In this lab session you will

- use graphs to check normality.
- generate random data to gain experience in assessing normality.
- use transformations to make a relationship more linear.
- examine outliers and influential observations in regression.
- use MINITAB to construct simultaneous confidence intervals and perform multiple comparisons.

1. **Checking normality:**  You may skip this section if your work for Assignment 3 covered it.

   The following data are the heights of the left frontal sinus in 14 fossils:
   
   42 27 25 40 33 31 42 34 35 25 29 30 29 35

   Enter the data into c1, and create a dotplot using the command
   
   ```
   MTB > dotplot c1
   ```

   This creates a dotplot in the session window. To create a dotplot in the Graph window, use the menu **Graph Dotplot**, which uses a built-in MINITAB macro. While you’re in the menu, create a histogram as well.

   Do the data appear normal? With a dotplot or histogram, it’s difficult to tell with only 14 points. Create a qq-plot using the commands
   
   ```
   MTB > nscore c1 c2
   MTB > plot c2*c1
   ```

   The qq-plot should be reasonably close to a straight line, suggesting that the assumption of normality is OK.

   To formally test the hypothesis that the data are normally distributed, you need to use the menu **Stat → Basic Statistics → Normality test**.

   The following data represents the October snow cover for Eurasia during 1970–79, in millions of square kilometers:
   
   6.5 12.0 10.0 10.7 7.9 21.9 12.5 14.5 9.2

   Test the hypothesis that these come from a normal distribution. You will notice that there are three tests for normality available in the **Normality test** menu. Try them all — the plot is the same in each case, but the *P*-value is different, because they involve different test statistics.

   In all three tests the hypothesis would be accepted at the 0.05 significance level. However, there are only 9 observations, so departure from normality would have to be substantial to be significant.

2. **Generating random data**

   It is sometimes difficult to judge whether a dotplot or histogram is consistent with a normal distribution, or whether a qq-plot is close enough to being straight. One way to gain experience is to use simulated data that you know comes from a normal distribution, and see what the graphs look like.

   Generate 20 observations from a normal distribution with mean 0 and standard deviation 1, using the command
   
   ```
   MTB > rand c1 20 normal(0,1)
   ```
MTB > random 20 c6

(Note that if you don’t want a standard normal distribution, or require another distribution such as F or t, the random command needs a subcommand.)

Now generate 4 more samples of 20 observations from N(0,1) by typing random 20 c7-c10.

Repeat this procedure (thus overwriting your samples), using the menu
Calc → Random Data → Normal

entering 20 in the top box and c6–c10 in the large box.

Create a normal qq-plot for each of the 5 samples, and examine it. You will probably find that some of the plots don’t look very straight at all! This shows that with 20 observations, a qq-plot which is not very straight may still be consistent with an assumption of normality. In one sense this is convenient, because you can assume normality and proceed with your statistical analysis; however, it is also the case that you haven’t proved the assumption of normality — you simply haven’t proved it to be false.

Now generate a sample of 100 observations from an $F_{2,16}$ distribution, using the menu. How many of them would you expect to be larger than the 0.95 quantile of $F_{2,16}$? Check how many are actually larger than $F_{2,16}(0.95)$; an easy way to do this is to sort them first — for example, if the sample is stored in c11, type

MTB > sort c11 c12
MTB > print c12

3. Transformations to create a linear relationship

Here are some measurements taken on coffee plants: $x =$ water potential, $y =$ leaf diffusive resistance.

<table>
<thead>
<tr>
<th>$x$</th>
<th>12</th>
<th>12</th>
<th>13</th>
<th>13</th>
<th>14</th>
<th>16</th>
<th>17</th>
<th>17</th>
<th>18</th>
<th>22</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>24</th>
<th>24</th>
<th>25</th>
<th>25</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>10</td>
<td>14</td>
<td>8</td>
<td>11</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>18</td>
<td>12</td>
<td>17</td>
<td>32</td>
<td>20</td>
<td>24</td>
<td>26</td>
<td>44</td>
<td>36</td>
<td>54</td>
<td>52</td>
</tr>
</tbody>
</table>

Enter the data, labelling the columns $x$ and $y$. Do a scatter plot of $y$ vs $x$. Compare the shape of the curve with the diagram on page 103 of the printed lecture notes; the curve should be straightened by transforming “$x$ up” and “$y$ down”. Recall the power transformation ladder:

UP ← cube square no transformation square root cube root log reciprocal → DOWN

We can begin by transforming $x$ up; create a column containing $x^2$ using either the command

MTB > let x2 = x**2

or the menu: Calc → Calculator.

Plot $y$ against $x^2$. It is still obviously curved, so calculate $x^3$ and plot $y$ against it. The relationship is still not linear, so it may be better to transform $y$ down. Try different transformations on $y$, starting with the weakest, $\sqrt{y}$ (type the command let sqrtym = sqrt(y) or similar). Make the transformation stronger until the relationship looks linear. (note: there may not be one obviously “best” transformation.)

Finally, try transformations on both $x$ and $y$ ($x$ up, $y$ down). Is there any improvement on transforming $y$ alone?

4. Outliers and influential observations in regression

In the following table, $x$ is the October-November snow cover in a particular region (in millions of square kilometers), and $y$ is the average December-February temperature (in degrees C) from 1969/70 to 1981/82.
Enter the data, and plot $y$ against $x$. Can you spot any outliers or influential points?

Fit a linear regression line, and graph the standardised residuals against $x$ (use the **Graphs** button in the menu). Is the large standardised residual (the one outside the range $(-2, 2)$) the outlier you spotted?

The observation in the bottom right-hand corner of the scatterplot is a considerable distance from the other points, and would be expected to be an influential point. Examine the MINITAB output from the regression, and you will see this point marked with an X under “Unusual Observations”. To see how influential this observation is, we will fit a regression line after removing it. A simple way of removing the point is to change either its $x$ value or $y$ value to * (missing value). Remove the point, and run the regression again.

Examine the output. Is the equation very different from the equation with the point included? Do a fitted line plot using the menu:

**Stat → Regression → Fitted Line Plot**

and look at the slope of the new line — it should be much flatter than before (check by doing a fitted line plot with the point included).

How well does a straight line fit without the influential point? What percentage of the variation in temperature is explained by the amount of snow cover? Examine the $P$-value for testing whether the slope is significantly different from 0.

Examine the value of $R^2$ and the $P$-value for the regression with the influential point included. It is evident that what appears to be a significant linear relationship, explaining over half the variation in the response variable, is highly influenced by one point. This means that conclusions based on this regression line could be quite shaky.

Now remove the influential observation again, and this time also remove the outlier ($x = 12.75, y = -15.7$). Run the regression again, and compare the equation with that produced by the entire data set. It is not very different! The outlier is probably a correct observation, but it would be interesting to investigate it to see if there was anything unusual about that particular winter, with such a small snow cover associated with a low average temperature.

Finally, for these data, snow cover could be validly seen as the response variable and temperature as the explanatory variable. For the entire data set, run a regression of $x$ on $y$, and compare the output with that resulting from the regression of $y$ and $x$. Why is $R^2$ the same? (think of the correlation coefficient, $r$) Should the fitted line be the same? Transpose the equation from the regression of $y$ on $x$, and compare it with the equation from the regression of $x$ on $y$. Why are they different?

### 5. One-way ANOVA

Students were randomly assigned to three study methods to determine the effect of study technique on learning. An assessment at the end of the trial gave the following test scores:

<table>
<thead>
<tr>
<th>Technique</th>
<th>Test score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read only</td>
<td>15 14 16 13 11 14</td>
</tr>
<tr>
<td>Read and underline</td>
<td>15 14 15 10 12 14</td>
</tr>
<tr>
<td>Read and take notes</td>
<td>18 18 18 16 18 20</td>
</tr>
</tbody>
</table>
• Enter the data into columns c1-c3 and carry out a one-way ANOVA.
• State $H_0$ and $H_1$ for this problem.

• What is the distribution of the $F$ statistic, if $H_0$ is true?

• What is the $P$-value, to the first significant (i.e. non-zero) digit?
  Note: In the ANOVA output, it is rounded to 0.000; to find the 4th decimal place, you need to use the menu Calc → Probability Distributions → F and enter the observed $F$ statistic as the Input Constant; alternatively, you can use the equivalent cdf command.

6. Simultaneous (“multiple comparison”) confidence intervals

MINITAB can construct simultaneous C.I.’s for the differences between all pairs of means, using the Fisher method or the Tukey method. The data need to be in stacked form.

Stack the data into c4, with the “subscripts” (i.e. the levels of the factor Technique in c5). Perform the ANOVA with the commands

MTB > oneway c4 c5;
SUBC> fisher.

Fisher method of multiple comparisons: Now run the ANOVA using the menu: Stat → ANOVA → One-way, clicking on the button Comparisons and ticking the box Fisher’s, individual error rate. Check that the output is identical to that produced by the commands.

• What is a 95% CI for $\mu_2 - \mu_3$? (note that if you label the factor levels “Read only”, etc., MINITAB will order them alphabetically when displaying CIs — this may not be in the order you want).

• MINITAB gives a “Critical value” of 2.131. Where does this critical value come from? (guess if you don’t know)

• What is the family error rate here? What does it mean?

• Draw a diagram to show the results of the multiple comparisons.

Tukey method of multiple comparisons: Now, repeat the above ANOVA using the subcommand Tukey in the session window, or use the menu and tick the box Tukey’s, family error rate.

• What is a 95% CI for $\mu_1 - \mu_2$?

• Where does the critical value of 3.67 come from?

• What is the individual error rate here? What does it mean?

• Change the family error rate from 5% to 10% by clicking on the button Comparisons, and in the box to the right of Tukey’s, family error rate, change the number 5 to 10. What is the individual error rate now?