Lecture 3

Various versions of the t-test

In this lecture, we will discuss various versions of the t-test for determining whether two means are equal or not. The theory behind the test was first developed by William Sealy Gosset who, working for Guiness brewery in Ireland, wrote under the pen name "Student" in order to safeguard Guiness's trade secrets. The pooled-variance t-test below is also called Student's t-test. The Student's t-test has a mathematical elegance about it that was essential at the time it was developed - in 1908, before modern calculators and computers. A more practical t-test, called Welch's or the Welch-Satterthwaite t-test, was developed in 1937 (by Welch) and 1946 (by Satterthwaite). This test is easily figured with a modern calculated and is standard in statistical software packages. The concept of a t-test should be familiar to you from your previous statistics course(s).

The situation we are going to consider in this lecture is comparing the means of two separate and independent groups - for instance, a treatment and a control group for a clinical trial where individuals were randomly assigned to one group or the other. The two populations need not be treatment and control - there could be two different treatments or two different categories. However, for the purpose of the discussion below, we will use the treatment and control terminology. The question is whether the treatment has a (positive) effect which could be measured with a two sided alternative (the means after the experiment are not equal) or either of the one-sided alternative (the treatment group's mean is greater than the control group's mean or the treatment group's mean is less than the control group's mean).

It is always a good idea to have in mind what the data looks like. In this case, the data will be two groups of numbers. The results for the treatment group will be one set of numbers: X_1, X_2, ..., X_{n_x} where n_x is the sample size for this group. The results for the control group will be another set of number: Y_1, Y_2, ..., Y_{n_y} where n_y is the sample size for the control group.

Let $\bar{X}$ be the average of the treatment data and $\bar{Y}$ be the average of the control data. Let s_x be the standard deviation of the treatment data and s_y be the standard deviation of the control data. The the test for whether the population means are equal or not (or that the treatment mean is smaller or that the treatment mean is larger) is based on the test statistic:

$$t = \frac{\bar{X} - \bar{Y}}{\text{S.E.}(\bar{X} - \bar{Y})}$$

There are two aspects of this formula that are omitted from the line above. One is obvious - the formula for calculating the denominator or *standard error* of the difference in the means. The difference of the two sample means is our estimate of the difference in the true, unobservable, population mean values, and the standard error is the standard deviation of this estimate. Means vary less than individuals, so the standard error for a mean is less than the standard deviation of the population.

The second is not necessarily obvious - the distribution of the t-test statistic depends on a number

called the *degrees of freedom*. Essentially, if you use a set of data to estimate one population quantity, say the mean, you lose one degree of freedom. That is, given $n-1$ data points and the sample mean, you can recover the n^{th} data point. In the case of the t-test, you use the data to estimate the population standard deviations that enter into the standard error formula, so you lose degrees of freedom there. You can also think of the degrees of freedom as follows: it is a way of taking into account your using the data twice: once to estimate standard deviations and then to test means.

Different versions of the t-test calculate the standard error and degrees of freedom differently. The *pooled* t-test assumes that the population standard deviation of the two groups are equal. It then calculates a pooled standard deviation using the data from the two groups as follows:

$$s_p = \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}}$$

The standard error and degrees of freedom for the t-test statistic are then given by

$$\text{S.E.}(\bar{X}-\bar{Y}) = s_p\sqrt{\frac{1}{n_x}+\frac{1}{n_y}} \qquad \text{d.f.} = n_x+n_y-2$$

This method has a nice mathematical elegance to it. Unfortunately, elegant mathematics often does not lead to practical statistics. A better version of the t-test, one that considers the possibility of different standard deviations in the different populations, is Welch's t-test. For this test, the standard error is calculated straightforwardly but the degrees of freedom formula is more complicated:

$$\text{S.E.}(\bar{X}-\bar{Y}) = \sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}} \qquad \text{d.f.} = \frac{\left(\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}\right)^2}{\frac{(s_x^2/n_x)^2}{n_x-1}+\frac{\left(s_y^2/n_y\right)^2}{n_y-1}}$$

When the two sample standard deviations are close, the degrees of freedom for Welch's test are close to the degrees of freedom for the pooled test (n_x+n_y-2). This is the maximum number of degrees of freedom possible. When one of the two sample standard deviations is much larger than the other, the degrees of freedom for Welch's test are closer to $n-1$ where n is the sample size for the group with the larger standard deviation.

A cheap alternative is often presented in elementary statistics texts: calculate the standard error as above for Welch's test but use the most conservative degrees of freedom possible ($\min(n_x,n_y)-1$). This test is pointless for any real researcher - it is only useful for the convenience of administering exams. Any decent statistical software package will have Welch's t-test pre-programmed for you to use and it is the t-test I recommend using in most circumstances. It is rare to be in a situation in which you are sure the two populations should have the same standard deviations. If you are in that situation but do not know it, Welch's t-test will likely give the degrees of freedom estimate close to that of the pooled t-test and the standard error it gives will likely be close to that given in the pooled t-test. If not, Welch's t-test provides an accurate way of taking into account the differences in standard deviations.

Significance level or Type I error

Tests have errors. There are two types of errors associated with hypothesis testing. The first is often set by the researcher. Before collecting data, most people have in mind a figure for the significance level (also called the type I error) of the test they are conducting. As discussed earlier, the most common value for the significance level of a test is 5% because Sir Ronald Fisher decided that a 1 out of 20 mistake was acceptable. The significance level is the probability you would reject the null hypothesis (here, of equal means) under the scenario the means actually are equal. The significance level corresponds to a cut-off value of the test. If your test statistic is larger, it means you think you are not so likely to have gotten it by chance so that you doubt the null hypothesis of equal means is true.

p-value

Most often, you will not report just whether your result is significant or not (unlikely to have happened by chance alone or somewhat likely to have happened by chance alone) but you will also report a p-value - the probability of your result or a more extreme result happening by chance alone given that the null hypothesis is true. If you set an arbitrary cut-off for significance (5% say), and your result has a p-value of 0.049, then it is significant while if it had a p-value of 0.051 it suddenly ceases to be significant. Why should this be so? Reporting the p-value allows the reader to assess your significance and make his own determination of the strength and/or suggestiveness of your results. Reporting p-values is standard practice in most fields.

Assessing significance and p-values is slightly different for one-sided versus two-sided alternatives. Here is an example. Suppose our alternative is that the mean of the treatment group is larger than the mean of the control group ($\mu_x > \mu_y$). Let's say we have a t-test statistic of $t = 2.0$ and the test has 10 degrees of freedom. The cut-off for significance at 5% is 1.81. Since our results is larger, our test is significant, leading us to believe that the two means are not equal and that, in fact, the mean of the treatment group is larger than the mean of the control group. Our p-value is $p(t(10d.f.) > 2.0) = 0.367$. The following is a picture of this situation:

In the case of a two-sided alternative, where, in advance we did not know whether the treatment would increase the measured variable or decrease it, we put half our significance in each tail. It is harder to reject the null hypothesis of no difference between the means in this case because, *a priori*, we did not know which side we should have been looking on. The 2-sided cut off for significance for a t-test with 10 degrees of freedom is 2.23 and our result of $t = 2.0$ is not significant. Our p-value is $p(|t(10d.f.)| > 2.0) = 2 \times 0.367 = 0.734$. We double the one-sided p-value because, *a priori* we did not know which tail to look at. The following is a picture of this situation:

Robustness

There is another issue - robustness. A statistician calls a test *robust* if it is still approximately valid when the assumptions on which it is based are not quite true.

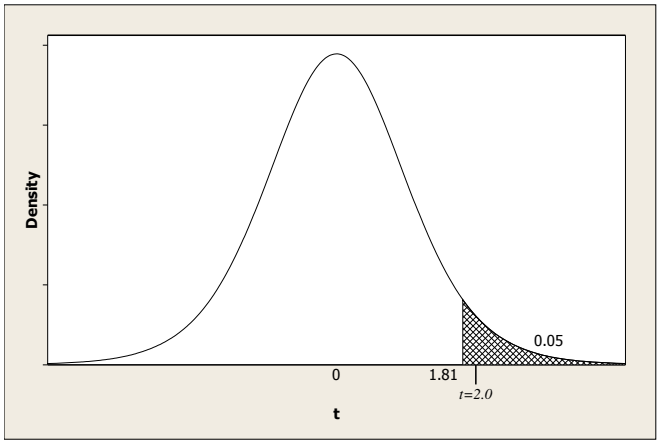

Figure 3.1: A pictorial representation of the rejection region for a one-sided t-test with 10 degrees of freedom at the 5% significance level. A theoretical result from data of t=2.0 is depicted.

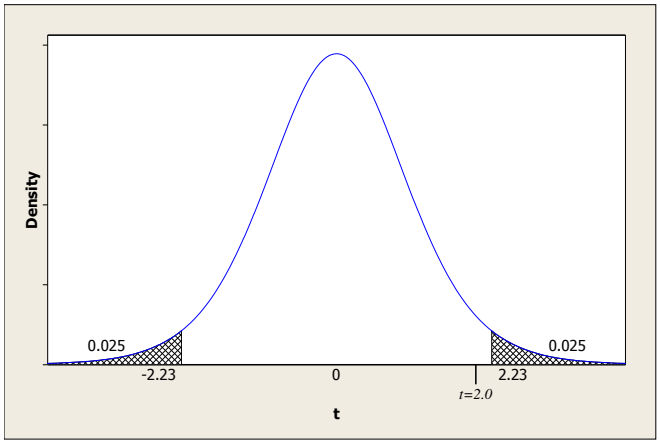

Figure 3.2: A pictorial representation of the rejection region for a two-sided t-test with 10 degrees of freedom at the 5% significance level. A theoretical result from data of t=2.0 is depicted.

A test of whether two means are equal or not is conducted at the 5% significance level, say. What this means is that, given all the assumptions of the test are true and the two means are indeed equal, the test will reject the null hypothesis of equal means 5% of the time. All tests make mistakes and the significance level is how often you allow this test to make a mistake.

A test is robust to a particular assumption if, when the assumption is invalid but all else holds, including that the two means are equal, the test still rejects the null hypothesis approximately 5% of the time - not much more or much less. This definition is vague - how invalid is the assumption and what is much more or much less than 5%? Clearly there is a sliding scale here.

The pooled t-test is somewhat robust to the assumption of equal population standard deviations but *only* when both sample sizes are the same ($n_x = n_y$). When the sample sizes are unequal and the standard deviations are unequal, the pooled t-test can give very misleading results. One prefers a robust test and Welch's t-test provides a robust alternative to the pooled t-test. The issue of robustness is explored using simulation, where you create data sets from known distributions over and over again, apply the test statistic of interest, and record the outcomes. See the homework problems for the next section.

Power

The significance level or type I error is the error associated with rejecting the null hypothesis (of equal means in this case) when the null hypothesis is true. Another error associated with a test is the type II error which is the error of accepting the null hypothesis when the null hypothesis is false. Since what it means for two means not to be equal exists on a sliding scale, the type II error is not only a function of the significance level of the test but also a function of how untrue the null hypothesis is (the true, non-zero, difference of the two means relative to the standard error of the test). The power of the test is one minus the type II error and is the probability of correctly rejecting the null hypothesis when it is false. All else equal, we want the most powerful test possible. The pooled t-test is slightly more powerful than Welch's t-test when it assumptions are met - but the difference is so slight that it is not, in my opinion, worth the risk of using the less robust pooled test.

Any decent software package will have a variety of t-tests pre-programmed for your use. Minitab has them under Statistics > Basic Statistics > 2-Sample t... I will demonstrate the use of this menu in class.

References and Readings

[1] W. S. Gosset. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

[2] F. E. Satterwaithe. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 6:110114, 1946.

[3] B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362, 1937.

Exercises for Lecture 3

1. –

2. –