8 Two-Sample Inferences for Means

SW Chapters 7 and 9

Comparing Two Sets of Measurements

Suppose you have collected data on one variable from two (independent) samples and you are interested in "comparing" the samples. What tools are good to use?

Example: Head Breadths

In this analysis, we will compare a physical feature of modern day Englishmen with the corresponding feature of some of their ancient countrymen. The Celts were a vigorous race of people who once populated parts of England. It is not entirely clear whether they simply died out or merged with other people who were the ancestors of those who live in England today. A goal of this study might be to shed some light on possible genetic links between the two groups.

The study is based on the comparison of maximum head breadths (in millimeters) made on unearthed Celtic skulls and on a number of skulls of modern-day Englishmen. The data are given below. We have a sample of 18 Englishmen and an independent sample of 16 Celtic skulls.

```
Row  ENGLISH  CELTS

  1      141    133
  2      148    138
  3      132    130
  4      138    138
  5      154    134
  6      142    127
  7      150    128
  8      146    138
  9      155    136
 10      158    131
 11      150    126
 12      140    120
 13      147    124
 14      148    132
 15      144    132
 16      150    125
 17      149
 18      145
```

What features of these data would we likely be interested in comparing? The centers of the distributions, the spreads within each distribution, the distributional shapes, etc.

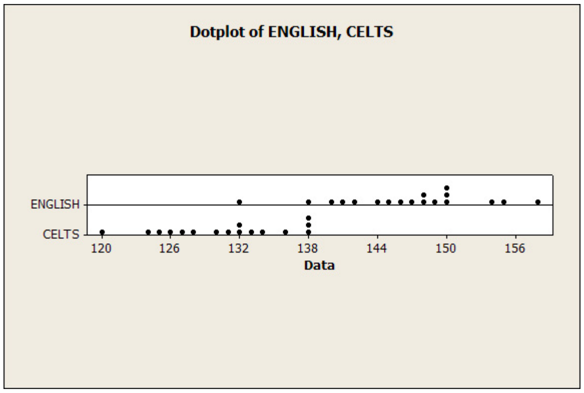
These data can be analyzed in Minitab as either STACKED data (1 column containing both samples, with a separate column of labels or **subscripts** to distinguish the samples) or UNSTACKED (2 columns, 1 for each sample). The form of subsequent Minitab commands will depend on which data mode is used. It is often more natural to enter UNSTACKED data, but with large data bases STACKED data is the norm (for reasons that I will explain verbally). It is easy to create STACKED data from UNSTACKED data and vice-versa. Graphical comparisons usually require the plots for the two groups to have the same scale, which is easiest to control when the data are STACKED.

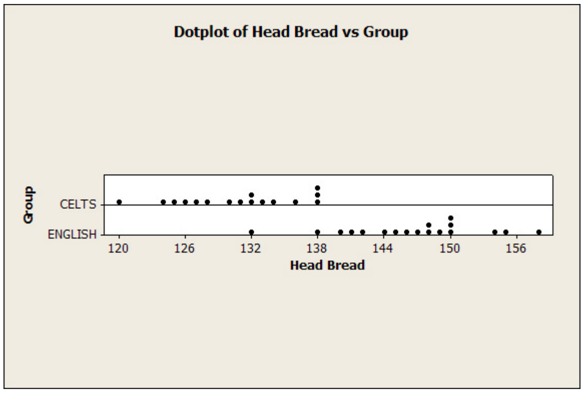
The head breadth data was entered as two separate columns, c1 and c2 (i.e. UNSTACKED). To STACK the data, follow: Data > Stack > Columns. In the dialog box, specify that you wish to stack the English and Celt columns, putting the results in c3, and storing the subscripts in c4. The output below shows the data in the worksheet after stacking the two columns.

```
Data Display

                           Head
Row  ENGLISH  CELTS  Bread  Group
  1      141    133    141  ENGLISH
  2      148    138    148  ENGLISH
  3      132    130    132  ENGLISH
  4      138    138    138  ENGLISH
  5      154    134    154  ENGLISH
  6      142    127    142  ENGLISH
  7      150    128    150  ENGLISH
  8      146    138    146  ENGLISH
  9      155    136    155  ENGLISH
 10      158    131    158  ENGLISH
 11      150    126    150  ENGLISH
 12      140    120    140  ENGLISH
 13      147    124    147  ENGLISH
 14      148    132    148  ENGLISH
 15      144    132    144  ENGLISH
 16      150    125    150  ENGLISH
 17      149           149  ENGLISH
 18      145           145  ENGLISH
 19                    133  CELTS
 20                    138  CELTS
 21                    130  CELTS
 22                    138  CELTS
 23                    134  CELTS
 24                    127  CELTS
 25                    128  CELTS
 26                    138  CELTS
 27                    136  CELTS
 28                    131  CELTS
 29                    126  CELTS
 30                    120  CELTS
 31                    124  CELTS
 32                    132  CELTS
 33                    132  CELTS
 34                    125  CELTS
```

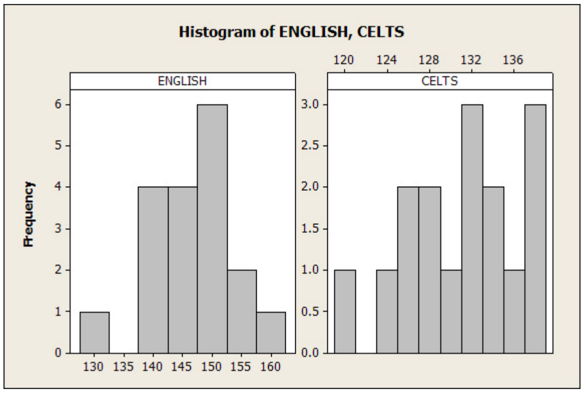
Plotting head breadth data:

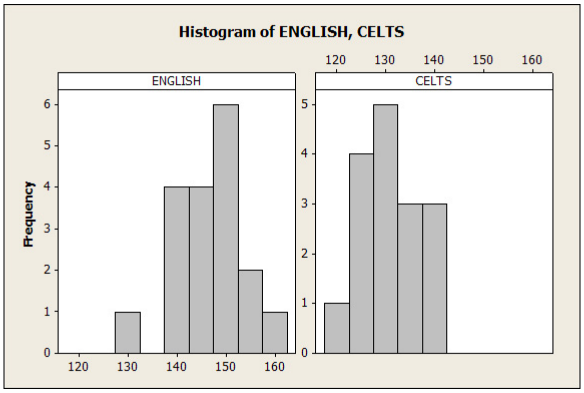
1. A dotplot with the same scale for both samples is obtained from the UNSTACKED data by selecting Multiple Y's with the Simple option, and then choosing C1 and C2 to plot. For the STACKED data, choose One Y With Groups, select c3 as the plotting variable and c4 as the Categorical variable for grouping. There are minor differences in the display generated – I prefer the Stacked data form. In the following, the Unstacked form is on the left, the stacked form on the right.



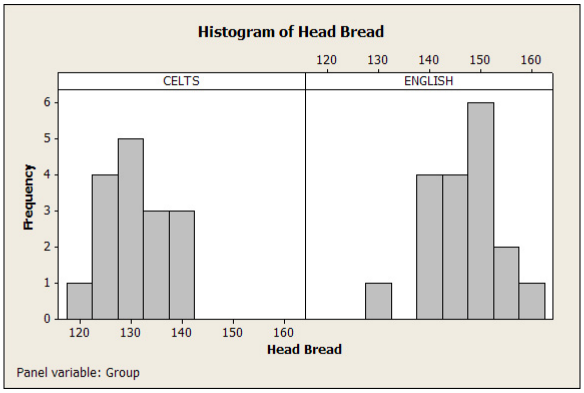


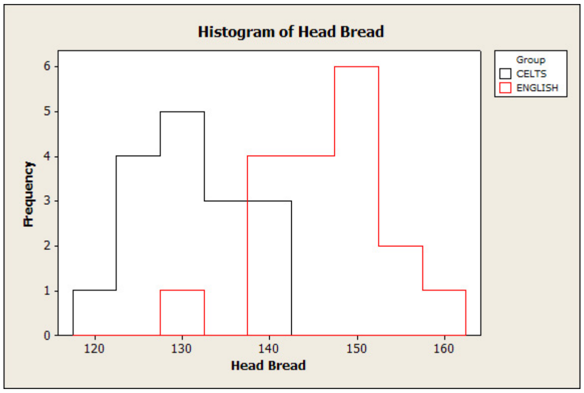
2. Histograms are hard to compare unless you make the scale and actual bins the same for both. Click on *Multiple Graphs* and check *In separate panels of the same graph*. That puts the two graphs next to each other. The left graph below is the unstacked form with only that option. Next check *Same X, including same bins* so you have some basis of comparison. The right graph below uses that option. Why is that one clearly preferable?





The stacked form is more straightforward (left graph below). Click on Multiple Graphs and define a By Variable. The Histogram With Outline and Groups is an interesting variant (right graph below).





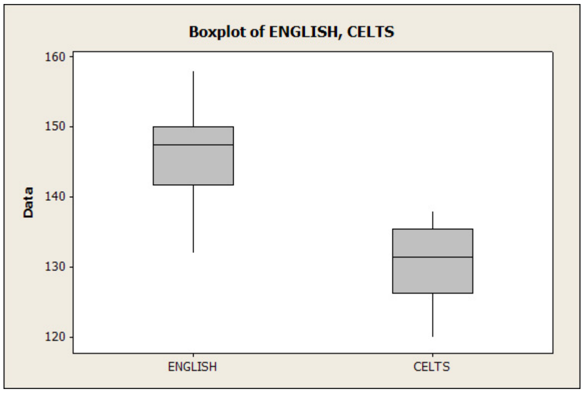
3. Stem-and-leaf displays in unstacked data can be pretty useless. The stems are not forced to match (just like with histograms). It is pretty hard to make quick comparisons with the following:

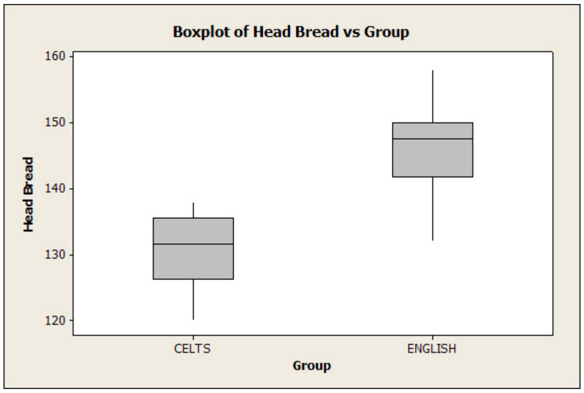
```
Stem-and-Leaf Display: ENGLISH, CELTS

Stem-and-leaf of ENGLISH  N  = 18
Leaf Unit = 1.0


 1   13  2
 2   13  8
 6   14  0124
(6)  14  567889
 6   15  0004
 2   15  58

Stem-and-leaf of CELTS  N  = 16
Leaf Unit = 1.0


 1  12  0
 1  12
 3  12  45
 5  12  67
 6  12  8
 8  13  01
 8  13  223
 5  13  4
 4  13  6
 3  13  888
```

Unfortunately, Minitab seems to be using an old routine for stem-and-leaf plots, and you cannot use stacked data with the Group variable we created. Minitab is wanting a numeric group variable in this case (their older routines always required numeric). Follow Data > Code > Text to Numeric in order to create a new variable in C5 with 1 for ENGLISH and 2 for CELTS. Now the stems at least match up:

```
Stem-and-Leaf Display: Head Bread

Stem-and-leaf of Head Bread  C5 = 1    N  = 18
Leaf Unit = 1.0


 1   13  2
 2   13  8
 6   14  0124
(6)  14  567889
 6   15  0004
 2   15  58

Stem-and-leaf of Head Bread  C5 = 2    N  = 16
Leaf Unit = 1.0


 2   12  04
 6   12  5678
(6)  13  012234
 4   13  6888
```

4. For boxplots, either Unstacked (Multiple Y's) or Stacked (One Y with Groups) works well. Again, I prefer the default from the stacked form, but it really doesn't matter much. Which is which below?





Many of the data summaries will work on either Unstacked or Stacked data. For the head breadth data, **descriptive statistics** output is given below, obtained from both the Stacked data (specifying data in c3 with c4 as a "by variable") and the Unstacked data (specifying data in separate columns c1 and c2).

```
Descriptive Statistics: ENGLISH, CELTS <<<<<<---------Unstacked

Variable   N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
ENGLISH   18   0  146.50     1.50   6.38   132.00  141.75  147.50  150.00
CELTS     16   0  130.75     1.36   5.43   120.00  126.25  131.50  135.50

Variable  Maximum
ENGLISH    158.00
CELTS      138.00

Descriptive Statistics: Head Bread  <<<<<<---------Stacked

Variable    Group     N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median
Head Bread  CELTS    16   0  130.75     1.36   5.43   120.00  126.25  131.50
            ENGLISH  18   0  146.50     1.50   6.38   132.00  141.75  147.50

Variable    Group        Q3  Maximum
Head Bread  CELTS    135.50   138.00
            ENGLISH  150.00   158.00
```

Salient Features to Notice

The stem and leaf displays and boxplots indicate that the English and Celt samples are slightly skewed to the left. There are no outliers in either sample. It is not unreasonable to operationally assume that the population frequency curves (i.e. the histograms for the populations from which the samples were selected) for the English and Celtic head breadths are normal.

The sample means and medians are close to each other in each sample, which is not surprising given the near symmetry and the lack of outliers.

The data suggest that the typical modern English head breadth is greater than that for Celts. The data sets have comparable spreads, as measured by either the standard deviation or the IQR (you need to calculate IQR or ask for it in the above summaries).

Two-Sample Methods: Paired Versus Independent Samples

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 The English and Celt head breadth samples are independent

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 observations 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 $(n_1,\ \overline{Y}_1,\ s_1)$ and $(n_2,\ \overline{Y}_2,\ s_2)$ be the sample sizes, means and standard deviations from the two samples.

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

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

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

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

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

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

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 $df = n_1 + n_2 - 2$. 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 $\sigma_1 = \sigma_2$. Formal tests of these assumptions will be discussed later.

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:

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

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 229-230.

Minitab does the pooled and Satterthwaite analyses, either on stacked or unstacked data. Follow the steps STAT > BASIC STATISTICS > 2 sample t. In the dialog box, specify the data to be analyzed, choose a CI level, and check if you wish to assume equal variances. The output will contain a p-value for a two-sided tests of equal population means and a CI for the difference in population means. If you check the box for assuming equal variances you will get the pooled method, otherwise the output is for Satterthwaite's method.

An important point to note: You can request individual values plots and side-by-side boxplots as an option in the dialog box - and the data need not be stacked.

Example: Head Breadths

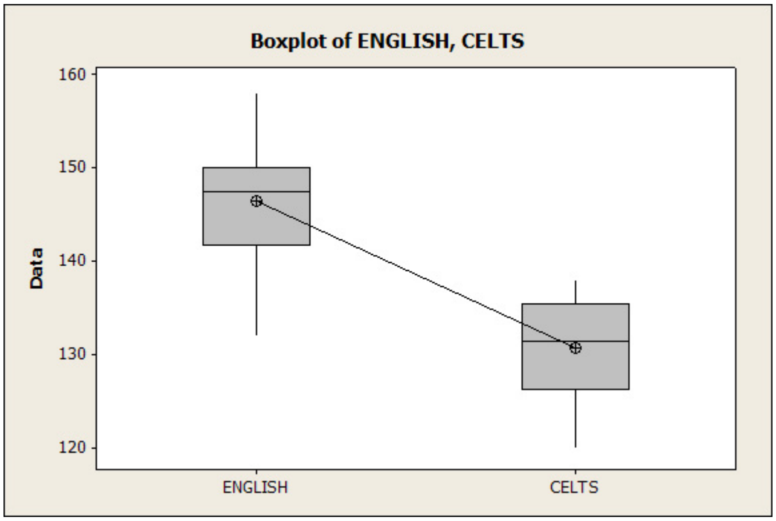
The English and Celts are independent samples. We looked at boxplots and stem and leaf displays, which suggested that the normality assumption for the *t*-test is reasonable. The Minitab output below shows the English and Celt sample standard deviations are fairly close, so the pooled and Satterthwaite results should be comparable. The pooled analysis is preferable here, but either is appropriate.

The form of the output will tell you which sample corresponds to population 1 and which corresponds to population 2. This should be clear from the dialog box if you use the UNSTACKED data, as I did. Here the CI tells us about the difference between the English and Celt population means, so I need to define μ_1 = population mean head breadths for all Englishmen and μ_2 = population mean head breadths for Celts.

```
Two-Sample T-Test and CI: ENGLISH, CELTS

Two-sample T for ENGLISH vs CELTS    <<<--------- Pooled

           N    Mean  StDev  SE Mean
 ENGLISH  18  146.50   6.38      1.5
 CELTS    16  130.75   5.43      1.4


 Difference = mu (ENGLISH) - mu (CELTS)
 Estimate for difference:  15.7500
 95% CI for difference:  (11.5809, 19.9191)
 T-Test of difference = 0 (vs not =): T-Value = 7.70  P-Value = 0.000  DF = 32
 Both use Pooled StDev = 5.9569


 Two-Sample T-Test and CI: ENGLISH, CELTS

 Two-sample T for ENGLISH vs CELTS   <<<--------- Satterthwaite

           N    Mean  StDev  SE Mean
 ENGLISH  18  146.50   6.38      1.5
 CELTS    16  130.75   5.43      1.4
```

```
Difference = mu (ENGLISH) - mu (CELTS)
Estimate for difference:  15.7500
95% CI for difference:  (11.6158, 19.8842)
T-Test of difference = 0 (vs not =): T-Value = 7.77  P-Value = 0.000  DF = 31
```

The boxplot, asked for optionally, is nice here – it show means, and connects them to emphasize the analysis being done.



Remarks: The $T =$ entry on the T-TEST line is t_{obs}, whereas $P =$ is the p-value.

The pooled analysis strongly suggests that $H_0 : \mu_1 - \mu_2 = 0$ is false, given the 2-sided p-value of .0000. We are 95% confident that $\mu_1 - \mu_2$ is between 11.6 and 19.9 mm. That is, we are 95% confident that the population mean head breadth for Englishmen (μ_1) exceeds the population mean head breadth for Celts (μ_2) by between 11.6 and 19.9 mm.

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 head breadth for Englishmen exceeds that for the Celts. Thus, the limits on the CI for $\mu_1 - \mu_2$ tells us how much larger the mean is for the English population (i.e. between 11.6 and 19.9 mm).

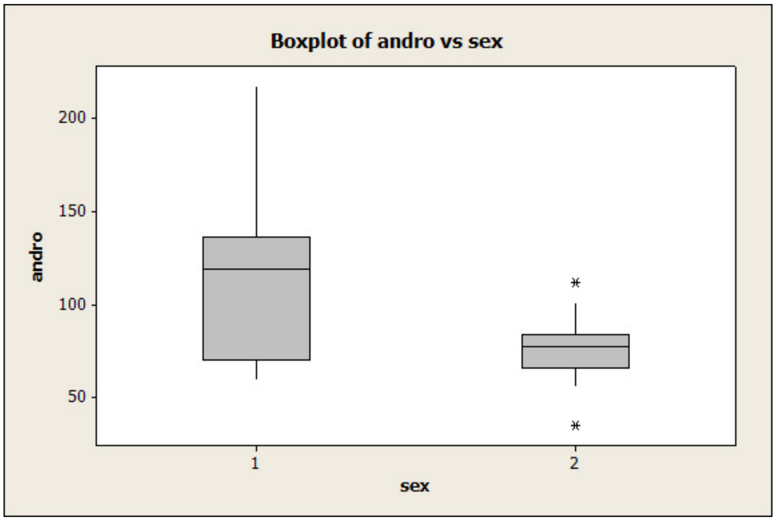
The interpretation of the analysis is always simplified if you specify the first sample in the dialog box (for an UNSTACKED analysis) to be the sample with the larger mean. Why?

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 - Mark McGwire's favorite dietary supplement). 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.

The raw data and Minitab output is given below. The boxplots suggest that the distributions are reasonably symmetric. However, the normality assumption for the women is unreasonable due to the presence of outliers. 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.

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.



```
Data Display

 Row    men  women  andro    sex

   1    217     84    217      1
   2    123     87    123      1
   3     80     77     80      1
   4    140     84    140      1
   5    115     73    115      1
   6    135     66    135      1
   7     59     70     59      1
   8    126     35    126      1
   9     70     77     70      1
  10     63     73     63      1
  11    147     56    147      1
  12    122    112    122      1
  13    108     56    108      1
  14     70     84     70      1
  15            80     84      2
  16           101     87      2
  17            66     77      2
  18            84     84      2
  19                   73      2
  20                   66      2
  21                   70      2
  22                   35      2
  23                   77      2
  24                   73      2
  25                   56      2
  26                  112      2
  27                   56      2
  28                   84      2
  29                   80      2
  30                  101     2
  31                   66     2
  32                   84     2

 Descriptive Statistics: men, women

 Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
 men       14   0  112.5     11.4   42.8     59.0   70.0   118.5  136.3    217.0
 women     18   0  75.83     4.06  17.24    35.00  66.00   77.00  84.00   112.00

 Stem-and-Leaf Display: andro

 Stem-and-leaf of andro  sex = 1    N  = 14
 Leaf Unit = 10


 1  0  5
 4  0  677
 5  0  8
 7  1  01
 7  1  2223
 3  1  44
 1  1
 1  1
 1  2  1


 Stem-and-leaf of andro  sex = 2    N  = 18
 Leaf Unit = 10


 1   0  3
 3   0  55
 (7)  0  6677777
 8   0  888888
 2   1  01
```

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 .008. 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 .002, whereas the 95% confidence limits are 14.1 and 59.2. That is, we are 95% confident that the population mean andro level for men exceeds that for women by at least 14.1 but by no more than 59.2. These results are qualitatively similar to the Satterthwaite conclusions.

```
Two-Sample T-Test and CI: men, women

Two-sample T for men vs women

        N   Mean  StDev  SE Mean
men    14  112.5   42.8       11 women  18   75.8   17.2      4.1


Difference = mu (men) - mu (women)
Estimate for difference:  36.6667
95% CI for difference:  (10.9577, 62.3756)
T-Test of difference = 0 (vs not =): T-Value = 3.02  P-Value = 0.008  DF = 16
```

```
Two-Sample T-Test and CI: men, women

Two-sample T for men vs women

        N   Mean  StDev  SE Mean
men    14  112.5   42.8       11 women  18   75.8   17.2      4.1


Difference = mu (men) - mu (women) Estimate for difference:  36.6667
95% CI for difference:  (14.1124, 59.2210)
T-Test of difference = 0 (vs not =): T-Value = 3.32  P-Value = 0.002  DF = 30
Both use Pooled StDev = 30.9914
```

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. One-sided two-sample tests are directly implemented in Minitab, by specifying the type of test in the dialog box. One-sided confidence bounds are given with the one-sided tests.

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 $\bar{d} = \overline{Y}_1 - \overline{Y}_2$ 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.

Minitab Analysis

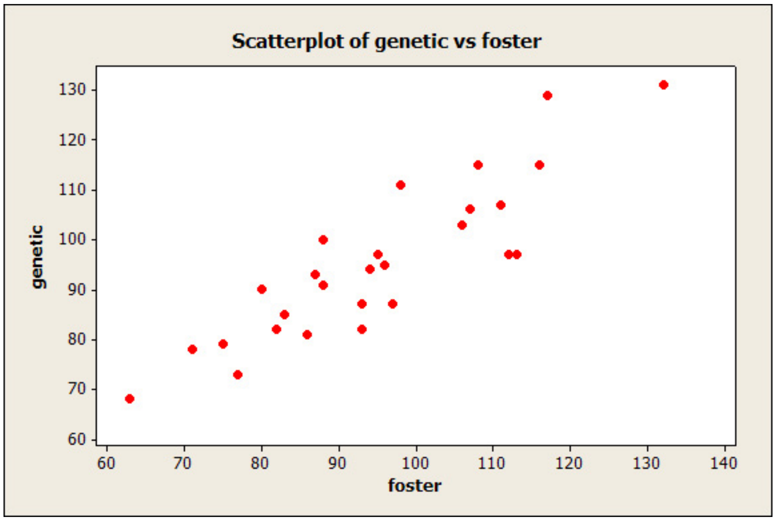
The most natural way to enter paired data is as two columns, one for each treatment group. At this point you can use the Minitab calculator to create a column of differences, and do the usual one-sample graphical and inferential analysis on this column of differences, or you can do the paired analysis directly without this intermediate step.

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.

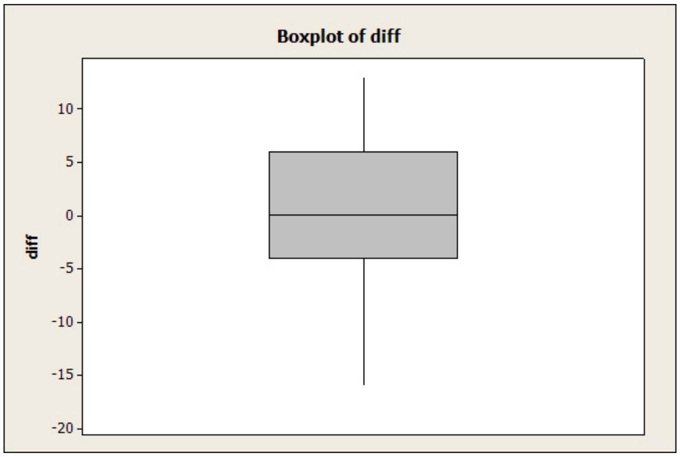
I created the data in the worksheet (c1=foster; c2=genetic), and computed the differences between the IQ scores for the children raised by the genetic and foster parents (c3=diff=genetic-foster). I also made a scatter plot of the genetic versus foster IQ scores.

```
Data Display

 Row  foster  genetic   diff

   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.

Given the sample of differences, I created a boxplot and a stem and leaf display, neither which showed marked deviation from normality. The boxplot is centered at zero, so one would not be too surprised if the test result is insignificant.



```
Stem-and-Leaf Display: diff

Stem-and-leaf of diff      N  = 27
Leaf Unit = 1.0


   2   -1 65
   4   -1 10
   6   -0 65
  (8)  -0 44311110
  13    0 02234
   8    0 5677
   4    1 0223
```

Given the sample of differences, I generated a one-sample CI and test (i.e. STAT > BASIC STATISTICS > 1-sample t). The hypothesis under test is $\mu_d = \mu_g - \mu_f = 0$. The p-value for this test is large. We do not have sufficient evidence to claim that the population mean IQs for twins raised apart are different. This is consistent with the CI for μ_d given below, which covers zero.

```
One-Sample T: diff

Test of mu = 0 vs not = 0

Variable   N      Mean     StDev   SE Mean           95% CI              T      P
diff      27  0.185185  7.736214  1.488835  (-2.875159, 3.245529)  0.12  0.902
```

Alternatively, I can generate the test and CI directly from the raw data in two columns, following: STAT > BASIC STATISTICS > paired-t, and specifying genetic as the first sample and foster as the second. This gives the following output, which leads to identical conclusions to the earlier analysis. If you take this approach, you can get high quality graphics in addition to the test and CI.

You might ask why I tortured you by doing the first analysis, which required creating and analyzing the sample of differences, when the alternative and equivalent second analysis is so much easier. (A later topic deals with non-parametric analyses of paired data for which the differences must be first computed.)

```
Paired T-Test and CI: genetic, foster

Paired T for genetic - foster

             N      Mean     StDev   SE Mean
genetic     27   95.2963   15.7353    3.0283
foster      27   95.1111   16.0823    3.0950
Difference  27  0.185185  7.736214  1.488835

95% CI for mean difference: (-2.875159, 3.245529)
T-Test of mean difference = 0 (vs not = 0): T-Value = 0.12
P-Value = 0.902
```

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

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.

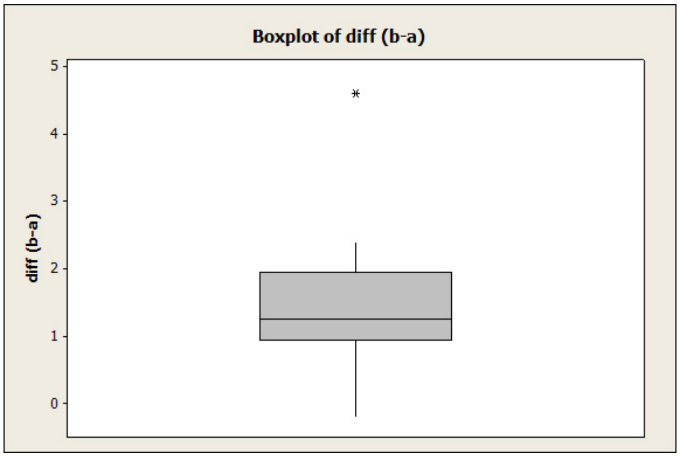
Let μ_A = population mean sleep gain (among troubled sleepers) on remedy A, and μ_B = population mean sleep gain (among troubled sleepers) on remedy B. Consider testing $H_0 : \mu_B - \mu_A = 0$ or equivalently $\mu_d = 0$, where $\mu_d = \mu_B - \mu_A$.

The observed distribution of differences between B and A is slightly skewed to the right, with a single outlier in the upper tail. The normality assumption of the standard one-sample t-test and CI are suspect here. I will continue with the analysis.

```
Data Display

                   diff
Row     a     b   (b-a)
  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
```

```
 Stem-and-Leaf Display: diff (b-a)

 Stem-and-leaf of diff (b-a)  N  = 10
 Leaf Unit = 0.10


 1   -0  2
 2   0   8
(6)  1   002338
 2   2   4
 1   3
 1   4   6
```

```
 One-Sample T: diff (b-a)

 Test of mu = 0 vs not = 0


 Variable     N     Mean    StDev  SE Mean          95% CI           T      P
 diff (b-a)  10  1.52000  1.27174  0.40216  (0.61025, 2.42975)  3.78  0.004
```



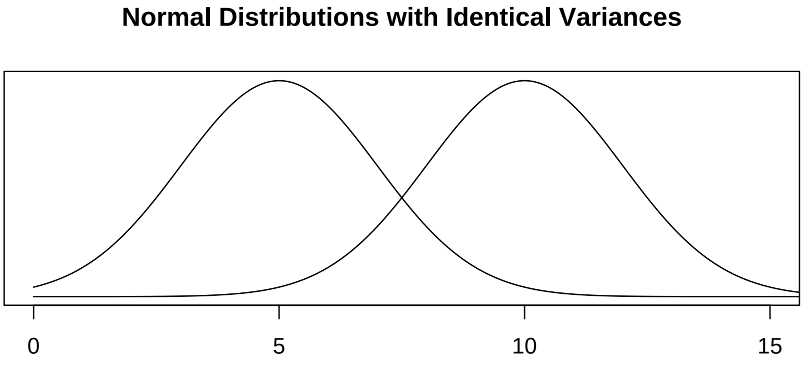
The p-value for testing H_0 is .004. We'd reject H_0 at the 5% or 1% level, and conclude that the population mean sleep gains on the remedies are different. We are 95% confident that μ_B exceeds μ_A by between .61 and 2.43 hours. Again, these results must be reported with caution, because the normality 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.

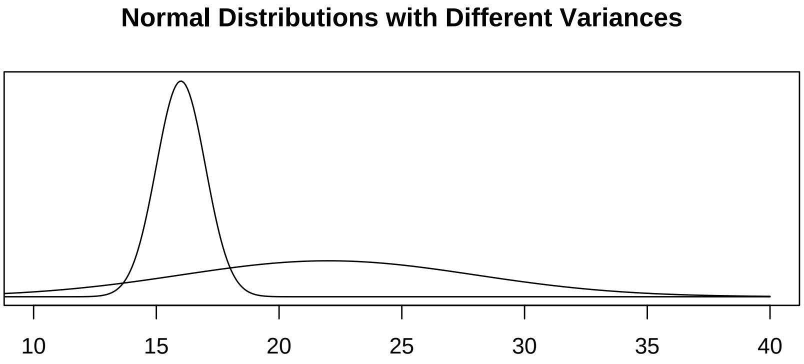
Query: In what order should the remedies be given to the patients?

Should You Compare Means?

The mean is the most common feature on which two distributions are compared. You should not, however, blindly apply the two-sample tests (paired or unpaired) without asking yourself whether the means are the relevant feature to compare. This issue is not a big concern when, as highlighted in the first graph below, the two (normal) populations have equal spreads or standard deviations. In such cases the difference between the two population means is equal to the difference between any fixed percentile for the two distributions, so the mean difference is a natural measure of difference.

Consider instead the hypothetical scenario depicted in the bottom pane below, where the population mean lifetimes using two distinct drugs for a fatal disease are $\mu_1 = 16$ months from time of diagnosis and $\mu_2 = 22$ months from time of diagnosis, respectively. The standard deviations under the two drugs are $\sigma_1 = 1$ and $\sigma_2 = 6$, respectively. The second drug has the higher mean lifetime, but at the expense of greater risk. For example, the first drug gives you a 97.7% chance of living at least 14 months, whereas the second drug only gives you a 90.8% chance of living at least 14 months. Which drug is best? It depends on what is important to you, a higher expected lifetime or a lower risk of dying early.





Nonparametric Procedures

Usually the biggest problems with assumptions of normality occur when we see extreme skewness and/or outliers. The first remedy most statisticians try in such cases is to transform the data using logs or another appropriate transformation to obtain approximate normality on the transformed scale. That often works well but does not handle nearly all problems. Nonparametric procedures are a set of methods designed as alternatives to procedures like t-tests and t-confidence intervals that can be applied even when sampling is not from a normal distribution. I will cover these in a very cursory fashion – this is actually a huge topic on its own.

Minitab implements some of the more popular methods if you follow the path `Stat > Nonparametrics`. The first three options are `1-Sample Sign`, `1-Sample Wilcoxon`, and `Mann-Whitney`. The Sign Test is an alternative to the 1-Sample t-test and makes no real assumption about the shape of the distribution sampled from; it focuses on the population *median* rather than the mean, however. The Wilcoxon Signed Rank test also is an alternative to the 1-Sample t-test; the only assumption about the distribution sampled from is that it is symmetric. The Mann-Whitney test is an alternative to the 2-Sample t-test. It focuses on differences in population medians, and assumes only that the two population distributions have the same general shape.

The Sign Test is pretty inefficient to use for data actually sampled from a normal distribution, but it protects against arbitrarily large outliers. The Wilcoxon and Mann-Whitney tests, if they are appropriate, are very efficient (just as powerful) relative to the t-test, and they also provide great protection against the bad effects of outliers.

Let's look at the Sign Test and Wilcoxon tests for the data on sleep remedies (paired data give rise to 1-Sample methods applied to the differences).

```
One-Sample T: diff (b-a)        <<<<<<<<< COMPARE WITH T

Test of mu = 0 vs not = 0


Variable     N     Mean    StDev  SE Mean         95% CI           T      P
diff (b-a)  10  1.52000  1.27174  0.40216  (0.61025, 2.42975)  3.78  0.004


Sign CI: diff (b-a)               <<<<<<<<< ASK FOR CI AND TEST SEPARATELY


Sign confidence interval for median

                                      Confidence
                           Achieved    Interval
             N  Median  Confidence  Lower  Upper  Position
diff (b-a)  10   1.250      0.8906  1.000  1.800         3
                            0.9500  0.932  2.005       NLI  <<-- USE THIS
                            0.9785  0.800  2.400         2


Sign Test for Median: diff (b-a)

Sign test of median =  0.00000 versus not = 0.00000

             N  Below  Equal  Above       P  Median
diff (b-a)  10      1      0      9  0.0215   1.250
```

```
Wilcoxon Signed Rank CI: diff (b-a)

                                          Confidence
                Estimated    Achieved      Interval
             N     Median  Confidence  Lower  Upper
diff (b-a)  10       1.30        94.7   0.80   2.70

Wilcoxon Signed Rank Test: diff (b-a)

Test of median = 0.000000 versus median not = 0.000000

                  N
                 for   Wilcoxon              Estimated
             N  Test  Statistic      P          Median
diff (b-a)  10    10       54.0  0.008           1.300
```

There is very little difference among these results. The sign test has the shortest CI (but it is for a population median, not mean). For real interpretation, though, your conclusions would not depend on which of these procedures you used. That at least makes you more comfortable if you go ahead and report the results of the t-test.

Let's go back to the androstenedione data set where we saw a problem with outliers. For purposes of illustration, we'll compare the Mann-Whitney to the 2-Sample t-test. Again, there is no real difference in a practical sense. I am uncomfortable with the Mann-Whitney here since the shapes do not really look the same.

```
Two-Sample T-Test and CI: men, women

 Two-sample T for men vs women

         N   Mean  StDev  SE Mean
 men    14  112.5   42.8       11
 women  18   75.8   17.2      4.1


 Difference = mu (men) - mu (women)
 Estimate for difference:  36.6667
 95% CI for difference:  (10.9577, 62.3756)
 T-Test of difference = 0 (vs not =): T-Value = 3.02  P-Value = 0.008  DF = 16

 Mann-Whitney Test and CI: men, women

         N  Median
 men    14  118.50
 women  18   77.00


 Point estimate for ETA1-ETA2 is 38.00
 95.4 Percent CI for ETA1-ETA2 is (3.99,56.01)
 W = 293.5
 Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0185
 The test is significant at 0.0183 (adjusted for ties)
```

Finally, to see that there really can be a difference, let's return to the income data from several lectures ago. The two large outliers pretty well destroy any meaning to the t-interval, but the sign-interval makes a lot of sense for a population median.

```
Data Display
```

```
Income
      7   1110      7      5      8     12      0      5      2      2     46
      7


One-Sample T: Income

Variable   N      Mean    StDev  SE Mean          95% CI
Income    12  100.917  318.008   91.801  (-101.136, 302.969)



Sign CI: Income


Sign confidence interval for median

                                      Confidence
                          Achieved     Interval
         N  Median  Confidence  Lower  Upper  Position
Income  12    7.00      0.8540   5.00   8.00         4
                        0.9500   2.79  10.95       NLI
                        0.9614   2.00  12.00         3

**** REMARK: NLI stands for non-linear interpolation
```

SW do discuss the Mann-Whitney test in Section 7.11.