4.2: Two-Sample t-Test

Learning Objectives

- To use Student's t-test for two samples when you have one measurement variable and one nominal variable, and the nominal variable has only two values. It tests whether the means of the measurement variable are different in the two groups.

Introduction

There are several statistical tests that use the t-distribution and can be called a t-test. One of the most common is Student's t-test for two samples. Other t-tests include the one-sample *t*–test, which compares a sample mean to a theoretical mean, and the paired *t*–test.

Student's t-test for two samples is mathematically identical to a one-way anova with two categories; because comparing the means of two samples is such a common experimental design, and because the t-test is familiar to many more people than anova, I treat the two-sample t-test separately.

When to use it

Use the two-sample t–test when you have one nominal variable and one measurement variable, and you want to compare the mean values of the measurement variable. The nominal variable must have only two values, such as "male" and "female" or "treated" and "untreated."

Null hypothesis

The statistical null hypothesis is that the means of the measurement variable are equal for the two categories.

How the test works

The test statistic, t_s, is calculated using a formula that has the difference between the means in the numerator; this makes t_s get larger as the means get further apart. The denominator is the standard error of the difference in the means, which gets smaller as the sample variances decrease or the sample sizes increase. Thus t_s gets larger as the means get farther apart, the variances get smaller, or the sample sizes increase.

You calculate the probability of getting the observed t_s value under the null hypothesis using the t-distribution. The shape of the t-distribution, and thus the probability of getting a particular t_s value, depends on the number of degrees of freedom. The degrees of freedom for a t-test is the total number of observations in the groups minus 2, or $n_1 + n_2 - 2$.

Assumptions

The t-test assumes that the observations within each group are normally distributed. Fortunately, it is not at all sensitive to deviations from this assumption, if the distributions of the two groups are the same (if both distributions are skewed to the right, for example). I've done simulations with a variety of non-normal distributions, including flat, bimodal, and highly skewed, and the two-sample t-test always gives about 5% false positives, even with very small sample sizes. If your data are severely non-normal, you should still try to find a data transformation that makes them more normal, but don't worry if you can't find a good transformation or don't have enough data to check the normality.

If your data are severely non-normal, *and* you have different distributions in the two groups (one data set is skewed to the right and the other is skewed to the left, for example), *and* you have small samples (less than 50 or so), then the two-sample t-test can give inaccurate results, with considerably more than 5% false positives. A data transformation won't help you here, and neither will a Mann-Whitney U-test. It would be pretty unusual in biology to have two groups with different distributions but equal means, but if you think that's a possibility, you should require a P value much less than 0.05 to reject the null hypothesis.

The two-sample t-test also assumes homoscedasticity (equal variances in the two groups). If you have a balanced design (equal sample sizes in the two groups), the test is not very sensitive to heteroscedasticity unless the sample size is very small (less than 10 or so); the standard deviations in one group can be several times as big as in the other group, and you'll get $P < 0.05$ about 5% of the time if the null hypothesis is true. With an unbalanced design, heteroscedasticity is a bigger problem; if the group with the smaller sample size has a bigger standard deviation, the two-sample t-test can give you false positives much too often. If your two

groups have standard deviations that are substantially different (such as one standard deviation is twice as big as the other), and your sample sizes are small (less than 10) or unequal, you should use Welch's *t*–test instead.

Example

In fall 2004, students in the $2p.m.$ section of my Biological Data Analysis class had an average height of 66.6 inches, while the average height in the $5p.m.$ section was 64.6 inches. Are the average heights of the two sections significantly different? Here are the data:

2 p.m.	5 p.m.
69	68
70	62
66	67
63	68
68	69
70	67
69	61
67	59
62	62
63	61
76	69
59	66
62	62
62	62
75	61
62	70
72	
63	

There is one measurement variable, height, and one nominal variable, class section. The null hypothesis is that the mean heights in the two sections are the same. The results of the t–test ($t = 1.29$, $32d.f.$, $P = 0.21$) do not reject the null hypothesis.

Graphing the results

Because it's just comparing two numbers, you'll rarely put the results of a t-test in a graph for publication. For a presentation, you could draw a bar graph like the one for a one-way anova.

Similar tests

Student's t-test is mathematically identical to a one-way anova done on data with two categories; you will get the exact same P value from a two-sample t-test and from a one-way anova, even though you calculate the test statistics differently. The t-test is easier to do and is familiar to more people, but it is limited to just two categories of data. You can do a one-way anova on two or more categories. I recommend that if your research always involves comparing just two means, you should call your test a two-sample t-test, because it is more familiar to more people. If you write a paper that includes some comparisons of two means and

some comparisons of more than two means, you may want to call all the tests one-way anovas, rather than switching back and forth between two different names (t-test and one-way anova) for the same thing.

The Mann-Whitney U-test is a non-parametric alternative to the two-sample t-test that some people recommend for non-normal data. However, if the two samples have the same distribution, the two-sample t-test is not sensitive to deviations from normality, so you can use the more powerful and more familiar t-test instead of the Mann-Whitney U-test. If the two samples have different distributions, the Mann-Whitney U-test is no better than the t-test. So there's really no reason to use the Mann-Whitney U-test unless you have a true ranked variable instead of a measurement variable.

If the variances are far from equal (one standard deviation is two or more times as big as the other) and your sample sizes are either small (less than 10) or unequal, you should use Welch's t-test (also know as Aspin-Welch, Welch-Satterthwaite, Aspin-Welch-Satterthwaite, or Satterthwaite t-test). It is similar to Student's t-test except that it does not assume that the standard deviations are equal. It is slightly less powerful than Student's t-test when the standard deviations are equal, but it can be much more accurate when the standard deviations are very unequal. My two-sample t-test spreadsheet will calculate Welch's t–test. You can also do Welch's t-test using this web page, by clicking the button labeled "Welch's unpaired t-test".

Use the paired t–test when the measurement observations come in pairs, such as comparing the strengths of the right arm with the strength of the left arm on a set of people.

Use the one-sample t–test when you have just one group, not two, and you are comparing the mean of the measurement variable for that group to a theoretical expectation.

How to do the test

Spreadsheets

I've set up a spreadsheet for two-sample t–tests twosamplettest.xls. It will perform either Student's t-test or Welch's t-test for up to 2000 observations in each group.

Web pages

There are web pages to do the t-test here and here. Both will do both the Student's t-test and Welch's t-test.

R

Salvatore Mangiafico's R *Companion* has a sample R programs for the two-sample t–test and Welch's test.

SAS

You can use PROC TTEST for Student's t-test; the CLASS parameter is the nominal variable, and the VAR parameter is the measurement variable. Here is an example program for the height data above.

```
DATA sectionheights;
INPUT section $ height @@;
DATALINES;
2pm 69 2pm 70 2pm 66 2pm 63 2pm 68 2pm 70 2pm 69
2pm 67 2pm 62 2pm 63 2pm 76 2pm 59 2pm 62 2pm 62
2pm 75 2pm 62 2pm 72 2pm 63
5pm 68 5pm 62 5pm 67 5pm 68 5pm 69 5pm 67 5pm 61
5pm 59 5pm 62 5pm 61 5pm 69 5pm 66 5pm 62 5pm 62
5pm 61 5pm 70
;
PROC TTEST;
CLASS section;
VAR height;
RUN;
```

The output includes a lot of information; the P value for the Student's t-test is under "Pr > |t| on the line labeled "Pooled", and the P value for Welch's t-test is on the line labeled "Satterthwaite." For these data, the P value is 0.2067 for Student's t-test and 0.1995 for Welch's.

Variable Method Variances DF t Value Pr > |t|

height Pooled Equal 32 1.29 0.2067

height Satterthwaite Unequal 31.2 1.31 0.1995

Power analysis

To estimate the sample sizes needed to detect a significant difference between two means, you need the following:

- the effect size, or the difference in means you hope to detect;
- the standard deviation. Usually you'll use the same value for each group, but if you know ahead of time that one group will have a larger standard deviation than the other, you can use different numbers;
- alpha, or the significance level (usually 0.05);
- beta, the probability of accepting the null hypothesis when it is false (0.50, 0.80, and 0.90 are common values);
- the ratio of one sample size to the other. The most powerful design is to have equal numbers in each group ($N_1/N_2 = 1.0$), but sometimes it's easier to get large numbers of one of the groups. For example, if you're comparing the bone strength in mice that have been reared in zero gravity aboard the International Space Station vs. control mice reared on earth, you might decide ahead of time to use three control mice for every one expensive space mouse ($N_1/N_2 = 3.0$).

The G*Power program will calculate the sample size needed for a two-sample t-test. Choose "t tests" from the "Test family" menu and "Means: Difference between two independent means (two groups" from the "Statistical test" menu. Click on the "Determine" button and enter the means and standard deviations you expect for each group. Only the difference between the group means is important; it is your effect size. Click on "Calculate and transfer to main window". Change "tails" to two, set your alpha (this will almost always be 0.05) and your power (0.5, 0.8, and 0.9 are commonly used). If you plan to have more observations in one group than in the other, you can make the "Allocation ratio" different from 1.

As an example, let's say you want to know whether people who run regularly have wider feet than people who don't run. You look for previously published data on foot width and find the ANSUR data set, which shows a mean foot width for American men of $100.6mm$ and a standard deviation of $5.26mm$. You decide that you'd like to be able to detect a difference of $3mm$ in mean foot width between runners and non-runners. Using G*Power, you enter $100mm$ for the mean of group 1, 103 for the mean of group 2, and $5.26$for the standard deviation of each group. You decide you want to detect a difference of $3mm$, at the $P < 0.05$ level, with a probability of detecting a difference this large, if it exists, of 90% ($1 - \text{beta} = 0.90$). Entering all these numbers in G*Power gives a sample size for each group of 66 people.
