Statistics for Chemists: a Basic Study Guide

Dr G Currell with support and encouragement from the Royal Society of Chemistry, UK.

This study guide is in the process of development - any comments and suggestions welcome

Introduction

This guide:

- identifies the key topics that provide an understanding of the basic statistics relevant to chemistry - you can then look up these topics, either on-line or in textbooks on basic statistics,
- provides hyperlinks to QVA tutorials (questions and video answers) that provide tutorial resources in the key topics supported by video feedback,
- provides hyperlinks to video tutorials on the use of MS Excel for data analysis,
- provides hyperlinks to other resources in the Royal Society of Chemistry's 'Discover Maths for Chemists' website.

Each QVA tutorial typically gives about five on-line questions with video answers. The QVA tutorial identified at any stage may contain some questions which relate to topics that you have not yet covered. This is due to the fact that there are a lot of *interconnections* between different aspects of statistics, and it will be useful for you to return to these QVA tutorials at a later time.

It is intended that this guide is flexible for different learning styles:

- some students may prefer to first assess their understanding of the key topics by using the hyperlinks to relevant QVA tutorials, which themselves might provide sufficient understanding,
- others may prefer to first study each topic using texts or the internet before testing themselves with the QVA tutorials.

Many of the electronic resources are available from the website for the book: Essential Mathematics and Statistics for Science, 2nd Edition, by Graham Currell and Antony Dowman, Wiley-Blackwell, 2009

Use of Microsoft Excel for data analysis

Excel is very convenient for managing experimental data - recording, sorting, presenting, etc. Excel also provides statistical functions that perform many of the basic statistical calculations that are needed for chemistry.

This guide provides links to video tutorials on the use of Excel in these areas, together with demonstration files that can be downloaded. Videos for basic calculations in Excel 2007, together with the procedures for drawing and analysing x-y graphs are given below. Links to more specialised techniques are given later in the appropriate sections.

Excel video tutorials:
Basic skills: Formatting, Editing, Calculations, Functions, Addressing
X-Y Graph: X-Y Graph, Trendline, Error Bars, Regression & correlation, Data Analysis Tools
Associated Excel files for download: Basic skills (2007), X-Y Graph (2007)

What does a chemist need to learn first about statistics?

It is a fact of life that experimental results always show some degree of random variation, and a chemist needs to be able to **handle and quantify the resultant uncertainties**.

For example, if we measure the concentration of a solution to be 23.00 mg dm^{-3}, it is very unlikely that the true value is exactly 23.00 mg dm^{-3} - so how do we describe the result in a way that conveys the inherent uncertainty?
The best that we can do might be to say that we are 95% confident that the true value lies between 22.86 mg dm^{-3} and 23.14 mg dm^{-3} - this is an example of a **95% confidence interval**.
To calculate this 'confidence interval' we need some simple mathematics to handle the random variation in experimental data - this is a function of the branch of mathematics called 'statistics'.

Where else does statistics help the chemist?

Statistics also allows a chemist to **make decisions**, based on probabilities, in a way that is understood and accepted universally by other scientists. This is the area of statistics called **hypothesis testing**, which provides a range of **tests** suitable for different problems.

For example, a number of replicate measures of a pollutant in industrial waste water might suggest that a regulatory level has been exceeded - but could the difference be due only to random uncertainty in the measurement itself? A *t*-test could be used to **decide**, with a **defined level of confidence**, whether the regulation has been broken.
Many other tests are available, e.g.
ANOVAs (Analysis of Variance) provide a range of tests on multiple sets of data that can pick out systematic changes.
Shewhard and CUSUM charts provide a statistical basis for monitoring the consistency/reliability of an on-going production process, and identify warning points when action needs to be taken.

The sources of experimental variation fall into TWO main *categories*:

- Variations in the **measurement process** itself, e.g. variations in the output from a spectrophotometer when exactly the same measurement is repeated.

- Variations in the **system being measured**, e.g. the emission from a radioactive isotope shows random fluctuations in addition to its long-term exponential decay.

We use statistics for TWO main *purposes*:

- To **describe** the data, by finding suitable numerical and graphical methods which convey a useful summary or 'picture' of the data.

- To **derive conclusions** based on a calculated probability of confidence

---

Describing experimental error and uncertainty

Many chemists use the term 'experimental error' to describe the inherent *uncertainty* created by the variations in experimental data - it is NOT intended to suggest that a *mistake* has been made in the experimental process.
We use a variety of terms to describe the relationship between the measured value and the unknown 'true' value:

Key topics: Accuracy, precision, random error, systematic error, bias
QVA tutorials: Accuracy and precision

Data samples and statistical populations

A statistical *sample* is a limited number of measurements that we use to *estimate* the more accurate result that we would get if we took all possible measurements (the *population*).
We use the values (statistics) measured from the *sample* to infer the 'true' values (parameters) for the *population*.
Note that the concentrations of a number of replicate *chemical* samples might be the data values for just one *statistical* sample!

Key topics: Statistical samples, statistical populations
QVA tutorials: Experimental errors and uncertainties

Describing a data set

When interpreting, or describing, experimental data, scientists agree on a limited number of numerical values (*statistics* for samples, and *parameters* for populations) which give good descriptions of the general form of the data.

Key topics:
Average value, mean, median - describe *where* the data set is on the value axis.
Standard deviation, interquartile range - describe the *spread* of the data along the axis.
Skewness, kurtosis - (more advanced) describe the *shape* of the data.
QVA tutorials:
Mean and standard deviation,
Non-parametrics: median and quartile,
Mean, median, standard deviation & variance - calculations
Skewness and kurtosis (advanced) - video feedback in preparation
Other resources:
Basic statistics glossary - M Seery, Dublin Institute of Technology
Basic statistical analysis - M Seery, Dublin Institute of Technology
Excel tutorial file for download: (video commentary in preparation)
Functions, Charts, Histograms**:** Statistics (2007)

Data distribution and confidence intervals

When we quantify uncertainties, it is very useful to know if the experimental variations follow some specific mathematical description.
In many experimental situations it is possible to assume that the variation in the data follows the probabilities of a Normal, or Gaussian, distribution.
Having made repeated (replicate) measurements of an unknown value, it is possible, assuming a normal distribution, to calculate a 95% confidence range for its true value.

Key topics:
Frequency graphs and histograms, Normal (or Gaussian) distribution
Central limit theorem, Confidence interval of the mean
QVA tutorials:
Frequency distribution of experimental data
Experimental uncertainty and the normal distribution
Confidence interval of replicate measurements
Excel videos:
Analysing replicate data: Functions, Data Analysis Tools, Random data demo
Associated Excel file for download: ExcelDataUncert01 (2007)
Excel tutorial file for download: (video commentary in preparation)
Distributions: Distributions (2007)

Statistics of straight line calculations - linear regression

It is very common to find a linear relationship between two variables, e.g. between *absorbance* and *concentration* in a spectrophotometric measurement. It is very convenient that statistics provides very easy ways of analysing these results, using the process of linear regression.

Key topics: Linear regression, slope, intercept, best-fit trendline, calibration line
QVA tutorials:
Study unit: Straight lines and linear regression
Excel videos:
X-Y data analysis: X-Y Graph, Functions, Data Analysis Tools, Random data and Error Bars
Beer's law example: Residuals, Best-estimate *x*-value, Uncertainty in *x*-value
Associated Excel files for download: ExcelDataUncert01 (2007), Beers Law

---

Assessing experimental data

Before applying any statistical analysis to experimental data, it is essential to know the type and characteristics of the data itself. Although experimental data often follows the normal distribution, other important distributions include the binomial and Poisson distributions.

Key topics:
Data types: ratio, interval, ordinal, nominal.
Normal (or Gaussian), Poisson, binomial distributions.
QVA tutorials:
Assessing experimental data
Testing for normality - video feedback in preparation

Propagation of errors / uncertainties

The final calculated value for the result of an experimental measurement is often affected by more than type of experimental variation, e.g. uncertainty in the judgement of the exact end-point of a titration as well as in the measurement of volume delivered.
The combination of these uncertainties is called the 'propagation of errors', and can be calculated easily, provided key rules are obeyed.

Key topics: **Combining uncertainties, propagation of errors**
QVA tutorials:
Combining uncertainties, propagation of errors
Errors and uncertainties in concentrations and dilutions

Hypothesis testing

A hypothesis test enables a chemist to set up a 'yes/no' type of question based on data which has inherent uncertainty, e.g. 'has the catalyst caused an increase in reaction rate?'. The typical hypothesis test provides the level of confidence with which you could say 'yes' - although it does not give a level of confidence in saying 'no'. A hypothesis test is good at 'proving' a positive, but not good at 'proving' a negative.

Key topics:
***t*-test** - tests for differences involving the mean (or means) of one or two data sets**.**
Chi-squared test - tests whether the numbers in categories differ from expected.
ANOVAs - tests for differences involving the means or more than two data sets**.**
Non-parametric tests - range of tests to perform similar functions to *t*-tests and ANOVAs for data that does not meet their requirements in type or distribution.
QVA tutorials: (in preparation)
t-tests
Chi-squared tests
ANOVAs
Non-parametric tests
Other resources:
One sample *t*-test - M Seery, Dublin Institute of Technology
Paired *t*-test - M Seery, Dublin Institute of Technology
One-way ANOVA - M Seery, Dublin Institute of Technology
Excel tutorial files for download: (video commentaries in preparation)
t-tests, *F*-tests & correlation: Parametric Tests (2007)
Chi-squared tests: Chi-Squared (2007)
ANOVAs: ANOVAs (2007)
