Contents

1 Chemometrics : Introduction to Module **2**
1.1 Quantitative nature of analytical chemistry 3
1.2 Learning Outcomes 5
1.3 Statement of Syllabus 5
1.4 Revision of Science Maths 3 6
1.5 Text Books 6

2 Introduction to R **7**
2.1 The R Project for Statistical Computing 7
2.2 Downloading and Installing R 9
2.3 Output Graphics from chemCal R package 10
2.4 Example of R Analysis: Titration experiment 11
2.5 Measures of Centrality and Dispersion 12
2.6 Output of R Procedure 13

3 Revision of Topics from MA4603 **14**
3.1 Statistical significance 14
3.2 Hypothesis testing: introduction 16

1 Chemometrics : Introduction to Module

- Kevin O'Brien
- email: kevin.obrien@ul.ie

1.1 Quantitative nature of analytical chemistry

- Modern analytical chemistry is overwhelmingly a quantitative science. A quantitative answer is much more valuable than a qualitative one.
- It may be useful for an analyst to claim to have detected some boron in a distilled water sample, but it is much more useful to be able to say how much boron is present.
- Often it is only a quantitative result that has any value at all.For example, almost all samples of (human) blood serum contain albumin; the only question is, how much ? Even where a qualitative answer is required, quantitative methods are used to obtain it.
- Quantitative approaches might be used to compare two soil samples. For example, they might be subjected to a particle size analysis, in which the proportions of the soil particles falling within a number say 10, of particle-size ranges are determined.
- Each sample would then be characterized by these 10 pieces of data, which could then be used to provide a quantitative assessment of their similarity.

The extremely rapid development of analytical techniques in biology and chemistry has left data analysis far behind, and as a result the statistical analysis and interpretation of the data has become a major bottleneck in the pipeline from measurement to information.
(Quote from "Chemometrics with R", R. Wehrens, Springer UseR! Series).

1.2 Learning Outcomes

- To give students a clear understanding of the importance of statistical methods in their work.
- To introduce students to the most widely used statistical techniques in the chemical process industries.
- To develop skills in the use of these techniques through actual case studies using statistical software packages

1.3 Statement of Syllabus

1. **Hypothesis testing** - type I and type II error, one and two-tailed tests, oc curves.
2. **Statistical process control** - various charts, mean/range, individuals/moving range, cusum charts.
3. **Capability studies** - capability indices.
4. **Correlation and Regression** - method of least squares, multiple regression, linear and non-linear models, regression analysis, analysis of residuals.
5. **Data Visualization** - Importance of plotting data.
6. **Design of experiments and analysis of variance** - one and two way ANOVA, interaction, factorial designs, responses and factors, Plackett-Burman design, response surface methodology.

1.4 Revision of Science Maths 3

- This module follows on from **Science Maths 3 - Introduction to Statistics MA4603** with Dr Joseph Lynch.
- If you could make some quick revisions of the course material there over the next three or four weeks that would be great.
- We will not be doing much in the way of pen-and-paper calculations (there will be a few questions here and there).
- For the first few classes, we will go back over the Normal Distribtion, Hypothesis Testing Correlation, Simple Linear Regression, P-values etc

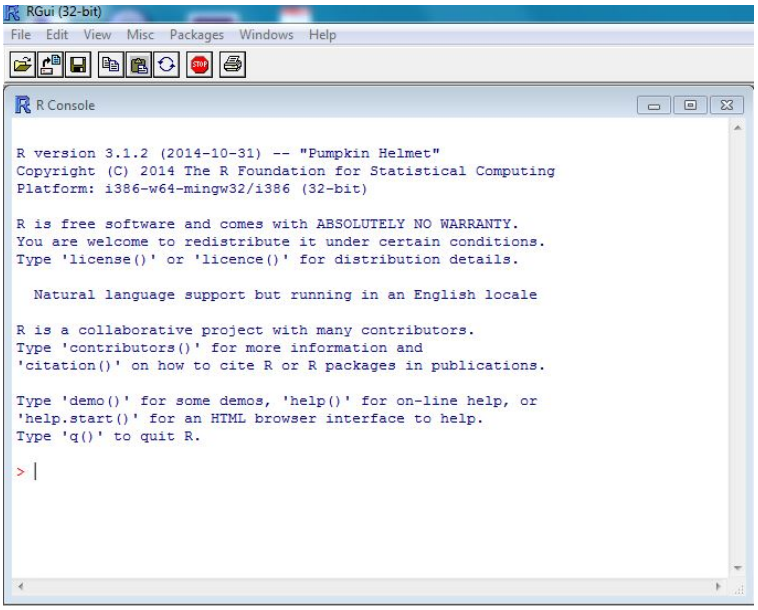
1.5 Text Books

My own notes would be suffficient. I will publish them on SULIS. You can have a look at the folowing publications too.

1. Statistical Analysis Methods for Chemists (Author : William P Gardiner)
2. simpleR - Using `R` for Introductory Statistics (Author : John Verzani)
3. An Introduction to `R` (Authors: The R Project)

2 Introduction to R

2.1 The R Project for Statistical Computing



R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools one can find implemented

- linear and nonlinear modelling,
- classical statistical tests,
- time-series analysis,
- classification,
- clustering,
- ...and many more.

One of R's strengths is the ease with which well-designed publication quality plots can be produced. including mathematical symbols and formulae where needed.

- `R` is a computing software for statistical analysis
- The package is available for all popular operating systems: Windows, Mac or Linux.
- It is free!
- Everyone (knowledgeable enough) can contribute to the software by writing a package.
- Packages are available for download through a convenient facility
- It is fairly well documented and the documentation is available either from the program help menu or from the web-site.
- It is the top choice of statistical software among academic statisticians but also very popular in industry specially among biostatisticians and medical researchers (mostly due to the huge package called Bioconductor that is built on the top of `R`).
- It is a powerful tool not only for doing statistics but also all kind of scientific programming.

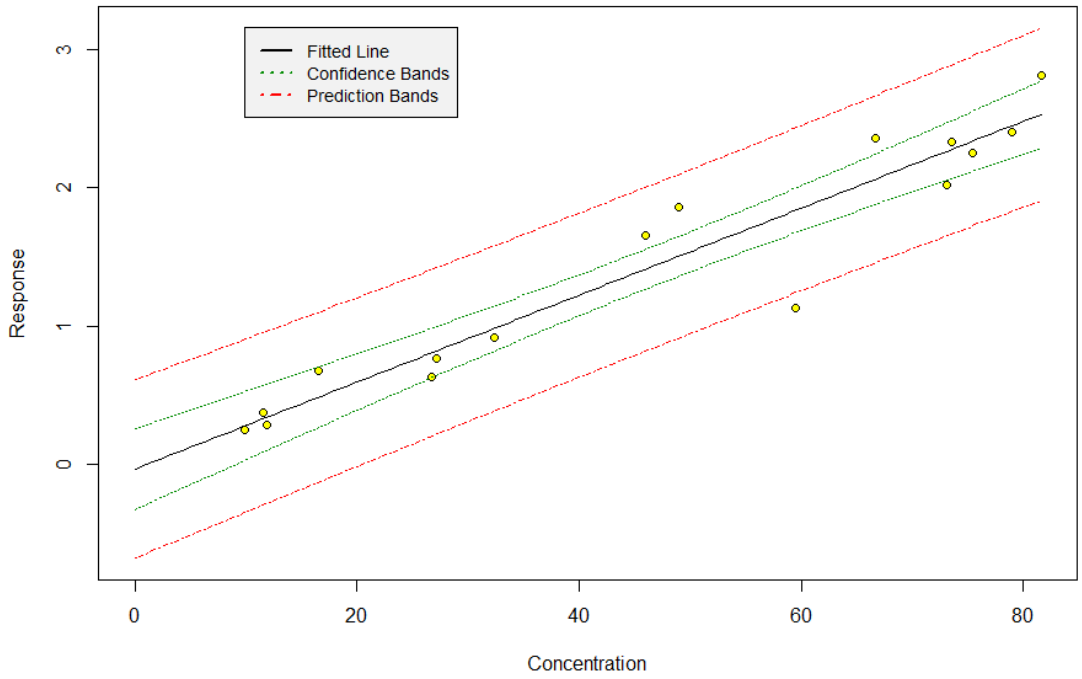
R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent. integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hard-copy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

2.2 Downloading and Installing R

- R can be downloaded from the CRAN website: *http://cran.r-project.org/*
- You may choose versions for windows, mac and linux.
- As per the instructions on the respective pages, you require the "base" distribution.
- Now you can download the installer for latest version of R , version 3.2.1.
- Select the default settings. Once you finish, the R icon should appear on your desktop.
- Clicking on this icon will start up the program.

2.3 Output Graphics from chemCal R package

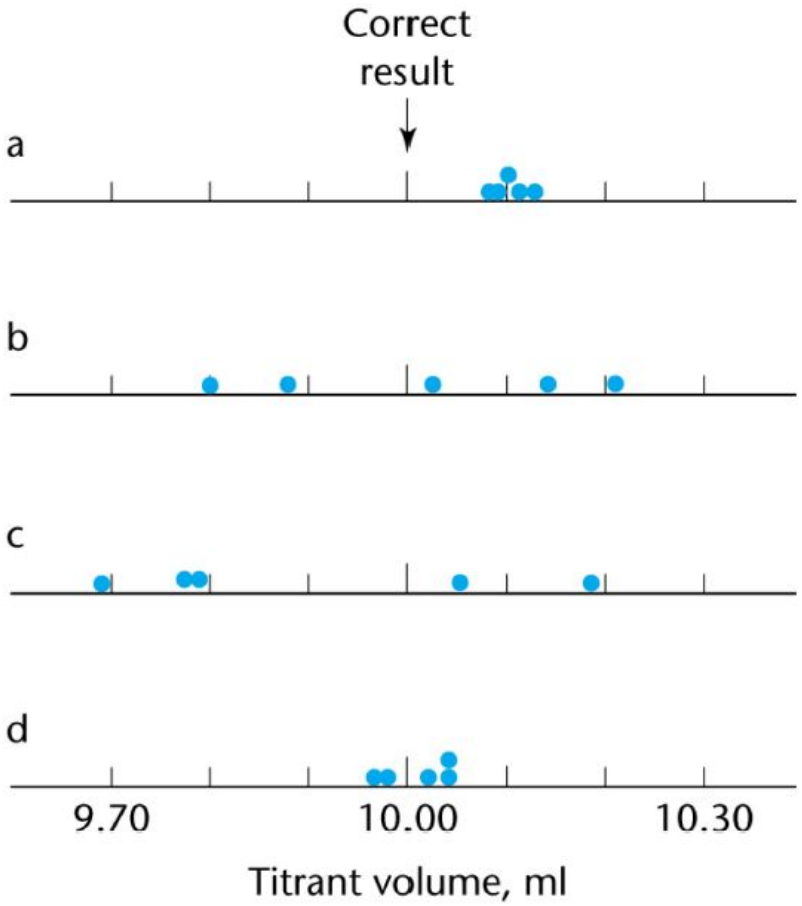


Remark Material covered in Simple Linear Regression will feature a lot.

2.4 Example of R Analysis: Titration experiment

Cosider the following titration experiment where 4 Students performing the same experiment five times, hence each yield 5 results.

Student	Results	(ml)				Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, biased
B	9.88	10.14	10.02	9.80	10.21	Imprecise unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.98	10.02	9.97	10.04	Precise, unbiased



2.5 Measures of Centrality and Dispersion

Two criteria were used to compare these results, the average value (technically know as a measure of centrality and the degree of spread (or dispersion).

- The average value used was the arithmetic mean (usually abbreviated to *the mean*), which is the sum of all the measurements divided by the number of measurements.
- The mean, $\bar{x}$, of n measurements is given by

$$\bar{x} = \frac{\sum x}{n}$$

- The dispersion (or spread) was measured by the difference between the highest and lowest values (i.e. the range).
- A more useful measure, which utilizes all the values, is the sample standard deviation, s, which is defined as follows:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

- Recall: variance is the standard deviation squared.
- The coefficient of variation is the relative standard deviation (in percentage)

2.6 Output of R Procedure

```
#Comuting means

apply(Titra,1,mean)

#       A      B      C       D
# 10.0950 9.9600 9.9300 10.0025

#and standard deviation

apply(Titra,1,sd)

#          A          B          C          D
#0.01290994 0.15055453 0.23036203 0.03304038
```

3 Revision of Topics from MA4603

3.1 Statistical significance

- **Statistical significance** is a mathematical tool used to determine whether the outcome of an experiment is the result of a relationship between specific factors or due to chance. Statistical significance is commonly used in the medical field to test drugs and vaccines and to determine causal factors of disease. Statistical significance is also used in the fields of psychology, environmental biology, and any other discipline that conducts research through experimentation.

- Statistics are the mathematical calculations of numeric sets or populations that are manipulated to produce a probability of the occurrence of an event. Statistics use a numeric sample and apply that number to an entire population.

- In a scientific study, a hypothesis is proposed, then data is collected and analyzed. The statistical analysis of the data will produce a number that is statistically significant if it falls below 5%, which is called the confidence level. In other words, if the likelihood of an event is statistically significant, the researcher can be 95% confident that the result did not happen by chance.

- Sometimes, when the statistical significance of an experiment is very important, such as the safety of a drug meant for humans, the statistical significance must fall below 3%. In this case, a researcher could be 97% sure that a particular drug is safe for

human use. This number can be lowered or raised to accommodate the importance and desired certainty of the result being correct.

- Statistical significance is used to reject or accept what is called the null hypothesis. A hypothesis is an explanation that a researcher is trying to prove. The null hypothesis holds that the factors a researcher is looking at have no effect on differences in the data.

- Statistical significance is usually written, for example,

$$...t = .02, p < .05...$$

  Here, "t" stands for the statistic test score and "p¡.05" means that the probability of an event occurring by chance is less than 5%. These numbers would cause the null hypothesis to be rejected, therefore affirming that the alternative hypothesis is true.

3.2 Hypothesis testing: introduction

> The process by which we use data to answer questions about parameters is very similar to how juries evaluate evidence about a defendant. *from Geoffrey Vining, Statistical Methods for Engineers, Duxbury, 1st edition, 1998.*

The objective of hypothesis testing is to access the validity of a claim against a counterclaim using sample data

- The claim to be proved is the alternative hypothesis(H_1).
- The competing claim is called the null hypothesis(H_0).
- One begins by assuming that H_0 is true.

If the data fails to contradict H_0 beyond a reasonable doubt, then H_0 is not rejected (recall - we would say "fail to reject" rather "accept").

However, failing to reject H_0 does not mean that we accept it as true. It simply means that H_0 cannot be ruled out as a possible explanation for the observed data.

- Hypothesis testing is a common practice in science that involves conducting tests and experiments to see if a proposed explanation for an observed phenomenon works in practice.
- A hypothesis is a tentative explanation for some kind of observed phenomenon, and is an important part of the scientific method. The scientific method is a set of steps that is commonly employed by those in scientific fields to give scientific explanations for various phenomena.
- One common method of hypothesis testing is known as **statistical hypothesis testing**, and typically deals with large quantities of data.
- Experiments and tests are conducted and the data is collected. If the data collected shows that it is unlikely that the results occurred by chance, it is considered statistically significant and can be used to support a hypothesis.
- The results of hypothesis tests are often expressed in terms of $p-$values.

Interpreting $p-$values

For the purposes of this module, we will use the following rules of thumb. If we do not have a significant $p-$value, we fail to reject the null hypothesis. If we have a significant result, we reject the Null Hypothesis.

- p-value is greater than 0.05 - Not Significant
- p-value is between 0.05 and 0.01 - Significant
- p-value is between 0.01 and 0.001 - Very Significant
- p-value is less than 0.001 - Highly Significant

Example 1

```
grubbs.test(x, two.sided=T)
        Grubbs test for one outlier
data:  x
G = 2.4180, U = 0.4202, p-value = 0.04811
alternative hypothesis: lowest value 3.51 is an outlier
```

Figure 1:

Example 2

```
ks.test(y,pnorm,mean(y),sd(y))
One-sample Kolmogorov-Smirnov test
data:  y
D = 0.1442, p-value = 0.9518
```

Figure 2: