Notes 5: t tests

1. The one sample t-test

(a) Formulas
Recall the Z test formula:

$$z_{\overline{X}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

The one sample t-test, which is very similar to the Z test, has the following formula:

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}} = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

where the only difference is SD versus σ in the Z test. That is, the standard error of the mean is now estimated by the formula:

$$s_{\overline{X}} = \frac{SD}{\sqrt{n}}$$

where the symbol, $s_{\overline{X}}$, is used to indicate that the standard error of the mean is being estimated with the sample SD. Recall that the standard error of the mean for the Z test was calculated as:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

(b) Hypotheses for the one sample t-test

Hypotheses for the one sample t-test are formulated in exactly the same manner as in the one sample Z test. Using SAT as an example (note, $\mu = 1000$), the one sample t-test non-directional hypothesis is symbolized as:

Non-directional:
H_1: $\mu \neq 1000$
H_0: $\mu = 1000$

Directional (one-tail tests):

Lower-tailed test
H_1: $\mu < 1000$
H_0: $\mu \geq 1000$ or
H_0: $\mu = 1000$ (this one is preferred)

Upper-tailed test
H_1: $\mu > 1000$
H_0: $\mu \leq 1000$ or
H_0: $\mu = 1000$ (this one is preferred)

Note that for the directional hypotheses, the alternative, H_1, states what one expects to find (as long as a relationship, or difference, is expected). For example, if one expects that a sample of students will have a higher than average IQ, then H_1: $\mu > 1000$. Similarly, if one expects that a given sample of students will have a lower than average IQ, the H_1: $\mu < 1000$.

(c) Critical t-values, (t_{crit})
Like the Z test, one may use critical values for hypothesis testing. Critical t-values are obtained from a t-table (see text). Note that t-values have distributions that are similar to normal distributions, but they are slightly fatter in the tails. Finding t-values in the t table is similar to the z table. To find the correct critical t-values (denoted as t_{crit}), one must first calculate the degrees of freedom (df or ν). For the one-sample t-test, degrees of freedom are defined as:

df (or ν) = n - 1

where, as before, n is the sample size. Degrees of freedom can be described as the amount of information available in the sample after certain mathematical restrictions are applied to the data.

(d) Statistical significance (α)
Next, one must determine the level of statistical significance for the analysis. As before, alpha is usually set at .10, .05, or .01. Once the alpha level is determined, critical values for the one sample t-test can be found.

(e) Decision rules
Deciding whether to reject or fail to reject the null can be determined by decision rules. The decision rules are:

Two-tailed tests:
If $t \leq -t_{crit}$ or $t \geq t_{crit}$, then reject H_0; otherwise, fail to reject H_0

One-tailed (upper-tailed) test
If $t \geq t_{crit}$, then reject H_0; otherwise, fail to reject H_0

One-tailed (lower-tailed) test
If $t \leq -t_{crit}$, then reject H_0; otherwise, fail to reject H_0

Note that $\mathbf{t}_{crit}$ symbolizes the critical t-value found in t tables and is different from t, which is the calculated t-ratio obtained from sample data.

(f) An example
A physical education teacher wishes to know whether his class of students is statistically above or below the national average in weight. The national average for eighth graders is $\mu = 100$. The student weights for his class are: 99, 98, 105, 110, 115, 103, 88, 125, 130, and 115. For this sample, n = 10, $\bar{X} = 108.8$, SD = 12.839, so

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{108.8 - 100}{12.839/\sqrt{10}} = \frac{8.8}{12.839/3.162} = \frac{8.8}{4.06} = 2.167.$$

The df (or ν) = n - 1 = 10 - 1 = 9, and set $\alpha = .05$.

The goal of the test is to determine whether the sample has an average weight that is statistically different from the national average. This calls for a non-directional test because the specific direction of the mean difference (higher or lower) was not indicated. Therefore,

H_0: $\mu = 100$, and
H_1: $\mu \neq 100$.

The critical value is $t_{crit} = 2.262$. The decision rule is:

If 2.167 $\leq$ -2.262 or 2.167 $\geq$ 2.262, then reject H_0; otherwise, FTR H_0.

Since 2.167 is neither less than -2.262 nor greater than 2.262, the null is not rejected (i.e., fail to reject) and one concludes that the sample does not have a statistically different mean from the national average.

What happens if instead it was hypothesized that the sample would have a <u>lower</u> than average weight, i.e.,

H_1: $\mu < 100$, and
H_0: $\mu = 100$.

This is a lower tailed test since the sample is expected to have a lower mean score. If $\alpha = .05$, the corresponding critical t-value is $t_{crit} = -1.833$. The decision rule states:

If 2.167 $\leq$ -1.833, then reject H_0; otherwise, fail to reject H_0.

Again, the null is not rejected, i.e., fail to reject.

Finally, had one hypothesized that the sample would be above average, then

H_1: $\mu > 100$, and
H_0: $\mu = 100$.

This is an upper-tailed test. The critical t-value, for an alpha of .05, is $t_{crit} = 1.833$, and this time the null is rejected, so the sample can be said to have a statistically higher average weight than normal.

It is also possible to perform hypothesis testing with the t-test without using critical t-values. Recall that the Z test had decision rules for p-values. Calculated probability values, p-values, are usually reported with statistical software, and the decision rules for the Z test also apply to the t test. See earlier notes for these decision rules. (This section to be discussed further in class.)

(g) <u>Assumptions</u>
The assumptions for the one sample t-test are identical to the Z test: normality and independence.

(h) Exercises

(1) Raw MAT scores are 31, 38, 27, 41, 39, and 36. Is this sample statistically different from the national average MAT of 30? Set $\alpha = .01$.

(2) Same scores, but $\alpha = .05$.

(3) Same scores and $\alpha = .05$, but this time hypothesize that the sample average will be greater than population average.

(4) Fifteen students have a sample MAT mean of 32.3 with a sample standard deviation of 4.73. Does this sample of students have a mean MAT score that is statistically different from the national average at the .01 level with a two tailed test? What about $\alpha = .10$ and a two tailed test? What about $\alpha = .05$ and an upper-tailed test?

(5) Suppose the following random sample of ITBS-math scores are observed in your middle school: 45, 58, 65, 63, 35, 43, 78, 55, 58, 69, 81, and 49. Is this evidence that your middle school has a student population above average in terms of mathematics skills (national average ITBS-math is 50)? Set alpha at .05.

(6) You wish to determine whether you are getting cheated every time you buy a bag of apples. The standard bag of apples that you buy states that it contains one pound (16 ounces) of apples. After you get home you notice that the bag only contains 15.5 ounces, not the stated 16 ounces. To determine whether or not the company is systematically cheating the consumer, you decide to buy every 16 ounce bag of apples in the three local grocery stores. After weighing each bag you find the following weights: 14.3, 15.5, 16.3, 17.0, 15.2, 15.9, 14.8, 15.0, 15.2, 15.9, 15.7, 15.6, and 16.1. Setting the significance level at .05, does it seem the company is systematically cheating the consumer? Which should you perform, an upper-, lower-, or two-tailed test? Why?

(7) Ford claims that its new car, Aspire, gets 39 mpg on the highway. Consumer Reports magazine wishes to test this claim, so they hire you for $1500 to perform the statistical testing. They buy 10 Aspires and road test each. They find the following mpg estimates for the cars: 32, 43, 39, 38, 34, 36, 35, 38, 39, and 36. Their question to you is: Does our sample of Aspires have an estimated mpg that is different from Ford's claim? Set alpha at .05 and give them an answer.

Computer output for exercises 1 through 7

Exercises 1 through 3

```
. ttest mat=30

  Variable |      Obs        Mean   Std. Dev.
  ---------+---------------------------------
       mat |        6    35.33333    5.316641

           Ho:  mean = 30
                    t = 2.46 with 5 d.f.
            Pr > |t| = 0.0574
```

<u>Exercise 4</u>

```
. ttesti 15 32.3 4.73 30

  Variable |       Obs        Mean    Std. Dev.
  ---------+---------------------------------
         x |        15        32.3        4.73

           Ho:  mean = 30
                   t = 1.88 with 14 d.f.
            Pr > |t| = 0.0806
```

<u>Exercise 5</u>

```
. ttest itbs=50

  Variable |       Obs        Mean    Std. Dev.
  ---------+---------------------------------
      ITBS |        12       58.25    13.93573

           Ho:  mean = 50
                   t = 2.05 with 11 d.f.
            Pr > |t| = 0.0649
```

<u>Exercise 6</u>

```
. ttest weight = 16

  Variable |       Obs        Mean    Std. Dev.
  ---------+---------------------------------
    weight |        13    15.57692    .7013722

           Ho:  mean = 16
                   t = -2.17 with 12 d.f.
            Pr > |t| = 0.0503
```

<u>Exercise 7</u>

```
. ttest mpg = 39

  Variable |       Obs        Mean    Std. Dev.
  ---------+---------------------------------
       mpg |        10          37    3.091206

           Ho:  mean = 39
                   t = -2.05 with 9 d.f.
            Pr > |t| = 0.0711
```

2. Confidence Intervals (CI) for Means

When estimating a parameter, one typically uses a point estimate like $\overline{X}$, s, or s^2. Using these point estimates, one may construct an interval which will show a possible interval range of values which might include the parameter being estimated.

A confidence interval (CI) for μ is found by:

$$(1 - \alpha)CI = \overline{X} \pm \left({}_{1-\alpha/2}t_{df}\right)\left(s_{\overline{X}}\right)$$

which, stated differently, is

(1 - α)CI = $\overline{X} \pm \left({}_{1-\alpha/2}t_{critical}\right)\left(s_{\overline{X}}\right)$

which is

(1 - α)CI = $\overline{X} \pm \left({}_{1-\alpha/2}t_{critical}\right)\left(\frac{SD}{\sqrt{n}}\right)$

or simply

(1 - α)CI = $\left(\overline{X} - {}_{1-\alpha/2}t_{critical}\left(\frac{SD}{\sqrt{n}}\right), \overline{X} + {}_{1-\alpha/2}t_{critical}\left(\frac{SD}{\sqrt{n}}\right)\right)$

This is 100(1 - α) confidence interval. That is, if α = .05, then this is a 100(1 - .05) = 100(.95) = 95% confidence interval, or .95CI. A .95CI means that one can be 95% confident that all intervals constructed like this for 100 random samples, in the long run, will contain the population value μ. This means that if 100 such intervals were constructed, on average the population value of μ would be correctly included in 95 of those intervals while would increase fail to include μ.

To calculate this CI, choose α, say at .05, then construct the interval by simply finding the critical value associated with α = .05, and filling in the rest of the formula.

Example:
Construct .95CI for a class of high school students (n = 12) with a mean IQ of 120 and a standard deviation of 16.5.

$$\left(120 - 2.201\frac{16.5}{\sqrt{12}},\ 120 + 2.201\frac{16.5}{\sqrt{12}}\right)$$

$$= \left(120 - 2.201 \times 4.763, 120 + 2.201 \times 4.763\right)$$

$$= \left(120 - 10.483, 120 + 10.483\right)$$

= (109.517, 130.483)

With such an interval, one may state that one is 95% confident that this interval contains the true μ for all students who are like the students in the particular high school class (apparently smart students).

Based upon this confidence interval, it seems that this high school class is quite different from the mean score typically found for IQ tests in the population. How does one know this?

The CI may also be used as a non-directional hypothesis test. If the hypothesized population value of μ is not within the CI, then H_0: μ = 100 may be rejected. Since the value 100 is not within the interval constructed, which ranges from 109.5 to 130.5, one may conclude that sample data appears to differ, statistically, from the hypothesized value of 100. In this particular case, the sample data such a mean that is higher than the expected value of 100.

Exercises

(1) Construct a .99CI for the following scores: 120, 123, 125, 101, 98, 101. Test the hypothesis, using the .99CI, that H_0: $\mu = 100$.

(2) Same as (1), but use a .95CI.

(3) Same as (1), but use a .90CI.

(4) Fifteen students have an SAT mean of 1200 with a standard deviation of 150. Does this sample of students have a mean SAT score statistically different from the population mean of 1000 at the $\alpha = .05$? Use a CI to answer this question. Is the mean statistically different if a .99CI is used?

3. The Two-Independent Samples t test (also called the Two Group t test)

(a) Situation

Both the Z test and the one sample t-test allow one to statistically comparing the mean of one sample of observations with a given population value (e.g., μ). If one is interested in comparing two independent groups, then the two independent sample t-test may be appropriate.

For example, suppose one is using a posttest only control group design to examine the effect of computer assisted learning in geography achievement among third graders. The control (or comparison) group is taught U.S. geography with the traditional methods using maps, textbooks, and workbooks. The experimental group uses the computer game Where in the U.S. is Carmen SanDiego. At the end of the lesson, both groups are given the same posttest. A two group independent t-test would be appropriate for determining statistical difference between the control and experimental groups.

(b) Hypothesis formulation:

One may formulate three different research hypotheses for the above example.

Non-directional:

The experimental and control group will have different levels of achievement in US geography.

H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$, or

H_0: $\mu_1 - \mu_2 = 0.00$ and H_1: $\mu_1 - \mu_2 \neq 0.00$

where μ_1 represents for group 1 (experimental group) and μ_2 represents group 2 (control group).

Directional (group 1 has higher mean than group 2):

The experimental group will show a higher level of achievement.

H_0: $\mu_1 \leq \mu_2$ and H_1: $\mu_1 > \mu_2$, or

H_0: $\mu_1 - \mu_2 \leq 0.00$ and H_1: $\mu_1 - \mu_2 > 0.00$

<u>Directional (group 2 has higher mean than group 1)</u>:
The experimental group will show a lower level of achievement.

H_0: $\mu_1 \geq \mu_2$ and H_1: $\mu_1 < \mu_2$, or

H_0: $\mu_1 - \mu_2 \geq 0.00$ and H_1: $\mu_1 - \mu_2 < 0.00$

(c) <u>Formulas for calculating the t ratio</u>
To test the above hypotheses, the two sample independent t statistic is calculated as:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}}$$

Since it is usually assumed that $\mu_1 - \mu_2 = 0.00$ (no difference in the population values), the t formula can be simplified to

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{SE_d}$$

where

$$SE_d = s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note that SE_d represents the <u>standard error of the difference</u>, which, like the <u>standard error of the mean</u>, represents the standard deviation of the sampling distribution for $\overline{X}_1 - \overline{X}_2$. The symbols s_1^2 and s_2^2 represent the variances for group 1 and group 2, respectively.

Recall that the sampling distribution of the sample mean has a known distribution that approaches the normal distribution when sample sizes are large. The sampling distribution for $\overline{X}_1 - \overline{X}_2$ also follows the central limit theorem. Note that the mean of the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is equal to $\mu_1 - \mu_2$. The standard error for $\overline{X}_1 - \overline{X}_2$ is $SE_d = s_{\overline{X}_1 - \overline{X}_2}$.

(d) <u>Degrees of Freedom</u>
Degrees of freedom for the two independent sample t-test are:

df (or ν) = $n_1 + n_2 - 2$

where the n_1 is the sample size for group 1 (experimental group) and n_2 is the sample size for group 2 (control group).

(e) Decision Rules

The decision rules are the same as for the one sample t-test.

Two-tailed test

If $t \leq -t_{crit}$ or $t \geq t_{crit}$, then reject H_0; otherwise, fail to reject H_0

One-tailed (upper-tailed, group 1 anticipated to have higher mean than group 2) test

If $t \geq t_{crit}$, then reject H_0; otherwise, fail to reject H_0

One-tailed (lower-tailed, group 1 anticipated to have lower mean than group 2) test

If $t \leq -t_{crit}$, then reject H_0; otherwise, fail to reject H_0

(f) Assumptions

The two independent samples t-test requires that the raw scores in both populations be normally distributed and independent. Also, the two populations should have equal (homogeneous) variances. The two group t-test is generally robust to non-normality and unequal variance (provided $n_1 \approx n_2$), but is not robust to dependence of observations.

(g) An Example

Recall the geography experiment. The scores for both groups are:

Experimental Group	Control Group
88	79
89	75
91	86
95	91
86	92
87	82
88	80
79	82
88	81
$\bar{X}_e$ = 87.889	$\bar{X}_c$ = 83.111
s = 4.256	s = 5.578
n = 9	n = 9

The experimental group has a mean of 87.889 and a standard deviation of 4.256, and the control group had a mean of 83.111 and a standard deviation of 5.578. There were 9 students in the experimental group and 9 students in the control group. So the two independent group t-test, with an $\alpha = .05$ and a non-directional test would be:

$$t = \frac{\bar{X}_e - \bar{X}_c}{s_{\bar{X}_e - \bar{X}_c}} = \frac{87.889 - 83.111}{\sqrt{\frac{18.114}{9} + \frac{31.114}{9}}} = \frac{4.778}{2.339} = 2.043$$

and the degrees of freedom are $df = n_1 + n_2 - 2 = 9 + 9 - 2 = 16$. The critical t is: $t_{crit} = 2.120$. The rejection regions are: $t \leq -2.120$, and $t \geq 2.120$, and the decision rule is:

If 2.043 ≤ -2.120 or 2.043 ≥ 2.120, then reject H_0; otherwise, FTR H_0

The correct decision is fail to reject H_0. One would therefore conclude the following:

> There is not a statistically significant difference in geography achievement between the experimental and control group for this sample at the .05 level of significance. This finding indicates achievement scores for geography students do not appear to differ between those who do and do not use the software Carmen SanDiego.

Note, however, what would happen if one hypothesized that the experimental group would have higher scores than the control group. If $\alpha = .05$, the critical value for an upper-tailed would be 1.746, so the decision rule would be:

If 2.043 ≥ 1.746, then reject H_0; otherwise, fail to reject H_0

Now H_0 is rejected, and one could conclude the following:

> The data indicate that students who learn with the computer program Carmen SanDiego show a statistically significant, at the .05 level, higher achievement score in U.S. geography. Thus, use of the software appears to benefit students.

(h) Confidence Intervals About Mean Differences
Recall the CI for a sample mean:

$$(1 - \alpha)\text{CI} = \bar{X} \pm \left({}_{1-\alpha/2}t_{critical}\right)\left(s_{\bar{X}}\right)$$

One may similarly compute a CI for the difference between two means. The formula is:

$$(1 - \alpha)\text{CI} = \left(\bar{X}_1 - \bar{X}_2\right) \pm \left({}_{1-\alpha/2}t_{critical}\right)\left(s_{\bar{X}_1 - \bar{X}_2}\right)$$

The .95CI for the above example is:

$$\begin{aligned}.95\text{CI} &= \left(\bar{X}_1 - \bar{X}_2\right) \pm \left({}_{.975}t_{critical}\right)\left(s_{\bar{X}_1 - \bar{X}_2}\right) \\ &= (87.889 - 83.111) \pm 2.12(2.339) \\ &= (4.778) \pm 4.959, \text{ or between -0.181 and 9.737}\end{aligned}$$

Since 0 is within this interval, H_0 will not be rejected.

Computer Analysis of Above Example

```
. ttest scores, by(group)

  Variable |      Obs        Mean   Std. Dev.
  ---------+---------------------------------
         0 |        9    83.11111    5.577734
         1 |        9    87.88889    4.255715
  ---------+---------------------------------
  combined |       18        85.5    5.404247

           Ho:  mean(x) = mean(y)  (assuming equal variances)
                      t = -2.04 with 16 d.f.
                Pr > |t| = 0.0579
```

(i) Strength of Association for Two Group t-test (effect size)
While a statistically significant t-test indicates that the two groups are probably not equal, the t-test does not indicate the strength of the association between the independent variable and the dependent variable. In the study just discussed, the independent variable (IV) is the presence or absence of the treatment, and the dependent variable (DV) is the posttest achievement score.

The question one may ask after rejecting H_0 is just how strong an impact does the treatment have on student achievement. One measure of the strength of the association between the treatment and the outcome is eta squared, η^2:

$$\eta^2 = \frac{t^2}{t^2 + df}$$

For example, the calculated t above was 2.043, so

$$\eta^2 = \frac{2.043^2}{2.043^2 + 16} = \frac{4.174}{4.174 + 16} = .207$$

The value obtained for η^2 may be interpreted in a manner identical to r^2, such as the variance explained or predicted in posttest scores by the treatment. In fact, if one calculates a Pearson's correlation between the two numerical variables listed in the table below (posttest scores and the indicator of treatment [1=treatment, 0=control]), the obtained r will be equal to .455 and the r^2 will be .207!

Posttest Scores	Indicator of Treatment	Treatment Condition
88	1	Experimental
89	1	Experimental
91	1	Experimental
95	1	Experimental
86	1	Experimental
87	1	Experimental
88	1	Experimental
79	1	Experimental
88	1	Experimental
79	0	Control
75	0	Control
86	0	Control
91	0	Control
92	0	Control
82	0	Control
80	0	Control
82	0	Control
81	0	Control

This should indicate to you that one may actually use a Pearson correlation to determine whether two groups are statistically different. For example, using the same experimental data, one could reproduce the same t value obtained from the two independent groups t-test using only the correlation r:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.455\sqrt{18-2}}{\sqrt{1-.207}} = \frac{1.82}{.891} = 2.043$$

In short, the two group independent t-test and the Pearson correlation coefficient provide identical inferential results. The two group t-test requires the calculation of η^2 in order to determine the strength of the relationship between the IV and DV.

(j) Effect Size (ES)

One may choose to relate to the reader the magnitude of the effect of the treatment by providing η^2. Another means of relaying this information, which is growing in importance in research today, is the standardized ES indicator.

ES, denoted in the researcher literature as d and/or Δ, may be calculated with one of two formulas. First, d is

$$d = \frac{\overline{X}_1 - \overline{X}_2}{SD_{within}}$$

where

$$\text{SD}_{\text{within}} = \sqrt{\frac{\sum\left(X-\overline{X}_1\right)^2+\sum\left(X-\overline{X}_2\right)^2}{\left(n_1-1\right)+\left(n_2-1\right)}}$$

$\text{SD}_{\text{within}}$ is essentially the average SD for the two groups.

Second, Δ is

$$\Delta = \frac{\overline{X}_1-\overline{X}_2}{\text{SD}_{\text{controlgroup}}}$$

where $\text{SD}_{\text{control group}}$ is simply the SD of the control group (if one is present).

Note that both d and Δ describe the magnitude of the difference between the two group means in standard deviation units. So, for example, if d or Δ = .2, then this indicates that the two group means differ by .2 standard deviations. The larger either d or Δ, the greater the difference between two groups, and, hence, the larger the effect of the treatment.

In the example used above the ES is

$$\text{SD}_{\text{within}} = \sqrt{\frac{\sum\left(X-\overline{X}_1\right)^2+\sum\left(X-\overline{X}_2\right)^2}{\left(n_1-1\right)+\left(n_2-1\right)}}$$

$$= \sqrt{\frac{144.908+248.913}{(9-1)+(9-1)}}$$

$$= \sqrt{\frac{393.821}{16}}$$

$$= \sqrt{24.614} = 4.961$$

$$\text{d} = \frac{\overline{X}_1-\overline{X}_2}{\text{SD}_{\text{within}}} = \frac{87.889-83.11}{4.961} = 0.963$$

If one wished to calculate Δ, then the corresponding ES is:

$$\Delta = \frac{\overline{X}_1-\overline{X}_2}{\text{SD}_{\text{controlgroup}}} = \frac{87.889-83.11}{5.578} = 0.857$$

Either ES is appropriate to use when an experimental group is compared to a control group. When two groups are compared and the two groups do not represent experimental and control (such as males vs. females), then one should use d as the measure of ES.

(k) Exercises

(1) Determine whether boys have a statistically different, at the 1% level, ITBS math score from girls. The mean math score for boys is 78 (s = 5.3) and the mean for girls is 73 (s = 6.1). There are 25 boys and 25 girls.

(a) What is the correct H_0 and H_1 in both written and symbol form?
(b) What are the critical and calculated t-values?

(2) Determine whether a statistical difference exists between men and women in weight:

Men:	156, 158, 175, 203, 252, 195
Women:	149, 119, 168, 123, 155, 126

(a) Test for a non-directional H_0 with $\alpha = .01$; what is the correct H_0, H_1?
(b) Test for a non-directional H_0 with $\alpha = .10$.
(c) Test the hypothesis that men will have lower weight, and set $\alpha = .10$. What is the correct H_0, H_1?

(3) Two classes of educational research were taught with two different methods of instruction, teacher guided (TG) and self paced (SP). Which had the better student achievement at the end of the quarter?

TG scores:	95, 93, 87, 88, 82, 92
SP scores:	78, 89, 83, 90, 78, 86

(a) Test for a non-directional H_0 with $\alpha = .01$; what is the correct H_0, H_1?
(b) Test for a non-directional H_0 with $\alpha = .10$.
(c) Test the hypothesis that TG will have higher scores, and set $\alpha = .05$. What is the correct H_0, H_1?

(l) Computer answers to exercises

Example 1

```
. ttesti 25 78 5.3 25 73 6.1

  Variable |       Obs        Mean    Std. Dev.
  ---------+-----------------------------------
         x |        25          78         5.3
         y |        25          73         6.1
  ---------+-----------------------------------
  combined |        50        75.5    6.193644

           Ho:  mean(x) = mean(y)  (assuming equal variances)
                      t = 3.09 with 48 d.f.
               Pr > |t| = 0.0033
```

Example 2

```
. ttest weight, by(sex)

  Variable |      Obs        Mean   Std. Dev.
---------+---------------------------------
       0 |        6         140    20.07984
       1 |        6    189.8333    35.89661
---------+---------------------------------
combined |       12    164.9167    38.02979

         Ho:  mean(x) = mean(y)  (assuming equal variances)
                    t = -2.97 with 10 d.f.
             Pr > |t| = 0.0141
```

Example 3

```
. ttest scores, by(groups)

  Variable |      Obs        Mean   Std. Dev.
---------+---------------------------------
       0 |        6          84     5.25357
       1 |        6        89.5    4.764452
---------+---------------------------------
combined |       12       86.75     5.57796

         Ho:  mean(x) = mean(y)  (assuming equal variances)
                    t = -1.90 with 10 d.f.
             Pr > |t| = 0.0867
```

4. Two Correlated Group t test (also called dependent samples t test)

The correlated t test allows the researcher to consider differences between two groups or sets of scores that are related to one-another. Under what conditions is one likely to find correlated or dependent samples or groups?

Condition 1
Before/After Studies; Multiple Measures on the Same Subject = This type of data occurs most often with pretest-treatment-posttest experimental designs. These types of designs are used to determine whether some treatment will change posttest scores relative to the pretest score. The pretest and posttest scores are related because the scores are taken from the same individuals, i.e., each person is measured twice.

Examples:
(a) A student takes the SAT, enrolls in an SAT enhancement class, and then retakes the SAT. Two scores from the same student exist.

(b) A teacher measured the reading performance of a third-grader, presented some treatment designed to increase reading performance, then remeasured the student's reading performance again (two scores from same individual).

(c) A PE teacher measures the vertical jumping ability of his class, provides his class a weight training program for one month, then remeasures vertical jumping ability of each student (two scores from same students).

Condition 2

Matched-Subjects = Two groups are involved in the study (experimental and control); and they are matched on some extraneous variable(s) that is likely to be related to the dependent variable being examined.

Examples:

(a) A teacher is interested in determining whether "Hooked on Phonics" increases third-grade students' reading performance. Using two groups of students, group A (the experimental group) will use "Hook on Phonics" for one month, and group B (the control) will be exposed to the usual reading lessons during the month. The teacher knows that IQ influences reading performance, so to control for the effects of IQ on the dependent variable (which is a posttest on reading performance), the researcher matches students in the two groups on their IQ levels in a fashion similar to the schematic below:

	Group A (treatment)	Group B (control)	
IQ score			IQ score
High (110+)	Beth and Sue	John and Ann	High (110+)
Middle (90-110)	Bob and Susan	Fred and Bill	Middle (90-110)
Low (<90)	Bryan and Bill	Josh and Walt	Low (<90)

In this scheme, students from both groups are matched according to their IQ levels. It is important to match on IQ since we would expect students with higher IQs to perform better on a reading test than students with lower IQs.

(b) As another example, one might make a comparison of faculty salary between men and women to determine whether sexual discrimination exists. It would be important to match men and women on academic rank since we know that assistant professors, on average, make less than associate and full professors.

Condition 3

Naturally occurring pairs = Natural pairs, such as husbands and wives, twins, brothers, sisters, brothers and sisters, parents and their children, etc. With naturally occurring pairs, one would expect the pairs to hold similar feelings, beliefs, attitudes, etc., so their scores will generally be related to one-another.

Examples:

(a) Determining whether husbands' attitudes toward politics are similar to their wives. Since people tend to marry others like themselves, one would expect that most husbands and wives to hold similar political views.

(b) Determining whether boys' IQ differs from girls' IQ. Since brothers and sisters are similar genetically, one might anticipate the two to have similar IQs, that is, their IQs are likely to be related; therefore, brothers and sisters need to be matched.

Hypothesis Formulation:

The hypothesis tested with the correlated t-test is the same as in the independent t-test.

For example, suppose one is in determining whether boys or girls get higher math scores on the ITBS. Clearly, intelligence plays an important part in determining mathematics performance, so this is a factor that needs to be controlled through matching. One may formulate several hypotheses, as demonstrated below.

Non-directional:
The average ITBS math scores will differ between boys and girls; their scores will differ on average.

H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$, or

H_0: $\mu_1 - \mu_2 = 0.00$ and H_1: $\mu_1 - \mu_2 \neq 0.00$

where μ_1 represents for group 1 (boys) and μ_2 represents group 2 (girls).

Directional (group 1 has higher mean than group 2):
Boys will score higher, on average, than girls.

H_0: $\mu_1 \leq \mu_2$ and H_1: $\mu_1 > \mu_2$, or

H_0: $\mu_1 - \mu_2 \leq 0.00$ and H_1: $\mu_1 - \mu_2 > 0.00$

Directional (group 1 has lower mean than group 2):
Boys will score lower, on average, than girls.

H_0: $\mu_1 \geq \mu_2$ and H_1: $\mu_1 < \mu_2$, or

H_0: $\mu_1 - \mu_2 \geq 0.00$ and H_1: $\mu_1 - \mu_2 < 0.00$

Theoretical Formula for Correlated t test
The t ratio for the correlated t test can be calculated as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{SE_d}$$

where $\overline{X}_1 - \overline{X}_2$ is the difference between the two sample means, and the denominator is the standard error of the difference, SE_d.

Note that this is identical to the formula for the two independent sample t test. The difference between the formulas for the independent and the correlated t test occurs in the calculation of the standard error of the difference.

For the correlated t test the standard error of the difference is calculated as:

$$SE_d = s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2(r_{12})\left(\frac{s_1}{\sqrt{n}}\right)\left(\frac{s_2}{\sqrt{n}}\right)}$$

but in the independent t test it is assumed that the groups are not related (scores between groups are not correlated), so the standard error looses the correlated term in the formula, i.e.:

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}-2(r_{12})\left(\frac{s_1}{\sqrt{n}}\right)\left(\frac{s_2}{\sqrt{n}}\right)} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}-2(0)\left(\frac{s_1}{\sqrt{n}}\right)\left(\frac{s_2}{\sqrt{n}}\right)} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}-0} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

If there is no correlation, then the SE_d formula reduces to the SE_d formula given in the independent samples t-test. In short, the primary difference between the two t tests is the calculation of the standard error of the difference, SE_d.

<u>Practical Formula for Correlated t test</u>
To calculate the correlated t statistic, the following formula is easier to use:

$$t = \frac{\bar{d}-\mu_d}{\sqrt{s_d^2/n}} = \frac{\bar{d}}{\sqrt{s_d^2/n}} = \frac{\bar{d}}{SE_d}$$

where $\bar{d}$ is the mean of the differences between pairs of scores, i.e.,

$$\bar{d} = \frac{\sum d}{n}$$

and SE_d is the standard error of the differences:

$$SE_d = \sqrt{s_d^2/n}$$

where s_d^2 is the variance of the difference scores, and is calculated like a regular variance, i.e.,

$$s_d^2 = \frac{\sum\left(d-\bar{d}\right)^2}{n-1}$$

In short, the correlated t test is may be viewed as the <u>mean of the difference,</u> $\bar{d}$, divided by the <u>standard error of the difference</u>, SE_d.

$$t = \frac{\bar{d}}{SE_d}$$

<u>Degrees of Freedom</u>:
The df for the correlated t test is calculated as:

df = n - 1

where n represents the <u>number of pairs</u> across the two groups.

<u>Decision Rules</u>:
The decision rules are the same as in the independent two-sample t test:

Two-tailed tests:
If $t \leq -t_{crit}$ or $t \geq t_{crit}$, then reject H_0; otherwise, fail to reject H_0

One-tailed (upper-tailed) test
If $t \geq t_{crit}$, then reject H_0; otherwise, fail to reject H_0

One-tailed (lower-tailed) test
If $t \leq -t_{crit}$, then reject H_0; otherwise, fail to reject H_0

Note that $\mathbf{t}_{crit}$ symbolizes the critical t-value found in t tables and is different from t, which is the calculated t-ratio obtained from sample data.

Example 1:
Suppose we are interested in determining whether salary differs between men and women faculty at GSU. When randomly selecting subjects for the study, it is important that we take into consideration their academic rank since full professors make more money than associate professors, and associates make more money than assistant professors, on average. Test the hypothesis of no difference between men and women, H_0: $\mu_1 = \mu_2$, at the 5% significance level.

Rank	Income Men	Difference	Income Women	Rank
Full	Bill = 48,000	- 3000	Beth = 51,000	Full
Full	Bob = 51,000	6000	Bertha = 45,000	Full
Associate	Billy = 43,000	- 1000	Bobby = 44,000	Associate
Associate	Burt = 38,500	2500	Bonnie = 36,000	Associate
Assistant	Brando = 24,500	- 500	Brenda = 25,000	Assistant
Assistant	Bart S. = 28,000	5000	Bette = 23,000	Assistant
Assistant	Brent = 33,000	7000	Beulah = 26,000	Assistant

$$\bar{d} = \frac{\sum d}{n} = \frac{16000}{7} = 2285.714$$

Difference D	Mean of Difference $\bar{d}$	Deviation $(d-\bar{d})$	Deviation Squared $(d-\bar{d})^2$
- 3000	- 2285.714	-5285.714	27938772.49
6000	- 2285.714	3714.286	13795920.49
- 1000	- 2285.714	-3285.714	10795916.49
2500	- 2285.714	214.286	45918.49
- 500	- 2285.714	-2785.714	7760202.49
5000	- 2285.714	2714.286	7367348.49
7000	- 2285.714	4714.286	22224492.49

$$SE_d = \sqrt{s_d^2 / n}$$

where s_d^2 is the variance of the difference scores, and is calculated like a regular variance, i.e.,

$$s_d^2 = \frac{\sum\left(d - \bar{d}\right)^2}{n-1}$$

$$\text{SE}_\text{d} = \sqrt{s_d^2 / n} = \sqrt{\frac{\left(\frac{\sum\left(d-\bar{d}\right)^2}{n-1}\right)}{n}} = \sqrt{\frac{\left(\frac{8992857143}{7-1}\right)}{7}}$$

$$= \sqrt{\left(\frac{8992857143}{6}\right)\Big/7} = \sqrt{\frac{1498809524}{7}} = 1463.269$$

so the t value will be:

$$t = \frac{\bar{d}}{\sqrt{s_d^2 / n}} = \frac{\bar{d}}{SE_d} = \frac{2285.714}{1463.269} = 1.562$$

The critical values at the .05 level for df = n - 1 = 6 are ± 2.447, so fail to reject H_0 and conclude that salaries do not appear to differ between men and women faculty at GSU even after controlling for academic rank.

What do you think would happen if an independent samples t test were used to analyze the above data?

Calculate the regular independent t test and see: M_{men} = 38000, SD_{men} = 10012.49, M_{women} = 35714.29, and SD_{women} = 11250.4.

Which is more powerful (recall that power represents the probability of rejecting a false H_0), the independent or correlated t test? Why?

<u>Example 2</u>:
A researcher wishes to discover whether or not the intake of orange juice increases the potassium level in the bloodstream. A group of 12 elderly patients are selected from those in a nursing home, where previous diet has been controlled. Potassium blood levels are measured for each subject. Next, each subject is given a quart of orange juice, and, two hours later, potassium levels are again measured. Test the difference in potassium levels at the 5% level. The data are as follows (the scaled scores represent potassium blood levels):

Subject	Before Potassium Level	After Potassium Level	Difference	Mean of Difference $\bar{d}$	Deviation $(d-\bar{d})$	Deviation Squared $(d-\bar{d})^2$
1	26	25	1	-2	3	9
2	25	28	-3	-2	-1	1
3	24	27	-3	-2	-1	1
4	23	26	-3	-2	-1	1
5	23	25	-2	-2	0	0
6	21	23	-2	-2	0	0
7	19	21	-2	-2	0	0
8	17	19	-2	-2	0	0
9	17	16	1	-2	3	9
10	16	19	-3	-2	-1	1
11	15	18	-3	-2	-1	1
12	14	17	-3	-2	-1	1

$$\bar{d} = \frac{\sum d}{n} = \frac{-24}{12} = -2.00$$

and the standard error of the difference is:

$$SE_d = \sqrt{s_d^2 / n} = \sqrt{\frac{\left(\frac{\sum (d-\bar{d})^2}{n-1}\right)}{n}} = \sqrt{\frac{\left(\frac{24}{12-1}\right)}{12}} = \sqrt{\frac{\left(\frac{24}{11}\right)}{12}} = \sqrt{\frac{2.182}{12}} = .426$$

so the calculated t value will be:

$$t = \frac{\bar{d}}{\sqrt{s_d^2 / n}} = \frac{\bar{d}}{SE_d} = \frac{-2}{.426} = -4.695$$

The hypothesis was that orange juice will increase potassium in the blood stream, i.e., the pretest scores will be lower than the posttest scores. This hypothesis indicates that a lower-tailed test is needed since H_0: $\mu_1 \geq \mu_2$ and H_1: $\mu_1 < \mu_2$.

The critical value at the .05 level for df = 12 - 1 = 11 is - 1.796, so we reject H_0, and conclude that orange juice does appear to increase the amount of potassium in the blood stream for elderly people.

<u>Exercises</u>:

(1) A researcher is interested in determining whether typing speed is affected by the kind of typewriter (electric versus manual) used. A group of student typists, equally experienced on both types of machines, are randomly selected and are matched on the basis of their typing speed (error-free words per minute). One group is then tested on an electric machine and the other group on a manual machine. Test H_0 at the 1% significance level. The data are as follows:

(a) What are the correct H_0 and H_1 in both written and symbolic form?
(b) What is (are) the critical value(s)?
(c) What is the obtained (calculated) t value?
(d) Did you reject or fail to reject H_0?
(e) Write your conclusion as if explaining the results to non-statisticians.

Pair	Typing Speed	Electric	Manual
1	High	50	42
2	High	65	60
3	Middle	72	65
4	Middle	90	85
5	Middle	48	50
6	Low	62	60
7	Low	75	60
8	Low	50	51
9	Low	68	59

(2) A psychologist wishes to look at the relationship between frustration and positive attitude. He hypothesizes that frustration affects attitude. Students are given an "Attitude Toward Psychologists" (ATP) instrument prior to taking their first exam in an introductory psychology course. After completing the ATP instrument, the students are then administered their course exam. The teacher, a psychologist, made the exam especially difficult in an attempt to frustrate his students. After completing the exam, all students were asked to fill out the ATP instrument again. High scores on the ATP instrument indicate more positive attitudes toward psychologist. The data are as follows:

Subject	Before Exam Scores	After Exam Scores
1	44	20
2	20	10
3	35	30
4	42	26
5	35	30
6	30	20
7	34	30
8	30	22
9	19	21
10	17	20
11	25	17
12	30	15
13	32	25
14	31	26
15	34	30
16	20	25
17	31	24
18	37	19
19	32	30
20	33	28
21	16	15

(a) What are the correct H_0 and H_1 in both written and symbolic form?
(b) What is (are) the critical value(s)?
(c) What is the obtained (calculated) t value?
(d) Did you reject or fail to reject H_0?
(e) Did frustration influence the students' attitude toward psychologists? Write your conclusion as if explaining the results to non-statisticians.

For additional examples, see chapter exercises in book and notes on course web page.