Lecture 8: Two-Sample Inferences on Means

SW Chapters 7 and 9

Suppose you have two populations of interest, say populations 1 and 2, and you are interested in comparing their (unknown) population means, μ_1 and μ_2. Inferences on the unknown population means are based on samples from each population. In practice, most problems fall into one of two categories.

1. **Independent samples**, where the sample taken from population 1 has no effect on which observations are selected from population 2, and vice versa. (SW Chapter 7)

2. **Paired** or dependent samples, where experimental units are paired based on factors related or unrelated to the variable measured. (SW Chapter 9)

The distinction between paired and independent samples is best mastered through a series of examples.

Example Suppose you are interested in whether the $CaCO_3$ (calcium carbonate) level in the Atrisco well field, which is the water source for Albuquerque, is changing over time. To answer this question, the $CaCO_3$ level was recorded at each of 15 wells at two time points. These data are paired. The two samples are the Times 1 and 2 observations.

Example To compare state incomes, a random sample of New Mexico households was selected, and an independent sample of Arizona households was obtained. It is reasonable to assume independent samples.

Example Suppose you are interested in whether the husband or wife is typically the heavier smoker among couples where both adults smoke. Data are collected on households. You measure the average number of cigarettes smoked by each husband and wife within the sample of households. These data are paired, i.e. you have selected husband wife pairs as the basis for the samples. It is reasonable to believe that the responses within a pair are related, or correlated.

Although the focus here will be on comparing population means, you should recognize that in paired samples you may also be interested, as in the problems above, in how obser-

vations compare within a pair. These goals need not agree, depending on the questions of interest. Note that with paired data, the sample sizes are equal, and equal to the number of pairs.

Two independent samples: CI and test using pooled variance

These methods assume that the populations have normal frequency curves, with equal population standard deviations, i.e. $\sigma_1 = \sigma_2$. Let $(n_1,\ \bar{Y}_1,\ s_1)$ and $(n_2,\ \bar{Y}_2,\ s_2)$ be the sample sizes, means and standard deviations from the two samples.

The standard CI for $\mu_1 - \mu_2$ is given by

$$\begin{aligned} Lower &= (\bar{Y}_1 - \bar{Y}_2) - t_{crit} SE_{\bar{Y}_1 - \bar{Y}_2} \\ Upper &= (\bar{Y}_1 - \bar{Y}_2) + t_{crit} SE_{\bar{Y}_1 - \bar{Y}_2} \end{aligned}$$

The t-statistic for testing $H_0 : \mu_1 - \mu_2 = 0 \quad (\mu_1 = \mu_2)$ against $H_A : \mu_1 - \mu_2 \neq 0 \quad (\mu_1 \neq \mu_2)$ is given by

$$t_s = \frac{\bar{Y}_1 - \bar{Y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}.$$

The standard error of $\bar{Y}_1 - \bar{Y}_2$ used in both the CI and the test is given by

$$SE_{\bar{Y}_1 - \bar{Y}_2} = s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Here the **pooled variance estimator**,

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

is our best estimate of the common population variance. The pooled estimator of variance is a weighted average of the two sample variances, with more weight given to the larger sample. If $n_1 = n_2$ then s^2_{pooled} is the average of s_1^2 and s_2^2.

The critical value t_{crit} for CI and tests is obtained in usual way from a t-table with $df = n_1 + n - 2$. For the test, follow the one-sample procedure, with the new t_s and t_{crit}.

The pooled CI and tests are sensitive to the normality and equal standard deviation assumptions. The observed data can be used to assess the reasonableness of these assumptions. You should look at boxplots and stem-and-leaf displays to assess normality, and should check whether $s_1 \approx s_2$ to assess the assumption $\sigma_1 = \sigma_2$.

Satterthwaite's Method

Satterthwaite's method assumes normality, but does not require equal population standard deviations. Satterthwaite's procedures are somewhat conservative, and adjust the SE and df to account for unequal population variances. Satterthwaite's method uses the same CI and test statistic formula, with a modified standard error:

$$SE_{\bar{Y}_1-\bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

and degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Note that $df = n_1+n_2-2$ when $n_1 = n_2$ and $s_1 = s_2$. The Satterthwaite and pooled variance procedures usually give similar results when $s_1 \approx s_2$.

SW use Satterthwaite's method for CI and tests, and only briefly touch upon the use of the pooled procedures. The df formula for Satterthwaite's method is fairly complex, so SW propose a conservative df formula that uses the minimum of $n_1 - 1$ and $n_2 - 1$ instead.

Examples: SW examples 7.7 and 7.8 pages 228-230.

JMP-IN Implementation

JMP-IN does the pooled t-test and CI in the **Fit Y by X Platform**. Although it is natural to enter the data into the spreadsheet as two columns, one for each sample, the data must then be stacked as shown in LAB. JMP-in also reports a test for equal population variances, and the results of a Welsh test, which for two samples is analogous to the Satterthwaite test of equal means. **JMP-IN** does not report a Satterthwaite CI for the difference in means. Kristina will show you how to set up the data and carry out the tests in Lab.

Example Androstenedione levels in diabetics

The data consist of independent samples of diabetic men and women. For each individual, the scientist recorded their androstenedione level (a hormone level). Let μ_1 = mean androstenedione level for the population of diabetic men, and μ_2 = mean androstenedione level for the population of diabetic women. We are interested in comparing the population means given the observed data.

ROW	men	women
1	217	84
2	123	87
3	80	77
4	140	84
5	115	73
6	135	66
7	59	70
8	126	35
9	70	77
10	63	73
11	147	56
12	122	112
13	108	56
14	70	84
15		80
16		101
17		66
18		84

I will compute a **pooled** 95% CI for $\mu_1 - \mu_2$ and the p-value for testing $H_0 : \mu_1 - \mu_2 = 0$ (or $\mu_1 = \mu_2$), using the **JMP-IN** summary statistics on the page after next. I will note that the Satterthwaite procedures are probably more reasonable here than the pooled methods, a point that we will return to when we examine the **JMP-IN** output. I chose the pooled method because the hand calculations are easier.

For the males, $n_1 = 14$, $\bar{Y}_1 = 112.50$ and $s_1 = 42.75$. For females $n_2 = 18$, $\bar{Y}_2 = 75.83$, and $s_2 = 17.24$. For a 95% CI, the $t-$critical value based on 30 degrees of freedom ($df = n_1 + n_2 - 2 = 14 + 18 - 2 = 30$) is $t_{.025} = 2.042$. The pooled variance estimate is

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(14 - 1) * 42.75^2 + (18 - 1) * 17.24^2}{30} = \frac{28811.01}{30} = 960.4$$

which gives $s_{pooled} = \sqrt{960.4} = 30.99$, and so

$$SE_{\bar{Y}_1 - \bar{Y}_2} = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 30.99\sqrt{\frac{1}{14} + 118} = 11.04.$$

Multiplying the critical value and the standard error gives

$$t_{crit} SE_{\bar{Y}_1 - \bar{Y}_2} = 2.042 * 11.04 = 22.55.$$

The difference in means is $\bar{Y}_1 - \bar{Y}_2 = 112.50 - 75.83 = 36.67$, so the CI has endpoints 36.67 ± 22.55. The lower limit is 14.12. The upper limit is 59.22. Thus, we are 95% confident that $\mu_1 - \mu_2$ is between 14.12 and 59.22. I will interpret this result after performing the t-test.

The t-statistic for testing H_0 is given by

$$t_s = \frac{\bar{Y}_1 - \bar{Y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}} = \frac{36.67}{11.04} = 3.32.$$

The p-value for the test is the area under the $t-$curve (with 30 *df*) outside ± 3.32, or twice the area to the right of 3.32. Looking at SW's table, the upper tail area is between .005 and .0005, so the p-value is between $2 * .0005 = .001$ and $2 * .005 = .01$. For a 5% test we would reject H_0 in favor of the alternative that μ_1 and μ_2 are not equal.

If the p-value is .05 or smaller, the 95% CI for $\mu_1 - \mu_2$ will not contain 0. Put another way, if 0 is not a plausible value for the difference between μ_1 and μ_2 then the only logical conclusion is that μ_1 and μ_2 differ.

The CI interpretation is made easier by recognizing that we concluded the population means are different, so the direction of difference must be consistent with that seen in the observed data, where the sample mean andro level for men exceeds that for women. Thus, the limits on the CI for $\mu_1 - \mu_2$ tells us how much larger the population mean is for men than for women. In particular, we are 95% confident the population mean andro level for diabetic men is between 14.12 and 59.22 greater than the population mean andro level for diabetic women. This general approach for unraveling the CI works even when the limits are both negative.

Moving on to the **JMP-IN** output we see that the boxplots suggest that the distributions are reasonably symmetric. However, the normality assumption is unreasonable due to the presence of outliers in each sample. The equal population standard deviation assumption also appears unreasonable. The sample standard deviation for men is noticeably larger than

the women's standard deviation, even with outliers in the women's sample. This is reinforced by the fairly small p-values associated with the various tests of equal variances.

I am more comfortable with the Satterthwaite analysis here than the pooled variance analysis. However, I would interpret all results cautiously, given the unreasonableness of the normality assumption. Using the Satterthwaite test, the data strongly suggest that the population mean androstenedione levels are different. In particular, the Welsh (Satterthwaite) p-value for testing $H_0 : \mu_1 - \mu_2 = 0$ is .0079. I will note that the 95% Satterthwaite CI for $\mu_1 - \mu_2$ extends from 11.0 to 62.4, which implies that we are 95% confident that the population mean andro level for diabetic men exceeds that for diabetic women by at least 11.0 but by no more than 62.4.

As a comparison, let us examine the output for the pooled procedure. The p-value for the pooled t-test is .0024, whereas the 95% confidence limits are 14.1 and 59.2. These results are qualitatively similar to the Satterthwaite conclusions, and agree with the hand calculations considered earlier.

One-Sided Tests

SW discuss one-sided tests for two-sample problems, where the null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$ but the alternative is directional, either $H_A : \mu_1 - \mu_2 < 0$ (i.e. $\mu_1 < \mu_2$) or $H_A : \mu_1 - \mu_2 > 0$ (i.e. $\mu_1 > \mu_2$). Once you understand the general form of rejection regions and p-values for one-sample tests, the one-sided two-sample tests do not pose any new problems. Use the $t-$statistic, with the appropriate tail of the $t-$distribution to define critical values and p-values. Unfortunately, one-sided two-sample tests are not directly implemented in $\mathbf{JMP-IN}$, but you can easily get p-values using a little thought.

For example, suppose in the androstenedione problem, theory suggests that you test $H_0 : \mu_1 - \mu_2 = 0$ against the alternative hypothesis that the population mean for men exceeds that for women: i.e. $H_A : \mu_1 - \mu_2 > 0$. For illustration, consider the pooled $t-$test output. The test statistic is positive, $t_s = 3.32$ and the two-sided test p-value is .0024. One-half this tabled p-value, or .0012, is the area under the $t-$distribution to the right of

the observed test statistic, so the p-value for the upper one-sided test must be .0012. With some thought, you can get one-sided confidence bounds as well.

Paired Analysis

With paired data, inferences on $\mu_1 - \mu_2$ are based on the sample of differences within pairs. By taking differences within pairs, two dependent samples are transformed into one sample, which contains the relevant information for inferences on $\mu_d = \mu_1 - \mu_2$. To see this, suppose the observations within a pair are Y_1 and Y_2. Then within each pair, compute the difference $d = Y_1 - Y_2$. If the Y_1 data are from a population with mean μ_1 and the Y_2 data are from a population with mean μ_2, then the d's are a sample from a population with mean $\mu_d = \mu_1 - \mu_2$. Furthermore, if the sample of differences comes from a normal population, then we can use standard one sample techniques to test $\mu_d = 0$ (i.e. $\mu_1 = \mu_2$), and to get a CI for $\mu_d = \mu_1 - \mu_2$.

Let $\bar{d} = \bar{Y}_1 - \bar{Y}_2$ be the sample mean of the differences (which is also the mean difference), and let s_d be the sample standard deviation of the differences. The standard error of $\bar{d}$ is $SE_{\bar{d}} = s_d/\sqrt{n}$, where n is the number of pairs. The paired $t-$test (two-sided) CI for μ_d is given by $\bar{d} \pm t_{crit} SE_{\bar{d}}$. To test $H_0 : \mu_d = 0 \quad (\mu_1 = \mu_2)$ against $H_A : \mu_d \neq 0 \quad (\mu_1 \neq \mu_2)$, use

$$t_s = \frac{\bar{d} - 0}{SE_{\bar{d}}}$$

to compute a p-value as in a two-sided one-sample test. One-sided tests are evaluated in the usual way for one-sample tests on means.

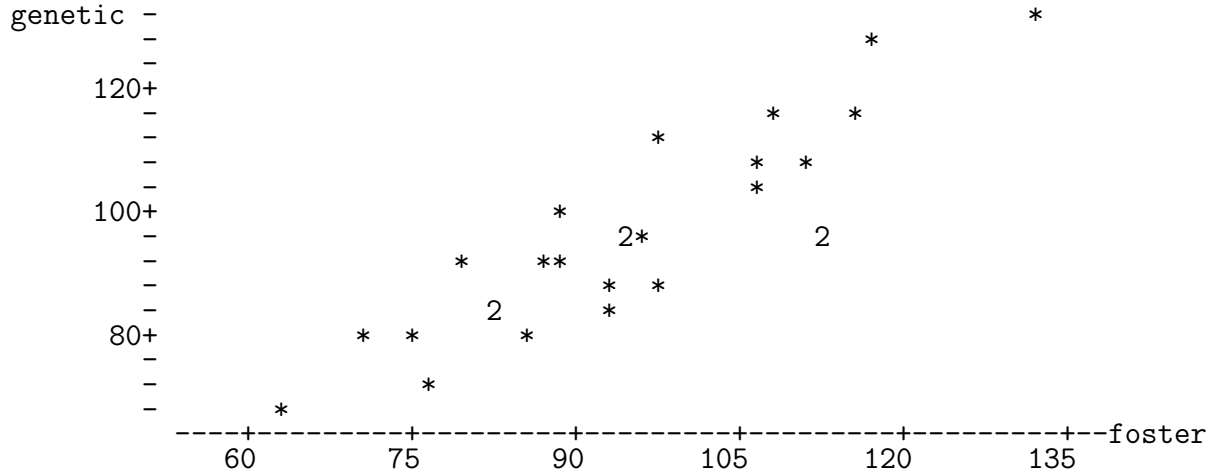
A graphical analysis of paired data focuses on the **sample of differences**, and not on the original samples. In particular, the normality assumption is assessed on the sample of differences.

I will do a simple hand calculation, then show you how to do the analysis in **JMP-IN** on another data set.

Example Paired analysis of data on twins

Burt (1966) presented data on IQ scores for identical twins that were raised apart, one by foster parents and one by the genetic parents. Assuming the data are a random sample of twin pairs, consider comparing the population mean IQs for twins raised at home to those raised by foster parents. Let μ_f=population mean IQ for twin raised by foster parents, and μ_g=population mean IQ for twin raised by genetic parents.

ROW	foster	genetic	genetic-foster
1	82	82	0
2	80	90	10
3	88	91	3
4	108	115	7
5	116	115	-1
6	117	129	12
7	132	131	-1
8	71	78	7
9	75	79	4
10	93	82	-11
11	95	97	2
12	88	100	12
13	111	107	-4
14	63	68	5
15	77	73	-4
16	86	81	-5
17	83	85	2
18	93	87	-6
19	97	87	-10
20	87	93	6
21	94	94	0
22	96	95	-1
23	112	97	-15
24	113	97	-16
25	106	103	-3
26	107	106	-1
27	98	111	13



This plot of IQ scores shows that scores are related within pairs of twins. This is consistent with the need for a paired analysis. A boxplot and stem and leaf display for the differences (not provided) show no marked deviation from normality. The boxplot is centered at zero, so one would not be too surprised if a test result is insignificant.

Let us compute a 95% CI for $\mu_g - \mu_f$. The sample mean, standard deviation and standard error of the differences (genetic minus foster) within twin pairs are $\bar{d} = .19$, $s_d = 7.74$, and $SE_{\bar{d}} = s_d/\sqrt{n} = 1.49$, respectively. There are $n = 27$ pairs, so the df for a one-sample t-CI or test are $27 - 1 = 26$. The critical value for a 95% CI is $t_{.025} = 2.056$. The CI endpoints are $.10 \pm 2.056 * 1.49$, or $.19 \pm 3.06$. The lower limit is -2.87 whereas the upper limit is 3.25. Rounding off to the nearest integer, we are 95% confident that the difference in population mean IQ scores for the two populations is between plus or minus 3 points. We would not reject $H_0 : \mu_g - \mu_f = 0$ at the 5% level. (Why?)

I will ask Kristina to show you the **JMP-IN** analysis of these data in your next Lab.

Remark: I could have defined the difference to be the foster IQ score minus the genetic IQ score. How would this change the conclusions?

JMP-IN Analysis

The most natural way to enter paired data is as two columns, one for each treatment group, and then to use the **JMP-IN** calculator to create a column of differences. Then, use the **Distribution of Y Platform** to do the analysis on the sample of differences. Alternatively, you can use the **Matched Pairs** platform directly (i.e. do not need to created differences within pairs), but this will not give a boxplot or stem and leaf display of the differences. I illustrate both approaches in the following example.

Example Paired comparisons of two sleep remedies.

The following data give the amount of sleep gained in hours from two sleep remedies, A and B, applied to 10 individuals who have trouble sleeping an adequate amount. Negative values imply sleep loss. In 9 of the 10 individuals, the sleep gain on B exceeded that on

A. Let μ_A = population mean sleep gain (among troubled sleepers) on remedy A, and μ_B = population mean sleep gain (among troubled sleepers) on remedy B. We are interested in whether the remedies lead to the same gain in sleep on average, so consider testing $H_0 : \mu_A - \mu_B = 0$ or equivalently $\mu_d = 0$, where $\mu_d = \mu_A - \mu_B$.

ROW	A	B	A-B
1	0.7	1.9	-1.2
2	-1.6	0.8	-2.4
3	-0.2	1.1	-1.3
4	-1.2	0.1	-1.3
5	0.1	-0.1	0.2
6	3.4	4.4	-1.0
7	3.7	5.5	-1.8
8	0.8	1.6	-0.8
9	0.0	4.6	-4.6
10	2.0	3.0	-1.0

JMP-IN output is included on the next page, first for the analysis on the created differences (A-B), then using the **Matched Pairs** platform. The observed distribution of differences between B and A is reasonably symmetric, but heavy tailed due to the presence of an outliers in the lower tail. The normality assumption of the standard one-sample t-test and CI are suspect here. I will continue with the analysis nonetheless.

The p-value for testing H_0 is .0044. We'd reject H_0 at the 5% or 1% level, and conclude that the population mean sleep gains on the remedies are different. We are 95% confident that $\mu_A - \mu_B$ is between -2.43 and -.61. That is, we are 95% confident that μ_B is between .61 and 2.43 hours greater than μ_A. Again, these results must be reported with caution, because the normality assumption is unreasonable. However, the presence of outliers tends to make the t-test and CI conservative, so we'd expect to find similar conclusions if we used the nonparametric methods discussed later in the semester. The matched pair output is minimal, but hopefully you see the agreement with the earlier output on the test statistic, p-value, and CI.

Query: In what order should the remedies be given to the patients?