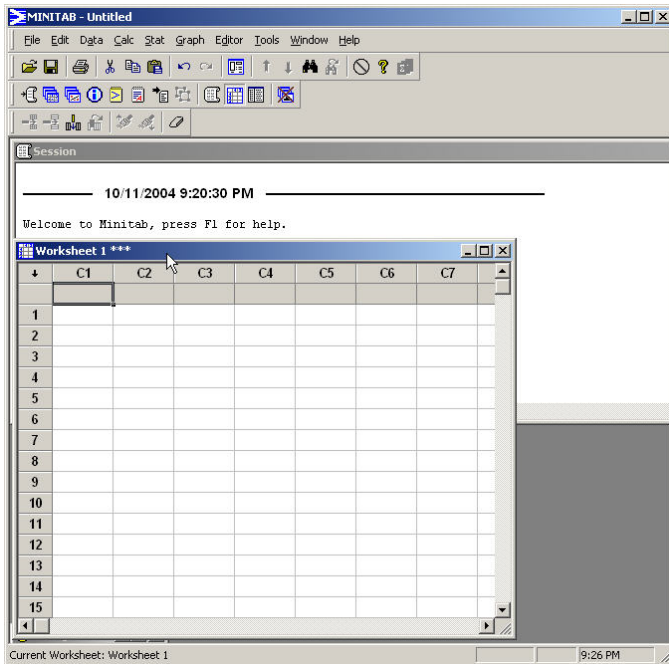


# 1 Introduction to Minitab

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](http://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`.

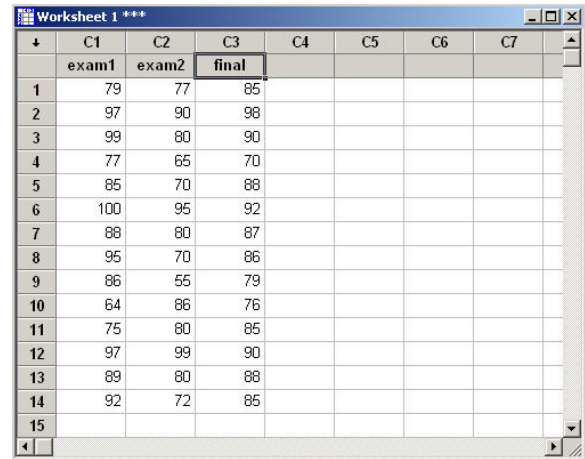
After you perform some tasks in Minitab with a particular worksheet (e.g. create graphs and invoke analyses) you can save all the work you have done by choosing **File -> Save Project as**. This will create a Minitab Project file in the form of `filename.mpj`. **NOTE:** If, after doing some work, you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



	C1	C2	C3	C4	C5	C6	C7
	<code>exam1</code>	<code>exam2</code>	<code>final</code>				
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).

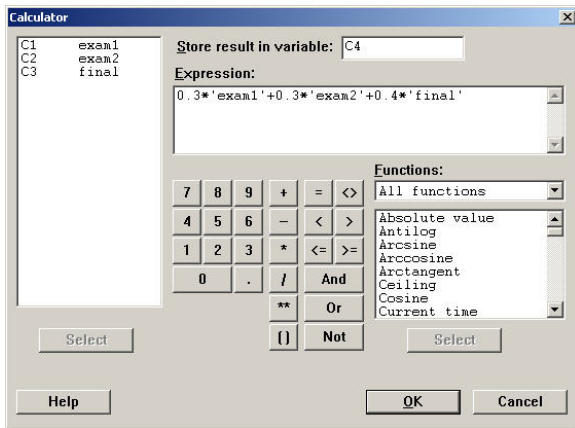
## 3 Manipulating Data in the Worksheet

### 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`:  $\text{composite} = 0.3 \cdot \text{exam1} + 0.3 \cdot \text{exam2} + 0.4 \cdot \text{final}$ . 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.



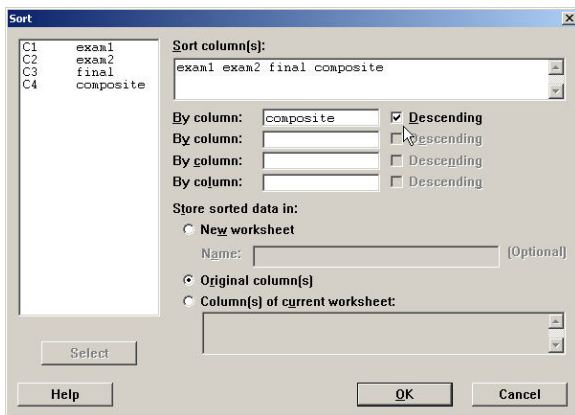
	C1	C2	C3	C4	C5	C6	C7
	exam1	exam2	final	composite			
1	79	77	85	80.8			
2	97	90	98	95.3			
3	99	80	90	89.7			
4	77	65	70	70.6			
5	85	70	88	81.7			
6	100	95	92	95.3			
7	88	80	87	85.2			
8	95	70	86	83.9			
9	86	55	79	73.9			
10	64	86	76	75.4			
11	75	80	85	80.5			
12	97	99	90	94.8			
13	89	80	88	85.9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

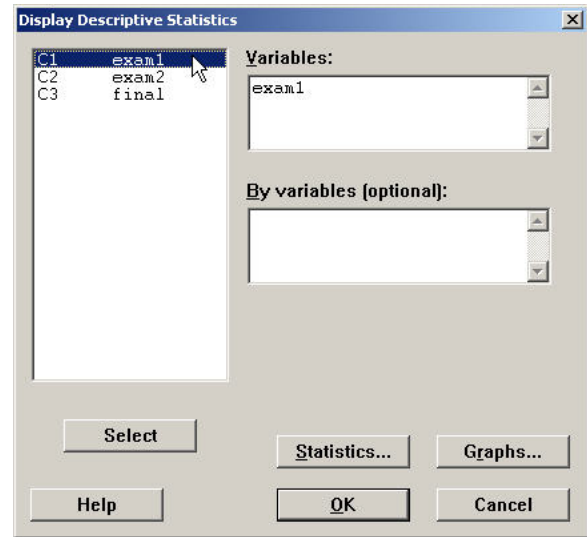


	C1	C2	C3	C4	C5	C6	C7
	exam1	exam2	final	composite			
1	100	95	92	95.3			
2	97	90	98	95.3			
3	97	99	90	94.8			
4	99	80	90	89.7			
5	89	80	88	85.9			
6	88	80	87	85.2			
7	95	70	86	83.9			
8	92	72	85	83.2			
9	85	70	88	81.7			
10	79	77	85	80.8			
11	75	80	85	80.5			
12	64	86	76	75.4			
13	86	55	79	73.9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

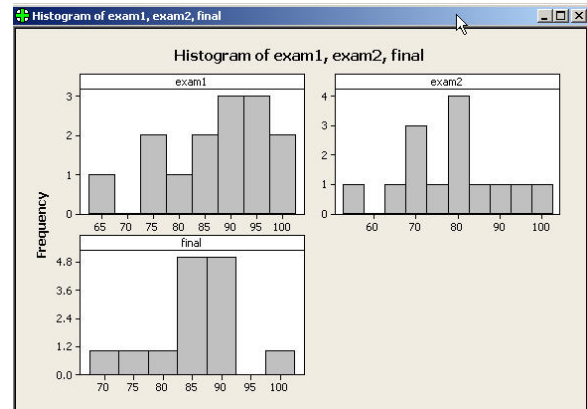
Note that the **stem** appears in the second column and the **leaves** appear in the last column. These last two columns are usually of most interest when looking at a stemplot.

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

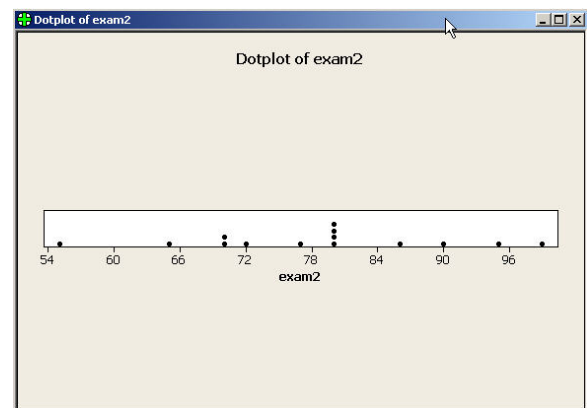
Create a histogram by clicking **Graph -> Histogram**. You will be given a choice of various histogram types (choose 'simple'). You will then see the familiar dialogue box where you specify the variable (or variables) you wish to create a histogram for. Note that you can create side-by-side or overlapping 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:

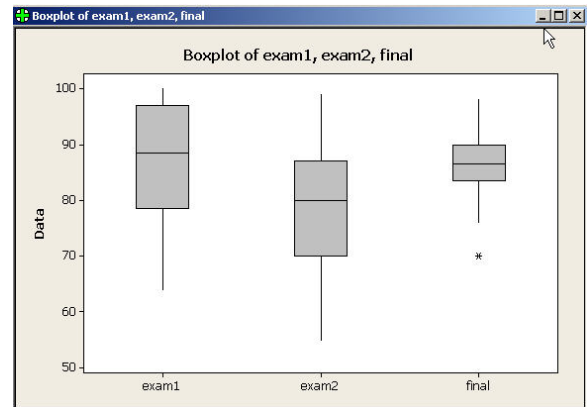


Notice that a dotplot is very similar to a histogram or stemplot.

## 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:

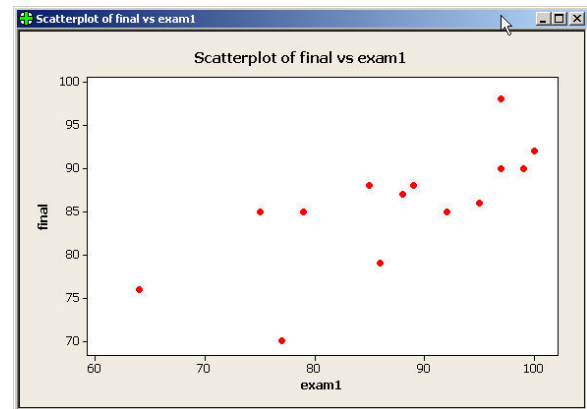
Notice that Minitab can identify extreme outliers (see boxplot for `final`).



## 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'.



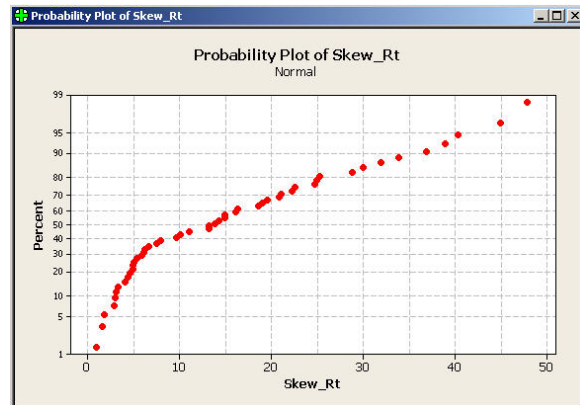
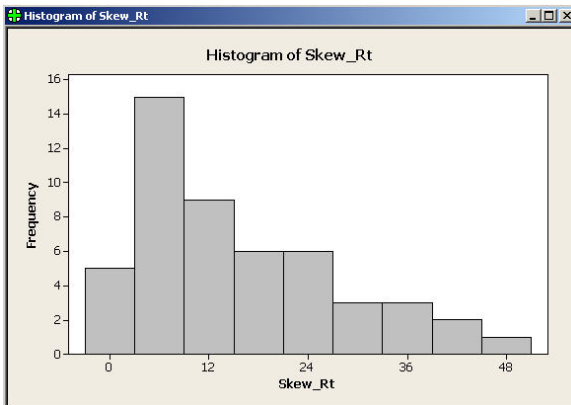
## 6.5 Probability Plot

Create a Probability Plot (or Quantile Plot) by clicking **Graph -> Probability Plot**. You will be given a choice of various dotplot types (choose 'simple'). Choose the appropriate variable of interest and click on 'Distribution'. For a Normal Probability Plot, select the Normal distribution with mean=0 and standard deviation=1. While in the 'Distribution' dialogue window, select the 'Data Display' tab and select the **Symbols Only** choice. Then click OK.

Keep in mind that probability plots are used to check the distributional assumption of a particular data set. It is often the case that we will assume normality for a given data set so a popular distribution choice is the normal. Although the given data set may not be based on a  $N(0, 1)$  distribution, we can use this distribution and search for linearity in the resulting probability plot. Deviations from linearity indicate the distributional assumption may be violated.

As an example, I have created a heavily right-skewed distribution under the label 'Skew\_Rt'. The

histogram for 'Skew\_Rt' is given below on the left. This skewed data set was used for creating a probability plot based under the assumption of normality. This figure is below on the right. Note the strong deviations from linearity, as what we would expect.



## 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

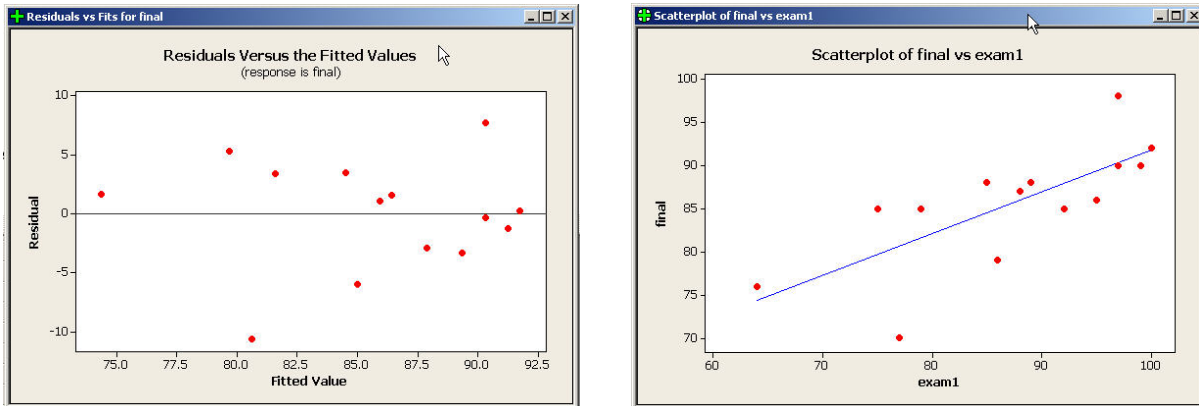
S = 4.89789    R-Sq = 54.1%    R-Sq(adj) = 50.3%

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,  $r$**

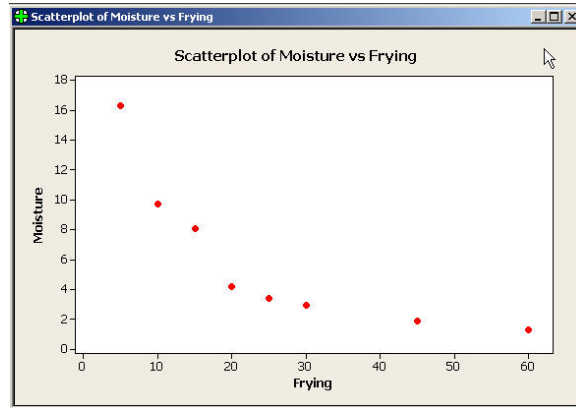
Determine the value of the correlation coefficient,  $r$ , by clicking `Stat -> Basic Statistics -> Correlation`. Select the two variables of interest from the left panel of the dialogue window. You may turn off the 'P-values' option. After clicking OK, you will see in the **Session** window the value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

**7.2 Nonlinear Relationships**

**7.2.1 Transforming Variables to Establish Linearity**

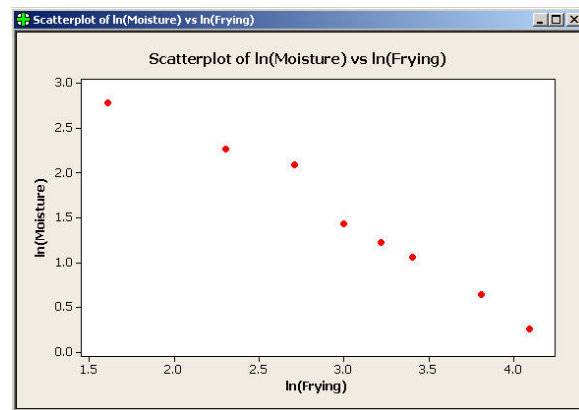
The data and corresponding scatterplot below exhibit a nonlinear relationship between the given two variables ( $X$ =frying time,  $Y$ =moisture):

	C1	C2	C3
	Frying	Moisture	
1	5	16.3	
2	10	9.7	
3	15	8.1	
4	20	4.2	
5	25	3.4	
6	30	2.9	
7	45	1.9	
8	60	1.3	
9			
10			



To establish linearity, we may modify the data by transforming one or both of the variables. For the given data set, it turns out that a natural logarithm transformation on both variables yields reasonable linearity. Remember, for instructions to modify/transform a given variable, revisit the earlier discussion under 'Creating New Variables' (click on Calc -> Calculator). The values of  $\ln(X)$  and  $\ln(Y)$  along with the corresponding scatterplot are shown below:

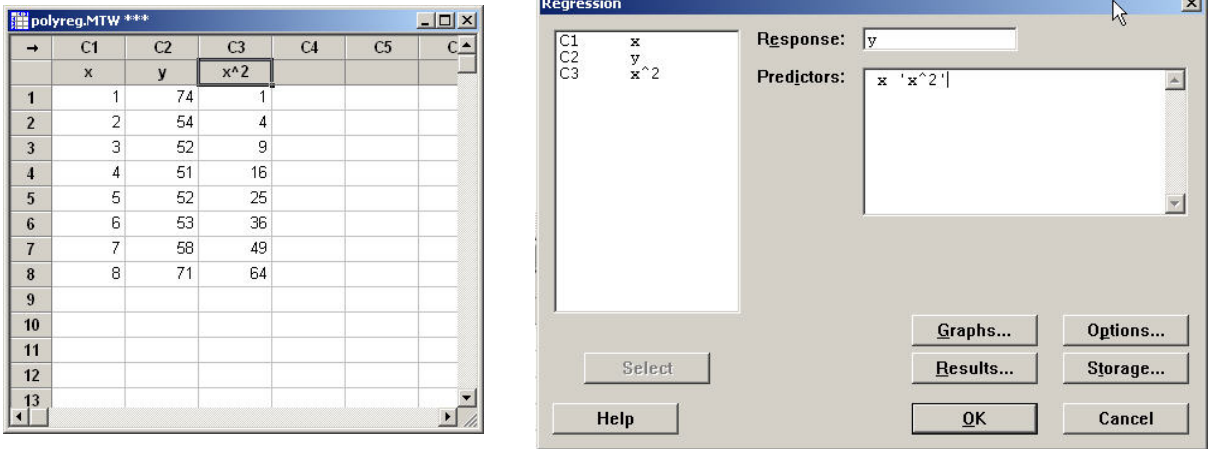
	C1	C2	C3	C4	C5
	Frying	Moisture	ln(Frying)	ln(Moisture)	
1	5	16.3	1.60944	2.79117	
2	10	9.7	2.30259	2.27213	
3	15	8.1	2.70805	2.09186	
4	20	4.2	2.99573	1.43508	
5	25	3.4	3.21888	1.22378	
6	30	2.9	3.40120	1.06471	
7	45	1.9	3.80666	0.64185	
8	60	1.3	4.0943445	0.26236	
9					
10					



### 7.3 Polynomial Regression

To propose a polynomial function as the regression model of interest, first you must modify the predictor variable and create new columns based on the function. For a quadratic fit, the model of interest would be  $\hat{y} = \beta_0 + \beta_1x + \beta_2x^2$  and so a column containing the squared values of  $x$  would be needed. Again, refer to the discussion under 'Creating New Variables' to see how this would be done (note:  $x^2$  would be typed in the 'Expression' area as  $x**2$ ).

Perform the least squares analysis by clicking Stat -> Regression -> Regression. In the 'Predictors' area, include the original  $x$  variable and the newly created  $x^2$  variable which is stored in the third column. These variables may be selected by double clicking on the choices which appear on the left panel. See below for the corresponding windows:



For the regression analysis, the following output appeared in the Session window:

The regression equation is  
 $y = 84.5 - 15.9 x + 1.77 x^2$

Predictor	Coef	SE Coef	T	P
Constant	84.482	4.904	17.23	0.000
x	-15.875	2.500	-6.35	0.001
x <sup>2</sup>	1.7679	0.2712	6.52	0.001

S = 3.51476    R-Sq = 89.5%    R-Sq(adj) = 85.3%

Analysis of Variance

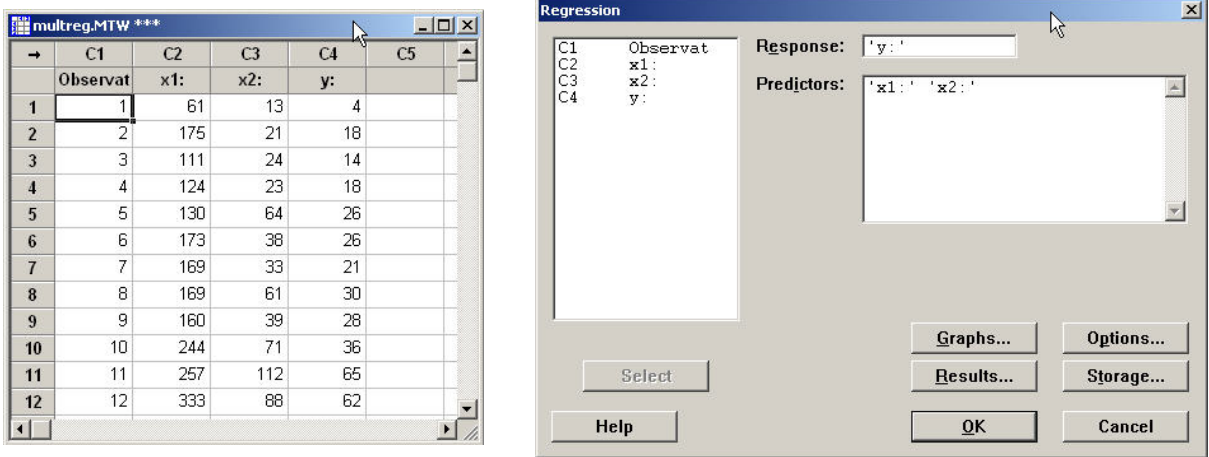
Source	DF	SS	MS	F	P
Regression	2	525.11	262.55	21.25	0.004
Residual Error	5	61.77	12.35		
Total	7	586.87			

Note that all the important information can be found above. The parameter estimates for the model can be found in the first table. Also, the **coefficient of multiple determination** is listed as 89.5%. Also note what we have defined as **SSResid**=61.77 and **SSTo**=586.87.

### 7.4 Multiple Linear Regression

Given multiple predictors, we can propose a more elaborate linear model. As an example, if the response variable is denoted as  $Y$  and if we have two explanatory variables  $X_1$  and  $X_2$ , a model we may consider is  $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ .

Perform the least squares analysis by clicking Stat -> Regression -> Regression. In the ‘Predictors’ area, include the  $x_1$  and  $x_2$  variables. These variables may be selected by double clicking on the choices which appear on the left panel. See below for the corresponding windows:



For the regression analysis, the following output appeared in the Session window:

The regression equation is  
 $y = - 7.35 + 0.113 x_1 + 0.349 x_2$ :

Predictor	Coef	SE Coef	T	P
Constant	-7.351	3.485	-2.11	0.061
x1:	0.11273	0.02969	3.80	0.004
x2:	0.34900	0.07131	4.89	0.001

S = 4.37937    R-Sq = 94.8%    R-Sq(adj) = 93.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	3529.9	1765.0	92.03	0.000
Residual Error	10	191.8	19.2		
Total	12	3721.7			

Again, note that all the important information can be found above. The parameter estimates for the model can be found in the first table. Also, the **coefficient of multiple determination** is listed as 94.8%. Finally, note what we have defined as **SSResid**=191.8 and **SSTo**=3721.7.

## 8 Working with Probability Distributions

### 8.1 Probability Density Function and Cumulative Distribution Function

Click on **Calc** -> **Probability Distributions** and select one of the distributions. You will then be able to specify the parameter(s) of the distribution and also find specific information about

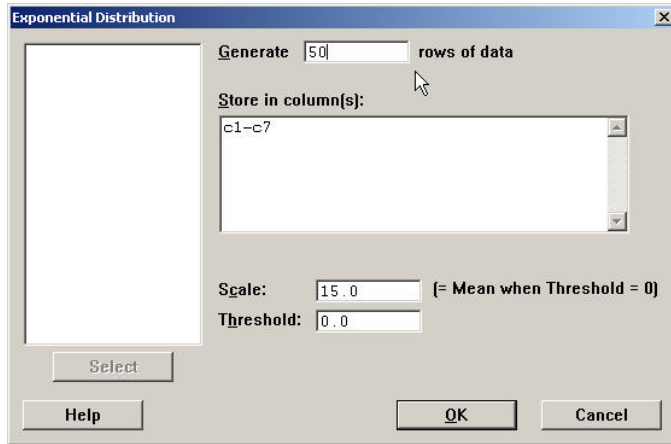
- the value of the corresponding probability density function  $f(x)$  at a particular setting of  $x$ .
- the value of the corresponding cumulative distribution function  $F(x)$  at a particular setting of  $x$ .
- the value of the inverse cumulative probability  $F^{-1}(p)$  at a particular setting  $p \in [0, 1]$ . As an example, you can find the value for  $x$  such that  $P(X \leq x) = 0.25$  is satisfied. In other words, this feature reports the  $p^{\text{th}}$  quantile or  $p^{\text{th}}$  percentile.

### 8.2 Simulating Data

You can use Minitab to simulate a random sample from your favorite distribution. Click on **Calc** -> **Probability Distributions**, select a distribution and specify the value(s) of the parameter(s). Indicate the number of observations you want in one sample (by specifying the number of rows of data). Finally, you can generate multiple samples by indicating multiple columns where the data are to be stored.

As an example, I used Minitab to generate a random sample of size  $n = 50$  from the exponential distribution with ‘scale’ parameter  $\beta = 15$  where the probability density function Minitab uses is given by  $f(x) = \frac{1}{\beta}e^{-x/\beta}$ . **Note** that this parameterization of the exponential pdf is slightly different from its other popular form:  $f(x) = \lambda e^{-\lambda x}$ . So, for our example, the value of  $\lambda$  would be  $\frac{1}{15}$ .

Seven total samples (each of size 50) were generated and stored in columns C1 through C7. The corresponding dialogue window and resulting simulated values are shown below:



	C1	C2	C3	C4	C5	C6	C7
1	7.932976	0.3579	81.8368	2.1618	3.6643	6.3538	51.0450
2	32.406	9.0334	9.6871	17.3452	2.3603	9.1863	12.6032
3	42.143	11.4695	3.0691	41.1698	14.9116	37.8144	5.0633
4	24.012	39.6957	9.2212	3.9373	0.7534	44.5950	8.3711
5	6.323	19.6888	8.7577	12.9589	7.4616	1.7478	17.1514
6	5.348	5.0955	38.1454	59.7682	32.3006	12.9277	6.4115
7	24.555	3.3645	0.0258	17.5379	12.2997	15.6212	2.1541
8	13.526	25.0718	31.5274	2.8310	1.1682	11.2742	1.7586
9	15.560	9.9006	6.4183	26.2543	1.5088	0.8422	16.9796
10	7.364	10.3681	1.6380	2.4457	0.6675	19.6417	44.3803
11	12.768	2.7709	2.1808	2.3632	3.8408	87.1984	57.5234
12	1.883	12.5525	1.1771	9.0488	27.9773	8.1839	12.0447
13	0.872	5.6034	16.8128	11.4933	27.7960	17.1389	8.8629
14	3.702	13.6365	11.9297	21.2361	23.7785	23.1058	14.7308
15	101.372	2.7442	8.9109	16.6107	17.3041	7.6381	11.6748
16	67.236	5.4079	29.2624	3.6386	13.6860	9.8522	2.8617

From the work above, you can compute and store the statistics from each sample by clicking Stat -> Basic Statistics -> Store Descriptive Statistics. Choose ‘Statistics’ and select the statistic of interest (such as the sample mean). After you click OK, and assuming you selected the sample mean as the statistic of interest, Minitab will store the values of the means in newly created columns.

To generate a histogram of these means, it would be convenient to have them stored as a column instead of in a row. You can do this by clicking Data -> Transpose Columns. Choose the columns that contains the means and select the location of where the newly created column will be stored (it may be convenient to choose ‘After last column in use’). This process of creating a histogram of statistics generated from simulated random samples would yield an approximation to the true **sampling distribution** of the statistic in question.

## 9 Confidence Intervals and Hypothesis Testing

### 9.1 One sample methods for a population mean: $\sigma$ is known (using the Z method)

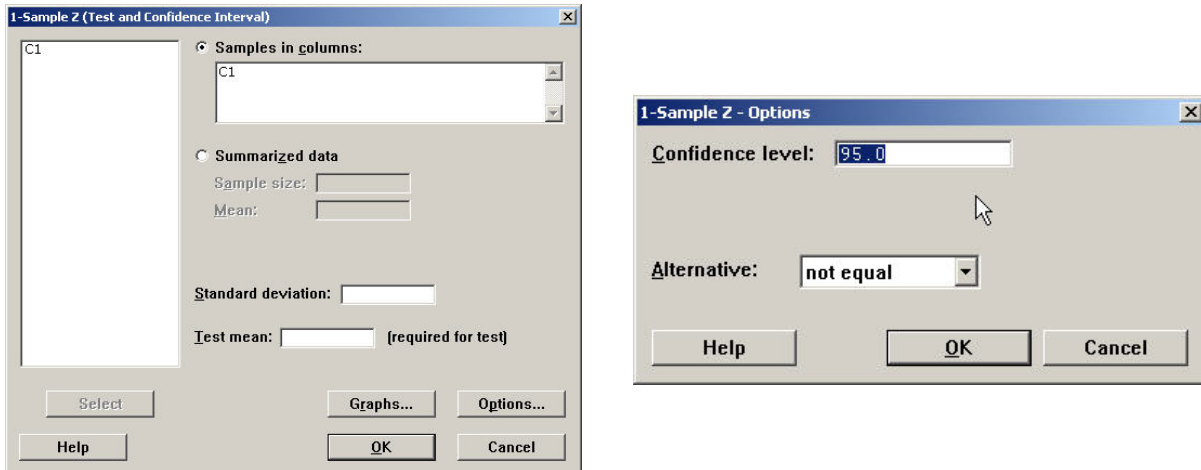
**Note:** This section is applicable when  $\sigma$  is unknown but when the sample size  $n$  is large.

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  $\sigma$ . If  $\sigma$  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].

### 9.1.1 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:



### 9.1.2 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  $\mu_0$  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.

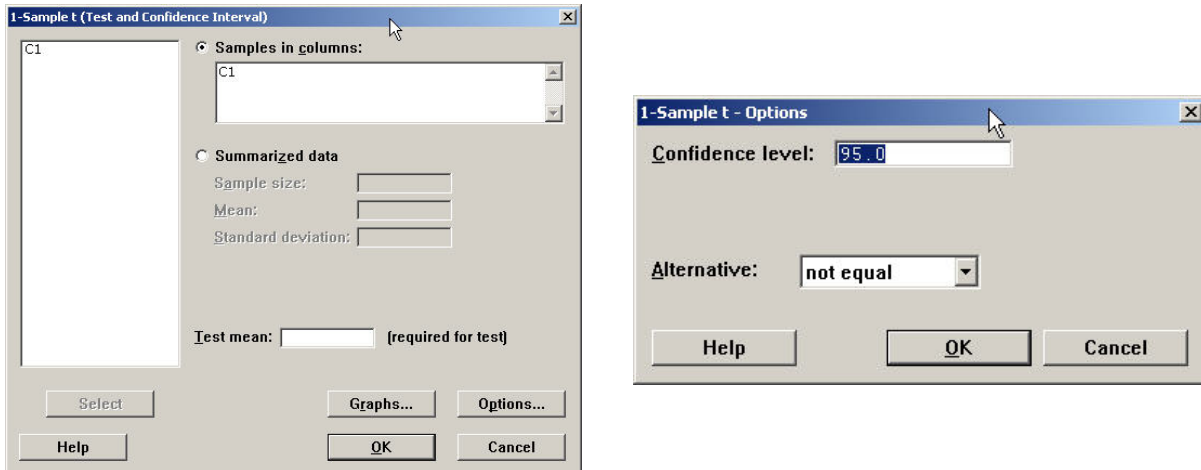
## 9.2 One sample methods for a population mean: $\sigma$ is unknown (using the $t$ method)

**Note:** This section is applicable when  $\sigma$  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’.

### 9.2.1 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:



### 9.2.2 Hypothesis testing, $H_0 : \mu = \mu_0$

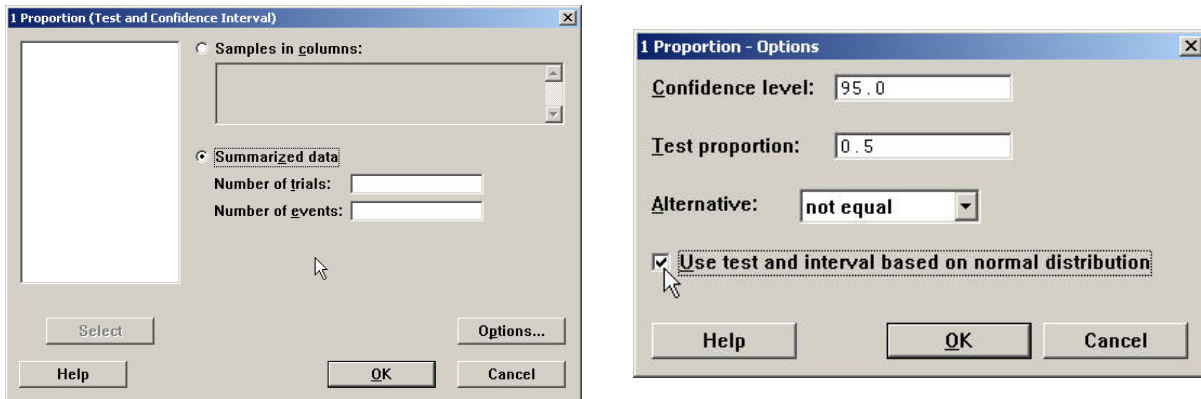
Click on Stat -> Basic Statistics -> One Sample t. If testing the hypothesis  $H_0 : \mu = \mu_0$ , ensure you input  $\mu_0$  into ‘Test mean’. Under ‘Options’, select the appropriate setting for ‘Alternative’ depending upon the form of  $H_a$  ( $\mu < \mu_0$ ,  $\mu > \mu_0$ ,  $\mu \neq \mu_0$ ). 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.

### 9.3 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).

### 9.3.1 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:



### 9.3.2 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  $\pi_0$ ) 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.

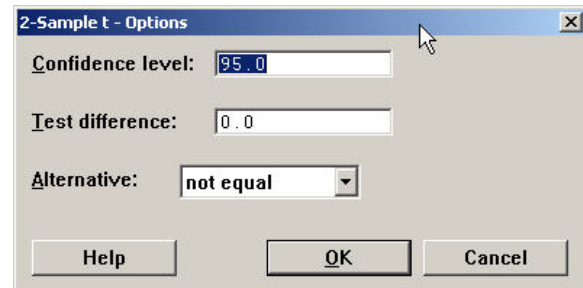
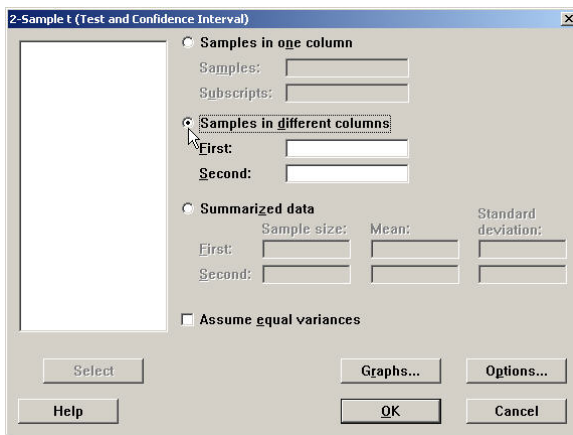
## 9.4 Two sample methods for population means: Independent samples (using the $t$ method)

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’.

### 9.4.1 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:



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

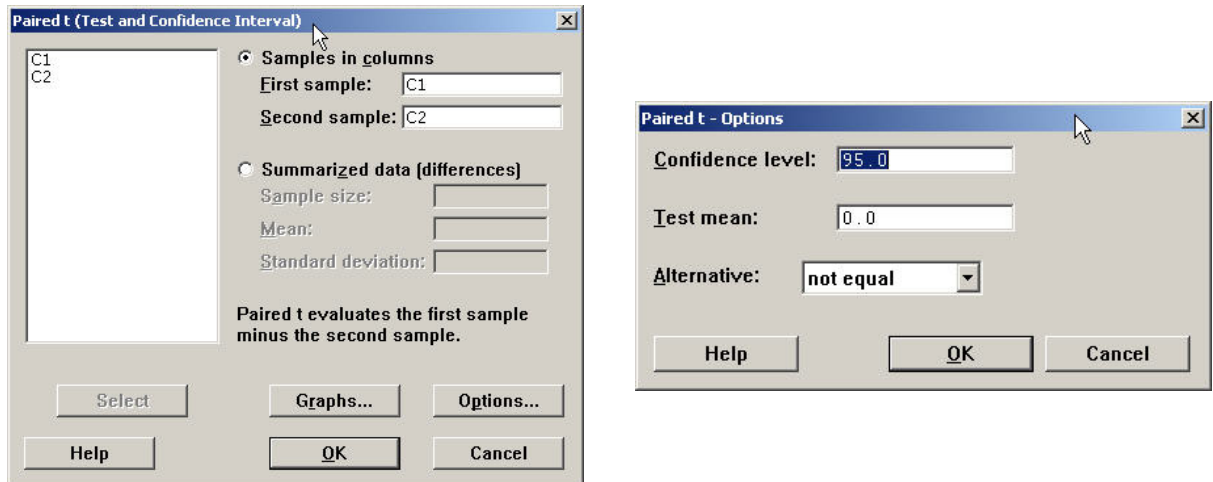
Click on Stat -> Basic Statistics -> Two Sample t. Under ‘Options’, if testing the hypothesis  $H_0 : \mu_1 - \mu_2 = \Delta = 0$ , ensure you input 0 into ‘Test difference’. Select the appropriate setting for ‘Alternative’ depending upon the form of  $H_a$  (one-sided or two-sided). Click on OK to generate 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.

### 9.5 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’.

#### 9.5.1 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:



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

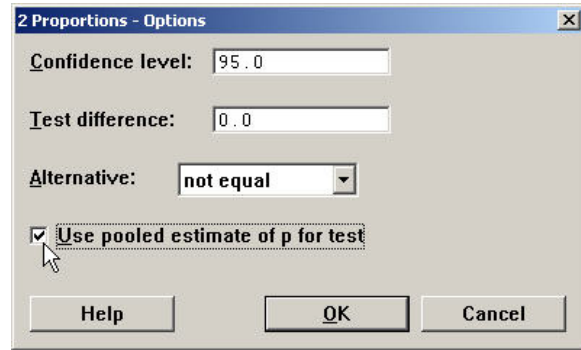
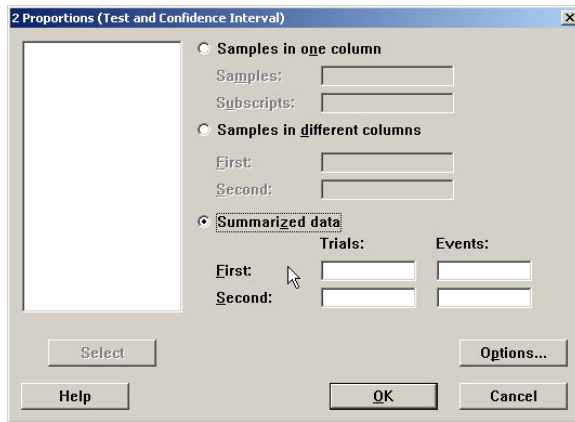
Click on Stat -> Basic Statistics -> Two Sample t. Under 'Options', if testing the hypothesis  $H_0 : \mu_1 - \mu_2 = \Delta = 0$ , ensure you input 0 into 'Test difference'. Select the appropriate setting for 'Alternative' depending upon the form of  $H_a$  (one-sided or two-sided). Click on OK to generate 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.

## 9.6 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).

### 9.6.1 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:



### 9.6.2 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.

## 9.7 Multi sample methods for population proportions: (using the $\chi^2$ method)

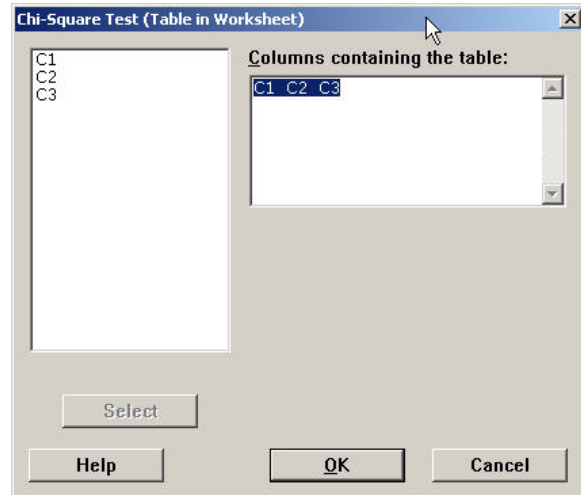
### 9.7.1 Testing for Homogeneity of Two Categorical Variables

Suppose we have data in the following form:

		Drug Usage: (Y)		
		Never	Rarely	Frequently
Political Views (X)	Liberal	479	173	119
	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:

	C1	C2	C3
1	479	173	119
2	214	47	15
3	172	45	85
4			



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.

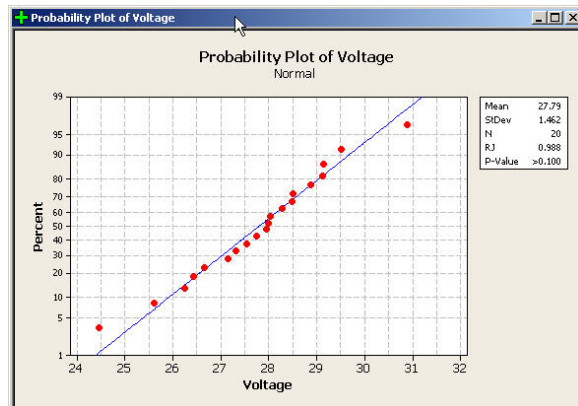
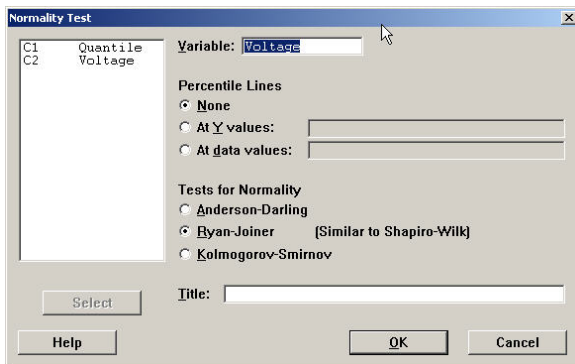
## 9.8 Testing for Normality

Input the data into column C1. The following test will create a probability plot and calculate the correlation coefficient. The corresponding hypotheses are:

$H_0$  : The data stems from a normal distribution

$H_a$  : The data does not stem from a normal distribution

Click on **Stat -> Basic Statistics -> Normality Test**. In the dialogue box which appears, select the appropriate column for 'Variable'. For 'Tests for Normality' check "Ryan-Joiner". Click on OK to generate the result of the hypothesis test. See below for the corresponding result for the given data set:



## 9.9 The Analysis of Variance (ANOVA)

### 9.9.1 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:

	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.

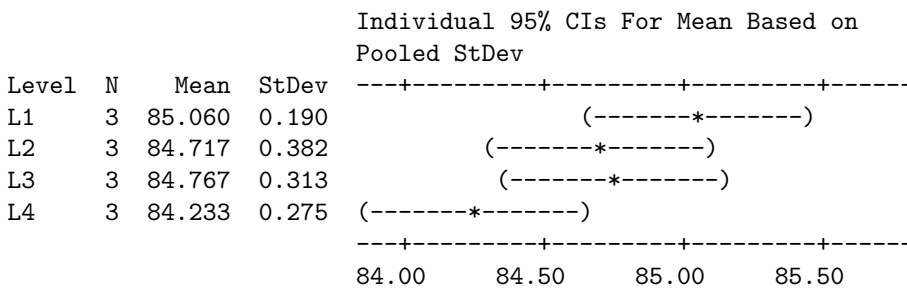
To perform the One-way ANOVA in Minitab for Method 2, click on **Stat** -> **ANOVA** -> **One Way (Unstacked)**. In the dialogue box which appears, select the appropriate column for 'Response' (C2). For 'Factor', select the appropriate column which contains the treatment labels (C1). Click on OK to generate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



Pooled StDev = 0.298

### 9.9.2 One-Way ANOVA: Multiple Comparisons

To apply a multiple comparisons analysis, follow the above directions for the one-way ANOVA method and select 'Comparisons' in the dialogue window. Choose from the available options (e.g. Tukey's, Dunnett's). If the overall level of  $\alpha$  is 5%, set the 'family error rate' to 5%.

### 9.9.3 Randomized Block Experiment

Suppose we have data on a response variable (power, in kW/hr) collected according to 5 treatments (Brand) and blocked according to 4 humidity levels (Hum). The data is summarized below:

Brand	Hum1	Hum2	Hum3	Hum4
1	685	792	838	875
2	722	806	893	953
3	733	802	880	941
4	811	888	952	1005
5	828	920	978	1023

This data will be stored in Minitab in the following method: input the levels of the treatment in one column (C1), input the levels of the blocking variable in another column (C2), and the corresponding values of the response variable in another column (C3). See below on how this was done in Minitab:

	C1	C2	C3
	Brand	Humidity	Power
1	1	1	685
2	2	1	722
3	3	1	733
4	4	1	811
5	5	1	828
6	1	2	792
7	2	2	806
8	3	2	802
9	4	2	888
10	5	2	920
11	1	3	838
12	2	3	893
13	3	3	880
14	4	3	952
15	5	3	978
16	1	4	875
17	2	4	953
18	3	4	941
19	4	4	1005
20	5	4	1023

Perform the Two-way ANOVA in Minitab by clicking on **Stat -> ANOVA -> Two Way**. In the dialogue box which appears, select the appropriate column for ‘Response’ (C3). For ‘Row factor’, select the appropriate column which contains the treatment levels (C1). For ‘Column factor’, select the appropriate column which contains the blocking levels (C2). Click on OK to generate the result of the hypothesis test. See below for the corresponding result for the given data set:

Two-way ANOVA: Power versus Brand, Humidity

Source	DF	SS	MS	F	P
Brand	4	53231	13307.8	95.57	0.000
Humidity	3	116218	38739.3	278.20	0.000
Error	12	1671	139.3		
Total	19	171120			

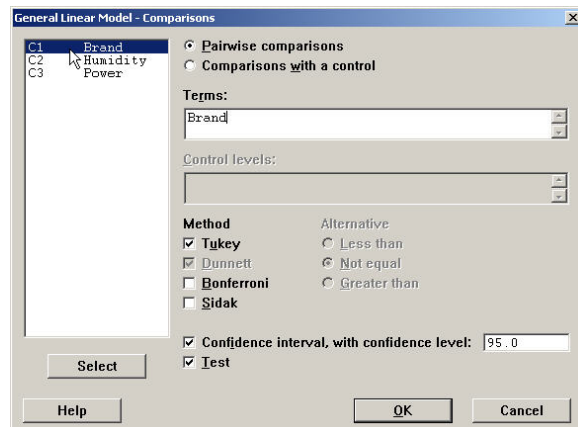
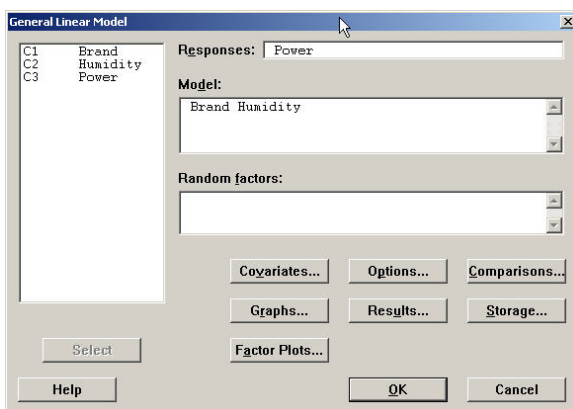
S = 11.80 R-Sq = 99.02% R-Sq(adj) = 98.45%

### 9.9.4 Randomized Block: Multiple Comparisons

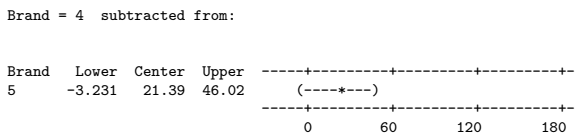
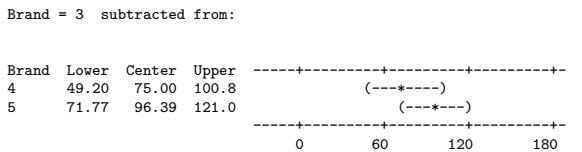
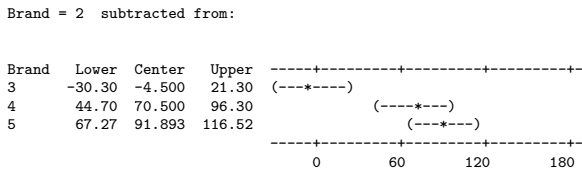
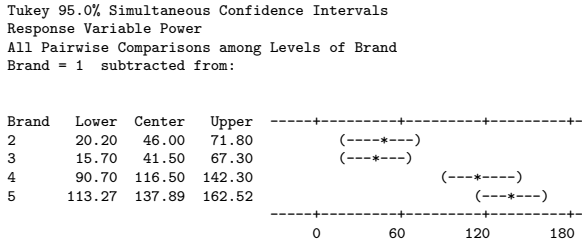
To apply a multiple comparisons analysis, we have to use a more general approach (called the **General Linear Model**) which will perform the same Two-Way ANOVA and also offer other analysis features.

Assuming the data have been typed into Minitab as shown above, click on **Stat -> ANOVA -> General Linear Model**. In the dialogue box which appears, select the appropriate column for ‘Response’ (C3). For ‘Model’, select the appropriate column which contains the treatment levels (C1) **and** the appropriate column which contains the blocking levels (C2). Click on OK to generate the result of the hypothesis test.

Select ‘Comparisons’ in the dialogue window. In the area for ‘Terms’, choose the column which contains the treatments for which you wish to perform comparisons (C1). Choose from the available options (e.g. Tukey’s, Dunnett’s). If the overall level of  $\alpha$  is 5%, set the ‘Confidence level’ to 95%. See below for the corresponding dialogue windows:



See below for the corresponding output:



Tukey Simultaneous Tests  
Response Variable Power  
All Pairwise Comparisons among Levels of Brand  
Brand = 1 subtracted from:

Brand	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
2	46.00	8.344	5.513	0.0010
3	41.50	8.344	4.974	0.0024
4	116.50	8.344	13.962	0.0000
5	139.75	8.344	16.748	0.0000

Brand = 2 subtracted from:

Brand	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
3	-4.500	8.344	-0.5393	0.9813
4	70.500	8.344	8.4490	0.0000
5	93.750	8.344	11.2354	0.0000

Brand = 3 subtracted from:

Brand	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
4	75.00	8.344	8.988	0.0000
5	98.25	8.344	11.775	0.0000

Brand = 4 subtracted from:

Brand	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
5	23.25	8.344	2.786	0.0978

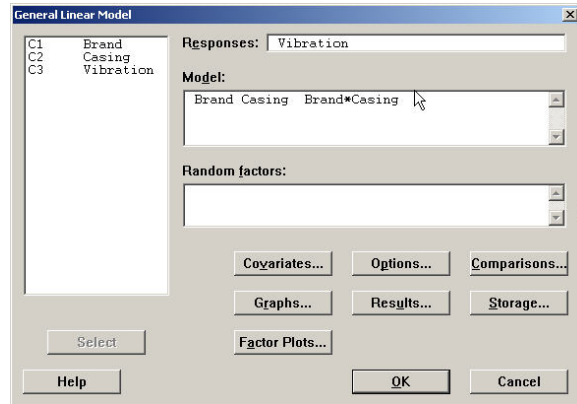
### 9.9.5 Two-Factor Designs

Suppose we have data on a response variable (vibration, in microns) collected according to two factors: factor A (brand) at 5 levels and factor B (casing) at 3 levels. At each of the  $5 \times 3 = 15$  treatment combinations, two observations were recorded. The data is summarized below:

	Cas1	Cas2	Cas3
B1	13.1, 13.2	15.0, 14.8	14.0, 14.3
B2	16.3, 15.8	15.7, 16.4	17.2, 16.7
B3	13.7, 14.3,	13.9, 14.3	12.4, 12.3
B4	15.7, 15.8	13.7, 14.2	14.4, 13.9
B5	13.5, 12.5	13.4, 13.8	13.2, 13.1

This data will be stored in Minitab in the following method: input the levels of the treatment in one column (C1), input the levels of the blocking variable in another column (C2), and the corresponding values of the response variable in another column (C3). See image on the left on how this was done in Minitab:

	C1	C2	C3
	Brand	Casing	Vibration
1	1	1	13.1
2	1	1	13.2
3	1	2	15.0
4	1	2	14.8
5	1	3	14.0
6	1	3	14.3
7	2	1	16.3
8	2	1	15.8
9	2	2	15.7
10	2	2	16.4
11	2	3	17.2
12	2	3	16.7
13	3	1	13.7



To perform the Two-way ANOVA in Minitab including an interaction analysis, click on **Stat -> ANOVA -> General Linear Model**. In the dialogue box which appears, select the appropriate column for 'Response' (C3). For 'Model', select the appropriate column which contains the treatment levels (C1) **and** the appropriate column which contains the blocking levels (C2). Also include the interaction term which, for this example, is the product of C1 and C2. See image on the right (above) for the corresponding dialogue window.

Click on OK to generate the result of the hypothesis test. In the output below, you will notice terms such as 'Seq SS', 'Adj SS', and 'Adj MS'. You may ignore the terms 'Adj' (short for adjusted) since these are the same sums of squares and mean squares we would obtain by hand calculations. See below for the corresponding output:

Analysis of Variance for Vibration, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Brand	4	36.6747	36.6747	9.1687	82.35	0.000
Casing	2	0.7047	0.7047	0.3523	3.16	0.071
Brand*Casing	8	11.6053	11.6053	1.4507	13.03	0.000
Error	15	1.6700	1.6700	0.1113		
Total	29	50.6547				

S = 0.333667    R-Sq = 96.70%    R-Sq(adj) = 93.63%

## 9.10 Power and Type II Error ( $\beta$ )

Consider the following example:

**Example:** A new process for mining copper is under investigation and it is supposed to produce an average of more than 50 tons of ore per day. Here are the 5-day data (in tons):

50    47    53    51    52

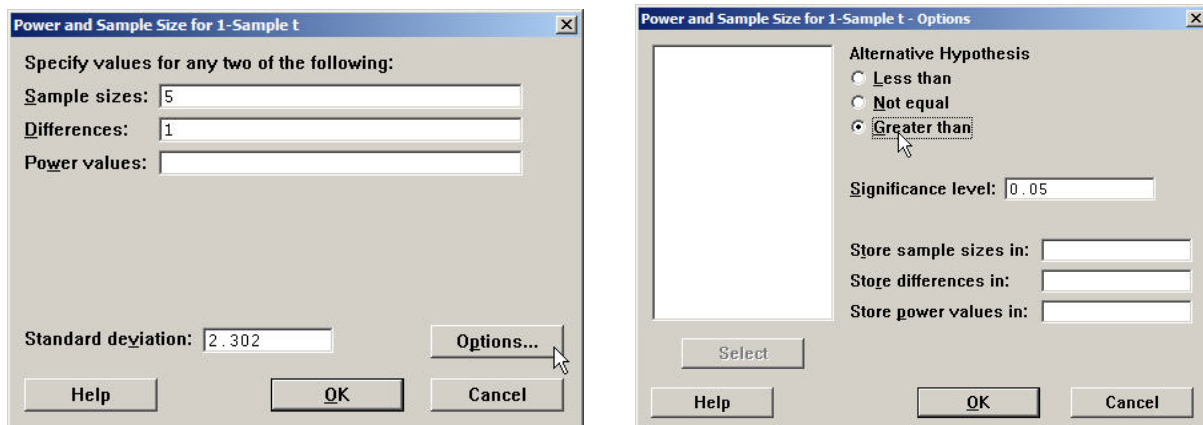
Do these figures present evidence at the 5% level that the process should be placed in full time operation?

Here, the appropriate hypotheses are  $H_0 : \mu = 50$  versus  $H_a : \mu > 50$ . Since the sample size is small and  $\sigma$  is unknown, the appropriate test statistic is the one-sample  $t$  statistic.

Note that the sample size is 5 and the sample standard deviation is 2.302. This will be important later.

To determine the power of the test assuming  $\mu = 51$ , click on **Stat** -> **Power and Sample Size** -> **1 Sample t**. Since  $H_0 : \mu = 50$ , the **difference** between the two  $\mu$  values is  $51 - 50 = 1$ . In the dialogue box which appears, enter the sample size (5), enter the **difference** (1), and the standard deviation (2.302). Click the “Options” button and select the appropriate form of the alternative hypothesis (in this case, it is ‘greater than’).

See below for the relevant windows:



The output shows that the power is 0.203233 when the difference is 1. Hence, the value of  $\beta$  = probability of the type II error =  $1 - 0.203233 = 0.796767$ . To see how the power changes for different settings of the difference, we performed the power analysis for the following difference values:  $\{1, 2, 3, 4, 5, 6, 7, 8\}$ . The output below shows the increasing power of the test for alternatives where  $\mu$  is further away from the hypothesized value  $\mu = 25$ . The following output was modified so that it could all fit efficiently. See below for the corresponding output:

## Power and Sample Size

## 1-Sample t Test

Testing mean = null (versus &gt; null)

Calculating power for mean = null + difference

Alpha = 0.05 Assumed standard deviation = 2.3

Difference	Sample Size	Power	Difference	Sample Size	Power
1	5	0.203233	5	5	0.987931
2	5	0.486412	6	5	0.998613
3	5	0.771631	7	5	0.999900
4	5	0.934030	8	5	1.00000