The t -test: Some details

In chapter 2 we saw that the test for one sample, testing whether a group mean was equal to a standard, was based on a ratio. In the numerator was the difference between the sample average and the hypothesized standard, and in the denominator the standard error of the sample average. The reference distribution for looking at this ratio was the *t* distribution. Now, for two groups, we'll see that for testing whether the two groups have the same mean, we'll look at a ratio again, and the ratio will again have a *t* distribution. But now the numerator has *two* random quantities in it, $\bar{y}_1$ and $\bar{y}_2$. The denominator is again a standard error, this time the standard error of the difference of these two averages. The formula for the denominator turns out to be:

$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ where s_1 and s_2 are the standard deviations estimated from each group. (Why it turns out to be that is the subject of the next section).

Unfortunately, though, because there are two standard deviations in this denominator, the resulting statistic is only *approximately* a *t*-statistic and its degrees of freedom are complicated. (Even worse, they're *fractional*!) The good news is that most software packages not only calculate the degrees of freedom, but look up the p-values for these as well. If you have to calculate these by hand, you can use the following formula (although I've never had to resort to using it myself):

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

Just to make the formula look even worse, we use the greek letter ν for the degrees of freedom. Now to look up the p-value in a table, you have to round down to the nearest whole number degrees of freedom. It turns out that the degrees of freedom will always be between the smaller of n_1-1 and n_2-1 (on the low side) and n_1+n_2-2 on the high side.

We can write the t-statistic as:

$$t \approx \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The squiggly lines indicate that this is only approximately true, but most statisticians, in practice, don't worry about this distinction.

A tale of two t-tests

The technical problem with this ratio is the two *different* standard deviations in the denominator. This causes more variability that the standard t-statistic can take. It makes the ratio only approximately distributed as a t (not a big deal) and drives down the degrees of freedom (sometimes a big deal). How could we fix this? Well, we could think like an economist and *assume* our way out of our problem. Let's assume that the standard deviations in the two groups is really the same in order to have only one randomly varying quantity in the denominator again.

But, don’t we have two different groups? We have a standard deviation for *each* of them. What do we do with the two different estimates? Since we don't want to throw any information away, we'll combine them. We could just average the two of them, but if one group has more observations, we’d like to give it more weight. For technical reasons, we actually take a weighted average of the squares of each standard deviation (the variances) and weight them by their degrees of freedom. We call the resulting combined variance the *pooled* estimate of the variance, and its square root the pooled estimate of the standard deviation. It’s been the bane of many introductory student’s existences for eons. But it’s not really that bad. Here’s what it looks like:

$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$. You can see that it really is a weighted average of the two variance estimates. And, of course, to get the common standard deviation, we just take the square root of s_p^2. Then, we replace *both* s_1^2 and s_2^2 by this same estimate in the denominator of the t-statistic. Now there’s only one random quantity in the denominator and the resulting ratio not only is distributed with a *t-distribution,* but it has n_1+n_2-2 degrees of freedom (at least under a certain set of assumptions we'll talk about soon):

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

This is known as the equal variance t-statistic, and historically this was the only t-test available for many years. Many books and software packages still treat this as the only two sample t-test. But the unequal variance t-test is more applicable and should be used in most cases. When the two sample variances are nearly equal, the unequal variance t-statistic will be essentially the same as the equal variance one. But, when the two sample variances are far apart, the unequal variance statistic will have *reduced* degrees of freedom and should be used instead.

Now, how will you know if the variances are equal? Which test should I use? Since we don't know whether the variances in each group are the same, should we ever use the equal variance t -test? Since the tests are nearly the same when the variances are close, and different when they are not, why not always use the unequal variance t -test? There are two schools of thought about this.

One school says to examine this assumption first. Look at the variance of the two groups, usually graphically (via box plots, for example). Some people advocate the use of *another* statistical test to test whether the variances are equal, but I don't recommend this. It's statistical overkill. If the variances don't appear to be different, use the equal variance test, otherwise use the unequal variance test. I prefer to be a little more conservative -- I always use the unequal variance test first. In the case where the two variances are close, it makes little practical difference. However, when the variances are not nearly equal, it is crucial to avoid the equal variance test. Most software packages provide both tests.

The other school says just to always use the unequal variance test. For historical reasons, though, the equal variance test is usually the default test, so beware. In fact, in many packages, it's the standard, and you will have to explicitly ask for the unequal variance test (also known as Welch's test).

So when will we ever use the equal variance test? Good question. The reason it's here is threefold.

- First, because it is the *standard* test in most software packages.
- Secondly, when we move to more than two groups, we'll have to assume that the variance is constant across the groups and we'll pool the variances there. So, the equal variance test has been good preparation for that.
- Finally, for data from experimental designs, it's not unreasonable to imagine that the variance might be unaffected by the treatments. For happenstance data, though, I'd be less inclined to believe that the variances in two groups necessarily are the same, and so I'd always investigate the unequal variance test. And in fact, even for data from an experiment, I'll always perform both tests, just to make sure that there's not a big difference.

Confidence intervals

In addition to knowing whether one group has a higher mean than another, one usually wants to know how *much* higher it is, as well. In order to answer this, we have to construct a confidence interval for the difference between the means of the two groups. But, after the last section, this is easy!

Remember that a confidence interval has the form:

$$\bar{y}_1 - \bar{y}_2 \pm t_{n,a/2}\ SE$$

where SE is the standard error of the difference $\bar{y}_1 - \bar{y}_2$. This formula is generic in that it can be used for either the equal or unequal variance assumption. For the equal variance case, ν would be $n_1 + n_2 - 2$, and for the unequal variance case, it's the mess we showed you earlier (but it's never more than $n_1 + n_2 - 2$).

So, being specific, if we assume that the variances in the two groups are equal, we can use:

$$\overline{y}_1 - \overline{y}_2 \pm t_{n_1+n_2-2,a/2}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ ,$$

while if we don’t we use:

$$\overline{y}_1 - \overline{y}_2 \pm t_{n,a/2}\ \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\ \text{, where}$$

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

The t-test: The Gory Details

This section is for those of you who don't trust anything until you ``get under the hood'' as Ross Perot used to say. It is not crucial to understanding the rest of the book, but neither is it really that difficult.

To start, let's look at the simple difference of the two group averages: $\overline{y}_1 - \overline{y}_2$. This difference will serve as the numerator of our test statistic. If this is very different from 0, we should reject the null hypothesis. The problem is that we need to express this in standard error units to know how far from 0 it is. So, we just need to derive the formula for the standard error of the difference between two averages.

The ratio that we eventually want is:

$$t = \frac{\overline{y}_1 - \overline{y}_2}{SE(\overline{y}_1 - \overline{y}_2)}\ .$$

It turns out that calculations are easier to perform on *variances* rather than standard deviations, so to calculate the standard error of the difference, we calculate its variance first.

To go any further, we'll need an important assumption about $\bar{y}_1$ and $\bar{y}_2$. In addition to assuming that the observations that make them up are normal, we'll need to assume that they are *independent.* In probability theory any two random variables X and Y are independent if the probability of X doesn't change as the value of Y changes. For two groups of observations, what does this mean? It means that how the observations in group 1 vary around their mean is not influenced by the corresponding values in the other group. Let's suppose group 1 is the control group and group 2 is the treatment group. Does knowing that engine 6 in group 1 got 3 mpg more than the average of all the engines in group 1 tell us anything about how engine 6 in group 2 did (or how any other car in group B did for that matter)? It *would* if engine 6 in both groups was the *same* engine. But, since we chose 20 engines at random in this case, it shouldn't. This is the *assumption of independence.*

Now independence is very powerful, because the variances of independent random variables add:

Var(X+Y) = Var(X) + Var(Y)

if X and Y are independent. Great, but what about Var(X-Y) ? We need the variance of the of the *difference* between two averages. It turns out that the variance of a difference (of two independent quantities) is the same:

Var(X-Y) = Var(X) + Var(Y).

There are several ways to see this. First, formally, we could just substitute -Y in for Y in the equation for the sum. Since, Var(Y) is the same as Var(-Y):

Var(X+(-Y)) = Var(X) + Var(-Y) = Var(X) + Var(Y).

Intuitively, many people assume that the variance of the difference is the difference, (not the sum) of the variances. Wouldn't this be nice? The world doesn't work like this, however. The variances add. If you don’t believe me, consider manufacturing two automobile parts, a ring and the piston that goes inside it. Suppose the rings designed to have inner diameter 3 inches, and the pistons are supposed to just fit inside them, so that they have diameter 3 inches on the average as well. Let's call the rings diameter X and the piston diamter Y . We want X-Y = 0 . If the mean of X is 3 and the mean of Y is also 3, then the *mean* of X-Y is 0 . But what about the standard deviation? What is SE(X-Y) ? Should it matter how precisely I make the parts? If the variance of the difference X-Y was the *difference* of their variances, then just setting the two standard deviations equal (to anything) would make the SE of the difference 0, which would mean that it had *no variation at all!* Doesn't it make more sense that we need to make the two standard deviations as small as possible to ensure that the difference won't vary too much? The answer is yes, because the variance of the difference is the *sum* of the variances, not their difference.

Well, back to our differences. This is all fine for X- Y, but we need the variance of the difference of two *averages.* If we just substitute our averages for X and Y in equation 3.?, we have

$$\mathrm{Var}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2) = \mathrm{Var}(\overline{\mathrm{y}}_1) + \mathrm{Var}(\overline{\mathrm{y}}_2)$$

Remember, the group averages satisfy the independence assumption because the units in the two groups have been randomly assigned to the two treatments. If we had used each engine twice -- giving each engine both RUX-7000 and Regular (perhaps a smarter strategy), we wouldn't have had independence. We'd have another design that we will talk about in a later chapter.

Now, taking the square root turns the variance into the standard error:

$$\mathrm{SE}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2) = \sqrt{\mathrm{Var}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2)} = \sqrt{\mathrm{Var}(\overline{\mathrm{y}}_1) + \mathrm{Var}(\overline{\mathrm{y}}_2)}$$

Remember that the standard error of an average is the standard deviation divided by the square root of the sample size($s\ /\sqrt{n}$) and so:

$$\mathrm{SE}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2) = \sqrt{\mathrm{Var}(\overline{\mathrm{y}}_1) + \mathrm{Var}(\overline{\mathrm{y}}_2)} = \sqrt{s_1^2 / n_1 + s_2^2 / n_2}$$

Now, of course, we don't know what the standard deviation of either group is, so we have to estimate them. When we do this we get the *estimate* of the standard error of the difference of the two averages:

$$\hat{\mathrm{S}}\mathrm{E}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2) = \sqrt{s_1^2 / n_1 + s_2^2 / n_2}$$

Now, we've got it!! This is our estimate of the standard error of the test statistic, the difference between the averages of the two groups. We compare the difference $\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2$ to this quantity. The ratio results in a test statistic that is approximately t –distributed. If this quantity is large, we **reject** the null hypothesis of equal means. Putting all this together, the ratio becomes:

$t \approx \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ which we saw before was the ***unequal variance t*** statistic.

The equal variance t statistic is obtained simply by using the pooled estimate of the for both s_1^2 and s_2^2.

An example of calculating the t-test

To have this make a little more sense, let's try this for the example from the beginning of the chapter.

Recall the data from Table 3.1:

	Regular	RUX-7000
	14.2	36.2
	21.9	11.5
	44.8	5.2
	9.5	10.1
	16.2	40.2
	38.7	17.1
	9.2	10.6
	5	45.5
	11.1	21.7
	35.4	16.1
Average	20.6	21.42
Std. Dev	14.067	14.161

The two standard deviations are *very close*, so it shouldn't matter which t-test we use here. Let's calculate both. To compute the *pooled variance*, we use:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(10-1)(14.067)^2 + (10-1)(14.161)^2}{10+10-2} = 199.2072$$

so $s_p = \sqrt{199.2072}=14.114$. This makes *sense* since the pooled standard deviation should always be between the two group standard deviations. We now calculate the standard error of $\overline{y}_1 - \overline{y}_2$:

$$\mathrm{SE}(\overline{y}_1 - \overline{y}_2) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 14.114\sqrt{1/10+1/10} = 6.312$$

This also makes sense and shows the problem with this experiment. The estimate of the difference of the averages from the two groups is over 6 mpg! We're going to need to see a difference of around 12 mpg just to start to see statistical significance. Since we saw only a 0.82 mpg difference, this is only 0.13 standard errors from 0 and will not be statistically significant: $t_{18}=\frac{0.82}{6.312}=0.13$.

Just to be complete, let's calculate the unequal variance t. Instead of pooling the variances we just use

$$\mathrm{SE}(\overline{y}_1 - \overline{y}_2) = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{14.067^2}{10}+\frac{14.161^2}{10}} = 6.312 .$$

Now, here it made *no* difference which formula we used because the two standard deviations were (unrealistically) close. This is evident as well in the fact that the degrees of freedom are essentially unchanged:

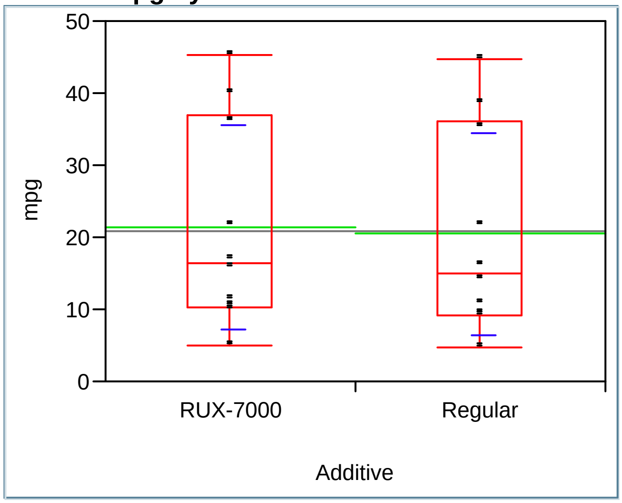
$$n = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{(n_1-1)}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{(n_2-1)}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{14.067^2}{10}+\frac{14.161^2}{10}\right)^2}{\frac{1}{(10-1)}\left(\frac{14.067^2}{10}\right)^2+\frac{1}{(10-1)}\left(\frac{14.161^2}{10}\right)^2} = 17.9993$$

So, in either case, we have a difference of 0.82 mpg with a standard error of 6.312. Under the assumption of the null hypothesis that there is no difference in the two means, this should have a t with 18 df distribution. Since the critical value at 0.025 for t_{18} is 2.10, we are no where near the critical value and we have no evidence against the null hypothesis.

The t-test in real life

Now, let's let the computer do the calculations for us and we'll interpret the results. Here are some output from this example:

t-test of mpg by Additive



t-Test (equal variance)

	Difference	t-Test	DF	Prob > |t|
Estimate	0.8200	0.130	18	0.8981
Std Error	6.3121			
Lower 95%	-12.4412			
Upper 95%	14.0812			

Assuming equal variances

Means and std dev for each group

Level	Number	Mean	Std. dev	Std Error	Lower 95%	Upper 95%
RUX-7000	10	21.4200	14.1607	4.4633	12.043	30.797
Regular	10	20.6000	14.0676	4.4633	11.223	29.977

Std Error uses a pooled estimate of error variance

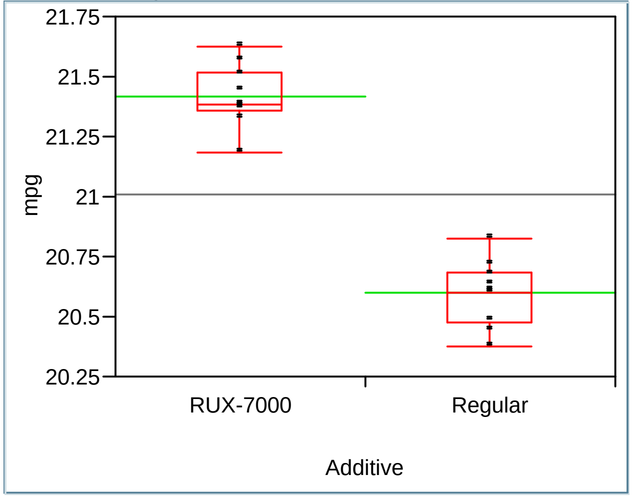
t-Test (unequal variance)

t-Test	DF	Prob>|t|
0.1299	17.999	0.8981

Here is everything we just calculated. Notice the standard error of the difference is 6.312 (using the pooled standard deviation) and the t-statistic is 0.13. When we use the unequal variance formula, it changes (very) slightly. The p-value of .8981 indicates that this is a very likely value to occur by chance when the null hypothesis is true. In other words, the difference in averages that we got was very typical from two groups that have no difference in treatment means.

For contrast, let's look at the analysis of the data from Table 3.2, where we controlled the engine type:

t-test of mpg By Additive



t-Test (equal variance)

	Difference	t-Test	DF	Prob > |t|
Estimate	0.820000	14.165	18	<.0001
Std Error	0.057889			
Lower 95%	0.698380			
Upper 95%	0.941620			

Assuming equal variances

Means for each group

Level	Number	Mean	Std Dev	Std Error	Lower 95%	Upper 95%
RUX-7000	10	21.4200	0.1261	0.04093	21.334	21.506
Regular	10	20.6000	0.1327	0.04093	20.514	20.686

Std Error uses a pooled estimate of error variance

t-Test (unequal variance)

t-Test	DF	Prob>|t|
14.1651	17.954	<.0001

What's different here? The first thing that should hit you is that the boxplots are very far apart. The two groups performances *look* different! The t-test (either one) just verifies this. Controlling the engine type has lowered the standard deviation in each group and in turn lowered the standard error of the difference between the averages. The t-ratio is now over 14, which has an astronomically low (can you say that?) p-value. There is *no way* these data came from two groups with equal means. We have to ***reject the null hypothesis***.

The t test is too hard

In 1951, an engineer named Duckworth went to the meeting of the Royal Statistical Society in London to complain about the test I've just described. Somewhat justifiably, he railed against the statistical community for coming up with such an arcane solution to such a simple problem. After all, looking at the boxplots, in figures 3.1 and 3.2, isn't it *obvious* what the conclusions should be? Of course, the answer is: not always, but John Tukey took Duckworth's criticisms to heart and came up with what he called ``A quick compact test to Duckworth's specifications" (1951).

Here's how (slightly simplified) the test works: You have two groups, A and B, with n_A and $n_{\mathbf{B}}$ observations respectively. For this test to work, one of the groups, say group A, has to have the smallest observation and the *other* group has to have the largest. (If one group has both the smallest and largest, the test doesn't work.) Now, count the observations in A that are smaller than *all* observations in B. Add to this the number of observations in B larger than *all* observations in A. Got it? Now, if this number is ≥ 7, the groups have different means at $\alpha = .05$. If the number is ≥ 10 it's significant at α=.01 and if it's ≥ 13, α=.001. It doesn't even matter what n_A and $n_{\mathbf{B}}$ are (although they should be at least 6 or so, and about equal).

Let's try it on our example. In Table 3.1, RUX-7000 has the largest observation, which is 45.5 mpg, and Regular has the smallest, 5.0 mpg, so we can apply the test. Now, it turns out that 45.5 is the *only* value in this group (call it B) larger than all of the first (this gives us 1 observation in B bigger than all in A), and 5.0 is the only value in the first smaller than all of the second (giving us 1 in A smaller than all in B). This gives us a grand total for our statistic of 2. Since 2 is less than 7 , we conclude that there is no evidence of differences in the group means (as we did with the t -test).

If we look at Table 3.2, however, all 10 mileages in the second group are larger than the first (giving us 10 so far) and all 10 in A are smaller than the smallest in B, (another 10) for a grand total of 20. Since $20 \geq 13$, the group means are different at $\alpha = .001$.

It doesn’t get much simpler than that! I always perform a Tukey test of this kind before plowing through the t -test!! But, you should always give the t-test results as well, when you go to publish your results, or try to convince other people at the committee meeting.

Another Example

Data have been collected from two hospitals on the gestation times of the last 100 live births. The question is whether the mean gestation time is the same in the two hospitals. For reference, the average gestation time in the U.S. is 266 days with a standard deviation of 16 days. The medical question is whether the women in each hospital are received adequate, or at least roughly equal, pre-natal care. A shorter gestation time, associated with lower birth weights and higher mortality risks would be indicative of inadequate care. Hospital A is located in an affluent suburb, while Hospital B is in the inner city. The data themselves are on the accompanying disk. Box plots of the gestation times from the two hospitals are shown below in figure 3.4:

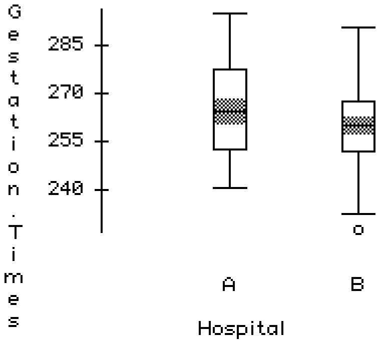


Figure 3.4 Gestation times (in days) from two hospitals.

In figure 3.4 we see that from the box plot alone there is perhaps *some* evidence that the mean gestation time in Hospital B is lower than in Hospital A. But it is not completely clear. Here is a case where we need the more precise information available by examining the t-test and other summary statistics. Looking at output of the t -test, we see that the difference between the average gestation times of the two hospitals is 6.10 days with a standard error of 1.90 days. This is 3.213 (the t -value) standard errors away from 0 with an associated (two-sided) p-value of 0.0015.

So, if the two hospitals had the same mean gestation time, data such as this would be highly unlikely (1.5 times out of 1000) to occur. This is evidence (**) against the null hypothesis, and so, we reject it with a p-value of 0.0015.

t-Test

```
              Difference    t-Test     DF      Prob > |t|
Estimate      6.10000       3.213      198     0.0015
Std Error     1.89864
Lower 95%     2.35584
Upper 95%     9.84416
Assuming equal variances
```

Means for Oneway Anova

```
Level   Number   Mean       Std Dev    Std Error   Lower 95%   Upper 95%
A       100      265.810    13.9477    1.3425      263.16      268.46
B       100      259.710    12.8820    1.3425      257.06      262.36
Std Error uses a pooled estimate of error variance
```

```
Welch's T-test testing Means Equal, allowing Std Devs Not Equal

t-Test    DF        Prob>|t|
3.2128    196.76    0.0015
```

We used a two-sided alternative here, assuming that we had *no prior* knowledge or disposition toward one direction or the other. If we had assumed that the suburban hospital (A) would have at least as great a mean gestation time as the inner city one, we could have used a one-sided null hypothesis. This would have cut our p-value in half.

We included both the so –called standard (which is really the more restrictive test assuming equal variances) and the “Unequal variance” t - test. I routinely look at both. Since they are so close (due to the fact that the standard deviations in each group are nearly identical), it doesn't matter which one we use. If they were different, I would look at the data to try to understand why, and most likely use the “unequal variance” results.

Finally, here is the Tukey test on the same data:

There are 3 Observations in Hospital A > all Observations in Hospital B and

7 Observations in Hospital B < all Observations in Hospital A.

Total: 10

Since the total (10) is ≥ 10, the difference is significant at $\alpha = .01$.

Assumptions

The assumptions for the t-test are that the errors are

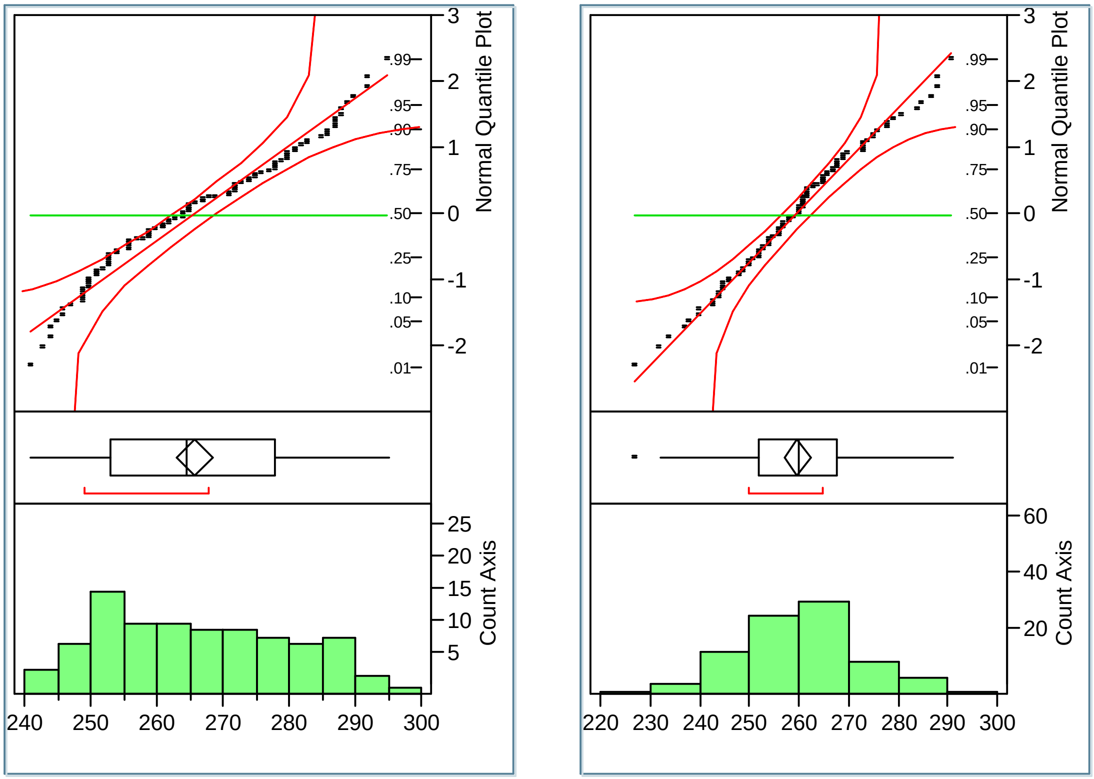
- normally distributed
- independent
- have the same variance

This last assumption means the same variance *within each group.* If you use the equal variance t-test you have the *additional* assumption that the variances within each group are also equal to each other.

The normality assumption is fairly crucial, especially if the sample sizes are small. Plotting the data from each group in a histogram and on a normal probability plot is a good idea. As long as the data are *symmetric*, the t-test is pretty robust. You should worry if the data are skewed to one side or if there are outliers.

Independence within a group usually has to do with something happening over time. If you have time data, it is a good idea to plot the data within each group against time and make sure there are no obvious trends or other patterns. You should also check to make sure the data aren’t getting more or less variable over time (the constant variance assumption).

Here are the plots for each hospital:



Hospital A

Hospital B

Figure 3.5 Plot of data from each hospital. Notice that the normal probability plots show that each is resonably normal since the data fall on a fairly straight line. There is one outlier in the data from Hospital B, but it's not terribly far from the bulk of the data.

Exercises

4. In the design of paper airplanes, many choices can effect both the stability and the duration of the flight. In this investigation, Andrew Speck looked at the effect

of changing the direction of the bend of the wing on the flight distance. In this investigation, Andrew Speck looked at the effect of changing the direction of the bend of the wing on the flight distance. He made 16 paper airplanes, 8 with wings bent up, and 8 with wings bent down. He randomized the order and flew them in a long corridor. The flight distances are recorded below:

Direction	Flight Length in inches	Order
Down	109	3
Down	161	4
Down	197.5	6
Down	191	12
Down	194.5	14
Down	165	15
Down	177	16
Down	123	9
Up	245.5	1
Up	168	2
Up	211	5
Up	214	7
Up	171	8
Up	262.5	10
Up	198	11
Up	214	13

Table 3.6 The distances in inches for 16 runs of a paper airplane.

(a) What is the null hypothesis? Is this a one or two-sided hypothesis?

(b) What do you conclude about the effect of wing bend on flight distance?

(c) Is there a difference in conclusions between the equal and unequal variance test?

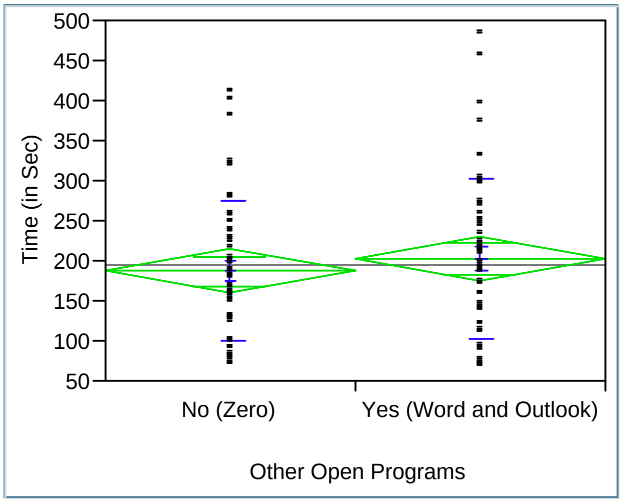
(d) Is a Tukey test appropriate here? If so, what does it say about the mean difference?

(e) Find a 99% confidence interval for the mean difference in flight distances for the two wing positions.

(f) What assumptions about the data did you make to answer the questions? Do they seem reasonable?

5. Downloading files over the Internet can be a frustrating experience, especially when the file is large. Of course, many factors effect the speed of the download including the traffic on the network, the type of computer you own, the browser you use and how many other programs you are running while trying to download. In order to see the effects of the latter, Josh Burns placed a 5 megabyte text file in a web site and downloaded it on 94 different occasions. He was the only person who knew of this file, so he could be fairly certain that no one else was trying to download it as well.

He used the same browser and computer for all 94 runs. In half of the runs, he downloaded the file just after booting up the machine, with no other programs running. In the other half, he had both Microsoft Word© and Microsoft Outlook© running. He randomized the order of the downloads, but ran all of them in the evening from 6 to 8 PM. The data from the 94 runs are available on the disk. Here are some summary statistics:



Means for Oneway Anova

Level	Number	Mean	Std Dev	Std Error	Lower 95%	U
No (Zero)	47	188.383	87.1434	13.595	161.38	2
Yes (Word and Outlook)	47	203.660	98.8917	13.595	176.66	2

Std Error uses a pooled estimate of error variance

(a) For which factors did Josh use control?

(b) Which factors did Josh randomize?

(c) What other factors did Josh not even consider? How might those affect the results?

(d) What is the null hypothesis? Is this a one or two sided hypothesis?

(e) Is there a statistically significant difference in mean download time between the two conditions? If no, does this mean that the null hypothesis is true?

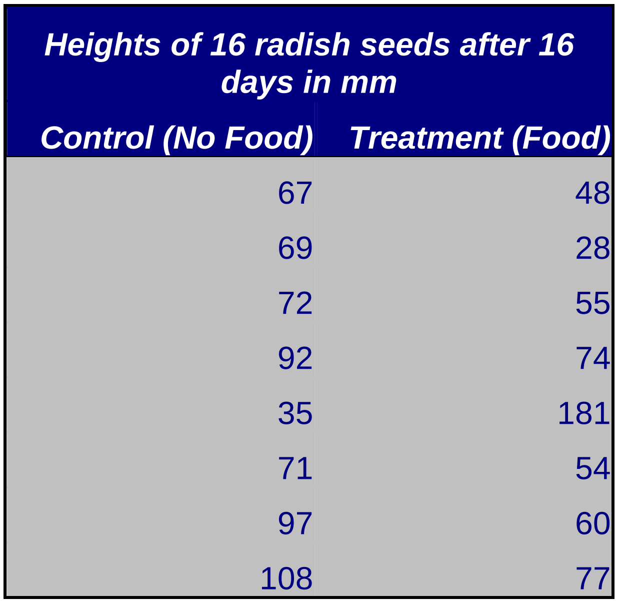
(f) What assumptions did you make about the data? Do they seem reasonable?

(g) Give a 95% confidence interval for the true mean difference. Does this interval contain 0? What does this mean?

(h) Suppose Josh considers a 20 second increase in download time from the average of 188 with no programs running to be annoying. Assuming that the standard deviation is 90 seconds, how many observations should he take to ensure a 90% chance of detecting this large a difference.

(i) What does the answer to the last question imply about the power of this test?

6. In an experiment to determine the effect of a commercially available plant food on plant growth, Stephen Lord grew radish seeds in sixteen different pots. Each dish received the same amount of light and water, but eight of the pots received plant food in the ratio of one tablespoon of plant food to each half cubic foot of soil. Ambient temperature was kept at 72° F as closely as possible. The seeds were allowed to germinate and grow for 10 days. The results are shown in the following table:



Heights of 16 radish seeds after 16 days in mm	
Control (No Food)	Treatment (Food)
67	48
69	28
72	55
92	74
35	181
71	54
97	60
108	77

(a) What is the null hypothesis? Do you think it's one sided or two sided?

(b) Is there a statistically significant difference between the two groups?

(c) Give a 95% confidence interval for the true difference in the means of the two groups? What is the best case scenario for the plant food? How much does it help? What is the worst case scenario? How much might it *hurt* young plant production?

(d) What assumptions about the data have you made? Do they seem reasonable? Is there an outlier in the Treatment group?

(e) In looking back at the original records it was discovered that the 5th plant in the treatment group actually grew to be 31 mm *not* 181 mm. There was a transcription error from the lab book to the computer. How does this change your answers to questions (a) – (d)?