Contents

1 Introduction to Analysis with R

1.1 Measures of Dispersion

Recall:

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

1.2 Titration experiment

Recall the titration experiment from the last class. 4 Students performing the same experiment five times, hence each yield 5 results.(Table 1.1 random and systematic errors).

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

Two criteria were used to compare these results, the average value (technically know

as a measure of location 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

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

In Chapter 1 the 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:

The standard deviation, s, of n measurements is given by

$$s = \sqrt{\frac{\sum(x - \bar{X})^2}{n-1}} (2.2)$$

1.3 Means and standard deviations using R

1.4 Bias and precision using mean and standard deviation

Classify bias and precision using means and standard deviation

of measurements.

```
#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
```

2 Testing Normality

An assessment of the normality of data is a prerequisite for many statistical tests as normal data is an underlying assumption in parametric testing. There are two main methods of assessing normality - graphically and numerically.

Importantly, Hypothesis tests are tests against normality. Correctly we are not determining whether a data set is normally distributed or not. What we are doing is testing to see if there is enough evidence to classify the data as non-normal. If not, we rely on the assumption of normality. Just because we can't demonstrate that the data is not normally distributed, does not mean that the data is normally distributed. Using multiple procedures in conjunction is advised for getting a proper assessment of the data.

- Shapiro Wilk Test
- Anderson-Darling Test
- Normal Probability Plot

3 Testing Normality

An assessment of the normality of data is a prerequisite for many statistical tests as normal data is an underlying assumption in parametric testing. There are two main methods of assessing normality - graphically and numerically.

3.1 Testing for normality of distribution

The chi-square test could be used for testing normality by dividing the range of data into bins and compare the count in each bin with the corresponding probabilities based on the normal distribution. Unfortunately, one needs relatively large data sample size in order to use chi-squared test (> 50), thus there is a need for a small sample size procedure.

3.2 Normaly Probability Plots

Normal probability plots One simple graphical way of comparing data to normal distributions is by plotting empirical quantile vs. corresponding normal quantile. Recall that the p-quantile for a given (empirical) distribution is the number below of which there is 100p(of the data). Consider data from Example 3.12.1 x=c(109,89,99,99,107,111,86,74,11
The normal probability plot is obtained in R using qqnorm(x)

3.3 KS test

Kolmogorov-Smirnov test in R

```
x=c(25.13,25.02,25.11,25.07,25.03,24.97,25.14,25.09)
```

```
ks.test(x,"pnorm",mean(x),sd(x))
One-sample Kolmogorov-Smirnov test
data: x
D = 0.1321, p-value = 0.995
alternative hypothesis: two-sided
```

3.4 Outliers

There may also be an outlier , or multiple outliers, present in the data. There are several formal tests to determine presence of an outlier.

- Grubbs' Test
- Dixon Test

In laboratory sciences, it is quite often the case that an outlieer measurement is the result of faulty or unclean equipment. Care must be take to assess that the measurement is an outlier, and not in fact an unuusal result that is in fact genuine. It is good practice not to remove outiers from an overall analysis. Although you may omit suspected outliers and run the analysis a second time.

3.5 Outliers in R

The tests for outliers come in a contributed package called outliers. In order to use it one has to download the package to the computer. It can be done for the command line by using install.package("outliers") or can by using a convenient interface of the software.

```
x=c(0.403,0.410,0.401,0.380)
grubbs.test(x)
# Grubbs test for one outlier
#data: x
#G = 1.4316, U = 0.0892, p-value = 0.09124
#alternative hypothesis: lowest value 0.38 is an outlier

dixon.test(x)
# Dixon test for outliers
#data: x
#Q = 0.7, p-value = 0.1721
#alternative hypothesis: lowest value 0.38 is an outlier
x=c(0.403,0.410,0.401,0.380,0.400,0.413,0.408)
grubbs.test(x)
dixon.test(x)
```

Bivariate Analysis

Correlation