot.In this case you have 6 observational units for each fertilizer, with 3 subsamples from each pot. You would want to analyze this with a nested t test. The "nested" factor in this case is the pots. It's important to note that we aren't interested in estimating the variability within each pot, we just want to take it into account.You might be tempted to run an unpaired samples t test here, but that assumes you have 6*3 = 18 replicates for each fertilizer. However, the three replicates within each pot are related, and an unpaired samples t test wouldn't take that into account.What if none of these sound like my experiment?If you're not seeing your research question above, note that t tests are very basic statistical tools. Many experiments require more sophisticated techniques to evaluate differences. If the variable of interest is a proportion (e.g., 10 of 100 manufactured products were defective), then you'd use z-tests. If you take before and after measurements and have more than one treatment (e.g., control vs a treatment diet), then you need ANOVA.How do I perform a t test using software?If you're wondering how to do a t test, the easiest way is with statistical software such as Prism or an online t test calculator.If you're using software, then all you need to know is which t test is appropriate (use the workflow here) and understand how to interpret the output. To do that, you'll also need to:Determine whether your test is one or two-tailedChoose the level of significanceIs my test one or two-tailed? Whether or not you have a one- or two-tailed test depends on your research hypothesis. Choosing the appropriately tailed test is very important and requires integrity from the researcher. This is because you have more "power" with one-tailed tests, meaning that you can detect a statistically significant difference more easily. Unless you have written out your research hypothesis as one directional before you run your experiment, you should use a two-tailed test.Two-tailed testsTwo-tailed tests are the most common, and they are applicable when your research question is simply asking, "is there a difference?"One-tailed testsContrast that with one-tailed tests, where the research questions are directional, meaning that either the question is, "is it greater than" or the question is, "is it less than". These tests can only detect a difference in one direction.Choosing the level of significanceAll t tests estimate whether a mean of a population is different than some other value, and with all estimates come some variability, or what statisticians call "error." Before analyzing your data, you want to choose a level of significance, usually denoted by the Greek letter alpha, . The scientific standard is setting alpha to be 0.05.An alpha of 0.05 results in 95% confidence intervals, and determines the cutoff for when P values are considered statistically significant.One sample t testIf you only have one sample of a list of numbers, you are doing a one-sample t test. All you are interested in doing is comparing the mean from this group with some known value to test if there is evidence, that it is significantly different from that standard. Use our free one-sample t test calculator for this.A one sample t test example research question is, "Is the average fifth grader taller than four feet?"It is the simplest version of a t test, and has all sorts of applications within hypothesis testing. Sometimes the "known value" is called the "null value". While the null value in t tests is often 0, it could be any value. The name comes from being the value which exactly represents the null hypothesis, where no significant difference exists. Any time you know the exact number you are trying to compare your sample of data against, this could work well. And of course: it can be either one or two-tailed.One sample t test formulaStatistical software handles this for you, but if you want the details, the formula for a one sample t test is:M: Calculated mean of your sampleμ: Hypothetical mean you are testing againsts: The standard deviation of your samplen: The number of observations in your sample.In a one-sample t test, calculating degrees of freedom is simple: one less than the number of objects in your dataset (you'll see it written as n-1).Example of a one sample t testFor our example within Prism, we have a dataset of 12 values from an experiment labeled "% of control". Perhaps these are heights of a sample of plants that have been treated with a new fertilizer. A value of 100 represents the industry-standard control height. Likewise, 123 represents a plant with a height 123% that of the control (that is, 23% larger).We'll perform a two-tailed, one-sample t test to see if plants are shorter or taller on average with the fertilizer. We will use a significance threshold of 0.05. Here is the output:You can see in the output that the actual sample mean was 111. Is that different enough from the industry standard (100) to conclude that there is a statistical difference?The quick answer is yes, there's strong evidence that the height of the plants with the fertilizer is greater than the industry standard (p=0.015). The nice thing about using software is that it handles some of the trickier steps for you. In this case, it calculates your test statistic (t=2.88), determines the appropriate degrees of freedom (11), and outputs a P value.More informative than the P value is the confidence interval of the difference, which is 2.49 to 18.7. The confidence interval tells us that, based on our data, we are confident that the true difference between our sample and the baseline value of 100 is somewhere between 2.49 and 18.7. As long as the difference is statistically significant, the interval will not contain zero.You can follow these tips for interpreting your own one-sample test.Graphing a one-sample t testFor some techniques (like regression), graphing the data is a very helpful part of the analysis. For t tests, making a chart of your data is still useful to spot any strange patterns or outliers, but the small sample size means you may already be familiar with any strange things in your data.Here we have a simple plot of the data points, perhaps with a mark for the average. We've made this as an example, but the truth is that graphing is usually more visually telling for two-sample t tests than for just one sample.Two sample t testsThere are several kinds of two sample t tests, with the two main categories being paired and unpaired (independent) samples.Paired samples t testIn a paired samples t test, also called dependent samples t test, there are two samples of data, and each observation in one sample is "paired" with an observation in the second sample. The most common example is when measurements are taken on each subject before and after a treatment. A paired t test example research question is, "Is there a statistical difference between the average red blood cell counts before and after a treatment?"Having two samples that are closely related simplifies the analysis. Statistical software, such as this paired t test calculator, will simply take a difference between the two values, and then compare that difference to 0.In some (rare) situations, taking a difference between the pairs violates the assumptions of a t test, because the average difference changes based on the size of the before value (e.g., there's a larger difference between before and after when there were more to start with). In this case, instead of using a difference test, use a ratio of the before and after values, which is referred to as ratio t tests.Paired t test formulaThe formula for paired samples t test is:Md: Mean difference between the samplessd: The standard deviation of the differencesn: The number of differencesDegrees of freedom are the same as before. If you're studying for an exam, you can remember that the degrees of freedom are still n-1 (not n-2) because we are converting the data into a single column of differences rather than considering the two groups independently.Also note that the null value here is simply 0. There is no real reason to include "minus 0" in an equation other than to illustrate that we are still doing a hypothesis test. After you take the difference between the two means, you are comparing that difference to 0.ExampleFor our example data, we have five test subjects and have taken two measurements from each: before ("control") and after a treatment ("treated"). If we set alpha = 0.05 and perform a two-tailed test, we observe a statistically significant difference between the treated and control group (p=0.0160, t=4.01, df = 4). We are 95% confident that the true mean difference between the treated and control group is between 0.449 and 2.47.Graphing a paired t testThe significant result of the P value suggests evidence that the treatment had some effect, and we can also look at this graphically. The lines that connect the observations can help us spot a pattern, if it exists. In this case the lines show that all observations increased after treatment. While not all graphics are this straightforward, here it is very consistent with the outcome of the t test. Prism's estimation plot is even more helpful because it shows both the data (like above) and the confidence interval for the difference between means. You can easily see the evidence of significance since the confidence interval on the right does not contain zero.Here are some more graphing tips for paired t tests.Unpaired samples t testUnpaired samples t test, also called independent samples t test, is appropriate when you have two sample groups that aren't correlated with one another. A pharma example is testing a treatment group against a control group of different subjects. Compare that with a paired sample, which might be recording the same subjects before and after a treatment.With unpaired t tests, in addition to choosing your level of significance and a one or two tailed test, you need to determine whether or not to assume that the variances between the groups are the same or not. If you assume equal variances, then you can "pool" the calculation of the standard error between the two samples. Otherwise, the standard choice is Welch's t test which corrects for unequal variances. This choice affects the calculation of the test statistic and the power of the test, which is the test's sensitivity to detect statistical significance. It's best to choose whether or not you'll use a pooled or unpooled (Welch's) standard error before running your experiment, because the standard statistical test is notoriously problematic. See more details about unequal variances here.As long as you're using statistical software, such as this two-sample t test calculator, it's just as easy to calculate a test statistic whether or not you assume that the variances of your two samples are the same. If you're doing it by hand, however, the calculations get more complicated with unequal variances. Unpaired (independent) samples t test formulaThe general two-sample t test formula is:M1 and M2: Two means you are comparing, one from each datasetSE: The combined standard error of the two samples (calculated using pooled or unpooled standard error)The denominator (standard error) calculation can be complicated, as can the degrees of freedom. If the groups are not balanced (the same number of observations in each), you will need to account for both when determining n for the test as a whole.ExampleAs an example for this family, we conduct a paired samples t test assuming equal variances (pooled). Based on our research hypothesis, we'll conduct a two-tailed test, and use alpha=0.05 for our level of significance. Our samples were unbalanced, with two samples of 6 and 5 observations respectively. The P value (p=0.261, t = 1.20, df = 9) is higher than our threshold of 0.05. We have not found sufficient evidence to suggest a significant difference. You can see the confidence interval of the difference of the means is -9.58 to 31.2.Note that the F-test result shows that the variances of the two groups are not significantly different from each other.Graphing an unpaired samples t testFor an unpaired samples t test, graphing the data can quickly help you get a handle on the two groups and how similar or different they are. Like the paired example, this helps confirm the evidence (or lack thereof) that is found by doing the t test itself.Below you can see that the observed mean for females is higher than that for males. But because of the variability in the data, we can't tell if the means are actually different or if the difference is just by chance. Nonparametric alternatives for t testsIf your data comes from a normal distribution (or something close enough to a normal distribution), then a t test is valid. If that assumption is violated, you can use nonparametric alternatives.T tests evaluate whether the mean is different from another value, whereas nonparametric alternatives compare either the median or the rank. Medians are well-known to be much more robust to outliers than the mean.The downside to nonparametric tests is that they don't have as much statistical power, meaning a larger difference is required in order to determine that it's statistically significant.Wilcoxon signed-rank testThe Wilcoxon signed-rank test is the nonparametric cousin to the one-sample t test. This compares a sample median to a hypothetical median value. It is sometimes erroneously even called the Wilcoxon t test (even though it calculates a "W" statistic).And if you have two related samples, you should use the Wilcoxon matched pairs test instead. The two versions of Wilcoxon are different, and the matched pairs version is specifically for comparing the median difference for paired samples. Mann-Whitney and Kolmogorov-Smirnov testsFor unpaired (independent) samples, there are multiple options for nonparametric testing. Mann-Whitney is more popular and compares the mean ranks (the ordering of values from smallest to largest) of the two samples. Mann-Whitney is often misrepresented as a comparison of medians, but that's not always the case. Kolmogorov-Smirnov tests if the overall distributions differ between the two samples. More t test FAQsWhat is the formula for a t test?The exact formula depends on which type of t test you are running, although there is a basic structure that all t tests have in common. All t test statistics will have the form:t: The t test statistic you calculate for your testMean1 and Mean2: Two means you are comparing, at least 1 from your own datasetStandard Error of the Mean: The standard error of the mean, also called the standard deviation of the mean, which takes into account the variance and size of your datasetThe exact formula for any t test can be slightly different, particularly the calculation of the standard error. Not only does it matter whether one or two samples are being compared, the relationship between the samples can make a difference too.What is a t-distribution?A t-distribution is similar to a normal distribution. It's a bell-shaped curve, but compared to a normal it has fatter tails, which means that it's more common to observe extremes. T-distributions are identified by the number of degrees of freedom. The higher the number, the closer the t-distribution gets to a normal distribution. After about 30 degrees of freedom, a t and a standard normal are practically the same.t-distribution by degree of freedomWhat are degrees of freedom?Degrees of freedom are a measure of how large your dataset is. They aren't exactly the number of observations, because they also take into account the number of parameters (e.g., mean, variance) that you have estimated.What is the difference between paired vs unpaired t tests?Both paired and unpaired t tests involve two sample groups of data. With a paired t test, the values in each group are related (usually they are before and after values measured on the same test subject). In contrast, with unpaired t tests, the observed values aren't related between groups. An unpaired, or independent t test, example is comparing the average height of children at school A vs school B. When do I use a z-test versus a t test?Z-tests, which compare data using a normal distribution rather than a t-distribution, are primarily used for two situations. The first is when you're evaluating proportions (number of failures on an assembly line). The second is when your sample size is large enough (usually around 30) that you can use a normal approximation to evaluate the means.When should I use ANOVA instead of a t test?Use ANOVA if you have more than two group means to compare.What are the differences between t test vs chi square?Chi square tests are used to evaluate contingency tables, which record a count of the number of subjects that fall into particular categories (e.g., truck, SUV, car). t tests compare the mean(s) of a variable of interest (e.g., height, weight).What are P values?P values are the probability that you would get data as or more extreme than the observed data given that the null hypothesis is true. It's a mouthful, and there are a lot of issues to be aware of with P values.What are t test critical values?Critical values are a classical form (they aren't used directly with modern computing) of determining if a statistical test is significant or not. Historically you could calculate your test statistic from your data, and then use a t-table to look up the cutoff value (critical value) that represented a "significant" result. You would then compare your observed statistic against the critical value.How do I calculate degrees of freedom for my t test?In most practical usage, degrees of freedom are the number of observations you have minus the number of parameters you are trying to estimate. The calculation isn't always straightforward and is approximated for some t tests.Statistical software calculates degrees of freedom automatically as part of the analysis, so understanding them in more detail isn't needed beyond assuaging any curiosity.Perform your own t testAre you ready to calculate your own t test? Start your 30 day free trial of Prism and get access to:A step by step guide on how to perform a t testSample data to save you timeMore tips on how Prism can help your researchWith Prism, in a matter of minutes you learn how to go from entering data to performing statistical analyses and generating high-quality graphs.