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

INSERT HERE

1.3 Means and standard deviations using R

```
#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
```

1.4 Bias and precision using mean and standard deviation

Classify bias and precision using means and standard deviation of measurements.

1.5 Measures of Dispersion

Recall:

- standard deviation = square root of variance

- variance = squared standard deviation
- coefficient of variation = relative standard deviation (in percentage)

1.6 The empirical distribution of repeated measurements

1) frequency table 2) histogram and dotchart - graphical representation of the empirical distribution

Nitrate ion concentration from Table 2.1

Table 2.1 Results of 50 determinations of nitrate ion concentration, in $\mu g\, ml^{-1}$

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

Also in file $Table2_1.txt$

```
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
```

The mean concentration

Reading data

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

Dotchart and histogram in R

```
#Dotchart
dotchart(x)
#Histogram and frequency table
Histogr=hist(x)
Histogr
```

1.7 The Normal Distribution - Theoretical distribution of repeated measurements

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

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

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 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.

Normal distribution 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)

#And plotting them one at the top of the other
par(mfrow=c(2, 1))
plot(z,dnorm(z),type="l",col="red",lwd=4)
plot(z,pnorm(z),type="l",col="red",lwd=4)

#Side by side
par(mfrow=c(1, 2))
plot(z,dnorm(z),type="l",col="red",lwd=4)
plot(z,pnorm(z),type="l",col="red",lwd=4)
```

1.8 Empirical vs Theoretical

The theoretical one can be compared with empirical by taking μ equal to the sample mean $\bar{X}$ and σ equal to sample standard deviation s. The following code compares empirical percentages with theoretical.

```
quantile(x,c(0.16,0.84))
qnorm(c(0.16,0.84),mean(x),sd(x))
```

Example 2.2.1 If repeated values of a titration are normally distributed with mean 10.15 ml and standard deviation 0.02 ml, find the proportion of measurements which lie between 10.12 ml and 10.20 ml.

Standardizing the first value gives

$$z = (10.12 - 10.15)/0.02 = -1.5.$$

From Table A.1, $F(-1.5) = 0.0668$.

Standardizing the second value gives

$$z = (10.20 - 10.15)/0.02 = 2.5.$$

From Table A.1, $F(2.5) = 0.9938$.

Thus the proportion of values between x = 10.12 to 10.20 (which corresponds to z =-1.5 to 2.5) is $0.9938 - 0.0668 = 0.927$.

No standardization needed in R

```
pnorm(c(10.12,10.20),10.15,0.02)
```

Not everything is normal, unfortunately - lognormal distribution

```
Concentr=scan("Figure2-5.txt")
hist(Concentr)
hist(Concentr,nclass=30)
```

Distribution of the sample mean

```
MatrConc=matrix(Concentr,ncol=4)
ConcM=rowMeans(MatrConc)
hist(ConcM)
MatrConc=matrix(Concentr,ncol=25)
ConcM=rowMeans(MatrConc)
hist(ConcM)
```

1.9 Two distributional effects of taking sample mean

- Reduction in standard deviation (increased precision)
- Distribution is becoming normal even if original is not.

For a sample of n measurements,standard error of the mean

$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}(2.5)$$

As expected, the larger n is, the smaller the value of the s.e.m. and consequently the smaller the spread of the sample means about n. The term 'standard error of the mean' might give the impression that o'/W gives the difference between mu and $\bar{X}$. This is not so: $\sigma/\sqrt{N}$ gives a measure of the variability of $\bar{X}$, as we shall see in the next section.

Another property of the sampling distribution of the mean is that, *even if the original population is not normal*, the sampling distribution of the mean tends to the normal distribution as n increases. This result is known as the ***central limit theorem***. This theorem is of great importance because many statistical tests are performed on the mean and assume that it is normally distributed.

Since in practice we can assume that distributions of repeated measurements are at least approximately normally distributed, it is reasonable to assume that the means of quite small samples (say $n > 5$) are normally distributed.

2 Confidence Intervals

2.1 Large sample distribution of sample mean

2.2 Confidence limits of the mean for large samples

Now that we know the form of the sampling distribution of the mean we can return to the problem of using a sample to define a range which we may reasonably assume includes the true value. (Remember that in doing this we are assuming systematic errors to be absent.) Such a range is known as a confidence interval and the extreme values of the range are called the confidence limits.

The term 'confidence' implies that we can assert with a given degree of confidence, i.e. a ' certain probability, that the confidence interval does include the true value.

The size of the confidence interval will obviously depend on how certain we want to be that it includes the true value: the greater the certainty, the greater the interval required.

Figure 2.6 shows the sampling distribution of the mean for samples of size n. If we assume that this distribution is normal then 95% of the sample means will lie in the range given by:

$$\mu - 1.96(\sigma/\sqrt{n}) < \mu < \mu + 1.96(\sigma/\sqrt{n})(2.6)$$

(The exact value 1.96 has been used in this equation rather than the approximate value, 2, quoted in Section 2.2. The reader can use Table A.1 to check that the pro- portion of values between z= -1.96 and z = 1.96 is indeed 0.95,)

2.3 Large Sample Confidence Intervals based on the sample mean

In practice, however, we usually have one sample, of known mean, and we require a range for μ, the true value.

Equation (2.6) can be rearranged to give this:

$$\bar{X} - 1.96(\sigma/\sqrt{n}) < \mu < \bar{X} + 1.96(\sigma/\sqrt{n}) (2.7)$$

Equation (2.7) gives the 95% confidence interval of the mean. The 95% confidence limits are $\bar{X} \pm 19.6\sigma/\sqrt{n}$ In practice we are unlikely to know σ exactly. However, provided that the sample is large, σ can be replaced by its estimate, s.

2.4 Example of computations using R

Finding confidence intervals for the mean for the nitrate ion concentrations in Table 2.1.

```
#reading data
x=scan("Table2_1.txt")
#setting the confidence level
CL=0.95
#computing confidence interval
n=length(x)
pm=sd(x)*c(qnorm(0.025),qnorm(0.975))/sqrt(n)
CI=mean(x)+pm
```

2.5 Small-Sample Case (n ¡ 30)

If the data have a normal probability distribution and the sample standard deviation s is used to estimate the population standard deviation s, the interval estimate is given by:

$$\bar{X} \pm t_{\alpha/2} s/\sqrt{n}$$

where $\alpha/2$ is the value providing an area of $\alpha/2$ in the upper tail of a Student's $t-$distribution with n - 1 degrees of freedom.

2.6 Confidence limits of the mean for small samples

As the sample size gets smaller, s becomes less reliable as an estimate of σ. This can be seen by again treating each column of the results in Table 2.2 as a sample of size five. The standard deviations of the 10 columns are 0.009, 0.015, 0.026, 0.021 0.013, 0.019, 0.013, 0.017, 0.010 and 0.018. We see that

the largest value of s is nearly three times the size of the smallest. To allow for this, equation (2.8) must be modified.

For small samples, the confidence limits of the mean are given by

$$\bar{X} \pm t_{n-1} s/\sqrt{n}$$