E6 DATA ANALYSIS TASK

Background

Before beginning this task, do the short ‘W6 Data Analysis 1’ and ‘W7 Data Analysis 2’ workshops in the ‘Laboratory Handbook’. These contain information on a number of the concepts described in this task as well as instructions on using Excel*. Although these tasks are set in the context of chemistry, they are applicable across any discipline where information is extracted from experimental measurements.

The Data

The instructions below assume that you are using Excel as this is the most common spreadsheet program used in science, commerce and business. You will also be able to complete the tasks using other spreadsheet programs but, in this case, you may need to refer to its manual or help facility.

Download the experimental data file from Sydney eLearning (WebCT) and open it in Excel. For convenience, it has been produced in ‘comma separated value’ (.csv) format which allows simple exchange of data between different spreadsheet programs. The first thing you should do is to save it as an Excel file to ensure that your work is preserved.

The file contains 5 columns of data. The first two, headed ‘Molar Mass from Flow Time’ and ‘Molar Mass from Density’, contain the raw experimental measurements entered during the flow-time and density experiments respectively.

In the next three columns, these data have been grouped to give the number (frequency) of measurements over data ranges. The first of these columns gives the bottom of the range and the next two columns give the frequencies of measurements for the two experiments. For example, in the figure below, 3 students recorded a molar mass between 41.36 – 42.24 g mol\(^{-1}\) and 1 student recorded a molar mass between 42.24 – 43.32 g mol\(^{-1}\) in the flow time experiment.

<table>
<thead>
<tr>
<th>Range</th>
<th>Frequency from Flow Time</th>
<th>Frequency from Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.36</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>42.24</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>43.12</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 1. Example data from columns D – E showing frequency of experimental data.

Assessment of this Task

The tasks below are assessed via a quiz on Sydney eLearning (WebCT). You will be able to attempt the quiz as many times as you like and it will record your final attempt.

As detailed below, the questions in the quiz depend on the data that is submitted during the week so you may need to go through the questions before working through the analysis in detail.

* These workshops and this task are best accomplished using Excel 2007 on a PC. Unfortunately, Excel 2008 on a Mac has a number of missing features, including the Solver add-in, which may make it difficult to complete parts of this task. It is therefore recommended that you use the PC version, on a university computer.
Task 1 – Data Spread

- Draw an XY scatter plot (with straight lines) containing the frequency data for both experiments†.

Excel will generate a graph according to a default template including line colours and styles. You can, and should, change these if you think they are not clear. If you are going to print in black and white, pick line styles that will be clear. In reports, simple, clear graphs are usually best. If you have not used Excel before, investigate the options to change the look of the graph.

Excel will also pick the scales used on the \( x \) and \( y \)-axes. You can, and should, change these if they are inappropriate. For example, if a student has mistakenly entered a molar mass which is several orders of magnitude too large, Excel will accommodate this resulting in a graph which is too squashed.

- The graph should contain a suitable label on each axis and a title which concisely describes its content. Use the ‘Help’ option in Excel if you cannot work out how to add these features.

- In the quiz, you will be asked to select the correct graph and to identify the peak in the distribution for both experiments.

Typical experimental data will show a spread of results with a Gaussian or normal distribution. If it does not, this is usually due to a mistake in the experimental design.

- Read the section in W6 on data spread.

Task 2 – Assessing Precision and Accuracy

- Read the section in W6 on the difference between precision and accuracy. They are not the same and you need to be aware of the difference between them for the next tasks.

Task 3 – Statistical Treatment of Data

The deviation of the average of a data set from the true value is an indicator of the accuracy of the data. The standard deviation is one indicator used to describe the spread of the measurements and describes their precision. Both the accuracy and the precision should be indicated when reporting data. Precision is very much like risk. When deciding on where to invest, a smart investor will want to know not only the average return but also the risk. An average profit of 50% sounds great but knowing that this actually reflects returns between -50% and 100% shows that the deal is actually very risky.

- Work out the average value of the molar mass obtained from the two experiments‡.

- Work out the standard deviation for each data set‡. The standard deviation has the same units as the measurement.

- In the quiz, you will be asked for the average values and the standard deviations for the two experiments. You will be asked to compare the accuracy and precision of the two experiments.

† If you do not know how to do this, read the instructions in the ‘W7 Data Analysis 2’ workshop in the ‘Laboratory Handbook’.
‡ If you do not know how to do this, read the instructions in the ‘W6 Data Analysis 1’ workshop in the ‘Laboratory Handbook’.
Task 4 – Reporting Results

The standard error, $s_M$, of the mean is a measure of the error in the final answer. It can be calculated from the standard deviation, $s$, and the number of measurements, $N$, using the formula:

$$s_M = \frac{s}{\sqrt{N}}$$

It is good practice when presenting results to use the following to indicate the error in the measurement:

$$value = mean \pm \frac{s}{\sqrt{N}}$$

- In the quiz, you will be asked for the molar mass for the two experiments in this format, including the units.

Task 5 - Outliers

The data set may contain points which are suspect – they may be due to data entry errors, a failure of equipment, incompetence or even fraud. Unfortunately, such problems are part of life (and hence science). There is no definitive method for assessing the reliability of individual measurements and several are in widespread use. The Q-test detailed in W6 is unsuitable for large data sets such as those generated in this experiment. Large data sets should fit a normal distribution and methods based on this are preferred. Here, you will use Chauvenet's criterion.

Chauvenet's criterion uses the mean and standard deviation of the observed data. Based on how many standard deviations the suspect point is from the mean, the normal distribution can be used to determine the probability that it is a normal measurement. This probability is multiplied by the number of data points taken. If the result is less than 0.50, the suspicious data point may be discarded.

For example, suppose a value is measured experimentally in several trials as 9, 10, 10, 10, 11, and 50. The mean is 16.7 and the standard deviation 16.34. The value of 50 looks suspicious. To test whether it should be discarded, the following steps are taken:

1. 50 differs from the mean of 16.7 by $50 - 16.7 = 33.3$.
2. This is slightly more than two standard deviations as $33.3 / 16.34 \sim 2.0$.
3. The normal distribution table in the appendix shows that 95% of the data lies within this number of standard deviations from the mean. The probability of taking data more than two standard deviations from the mean is roughly $100 - 95 = 5\%$ or 0.05.
4. Six measurements were taken, so the statistic value (data size multiplied by the probability) is $6 \times 0.05 = 0.3$.
5. Because $0.3 < 0.50$, according to Chauvenet's criterion, the measured value of 50 can be discarded.
6. Removal of this point leads to a new mean of 10 and a reduced standard deviation of 0.7.
If such data can be identified, they can be eliminated from the calculations leading to more accurate and more precise data. However, great care and professionalism\(^\text{\textsuperscript{\textsection}}\) are needed when doing this as it effectively introduces a bias which may be worse than accepting the lack of precision.

- In the quiz, you will be provided with *one* suspect data point and you will be asked to follow this method to work out whether Chauvenot’s criterion would suggest this point is an outlier.

**Appendix – Normal Distribution**

Percentage of data within the normal distribution curve.

<table>
<thead>
<tr>
<th>number of standard deviations from mean</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>percentage of data within normal distribution</td>
<td>68.4%</td>
<td>95.2%</td>
<td>99.3</td>
<td>99.7%</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

1. If the measurement is more than 5 standard deviations from the mean, use 99.9%

2. Where the number of standard deviations is not close to an integer, interpolate between the values. For example, suppose a point is 2.5 standard deviations from mean.
   (i) 95.2% of the data is within 2 standard deviations
   (ii) 99.3% of the data is within 3 standard deviations
   (iii) Approximately 97.3% of the data is within 2.5 standard deviations.

\(\text{\textsection}\) You can certainly improve the accuracy and precision of your data by removing outliers. Is it ethical and in the interest of others to remove the point, or merely to make you look better?