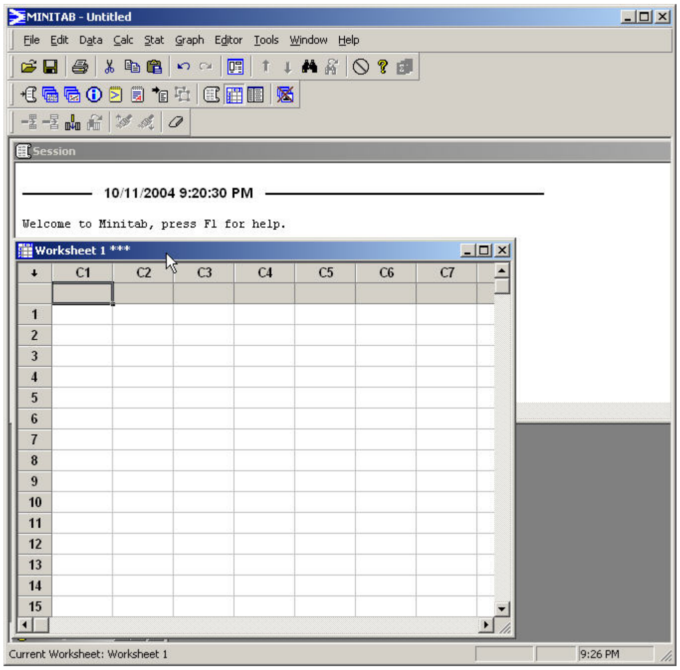
1 Introduction to Minitab

Minitab is a statistical analysis software package. The software is freely available to all students and is downloadable through the 'Technology Tab' at `my.calpoly.edu`. When you first launch Minitab, you will see the following screen:



Notice that the active area contains two important regions: the **Worksheet** and the **Session** windows. The **Worksheet** is a spreadsheet interface where you can input, sort, and otherwise manipulate data.

The **Worksheet** window in the adjacent picture has been resized and repositioned for the purpose of this tutorial. You may choose to resize/move the various windows for your convenience.

The **Session** window will contain a copy of the commands you invoke along with any statistical analyses you may perform.

1.1 Worksheet Format `.mtw` versus Project Format `.mpj`

Once you input data, you may store the worksheet by clicking on `File -> Save Current Worksheet as`. This will create a Minitab Worksheet file in the form of `filename.mtw`.

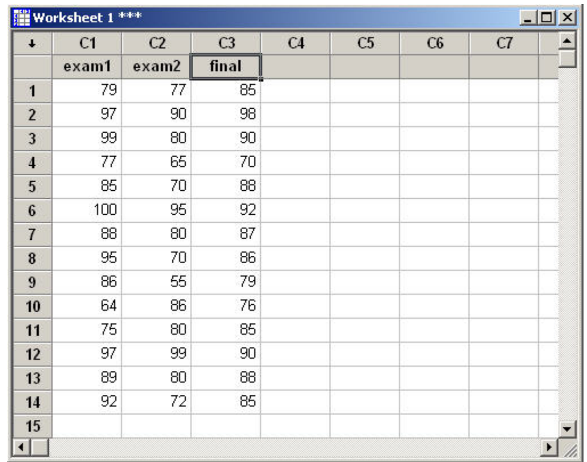
After you perform some tasks in Minitab with a particular worksheet (e.g. create graphs and invoke analyses) you can save all the work you have done by choosing `File -> Save Project as`. This will create a Minitab Project file in the form of `filename.mpj`. **NOTE:** If, after doing some work, you choose to save the file as a Minitab Worksheet, you will only save the information created in the spreadsheet but will lose **all** other work performed.

2 Inputing Data into the Worksheet

Data can be placed into Minitab in various ways. The most convenient way is to open a Minitab Worksheet file that already contains the data. Many textbook data sets are available as `.mtw` or `.mpj` files found on an accompanying textbook CDROM and/or accompanying website.

Typing data directly into Minitab is simple. Point the cursor to the first column and type in the first data entry in row '**1**' of column '**C1**'. You may also choose to create a column/variable label by typing the label in the grey cell just above row '**1**' of the column of interest.

After typing in some data values for three variables (`exam1`, `exam2`, and `final`) we have the following picture:



Worksheet 1 ***

↓	C1	C2	C3	C4	C5	C6	C7
	exam1	exam2	final				
1	79	77	85				
2	97	90	98				
3	99	80	90				
4	77	65	70				
5	85	70	88				
6	100	95	92				
7	88	80	87				
8	95	70	86				
9	86	55	79				
10	64	86	76				
11	75	80	85				
12	97	99	90				
13	89	80	88				
14	92	72	85				
15							

If you have data in another file or program (such as Excel), you can easily cut/paste the data into the worksheet. After copying the data, go to the Minitab worksheet window, select the first row of a particular column, and select 'Paste' from the `Edit` menu (or select 'Paste' with a right click).

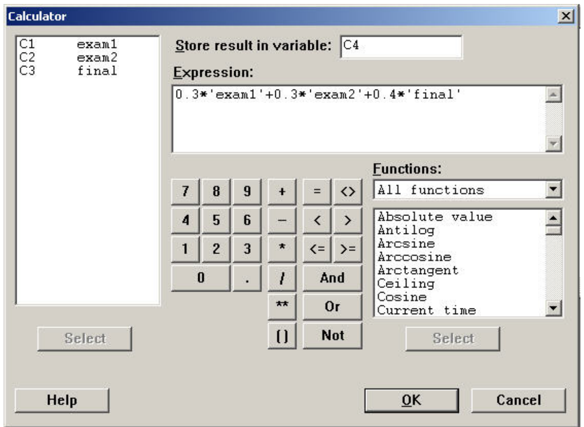
3 Manipulating Data in the Worksheet

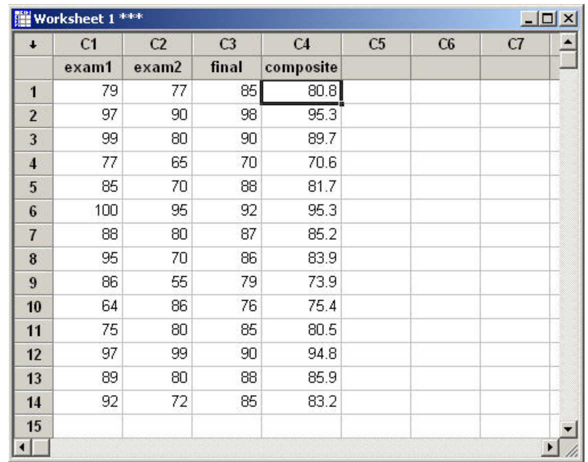
3.1 Creating New Variables

There are many instances when you may want to perform mathematical manipulations of variables and store them elsewhere. This becomes especially important in the discussion under 'Least Squares Regression'.

As an example, suppose we wanted to construct a composite weighted average of `exam1`, `exam2`, and `final`: $\texttt{composite} = 0.3 \cdot \texttt{exam1} + 0.3 \cdot \texttt{exam2} + 0.4 \cdot \texttt{final}$. We will store the composite scores in a new column.

Click on `Calc -> Calculator`. In the dialogue box which appears, type in the name of the column where the results are to be stored. In the 'Expression' box, type the mathematical expression of interest. When a variable name is to be inserted into the expression, you can double click on the names of variables found on the left panel.





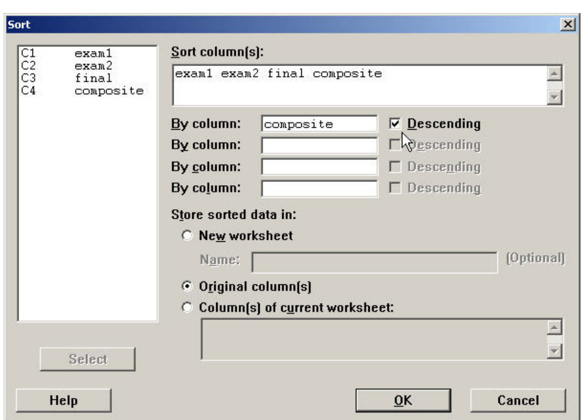
	C1 exam1	C2 exam2	C3 final	C4 composite	C5	C6	C7
1	79	77	85	80.8			
2	97	90	98	95.3			
3	99	80	90	89.7			
4	77	65	70	70.6			
5	85	70	88	81.7			
6	100	95	92	95.3			
7	88	80	87	85.2			
8	95	70	86	83.9			
9	86	55	79	73.9			
10	64	86	76	75.4			
11	75	80	85	80.5			
12	97	99	90	94.8			
13	89	80	88	85.9			
14	92	72	85	83.2			
15							

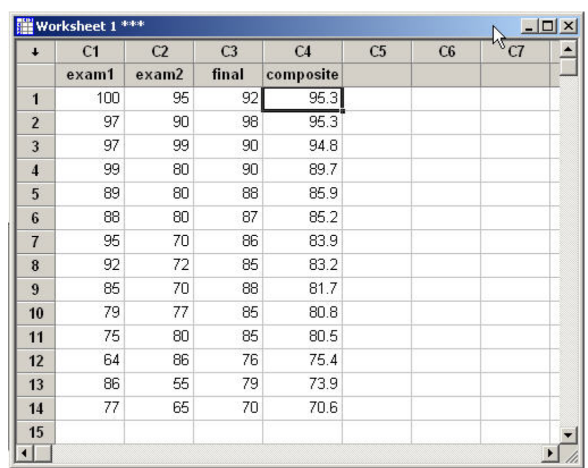
3.2 Sorting Variables

Using the newly constructed `composite` variable, suppose we want to sort the entire worksheet based on this variable in ascending (or descending) order.

To sort the worksheet based on a particular variable, click on `Data -> Sort`. In the resulting dialogue box, under the 'Sort Columns' area include all variables by double clicking on each name from the left panel (assuming you wish to sort the entire worksheet simultaneously, which is usually the case). Under 'By Column', choose the appropriate variable(s) you want to use as the sorting criteria. You may have the sorted data appear in a new worksheet or in new columns of the current worksheet. The sorted data may also be overwritten on the original columns. In the figures below, dialogue window appears on the left and the sorted data (overwritten on the original columns, in descending order) appear in the worksheet on the right.

Below, the figure on the left shows that the values of the expression will be stored in column C4. The figure on the right shows the resulting information created in the worksheet. Compare with the unsorted worksheet above.

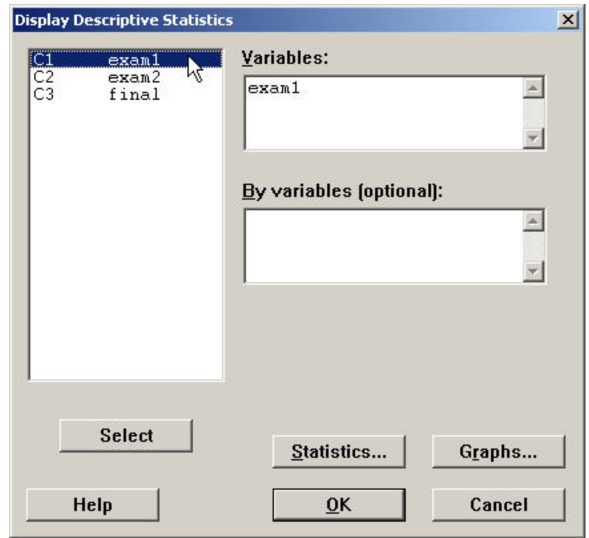




	C1 exam1	C2 exam2	C3 final	C4 composite	C5	C6	C7
1	100	95	92	95.3			
2	97	90	98	95.3			
3	97	99	90	94.8			
4	99	80	90	89.7			
5	89	80	88	85.9			
6	88	80	87	85.2			
7	95	70	86	83.9			
8	92	72	85	83.2			
9	85	70	88	81.7			
10	79	77	85	80.8			
11	75	80	85	80.5			
12	64	86	76	75.4			
13	86	55	79	73.9			
14	77	65	70	70.6			
15							

4 Basic Statistical Analysis

Now that we have some data in the worksheet, let us perform some Minitab functions. The first will be to compute basic statistics from the data. This can be done by clicking `Stat -> Basic Statistics -> Display Descriptive Statistics`. You will then see the following dialogue window open:



To choose the appropriate column/variable, simply select from the left panel which should list all available variables in the current worksheet and double click on one (or more) choices.

For the given data set above, the corresponding output for `exam1` (which will be stored in the **Session** window) is given below:

```
Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
exam1     14   0  87.36     2.82  10.55    64.00  78.50   88.50  97.00   100.00
```

5 Stem and Leaf Plot (Stemplot)

Create a stemplot (stem-and-leaf plot) by clicking `Graph -> Stem-and-Leaf`. You will then see a dialogue box similar to what was discovered above (see **Basic Statistical Analysis**). Keep in mind that you can choose to trim or include outliers within this dialogue box (look for the check box). The resulting stemplot for the `exam1` variable is given here:

```
  1   6   4
  1   6
  1   7
  4   7   579
  4   8
(4)   8   5689
  6   9   2
  5   9   5779
  1  10   0
```

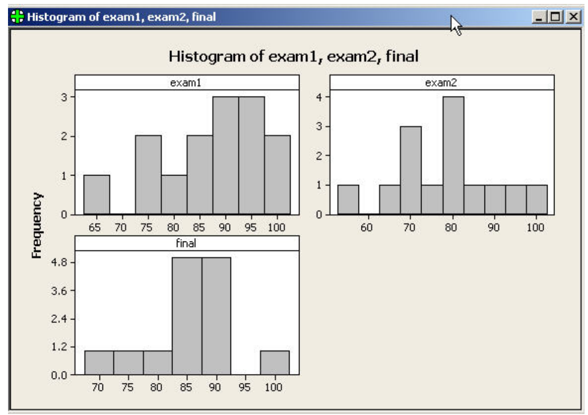
Note that the **stem** appears in the second column and the **leaves** appear in the last column. These last two columns are usually of most interest when looking at a stemplot.

The first column contains 'cumulative counts'. If the median value for the sample is included in a row, the count for that row is enclosed in parentheses. The values for rows above and below the median are cumulative. The count for a row above the median represents the total count for that row and the rows above it. The value for a row below the median represents the total count for that row and the rows below it.

6 Basic Graphs

6.1 Histogram

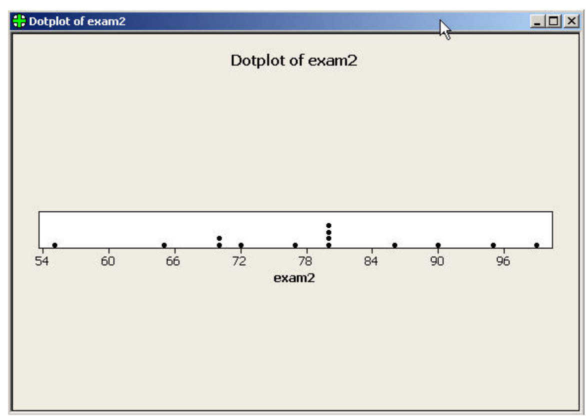
Create a histogram by clicking `Graph -> Histogram`. You will be given a choice of various histogram types (choose 'simple'). You will then see the familiar dialogue box where you specify the variable (or variables) you wish to create a histogram for. Note that you can create side-by-side or overlapping histograms by choosing the **Multiple Graphs** button. A side-by-side histogram array was created for all three variables (`exam1, exam2, final`) from our data set:



To adjust the number of classes and/or the class widths in the created histogram, left click on the x-axis of the histogram and then right click to obtain a menu. Select **Edit x-scale** and select **Binning tab**. Now you can change the interval definition to the number of intervals you would like and/or specify the midpoints of those intervals. If you wanted to have the midpoints at $\{60, 70, 80, 90, 100\}$ then simply type "60 70 80 90 100" in the area under 'Midpoint/Cutpoint positions' (note that you do not include commas to separate the values).

6.2 Dotplot

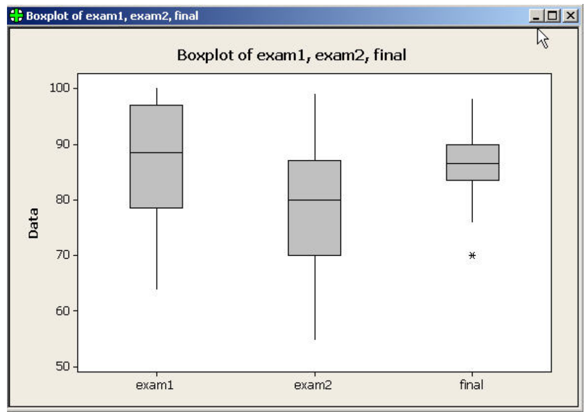
Create a dotplot by clicking `Graph -> Dotplot`. You will be given a choice of various dotplot types (choose 'simple'). Choose the appropriate variables in the subsequent window. The dotplot for `exam2` has been created on the right:



Notice that a dotplot is very similar to a histogram or stemplot.

6.3 Boxplot

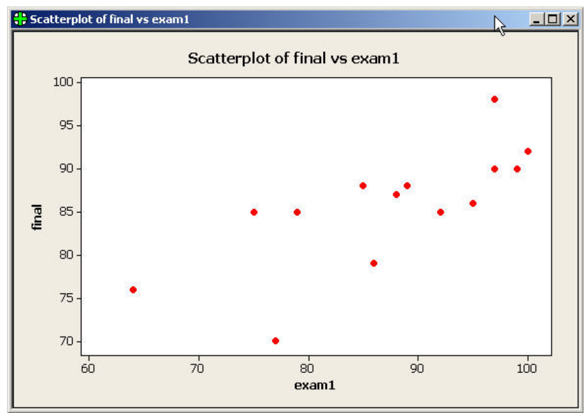
Create a boxplot by clicking `Graph -> Boxplot`. You will be given a choice of various dotplot types. If you want to create one boxplot for just one variable choose 'simple'. Since our data set contains three variables, let us choose "Multiple Y's Simple". Choosing all three variables to be placed one by one in the subsequent window the following side-by-side boxplots were constructed:



Notice that Minitab can identify extreme outliers (see boxplot for `final`).

6.4 Scatterplot

Create a scatterplot by clicking `Graph -> Scatterplot`. You will be given a choice of various dotplot types (choose 'simple'). Selecting `final` as the Y variable and `exam1` as the X variable, the following scatterplot was created:



Notice that there seems to be a linear trend between the two variables. It may be reasonable to consider fitting a line to this data. This will be discussed later under 'Least Squares Regression'.

7 Least Squares Regression

7.1 Simple Linear Regression and Correlation

Under the discussion for scatterplots, we noticed a linear relationship between `final` (as the Y variable) and `exam1` (as the X). Since linear regression is useful in the context of an explanatory and response variables, we will consider `exam1` as the predictor or final exam performance.

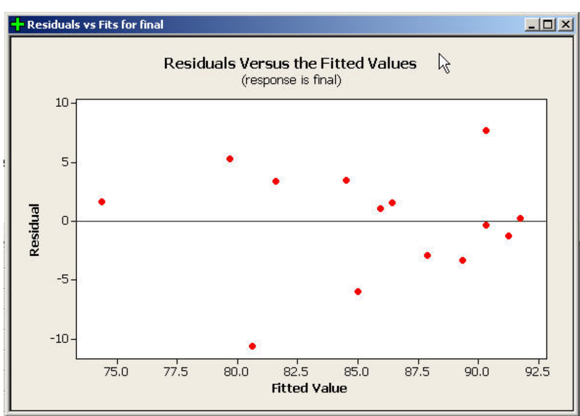
Determine the best fitted least squares line by clicking `Stat -> Regression -> Regression`. Select the appropriate response (Y) variable and the appropriate predictor (X) variable. To create a **Residual Plot**, click the "Graphs" button and choose "Residuals versus fits" to create the residual plots we've discussed in class.

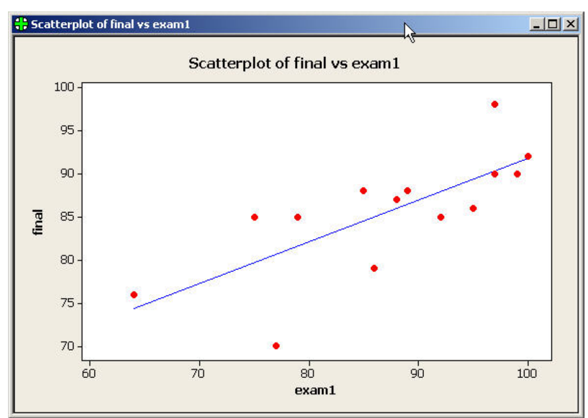
For the regression of `final` on `exam1`, the following output appeared in the Session window:

```
Regression Analysis: final versus exam1

The regression equation is
final = 43.3 + 0.484 exam1


Predictor     Coef  SE Coef     T      P
Constant     43.34    11.32  3.83  0.002
exam1       0.4842   0.1287  3.76  0.003


S = 4.89789   R-Sq = 54.1%   R-Sq(adj) = 50.3%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  339.34  339.34  14.15  0.003
Residual Error  12  287.87   23.99
Total           13  627.21
```

Note that all the important information can be found above. The slope and intercept for the least squares line can be found as the coefficient for `exam1` and `Constant` respectively. Also, the **coefficient of determination** is listed as 54.1%. Also note what we have defined as **SSResid**=287.87 and **SSTo**=627.21.

Plot the best fitted line on the scatterplot by clicking `Graph -> Scatterplot` and choose the 'With Regression' type. This is shown below on the right whereas the residual plot appears on the left:





Correlation, r

Determine the value of the correlation coefficient, r, by clicking `Stat -> Basic Statistics -> Correlation`. Select the two variables of interest from the left panel of the dialogue window. You may turn off the 'P-values' option. After clicking OK, you will see in the **Session** window the value of the Pearson Correlation Coefficient (Pearson created this statistic). If you have already done a least squares linear regression fit (as shown above), simple take the value of `R-Sq` (which is r^2) and take the square root of its value.

8 Confidence Intervals and Hypothesis Testing

8.1 One sample methods for a population mean: σ is known (using the Z method)

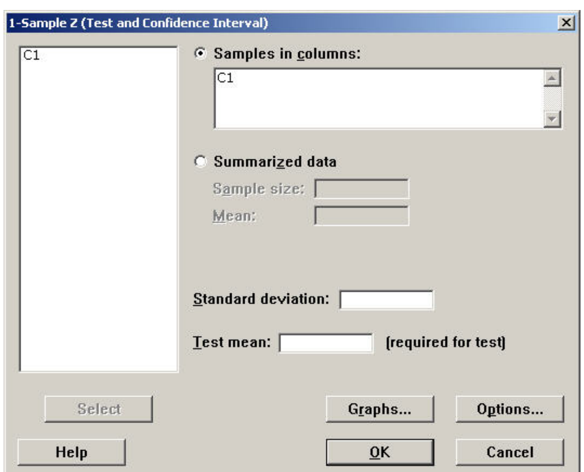
Note: This section is applicable when σ is unknown but when the sample size n is large.

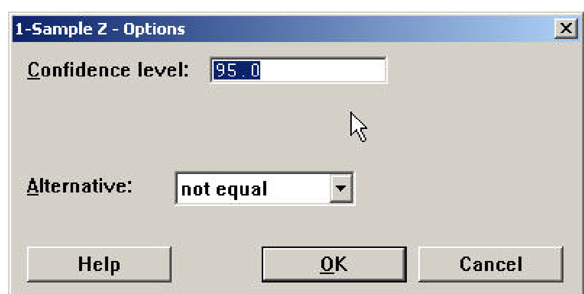
Input the data into a column in the worksheet (such as C1). Click on `Stat -> Basic Statistics -> One Sample Z`. In the dialogue box which appears, select the appropriate column for 'Samples in columns'.

For 'Standard deviation', input the known value of σ. If σ is unknown, but n is large, then input the value of the **sample standard deviation**. You may need to use Minitab to find the value of the sample standard deviation [`Stat -> Basic Statistics -> Display Descriptive Statistics`].

8.1.1 Confidence interval

Under 'Options', you can determine the two-sided corresponding 95% confidence interval by setting the 'Confidence level' to 95% and by setting 'Alternative' to 'not equal'. To create a one-sided interval, set the 'Alternative' to one of the other settings. Click on OK to generate the confidence interval which will be reported in the **Session** window. See below for the relevant windows:





8.1.2 Hypothesis testing, $H_0 : \mu = \mu_0$

Click on `Stat -> Basic Statistics -> One Sample Z`. If testing the hypothesis $H_0 : \mu = \mu_0$, ensure you input μ_0 into 'Test mean'. Under 'Options', select the appropriate setting for 'Alternative' depending upon the form of H_a ($\mu < \mu_0$, $\mu > \mu_0$, $\mu \neq \mu_0$). Click on OK to generate the result of the hypothesis test. Note that the value of the test statistic and the corresponding p-value will be reported in the **Session** window.

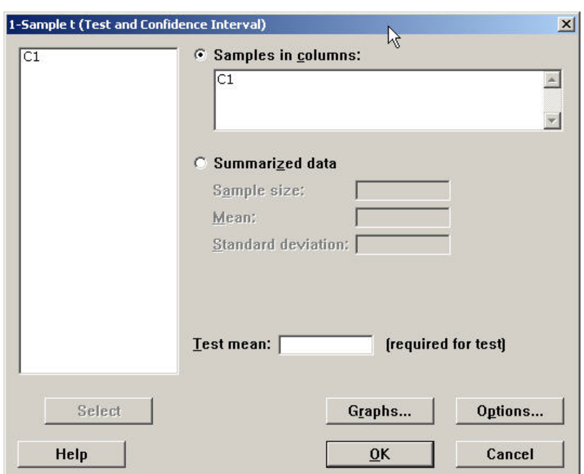
8.2 One sample methods for a population mean: σ is unknown (using the t method)

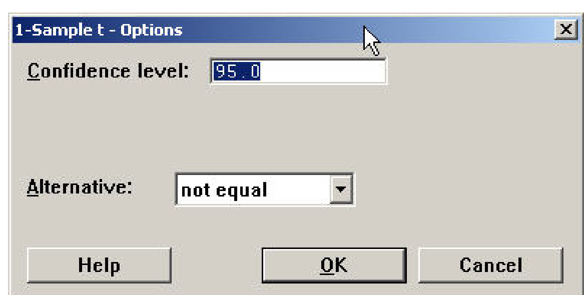
Note: This section is applicable when σ is unknown and when the sample size n is small.

Input the data into a column in the worksheet (such as C1). Click on `Stat -> Basic Statistics -> One Sample t`. In the dialogue box which appears, select the appropriate column for 'Samples in columns'.

8.2.1 Confidence interval

Under 'Options', you can determine the two-sided corresponding 95% confidence interval by setting the 'Confidence level' to 95% and by setting 'Alternative' to 'not equal'. To create a one-sided interval, set the 'Alternative' to one of the other settings. Click on OK to generate the confidence interval which will be reported in the **Session** window. See below for the relevant windows:





8.2.2 Hypothesis testing, $H_0 : \mu = \mu_0$

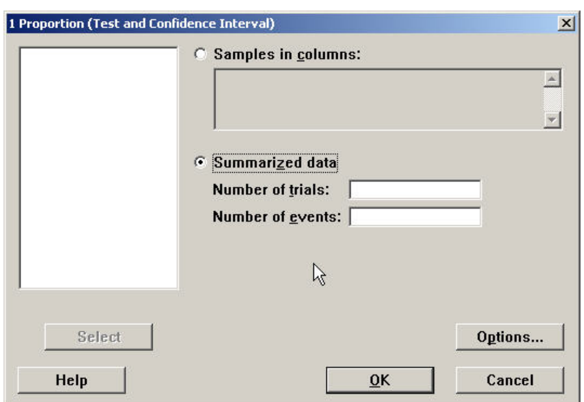
Click on `Stat -> Basic Statistics -> One Sample t`. If testing the hypothesis $H_0 : \mu = \mu_0$, ensure you input μ_0 into 'Test mean'. Under 'Options', select the appropriate setting for 'Alternative' depending upon the form of H_a ($\mu < \mu_0$, $\mu > \mu_0$, $\mu \neq \mu_0$). Click on OK to generate the result of the hypothesis test. Note that the value of the test statistic and the corresponding p-value will be reported in the **Session** window.

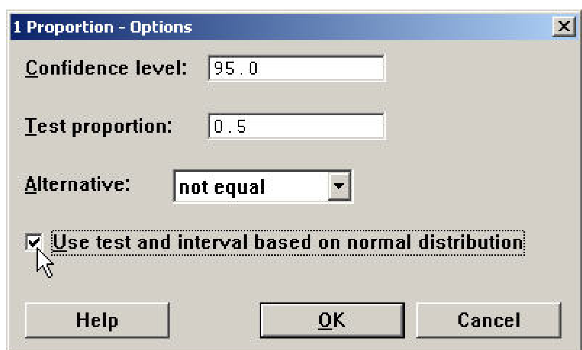
8.3 One sample methods for a population proportion: (using the Z method)

Click on `Stat -> Basic Statistics -> 1 Proportion`. In the dialogue box which appears, select 'Summarized data' and, for 'Number of trials' input the sample size n, for 'Number of events' input the number of *successes* that were observed out of n from the sample. If the data is formatted in a column with 0s and 1s (where 0 represents a failure, 1 represents a success), then you may select 'Samples in columns' and indicate the appropriate column (such as C1).

8.3.1 Confidence interval

Under 'Options', you can determine the two-sided corresponding 95% confidence interval by setting the 'Confidence level' to 95% and by setting 'Alternative' to 'not equal'. To create a one-sided interval, set the 'Alternative' to one of the other settings. **Be sure to select the 'Use test and interval based on normal distribution' option**. Click on OK to generate the confidence interval which will be reported in the **Session** window. See below for the relevant windows:





8.3.2 Hypothesis testing, $H_0 : p = p_0$

Click on `Stat -> Basic Statistics -> 1 Proportion`. Under 'Options', if testing the hypothesis $H_0 : p = p_0$ (or $H_0 : \pi = \pi_0$ depending on the author), ensure you input p_0 (or π_0) into

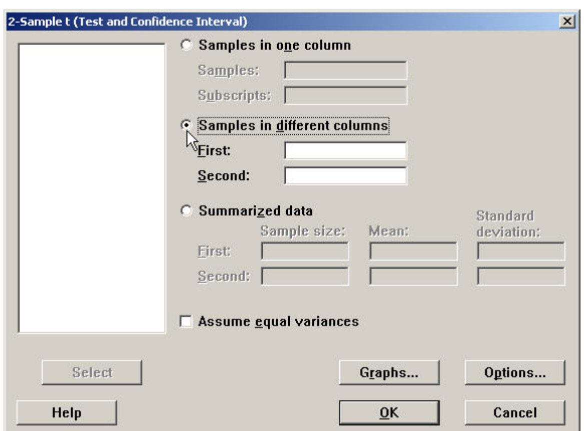
'Test proportion'. Select the appropriate setting for 'Alternative' depending upon the form of H_a (one-sided or two-sided). **Be sure to select the 'Use test and interval based on normal distribution' option**. Click on OK to generate the result of the hypothesis test. Note that the value of the test statistic and the corresponding p-value will be reported in the **Session** window.

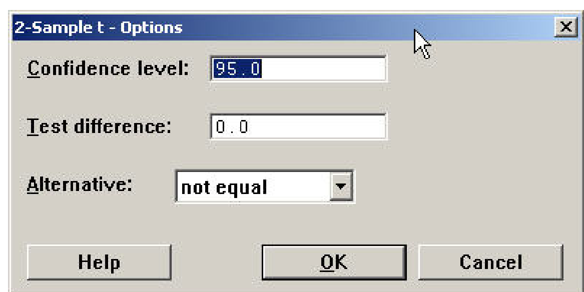
8.4 Two sample methods for population means: Independent samples (using the t method)

Input the data into two columns in the worksheet (such as C1 and C2). Click on `Stat -> Basic Statistics -> Two Sample t`. In the dialogue box which appears, select the appropriate column for 'Samples in different columns'.

8.4.1 Confidence interval

Under 'Options', you can determine the two-sided corresponding 95% confidence interval by setting the 'Confidence level' to 95% and by setting 'Alternative' to 'not equal'. To create a one-sided interval, set the 'Alternative' to one of the other settings. Click on OK to generate the confidence interval which will be reported in the **Session** window. See below for the relevant windows:





8.4.2 Hypothesis testing, $H_0 : \mu_1 - \mu_2 = \Delta$

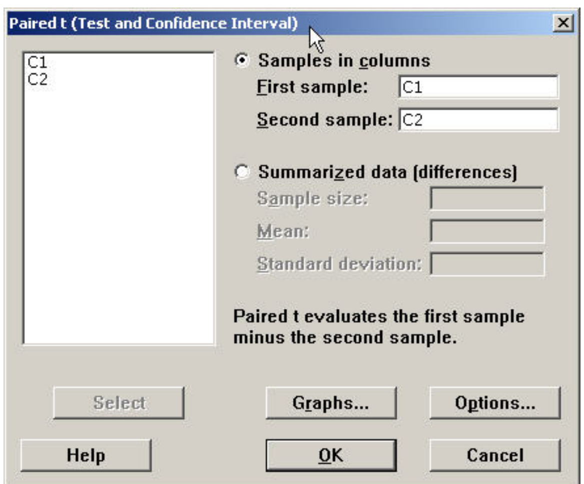
Click on `Stat -> Basic Statistics -> Two Sample t`. Under 'Options', if testing the hypothesis $H_0 : \mu_1 - \mu_2 = \Delta = 0$, ensure you input 0 into 'Test difference'. Select the appropriate setting for 'Alternative' depending upon the form of H_a (one-sided or two-sided). Click on OK to generate the result of the hypothesis test. Note that the value of the test statistic and the corresponding p-value will be reported in the **Session** window.

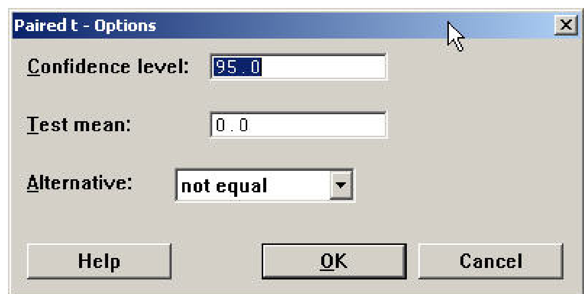
8.5 Two sample methods for population means: Dependent samples (using the paired-*t* or matched pairs method)

Input the data into two columns in the worksheet (such as C1 and C2). Click on `Stat -> Basic Statistics -> Paired t`. In the dialogue box which appears, select the appropriate column for 'Samples in columns'.

8.5.1 Confidence interval

Under 'Options', you can determine the two-sided corresponding 95% confidence interval by setting the 'Confidence level' to 95% and by setting 'Alternative' to 'not equal'. To create a one-sided interval, set the 'Alternative' to one of the other settings. Click on OK to generate the confidence interval which will be reported in the **Session** window. See below for the relevant windows:





8.5.2 Hypothesis testing, $H_0 : \mu_1 - \mu_2 = \Delta$

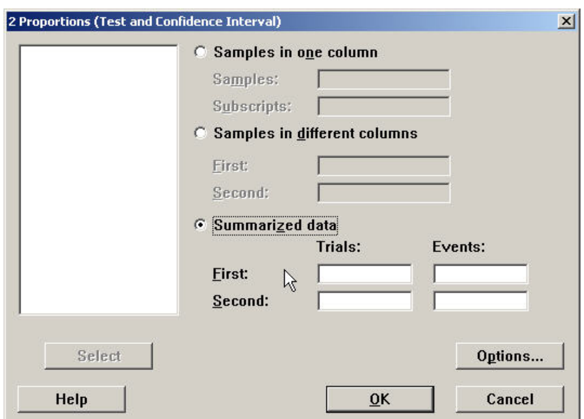
Click on `Stat -> Basic Statistics -> Two Sample t`. Under 'Options', if testing the hypothesis $H_0 : \mu_1 - \mu_2 = \Delta = 0$, ensure you input 0 into 'Test difference'. Select the appropriate setting for 'Alternative' depending upon the form of H_a (one-sided or two-sided). Click on OK to generate the result of the hypothesis test. Note that the value of the test statistic and the corresponding p-value will be reported in the **Session** window.

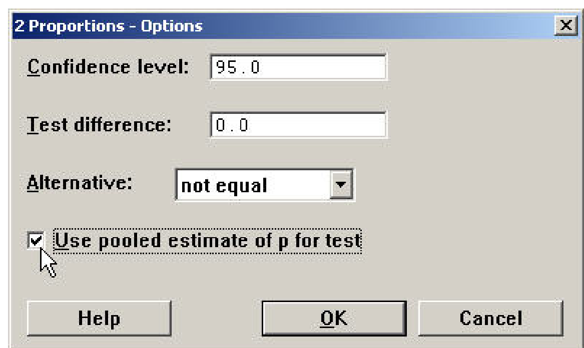
8.6 Two sample methods for population proportions: (using the Z method)

Click on `Stat -> Basic Statistics -> 2 Proportions`. In the dialogue box which appears, select 'Summarized data' and, for 'Trials' input the sample size n_1 and n_2 under 'First' and 'Second' respectively. For 'Events' input the number of *successes* that were observed in group 1 under 'First' and input the number of *successes* that were observed in group 2 under 'Second' . If the data is formatted in 2 columns each with 0s and 1s (where 0 represents a failure, 1 represents a success), then you may select 'Samples in different columns' and indicate the appropriate column (such as C1 and C2).

8.6.1 Confidence interval

Under 'Options', you can determine the two-sided corresponding 95% confidence interval by setting the 'Confidence level' to 95% and by setting 'Alternative' to 'not equal'. To create a one-sided interval, set the 'Alternative' to one of the other settings. **Be sure to select the 'Use pooled estimate of p for test' option**. Click on OK to generate the confidence interval which will be reported in the **Session** window. See below for the relevant windows:





8.6.2 Hypothesis testing, $H_0 : p_1 - p_2 = \Delta$

Click on `Stat -> Basic Statistics -> 1 Proportion`. Under 'Options', if testing the hypothesis $H_0 : p_1 - p_2 = \Delta = 0$ (or $H_0 : \pi_1 - \pi_2 = \Delta = 0$ depending on the author), ensure you input 0 into 'Test difference'. Select the appropriate setting for 'Alternative' depending upon the form of H_a (one-sided or two-sided). **Be sure to select the 'Use pooled estimate of p for test' option**. Click on OK to generate the result of the hypothesis test. Note that the value of the test statistic and the corresponding p-value will be reported in the **Session** window.

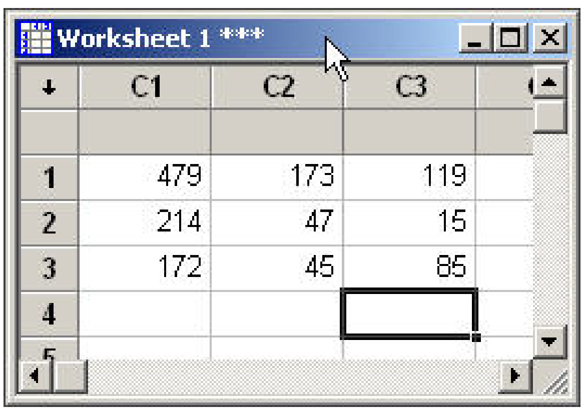
8.7 Multi sample methods for population proportions: (using the χ^2 method)

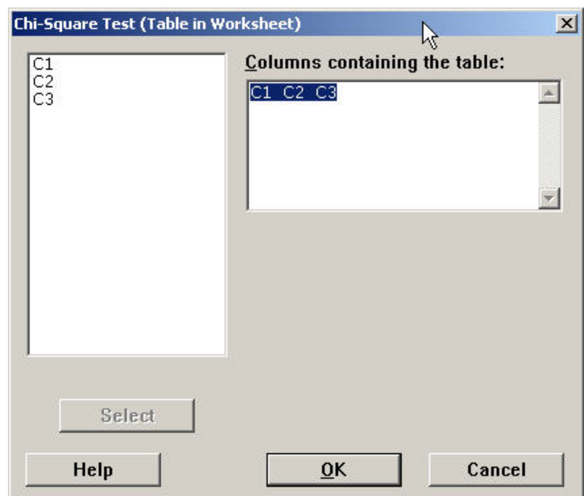
8.7.1 Testing for Homogeneity of Two Categorical Variables

Suppose we have data in the following form:

		Drug Usage: (Y)		
		Never	Rarely	Frequently
	Liberal	479	173	119
Political Views (X)	Conserv.	214	47	15
	Other	172	45	85

Input the data into columns C1 through C3 as shown below. Also note the corresponding dialogue window for the Chi-Square test:





Click on `Stat -> Tables -> Chi-Square test`. In the dialogue box which appears, select the appropriate columns for 'Columns containing the table'. Click on OK to generate the result of the hypothesis test. See below for the corresponding result for the given data set:

```
Chi-Square Test: C1, C2, C3

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

           C1      C2      C3  Total
    1     479     173     119    771
       494.38  151.46  125.17
```

```
          0.478   3.064   0.304

     2      214      47      15    276
         176.98   54.22   44.81
          7.746   0.961  19.828

     3      172      45      85    302
         193.65   59.33   49.03
          2.420   3.459  26.394

Total       865     265     219   1349

Chi-Sq = 64.654, DF = 4, P-Value = 0.000
```

In the output, for a particular cell from the table, note that the first entry corresponds to the observed count. The second entry corresponds to the expected count. And the third entry corresponds to the Chi-square contribution. Adding all the Chi-square contributions will yield the overall Chi-square statistic which is given in the last row of information. Note that the corresponding degrees of freedom and p-value are given as well. This is all reported in the **Session** window.

8.8 The Analysis of Variance (ANOVA)

8.8.1 One-Way or Single-Factor ANOVA

Suppose you have data according to the following four treatments (`L1 -- L4`):

```
L1          85.06   85.25   84.87
L2          84.99   84.28   84.88
L3          84.48   84.72   85.10
L4          84.10   84.55   84.05
```

Suppose we input this data in one of the following ways:

For Method 1, we input the levels of the treatment in one column (C1) and the corresponding values of the response variable in another column (C2). This is called the **stacked** method in Minitab.

For Method 2, we input the response values of a given treatment in a unique column. The data for treatments `L1` through `L4` are stored in columns C4 through C7 respectively. See image below:

X9-2-15.MTW ***

→	C1-T	C2	C3	C4	C5	C6	C7
				L1	L2	L3	L4
1	L1	85.06		85.06	84.99	84.48	84.10
2	L1	85.25		85.25	84.28	84.72	84.55
3	L1	84.87		84.87	84.88	85.10	84.05
4	L2	84.99					
5	L2	84.28					
6	L2	84.88					
7	L3	84.48					
8	L3	84.72					
9	L3	85.10					
10	L4	84.10					
11	L4	84.55					
12	L4	84.05					

To perform the One-way ANOVA in Minitab for Method 1, click on `Stat -> ANOVA -> One Way`. In the dialogue box which appears, select the appropriate column for 'Response' (C2). For 'Factor', select the appropriate column which contains the treatment levels (C1). Click on OK to generate the result of the hypothesis test.

To perform the One-way ANOVA in Minitab for Method 2, click on `Stat -> ANOVA -> One Way (Unstacked)`. In the dialogue box which appears, select the appropriate column for 'Response' (C2). For 'Factor', select the appropriate column which contains the treatment labels (C1). Click on OK to generate the result of the hypothesis test.

Both methods will, of course, yield the same output. See below for the corresponding result for the given data set:

```
One-way ANOVA: C2 versus C1

Source  DF      SS      MS     F      P
C1       3  1.0559  0.3520  3.96  0.053
Error    8  0.7114  0.0889
Total   11  1.7673

S = 0.2982   R-Sq = 59.75%   R-Sq(adj) = 44.65%

                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level  N    Mean  StDev  ---+---------+---------+---------+------
L1     3  85.060  0.190                 (-------*-------)
L2     3  84.717  0.382          (-------*-------)
L3     3  84.767  0.313           (-------*-------)
L4     3  84.233  0.275  (-------*-------)
                         ---+---------+---------+---------+------
                         84.00     84.50     85.00     85.50
Pooled StDev = 0.298
```