Contents

1 About this Module

1.1 Rationale And Purpose Of The Module

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).

- 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

1.2 Syllabus

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

1.3 Recommended Text

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. Rprovides 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 2.17.
- 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 Statistical Tables using `R`

The following is a fragment of the tables of the values of $F(x)$ for the standard normal ('Z') cumulative distribution function from page 254 of the main textbook.

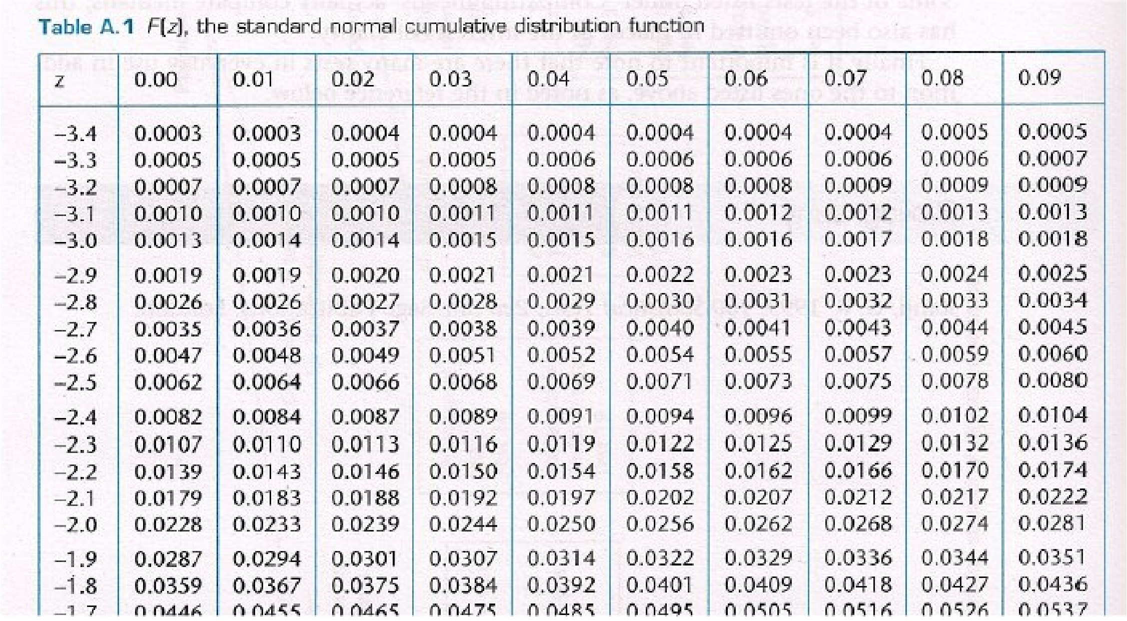


Table A.1 $F[z]$, the standard normal cumulative distribution function

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.4	0.0003	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005
−3.3	0.0005	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007
−3.2	0.0007	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
−3.1	0.0010	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013
−3.0	0.0013	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
−2.9	0.0019	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
−2.8	0.0026	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
−2.7	0.0035	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
−2.6	0.0047	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
−2.5	0.0062	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080
−2.4	0.0082	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
−2.3	0.0107	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
−2.2	0.0139	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
−2.1	0.0179	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
−2.0	0.0228	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
−1.9	0.0287	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
−1.8	0.0359	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
−1.7	0.0446	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537

```
# Segment 1A-1
# Preceding line with the symbol # makes it a comment in R
# The following line produce a single value of the standard normal cumulative
# function. It is the value corresponding to the first value in the table

pnorm(-3.4)

#[1] 0.0003369293
#Then the first row of the table

z=seq(-3.4,-3.31,by=0.01)
pnorm(z)

# [1] 0.0003369293 0.0003494631 0.0003624291 0.0003758409 0.0003897124
# [6] 0.0004040578 0.0004188919 0.0004342299 0.0004500872 0.0004664799
# And all values from the table

z=seq(-3.4,3.4,by=0.01)
pnorm(z)

#  [1] 0.0003369293 0.0003494631 0.0003624291 0.0003758409 0.0003897124
#  [6] 0.0004040578 0.0004188919 0.0004342299 0.0004500872 0.0004664799
# [11] 0.0004834241 0.0005009369 0.0005190354 0.0005377374 0.0005570611
# [16] 0.0005770250 0.0005976485 0.0006189511 0.0006409530 0.0006636749
# [21] 0.0006871379 0.0007113640 0.0007363753 0.0007621947 0.0007888457
# [26] 0.0008163523 0.0008447392 0.0008740315 0.0009042552 0.0009354367
# [31] 0.0009676032 0.0010007825 0.0010350030 0.0010702939 0.0011066850
```

There is more than meets the eye in the table. It is not only the table values that can be explored for the standard normal distribution using R. Recall that the normal distribution is defined by the density function:

$$f(z) = \frac{1}{\sqrt{(2\pi)}} e^{-Z^2/2}.$$

The density represents distribution of probability for a random variable associated with it. The area under the density represents the probability so the that the total area under it is equal to one. The area accumulated up to certain value z_o represents probability that a corresponding random variable takes value smaller than z and this probability defines the cumulative distribution function $F(z)$ which is tabularized.

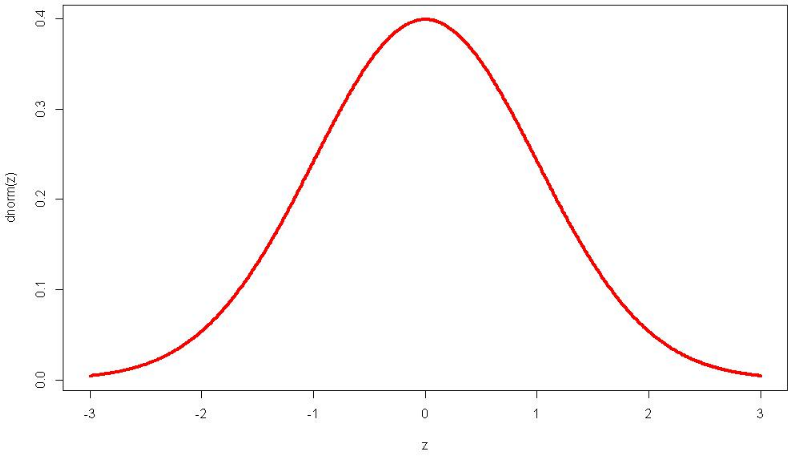
All this can be seen in R. The following code explores various aspects of the standard normal distribution:

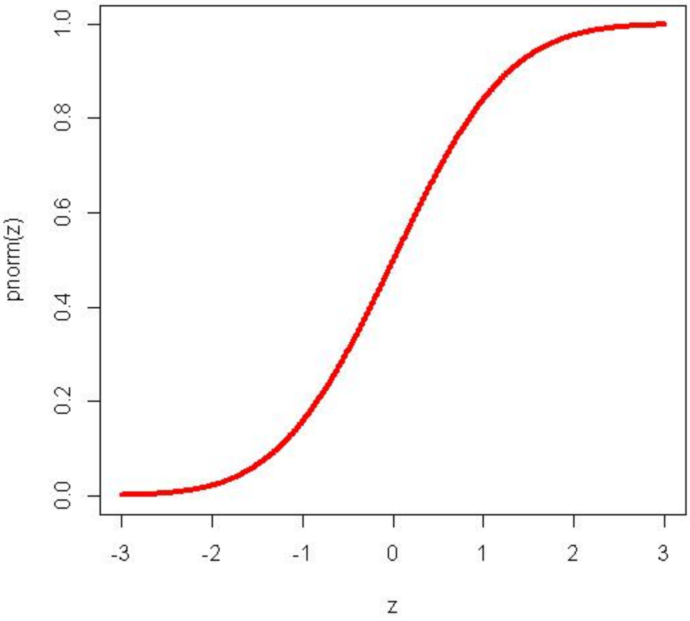
```
#Plotting the density function of the standard normal variable
z=seq(-3,3,by=0.01)
plot(z,dnorm(z),type='l',col="red",lwd=4)

#Plotting the cumulative distribution function (that one from the table)
plot(z,pnorm(z),type='l',col="red",lwd=4)
```

The R code results in the following plots.

- The probability density function.

- The cumulative density function.





2.4 Data Analysis with R

Data from Table 1.1 of the textbook

Table 1.1 Random and systematic errors

Student	Results	(ml)				Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, unbiased
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

This is also given in the text file Table1 − 1.txt, the contents of which is given below:

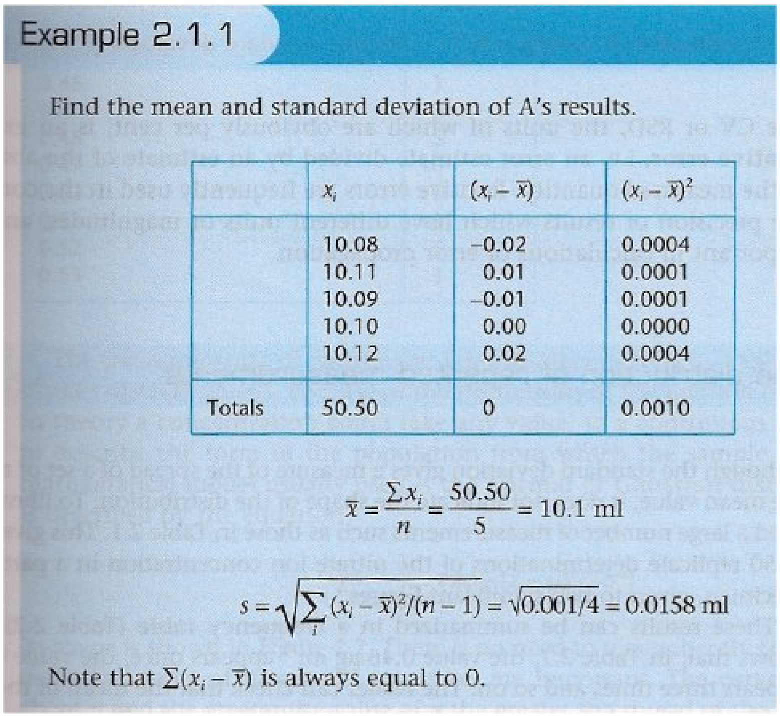
```
A 10.08 10.11 10.09 10.10

B 9.88 10.14 10.02 9.80

C 10.19 9.79 9.69 10.05

D 10.04 9.98 10.02 9.97
```

Reading data from a file to R:

```
#Reading the data from

Titra=read.table("Table1-1.txt", row.names = 1)

Titra

# V2 V3 V4 V5

#A 10.08 10.11 10.09 10.10

#B 9.88 10.14 10.02 9.80

#C 10.19 9.79 9.69 10.05

#D 10.04 9.98 10.02 9.97

#Listing the first row

Titra[1,]

#and the fourth column

Titra[,4]
```

Means and standard deviations

Find the mean and standard deviation of A's results.



Example 2.1.1

Find the mean and standard deviation of A's results.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	10.08 10.11 10.09 10.10 10.12	−0.02 0.01 −0.01 0.00 0.02	0.0004 0.0001 0.0001 0.0000 0.0004
Totals	50.50	0	0.0010

$$\bar{x} = \frac{\sum x_i}{n} = \frac{50.50}{5} = 10.1 \text{ ml}$$

$$s = \sqrt{\sum_i (x_i - \bar{x})^2/(n-1)} = \sqrt{0.001/4} = 0.0158 \text{ ml}$$

Note that $\sum(x_i - \bar{x})$ is always equal to 0.

Means and standard deviations much faster and better

```
#Comuting means

rowMeans(Titra)

# 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
```

Nitrate ion concentration from Table 2.1

Table 2.1 Results of 50 determinations of nitrate ion concentration, in μg ml^{-1} (Also in the file Table $2-1$.txt.)

0.51	0.51	0.51	0.50	0.51	0.49	0.52	0.53	0.50	0.47
0.51	0.52	0.53	0.48	0.49	0.50	0.52	0.49	0.49	0.50
0.49	0.48	0.46	0.49	0.49	0.48	0.49	0.49	0.51	0.47
0.51	0.51	0.51	0.48	0.50	0.47	0.50	0.51	0.49	0.48
0.51	0.50	0.50	0.53	0.52	0.51	0.50	0.50	0.51	0.51

```
0.51 0.51 0.51 0.50 0.51 0.49 0.52 0.53 0.50 0.47
0.51 0.52 0.53 0.48 0.49 0.50 0.52 0.49 0.49 0.50
0.49 0.48 0.46 0.49 0.49 0.48 0.49 0.49 0.51 0.47
0.51 0.51 0.51 0.48 0.50 0.47 0.50 0.51 0.49 0.48
0.51 0.50 0.50 0.53 0.52 0.51 0.50 0.50 0.51 0.51
```

Compute the mean concentration, and the standard deviation:

```
#Getting data in a vector
x=scan('Table2_1.txt')
mean(x)
#[1] 0.4998
sd(x)
#[1] 0.01647385
```

2.5 Bootstrap Methods

If we would repeat our experiment of collecting 50 samples of nitrate concentrations many times we would see the range of error. But it would be a waste of resources and not a viable method.
Instead we re-sample 'new' data from our data and use so obtained new samples for assessment of the error. The following R code does the job.

```
#Getting data in a vector
m=mean(x)
bootstrap=vector('numeric',500)
for(i in 1:500)
 {
 bootstrap[i]=mean(sample(x,replace=T))-mean(x)
 }
#The distribution of estimation error
hist(bootstrap)
```

The conclusion of this procedure is that the nitrate concentration is 4999 ± 0.005. We are specifically interested in how R was easily able to implement a solution for this.

3 Introduction - systematic vs. random errors

3.1 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.

3.2 Errors in quantitative analysis

Since quantitative studies play a dominant role in any analytical laboratory, it must be accepted that the errors that occur in such studies are of supreme importance. No quantitative results are of any value unless they are accompanied by some estimate of the errors inherent in them!

Example 1 - detecting a new analytical reagent

- A chemist synthesizes an analytical reagent that is believed to be entirely new.
- The compound is studied using a spectrometric method and gives a value of 104.
- The chemist finds that no compound previously discovered has yielded a value of more than 100.
- Has the chemist really discovered a new compound?
- The answer lies in the degree of reliance to experimental value of 104.
- If the result is correct to within 2 (arbitrary) units, i.e. the true value probably lies in the range 102 ± 2, then a new material has probably been discovered.
- If, however, investigations show that the error may amount to 10 units i.e. 104 ± 10, then it is quite likely that the true value is actually less than 100, in which case a new discovery is tar from certain.
- A knowledge of the experimental errors is crucial!!

Example 2 - replicates in a titrimetric experiment

- Analysts commonly perform several replicate determinations in the course of a single experiment.
- An analyst performs a titrimetric experiment four times and obtains values of 24.69,24.73,24.77 and 25.39 ml.
- All four values are different, because of the variations inherent in the measurements
- The fourth value (25.39 ml) is substantially different from the other three.
- Can it be safely rejected, so that (for example) the mean titre is reported as 24.73 ml, the average of the other three readings?

3.3 Systematic and random errors

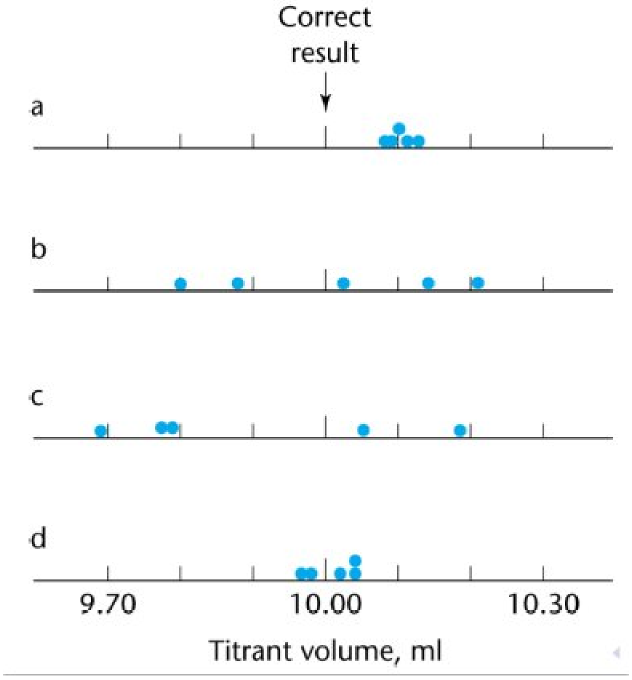
Experimental scientists make a fundamental distinction between ***random***, and ***systematic*** errors. To distinguish between random and systematic errors let us consider a real experiment.

Four students (A-D) each perform an analysis in which exactly 10.00 *ml* of exactly 0.1 M sodium hydroxide is titrated with exactly 0.1 NI hydrochloric acid. Each student performs five replicate titrations, with the results shown in Table 1.1.

Student	Results	(ml)				Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, unbiased
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

Graphical illustration

The results of experiment represented by dot-plots. (The true value is 10.00).



Systematic error and bias

Systematic error is a deviation of all measurements in one direction from the true value. It is well represented by the difference between the average value of the determined values and the true value of the measured quantity. This difference is called the bias of measurements.

Random error and precision

Random error is a deviation of a measurement from the average of measured values. It is well represented by the standard deviation of measurements. This value is often called precision of measurements.

Combined error vs. accuracy

Accuracy is in inverse relation to the total deviation of a single measurement from the true value.