
DATA HANDLING SKILLS

BSc Biochemistry
BSc Biological Science
BSc Bioinformatics

March 22, 2003

Olaf Wolkenhauer
Department of Biomolecular Sciences and
Department of Electrical Engineering and Electronics
Address: Control Systems Centre
UMIST
Manchester M60 1QD, UK
Tel/Fax: +44 (0)161 200 4672
e-mail: o.wolkenhauer@umist.ac.uk
<http://www.systemsbiology.umist.ac.uk>

Contents

1	Introduction	3
2	Visualising and Organising Data	4
3	Descriptive Statistics	9
4	The Normal Distribution	13
5	Sampling Errors	19
6	Testing for Differences: The t-Test	22
7	Categorical Data: The Chi-Square Test	30
8	Finding Associations: Correlation	33
9	Modelling Relationships: Linear Regression	33
10	More Exercises	34
11	Solutions to Exercises in Section 10	37
12	Symbols and Notation	38
13	Further Reading	39

1 Introduction

In this course module we explore how mathematics supports the analysis of experimental data using techniques, developed in the area of *statistics*. The reason why statistical techniques are required is simply that most biological measurements usually contain *non-biological* variation. The purpose of statistics is then to allow reasoning in the presence of uncertainty. For most statistics courses in the life sciences there is not enough time to fully cover the material which is necessary and useful for your degree, projects, and career. I therefore strongly recommend the purchase a more comprehensive treatment of statistics. Section 13 provides you with a list of recommended books.

We use the term **variable** to refer to the quantity or object that is being observed or measured. The term **observation**, *observed value* or *value* for short, is used for the result of the measurement. We use the terms observation and measurement interchangeably. Examples for typical variables, are measures such as ‘weight’, ‘length’, or the ‘count’, say the number of colonies on a Petri dish. If the value of a variable is subject to uncertainty, then the variable is called a **random variable**.

There are different types of data: **categorical** data, for example a colour or type of object observed; **discrete** data are numbers that can form a list:

1, 2, 3, 4, 5, ...

0, 0.2, 0.4, 0.6, 0.8, ...

Continuous data are ‘real numbers’, numerical values such as height, weight, and time. Because we usually take measurements with devices that have limited accuracy, continuous values are usually recorded as discrete values. For example, the length may only be recorded to the nearest millimeter.

Number	Sample A	Sample B
1	123	54
2	56	202
3	1283	232
4	31	90
5	329	982
⋮	⋮	⋮

Time [h]	Measurement
1	34
2	35
3	67
4	84
5	25
⋮	⋮

The two tables above illustrate two different kinds of analysis. The table on the left gathers data from repeated experiments. Let us assume that the rows are **replicate observations** while the columns (Sample A and B) are two different experiments. Sample A and B may therefore represent a **repeated experiment** under different conditions. For example, sample A could be our reference sample, say “without treatment” and sample B is used to see whether any biological changes have occurred from the treatment. For the table on the left, the order of the rows doesn’t matter. For each sample, the values in the rows are **repeated observations** (or repeated measurements).

Variables and observations

Types of data

The reason to take repeated measurements is simply because we expect some *uncertainty* through *variation*. In other words, the reason to use statistical techniques is that in addition to biological variation (which we investigate) and non-biological variation such as measurement errors and other, often technical problems during the experiment. Repeating a measurement of the same variable, under the same condition, we would expect the values to be the same; in practice they are not, and an important task is to identify typical values and to quantify the variability of the data.

In the table on the right, the order of the rows matters and there are no replicate observations, the left column denotes time and the data are referred to as a time series. In studying data from time course experiments we typically want to answer questions about the trend.

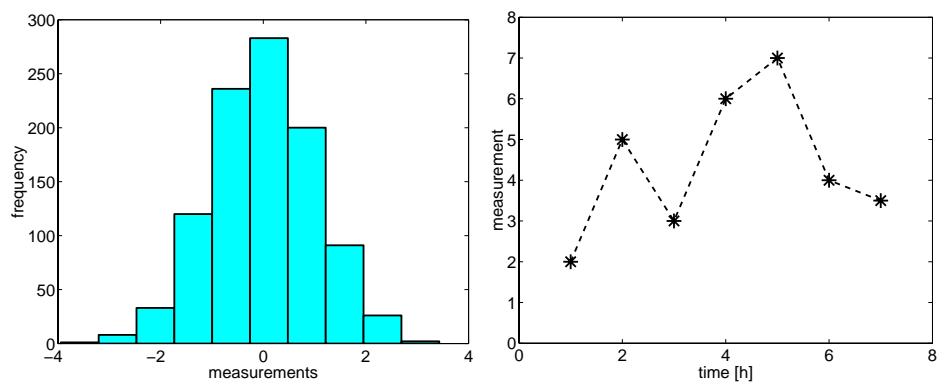


Figure 1.1: Typical graphs used to visualise data: histogram (left) and time-series (right).

The graphs in Figure 1.1 are typical visualisations of the two kinds of data shown in the tables above. The histogram on the left is used for data of the kind in the left table, while the time-series plot is a visualisation of the kind of data listed in the table on the right. Large tables of numbers are not very revealing and diagrams play therefore an important role in detecting patterns. For the picture on the left, typical characteristics we are going to look for in such data sets are the variability (**spread**) of the data and whether they cluster around a particular point (**central tendency**).

2 Visualising and Organising Data

The data we deal with are often *repeated observations* of the same variable. For example, we count the number of colonies on 24 plates:

0, 3, 1, 2, 4, 0, 6, 2, 1, 1, 0, 6
1, 2, 2, 6, 0, 2, 1, 3, 3, 2, 1, 1

We refer to the data in this table as **raw data** since they have not been re-organised or summarised in any way. The **tally chart** is a simple summary, counting for each possible outcome the **frequency** of that outcome.

Tally chart

PRACTICE. Complete the tally chart for the data in the table above:

score	tallies
0	
1	
2	
3	
4	
5	
6	

The tally count in each row gives the frequency of that outcome. A table which summarises the frequencies is called **frequency distribution** and is simply obtained by turning the tally count into a number. A frequency distribution for categorical data can be visualised by a **bar chart**. Take for example the colour of a colony on a Petri dish and let us classify any particular colony into either ‘blue’, ‘white’, or ‘brown’. Given the following frequency distribution, the bar chart is shown in Figure 2.1.

Bar chart

blue	white	brown
18	12	23

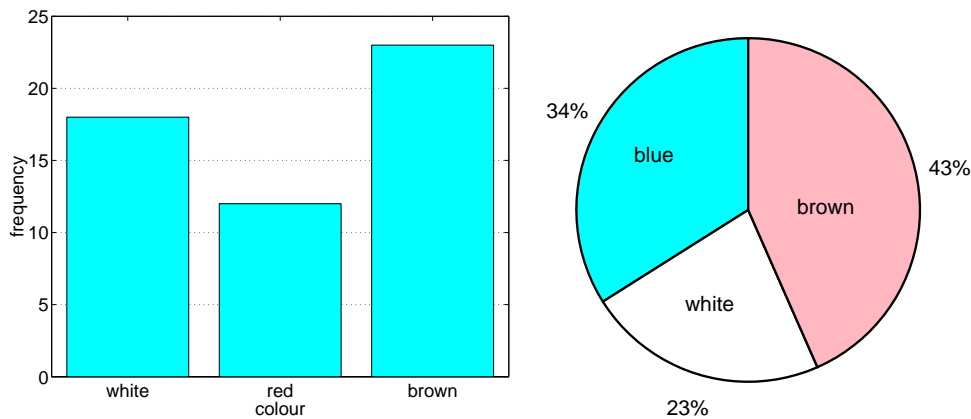


Figure 2.1: Left: Bar chart of counts of colonies on a Petri dish for different colours. Notice that in a bar chart the columns are separated. Right: Pie chart as an alternative representation of the same information.

Note: It is important to make the meaning of charts and plots clear by labelling the axes.

For the next example we consider the measurement of the length of some 100 objects. The difference between the largest value and smallest value in a sample is called the **range** of the data set. When summarising a large number of continuous data, it is often useful to group the range of measurements into **classes**, and to determine the number of objects belonging to each class (the **class frequency**). Table 2.1 summarises the data in a format which is called **frequency distribution**. Figure 2.2, visualises the information in Table 2.1. Bar charts are not appropriate for data with grouped frequencies for ranges of values. What is shown

Range

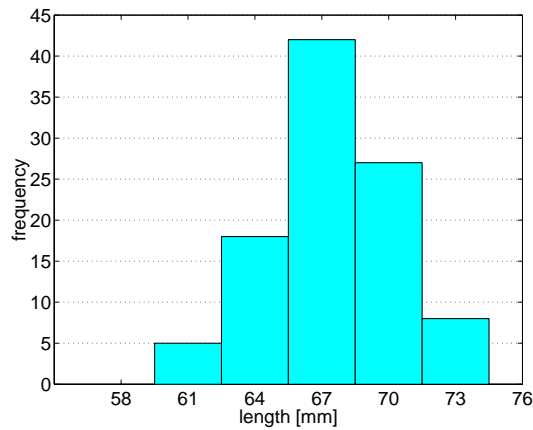


Figure 2.2: Frequency histogram for the data in Table 2.1. Notice the difference between a bar chart and a histogram.

in Figure 2.2, is called a **frequency histogram** and is one of the most important graphs for us. The importance difference to the bar chart is that the width of the bars matters. The key feature of a histogram is that

the **area** is proportional to the *class frequency*.

If, as is the case in Figure 2.2, the class width are equal, the area is not only proportional to the frequency but also the height is proportional to the frequency.

Table 2.1: Frequency distribution for length measurement of 100 objects, recorded to the nearest millimeter.

Class interval	Number of objects
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8

The **relative frequency** of a class is the frequency of the class divided by the total frequency (the total number of objects measured). (The relative frequency is often expressed as a percentage (“out of hundred”).) The graph visualising the relative frequency of occurrences of values in a sample is referred to as the **relative-frequency histogram**. There are at least two reasons to use a relative-frequency distribution: percentages are very intuitive and secondly the relative-frequency distribution allows us to compare two samples with each having a different total number of objects.

Notice that for the histogram above, the class intervals have equal length (3 mm) and partition the range of values into equally sized groups. For the class (interval) 63–65, the values 63 and 65 are called **class limits**. If values are measured to the nearest millimeter, the class interval 60–62 includes all measurements from 59.5mm to 62.5mm. These numbers are called **class boundaries**.

Histograms

Table 2.2: Recorded weights, measured to the nearest gram of 100 1kg objects.

1038	1018	1016	1017	1010	1019	1013	1012	1020	1021
1021	1011	1019	1021	1013	1000	1020	1026	1018	1003
1020	1014	1019	1005	1020	1023	1015	1007	1014	1012
1024	1019	1013	1015	1022	1016	1031	1020	1010	999
1005	1016	1019	1017	1029	1018	1020	1023	1014	1022
1020	1018	1020	1000	1020	1033	1010	1013	1030	1005
1013	1019	1021	1016	1012	1017	999	1021	1014	1009
1035	1001	1040	1011	1026	1005	1019	1018	1009	1022
1027	1016	1026	1006	1013	1018	1032	1019	1029	1020
1021	1036	1017	1025	1022	998	1021	1008	1003	1015

The **class mark** is the midpoint of the class interval and is obtained by adding the lower and upper class limit and dividing by two. As you can imagine for some data sets, equal class sizes are not appropriate and the best number of class intervals is often not obvious. Therefore, while the histogram can reveal some basic characteristics of the data, which are usually not obvious from the table of measurements, there is also a ‘loss’ of information for values within class intervals.

PRACTICE. For the raw data in Table 2.2,

1. Determine the range.
2. Construct an un-grouped frequency distribution table:

<i>weight [g]</i>	<i>tally count</i>	<i>frequency</i>
998		1
999		2
1000		2
⋮	⋮	⋮

3. Construct a grouped frequency distribution table using a class width of 5g:

<i>class interval</i>	<i>class boundary</i>	<i>class mark</i>	<i>frequency</i>
998 - 1002	997.5 - 1002.5	1000	6
1003 - 1007	1002.5 - 1007.5	1005	8
⋮	⋮	⋮	⋮

4. Construct a grouped frequency distribution table using a class width of 10g
5. Draw a relative frequency histogram for the data (class width 10g).

An alternative way to represent the information of the frequency distribution is to answer the question “what proportion of the data have values less than x ?”. Such a diagram is referred to as the **cumulative**

frequency distribution and *relative* cumulative frequency distribution. The term **cumulative distribution function** (cdf) is used in general to describe a cumulative distribution and is denoted $F(x)$. Consider the raw data in Table 2.3, Figure 2.3 shows the cumulative frequency diagram.

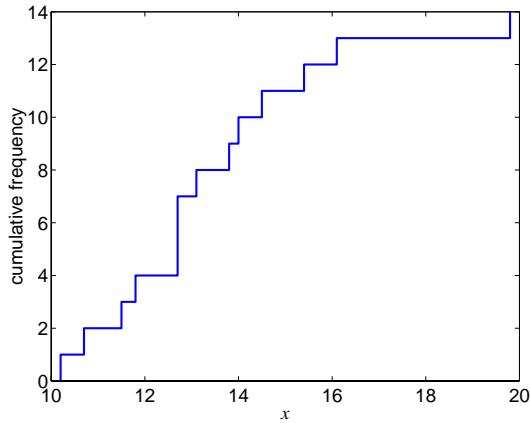


Figure 2.3: Cumulative frequency distribution for the data in Table 2.3. Notice that for any $x \geq 19.8$ the relative cumulative frequency equals the total number of observations (14).

PRACTICE. For the raw data given in Table 2.3:

1. Determine the cumulative frequency at the following points:

$x < 10.2$	$x < 11.2$	$x < 12.2$	$x < 13.2$	$x < 14.2$	$x < 15.2$
$x < 16.2$	$x < 17.2$	$x < 18.2$	$x < 19.2$	$x < 20.2$	

2. Calculate the relative cumulative frequency in percent.
3. Draw the the relative cumulative frequency distribution.
4. Determine in a table the relative frequencies for the following class intervals:

$x < 10.2$	$10.2 \leq x < 11.2$	$11.2 \leq x < 12.2$
$12.2 \leq x < 13.2$	$13.2 \leq x < 14.2$	$14.2 \leq x < 15.2$
$15.2 \leq x < 16.2$	$16.2 \leq x < 17.2$	$17.2 \leq x < 18.2$
$18.2 \leq x < 19.2$	$19.2 \leq x < 20.2$	

5. Draw the relative frequency histogram.

Note: The symbol \leq means “less or equal”, while $<$ means “less than”.

Table 2.3: Raw data set. See also Figure 2.3.
 12.7, 14.5, 15.4, 11.8, 19.8, 12.7, 11.5
 10.2, 12.7, 10.7, 14.0, 13.1, 13.8, 16.1

Remark: In practice one would rarely draw histograms and distribution functions by hand. Since there are various interpretations of histograms and distribution functions, it is therefore important to check the scale of the ordinate axis and to provide a clear label. When using a software tool, such as Minitab, MS Excel, or Matlab, it is important that you try to understand what is plotted and not just accept the result only because it looks similar to what you expected.

3 Descriptive Statistics

Descriptive statistics help us to summarise information burried in the data and quantifies some of the properties of the diagrams we have used before. The purpose is therefore to extract essential information from the raw data, not in a diagram but in form of numbers. For reasons that will become clear later, we refer to a given set of data, as the **sample**. We first consider two descriptive statistics of a sample: a **measure of central tendency** ('measure of location') and a **measure of variability** ('measure of spread'), see Figure 2.2. The former comes in three variations: the **sample mode**, **sample median**, and the **sample mean**.

Sample

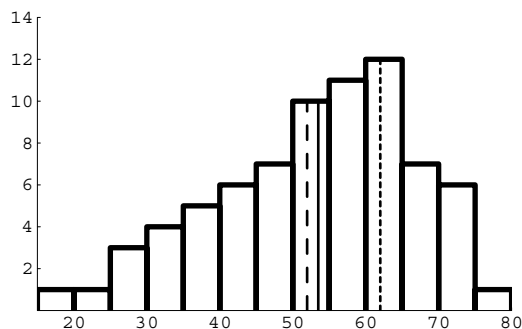


Figure 3.1: A skewed distribution. The dotted line denotes the mode, the solid line the median, and the dashed line the mean.

The **sample mode** of discrete data is the most frequent value. For the data in Table 2.2, the mode is therefore 1040. The mode is found in the histogram from the highest bar. This is a simple measure but may not be unique as there may be more than one bar with the same frequency. In this case, the histogram has more than one peak. For two such outcomes we speak of a **bimodal** distribution or in general from a **multimodal** distribution.

Mode

The **sample median** describes the 'middle' of the data set and splits therefore the sample into two halves. For the following sample (arranged in the order of magnitude!):

Median

1, 4, 6, 8, 9, 11, 17

The median value is 8. For an even number of observations we find two middle values and by definition, we calculate the median as their average. For example, for the following sample

1, 4, 6, 8, 9, 11, 17, 20

the median value is $(8 + 9)/2 = 8.5$.

The **sample mean** is usually calculated as the **average** of the values in the sample and is therefore often called the **arithmetic mean**. For unimodal distributions, the sample mean gives us a measure of central tendency, a value around which the other values tend to cluster. Let us denote the sample by X , with individual observations denoted x_i . For the sample above, we therefore have

$$X = \{1, 4, 6, 8, 9, 11, 17, 20\}$$

where index i , ranges from $i = 1, \dots, n$ and $n = 8$, denotes the sample size. For example $x_3 = 6$. The curly brackets $\{\}$ are used to denote an (unordered) list. The sample mean is commonly denoted with a bar over the symbol used to denote the sample, \bar{X} , and is calculated as

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} \quad (3.1)$$

The sample mean for the sample $X = \{1, 4, 6, 8, 9, 11, 17, 20\}$ is

$$\bar{X} = \frac{1 + 4 + 6 + 8 + 9 + 11 + 17 + 20}{8} = \frac{76}{8} = 9.5$$

If the *distinct* values x_1, x_2, \dots, x_m occur f_1, f_2, \dots, f_m times, respectively, the sample mean can also be calculated by the following formula:

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_m x_m}{f_1 + f_2 + \dots + f_m} = \frac{\sum_{i=1}^m f_i x_i}{\sum_{i=1}^m f_i} = \frac{\sum_{i=1}^m f_i x_i}{n} . \quad (3.2)$$

Note the difference in the subscripts n respectively m and that the x_i correspond to class marks. For example, if 5, 8, 6, and 2 occur with frequencies 3, 2, 4, and 1 respectively, the sample mean is

$$\bar{X} = \frac{3 \cdot 5 + 2 \cdot 8 + 4 \cdot 6 + 1 \cdot 2}{3 + 2 + 4 + 1} = \frac{15 + 16 + 24 + 2}{10} = 5.7$$

Note that if the distribution function is symmetric and unimodal, the mode, mean and median coincide (Create an example that proves this!). The mean is the most frequently used statistic for a central tendency in samples but is also more affected by *outliers* than is the median. An **outlier** is an abnormal, erroneous, or mistaken value. Such extreme values can distort the calculation of the centre of the distribution.

Next, we come to a measures of dispersion or spread in the data. The previously introduced **range** gives a basic idea of the spread but is only determined by the extreme values in the sample. The variability of the data corresponds in the histogram to its width around the centre. A natural measure of the spread is therefore provided by the sum of squares of the deviations from the mean:

$$\sum_{i=1}^n (x_i - \bar{X})^2 = (x_1 - \bar{X})^2 + \dots + (x_n - \bar{X})^2$$

Sample mean \bar{X}

Sample size n

Outliers

Range

We square the differences to avoid negative differences which could distort the measure (Why or how?). This is a measure of variation but is very much dependent on the sample size n . To calculate the variation within a sample, the average squared deviation from the mean, denoted σ_n^2 , is called the **sample variance**:

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 \quad (3.3)$$

This is not the only possible measure of variance, and in fact there are good reasons to use a slight variation of (3.3), called **unbiased estimate of the population variance**, denoted s^2 :

$$s^2 = \sigma_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2 \quad (3.4)$$

The only difference to (3.3) is that we divide by $n-1$ instead of n , and for a large n the difference seems irrelevant. However, once we have introduced the concept of a **population**, it turns out that equation (3.3) would provide an accurate measure only of the variability in the sample but is a biased estimate of the population variance. Note that the subscript n in equation (3.3) is important to clarify that this is an estimate based on n values. As we will find later, there is a difference between “the mean value” (of a population) and “the sample mean” (Which mean we mean by talking about “the mean”, will usually be clear from the context).

Calculating the variance without using a software tool or calculator with statistical functions, formulas (3.3) and (3.4) are awkward. However, we can simplify the calculation as follows. Since

$$\sum (x_i - \bar{X})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n},$$

hence:

$$\sigma_n^2 = \frac{\sum x_i^2}{n} - \frac{(\sum x_i)^2}{n^2},$$

which leads to the more convenient, equivalent formula:

$$s^2 = \sigma_{n-1}^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) = \frac{n}{n-1} \sigma_n^2 \quad (3.5)$$

A ‘problem’ with the equations for variance above is that they report the variability not in the same units of the data but squared. To obtain a measure of variation in the same units of the data one takes the square root of the variance, leading to what is called the **sample standard deviation**:

$$\sigma_n = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n^2}} \quad (3.6)$$

PRACTICE. 1. For the following sample (set of raw data)

$$X = \{3.9, 23.3, 4, 7.6, 25.2, 17, 22, 21.2\}$$

Determine the sample mean, the sample standard deviation and the unbiased estimate of the population variance.

2. For the data in Table 2.3 calculate the sample mean (Solution: 13.5) and the unbiased estimate of the population standard deviation (Solution: 2.377). Note: do not use the statistical functions of your calculator.

When data are summarised by a frequency distribution, i.e., in the form “value x_i occurs with frequency f_i ”, we can use different equations. Let m denote the number of distinct values of x in the sample, the formulas for the sample variances become:

$$\sigma_n^2 = \frac{1}{n} \left[\sum_{i=1}^m f_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^m f_i x_i \right)^2 \right]$$

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^m f_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^m f_i x_i \right)^2 \right]$$

where n is the total number of frequencies, the sample size $n = \sum_{i=1}^m f_i$. Take care of the difference between n and m . As before, the standard deviation is simply obtained by taking the square root of the variance. These formulas can also be used for class frequencies (cf. Table 2.1). In this situation, x_i denotes the class mark and since we replace the data in any particular class (or bin of the histogram) by the class mark, we have to remember that this is only an approximate calculation.

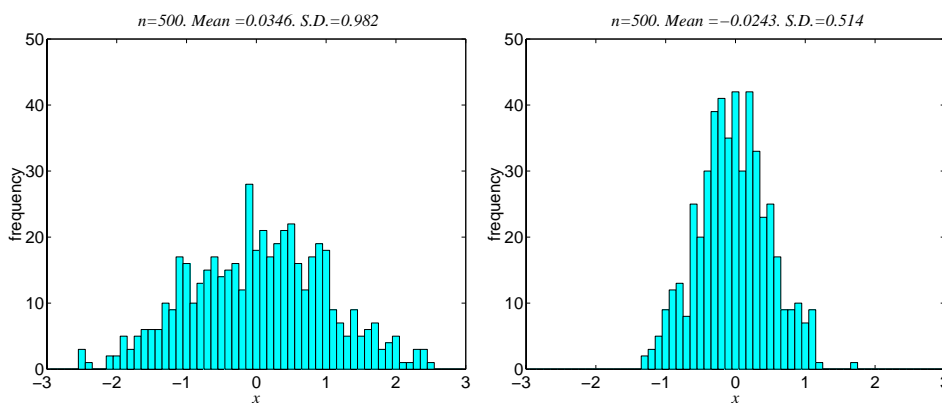


Figure 3.2: Frequency histograms for two samples with different standard deviations. The sample means are nearly the same while the data in the histogram on the left have a greater spread (greater standard deviation). Most software programmes will adjust the scale for the axes automatically. Always check the scales as otherwise the comparison of distributions can be misleading.

PRACTICE. Try the following exercises using the equations above and without using statistical functions of your calculator. Next, try the same exercise again using the statistical mode of your calculator. Note that you will need to do many more exercises to become confident with the formulas and to remember how to use the calculator. (More exercises can be found in Section 10)

1. For the data in Table 2.1, determine the sample variance.
2. For the data in the table below, determine the sample mean and sample standard deviation:

class mark	70	74	78	82	86
frequency	4	9	16	28	45

Note: There are a number of other concepts, we have not dealt with but which you can find explained in the literature:

1. *Weighted Arithmetic Mean:* As (3.1) but each value is weighted for its relevance or importance.
2. *Harmonic Mean, Geometric Mean:* The geometric mean is used when data are generated by an exponential law. The harmonic mean is the appropriate average to use when we are dealing with rates and prices.
3. *Quartiles, Deciles, and Percentiles:* Like the median splits the data into halves, these divide the data in different parts. So called Box-Whisker diagrams (“box-plots”) are frequently used compare distributions.
4. *Moments, Skewness, and Kurtosis:* In our examples we have somehow implicitly assumed that the distributions are uni-modal and symmetric. These measures give additional information about the shape and appearance of the distribution.

Remark: A note of caution is due with regard to notation. Although there are few commonly used symbols to denote statistical measures, their use varies from book to book.

4 The Normal Distribution

In the previous section we considered a *sample* of data. The values in a sample were obtained by repeated experiments, observations, or measurements. We collected more than one value because we expected some variation in the data and we determined some characteristic value (the sample mean) and the variation around this typical value (the sample standard deviation). As the term suggests, a sample itself is characteristic of something more general - the **population**. By testing a sample of cultures from *E. coli* we wish our results to apply to *E. coli* cultures in general. The concepts of sampling a population is most intuitive in the context of polls before an election. To infer what the population is going to vote, a selected group (sample) of voters is studied. Drawing



Population

conclusions about the population from the analysis of a sample, is referred to as *inductive statistics* or **inferential statistics**. Because these conclusions cannot be absolute certain, conclusions are stated in terms of probabilities.

For the sample to be representative we have to take great care. It is very important to state what population is meant and how the sampling was conducted. As you can imagine, the sampling process is often the basis for the misuse of statistics.

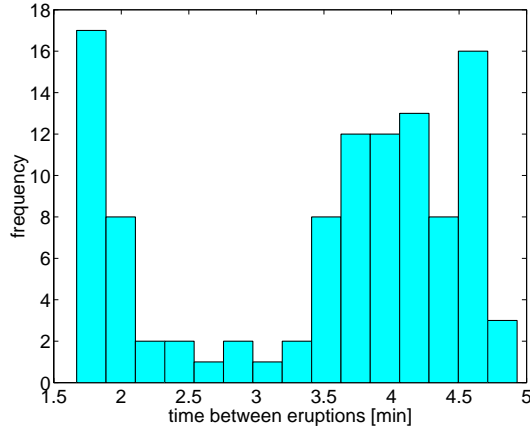


Figure 4.1: Frequency histogram of the time intervals between eruptions of the Old Faithful geyser.

Bell-shaped frequency histograms for continuous variables, like those in Figure 3.2, are very common in all areas of science, engineering, economics, ... and quite independent of the kind of experiments conducted. A histogram will usually help us to decide whether this is indeed the case and to prove the point that this is not always the case, consider Figure 4.1. The frequency histogram shows the recorded time intervals between eruptions of the ‘Old Faithful’ geyser in Yellowstone National Park in Wyoming, USA. The distribution is clearly bi-modal.

In Figure 4.2 in the top left figure we show the *relative* frequency histogram of Figure 3.2 (left). In upper right figure, we changed the vertical scale to **relative frequency density** so that the total area sum of all areas of the bars equals 1. This is done by dividing the relative frequency by the class width (0.5 in the figure). The two lower distributions shown in Figure 4.2, demonstrate what happens to the relative frequency density of a continuous random variable as the sample size increases. While the area remains fixed to one, the relative frequency density function approaches gradually a curve, called **probability density function**, and denoted $p(x)$.

For many random variables, the probability density function is a specific bell-shaped curve, called the **normal distribution** or **Gaussian distribution**. This is the most common and most useful distribution, which we assume to represent the probability law of our population. It is defined by the equation

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (4.1)$$

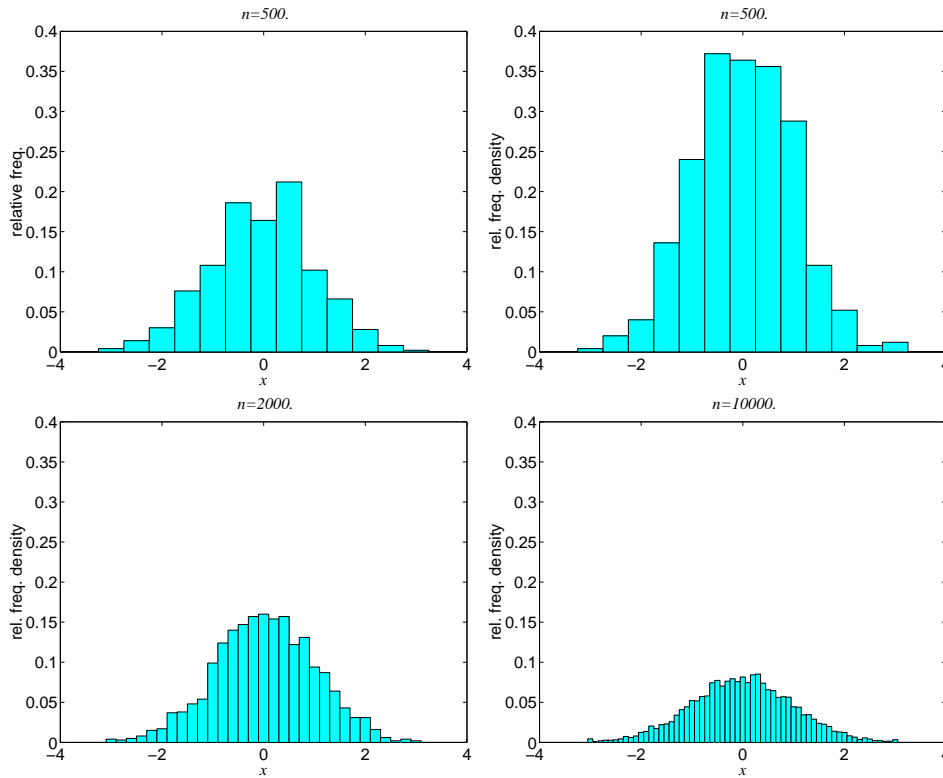


Figure 4.2: Top Left: Relative frequency histogram of Figure 3.2 (left). Top Right: Relative frequency density. Bottom Left and Right: as n increases the relative frequency density approaches an exponential distribution which does not change as n increases.

where μ denotes the **population mean**, and σ^2 is the **population variance**. The constants $\pi = 3.14159 \dots$ and $e = 2.71828 \dots$ make an equally impressive appearance in statistics as they do in mathematics.

If we assume our population follows the Gaussian distribution, the sample statistics, \bar{X} (3.1), and s^2 (3.4) are considered to be *estimates* of the real μ and σ^2 respectively. In biological experiments we often repeat measurements (**replicate measurements**) and then average the sample to obtain a more accurate value. To guess how many replicates we may need, in Figure 4.3 we have randomly selected 50 values from a process that follows a normal distribution with zero mean and unit variance. The histogram is shown on the left. We then took 2, 3, \dots , 50 values to calculate the sample mean. Since the population mean is zero, the sample mean calculated by equation (3.1) should be around zero. As the graph on the right shows, only for more than 30 replicates we get reliable estimates of the real mean value. The sample mean is therefore dependent on the sample size n and subject to variations. This problem is further discussed in Section 5.

The simplest of the normal distributions is the **standard normal** distribution. It has zero mean and unit variance. As shown in Figure 4.4, the area plus/minus one standard deviations from the mean captures 68.27% of the area. Since the total area equals 1, we can say that, the probability that an observation is found in the interval $[-\sigma, \sigma]$ is 0.68. In general, for any interval $[a, b]$ in X , the probability $P(a < x < b)$ is calculated by the area under the curve. It is useful to remember some of

Population mean μ
Population variance σ

Estimates of μ and σ .

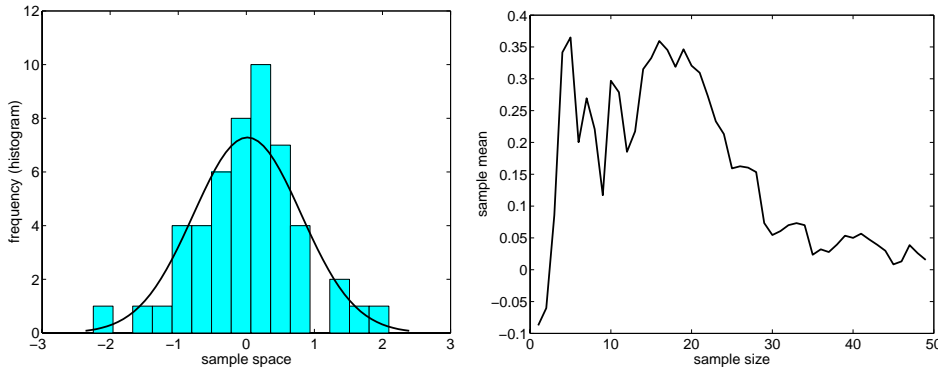


Figure 4.3: Estimation of the mean value for increasing sample sizes (from 2 to 50). The data were randomly sampled from a standard normal distribution with zero mean and unit variance.

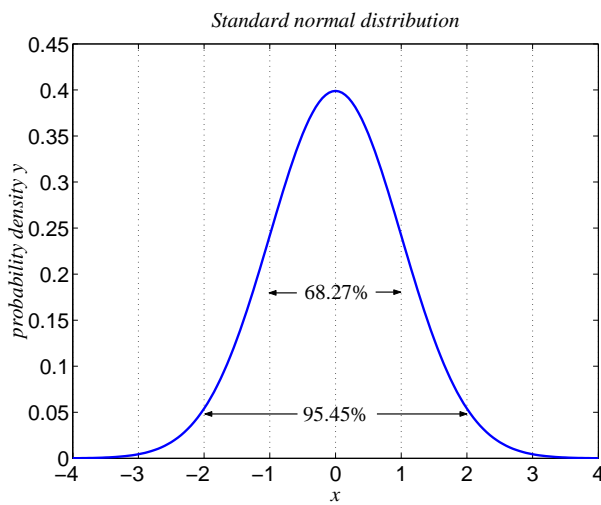


Figure 4.4: Standard normal distribution with zero mean and unit variance. The values for x are in standard units z .

the typical values for the normal distribution:

50% of the observations fall between $\mu \pm 0.674\sigma$.

95% of the observations fall between $\mu \pm 1.960\sigma$.

99% of the observations fall between $\mu \pm 2.576\sigma$.

It is often convenient to ‘translate’ an arbitrary Gaussian distribution to **standard units** by subtracting the mean and dividing by the standard deviation

$$z = \frac{x - \mu}{\sigma} . \quad (4.2)$$

Equation (4.1) is then replaced by the so called *standard form*:

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} , \quad (4.3)$$

where the constant $1/(\sqrt{2\pi})$ ensures that the area under the curve is equal to unity. Since we can easily translate forth and back between the

Standard units

actual distribution in question and the standard form, statistical tables and software programmes will usually only provide information about z -values. The main reason to use tables is however that formula (4.3) is too complicated to integrate the area under the curve. Statistical tables are therefore used to help calculate the probability of observations falling into certain regions. Statistical tables vary considerably from book to book and you should make sure that you are familiar with the table used in your examination.

PRACTICE. Try answering the following questions from the curve in Figure 4.4:

1. What percentage of the observations will be at least one but less than two standard deviations below the mean?
2. What percentage of the observations will be more than two standard deviations away from the mean?
3. Mark the plus/minus 3 standard deviation region; about what percentage of the observations would fall within three standard deviations of the mean?

Virtually all tables quote probabilities corresponding to *one* tail of the distribution only. This will be either

- a) the area between the mean and a positive z -value,
- b) the area between positive z -value and infinity.

Case b) gives the standard normal, cumulative probability in the right-hand tail. In other words, for a given value z_0 , the table provides information about the area that corresponds to the probability $P(z \geq z_0)$. This situation is for example the case for tables in [1] where areas in the tail of the normal distribution are tabulated as $1 - \Phi(z)$, and $\Phi(z)$ is the cumulative distribution function of a standardized Normal variable z . Thus $1 - \Phi(z) = 1/(\sqrt{2\pi}) \int_z^\infty e^{-x^2/2} dx$ is the probability that a standardized Normal variable z selected at random will be greater than the value $z_0 (= (x - \mu)/\sigma)$.

Example: Suppose we assume a normal distribution for an experiment with $\mu = 9.5$ and standard deviation $\sigma = 1.4$. We wish to determine the probability of observations greater than 12. Using the information in Table 4.1, we first must standardize the score $x = 12$ from equation (4.2)

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{12.0 - 9.5}{1.4} = \frac{2.5}{1.4} = 1.79 \end{aligned}$$

Using a statistical table we obtain

$$P(x > 12) = P(z > 1.79) = .037 \approx 4\%$$

Can you see how one can determine the probability for any interval $[a, b]$ from the same table?

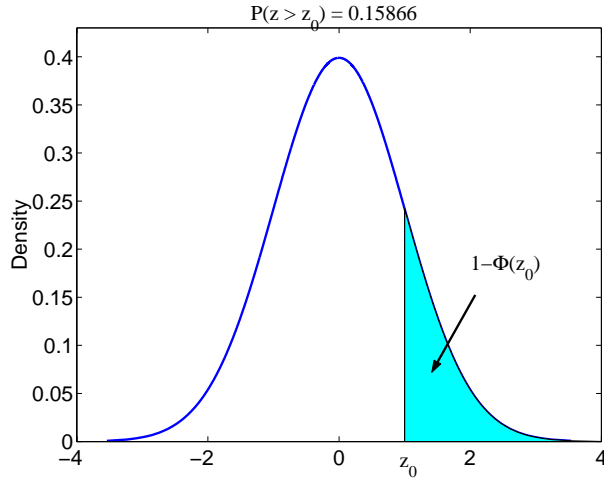


Figure 4.5: Standard normal distribution. The values of the shaded area are listed in Table 4.1.

Table 4.1: Extract of a statistical table for the the standard normal, cumulative probability in the right-hand tail (see Figure 4.5). The column on the left defines the given value z_0 and the columns to the right give the probability $P(z \geq z_0)$ for $0, 1, \dots, 9$ decimal places.

z_0	next decimal place of z_0							
	0	1	2	...	6	7	8	9
0	.500	.496	.492476	.472	.468	.464
\vdots								
0.3	.382	.378	.374359	.356	.352	.348
\vdots								
1.7	.045	.044	.043039	.038	.038	.037
\vdots								\vdots

PRACTICE. Answer the following questions using Table 4.1.

1. Assuming a normal distribution with $\mu = 9.5$ and standard deviation $\sigma = 1.4$, determine the probability of observations being greater than 10.
2. As before but determine the probability of observations being greater than -12 .
3. Calculate the probability of values falling in between 10 and 12.

Note: There are a number of important concepts we have not dealt with and you are encouraged to study one of the books recommended in Section 13. In particular the following two distributions are important:

1. *Binomial Distribution:* For a fixed number of independent trials, in which each trial has only two possible outcomes. The probabilities for the two possible outcomes are the same in each trial.
2. *Poisson Distribution:* To describe temporal or spatial processes the Poisson distribution is often a good model. Both, the binomial

and the Poisson distributions are *discrete* distributions, while the Normal distribution is continuous.

5 Sampling Errors

In the previous section we introduced the Normal distribution as a model for a population. Given a sample of data it is natural to think of the sample mean \bar{X} as an estimate of the population mean μ and the sample variance s^2 as an estimate of the population variance σ^2 . However, if we were to repeat taking samples from the same population, we would find that the sample mean itself is subject to variations as illustrated in Figure 4.3. Comparing samples by comparing the sample means requires therefore careful consideration of the uncertainty involved in such decisions. Tests, comparing samples are introduced in the next section and in this section we are going to estimate the error that can occur as a result of the variability of samples.

If we were able to take an infinite number of samples from a population with mean μ and standard deviation σ , the sample means \bar{X} would also be normally distributed, with mean μ and **standard error** SE. The standard error is calculated as

$$SE = \frac{\sigma}{\sqrt{n}} . \quad (5.1)$$

Note the dependency of the standard error on the sample size. The bigger the sample size n , the smaller the standard error and the better is our estimate.

Like for the standard deviation of the population model, 95% of the samples would have a sample mean within 1.96 times the standard error; 99% of the sample means would fall within 2.58 times the standard error, and 99.9% within 3.29 times the standard error. The 95%, 99%, and 99.9% limits can be used to describe the quality of our estimate and are referred to as **confidence intervals**.

Unfortunately, there is a problem with the calculation of the SE using equation (5.1): we do not know σ ! However, we have an estimate of the standard deviation in form of s and we can *estimate* the standard error therefore as follows:

$$\overline{SE} = \frac{s}{\sqrt{n}} . \quad (5.2)$$

Because the standard error is only estimated, the sample mean, \bar{X} will have a distribution with a wider spread than the normal distribution. In fact, it will follow a distribution, known as the ***t*-distribution**. The shape of this distribution will naturally depend on the sample size n . Actually one says, it is dependent on the “**degrees of freedom**”, which is in this case equal to $(n - 1)$. Figure 5.1, illustrates the difference between the Normal distribution and the *t*-distribution.

Confidence limits for the sample mean can be calculated using a table of **critical values** of the *t*-statistic. The critical *t* value $t_{(n-1)}(5\%)$ is the number of (estimated) standard errors \overline{SE} away from the estimate

Standard error

Confidence intervals

t-distribution

Degrees of freedom

Critical values

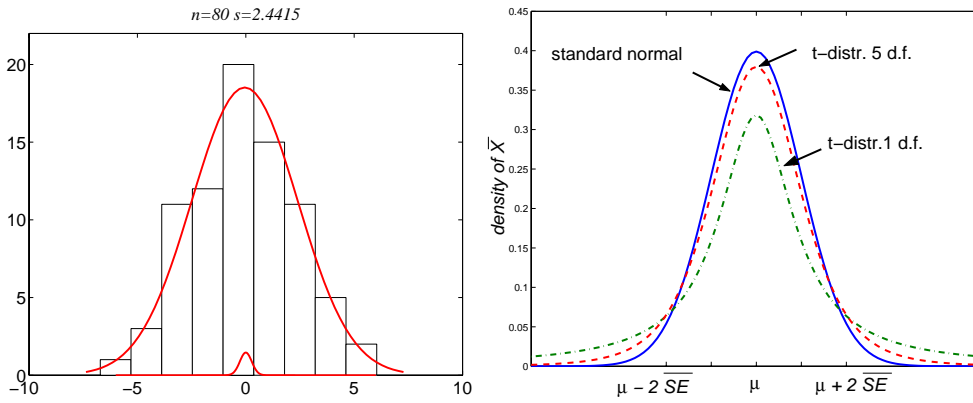


Figure 5.1: Left: The distribution of sample means is much narrower than the distribution of the population. Right: The distribution of sample means \bar{X} follows the t -distribution. This distribution is dependent on the sample size n (expressed as the ‘degrees of freedom’ $(n - 1)$). The greater the degrees of freedom, the narrower the distribution becomes and the closer the t -distribution approaches a Normal distribution.

of population mean \bar{X} , within which the real population mean μ will be found 95 times out of hundred (... with probability 0.95). Why this is called a *critical* value will become clearer in the next section on testing differences. The 95% limits define the 95% confidence interval (95% CI), which we calculate as follows

$$95\% \text{ CI}(\text{mean}) = \bar{X} \pm (t_{(n-1)}(5\%) \times \overline{\text{SE}}) \quad (5.3)$$

where $(n - 1)$ is the number of degrees of freedom. Similar one can determine the 99% and 99.9% confidence intervals for the mean by substituting the critical t values for 1% and 0.1% into equation (5.3), respectively. Table 5.1 shows an extract from a table with critical values for the t -statistic.

Table 5.1: Critical values of t at the 5%, 1%, and 0.1% significance levels. Reject the null hypothesis if the absolute value of t is larger than the tabulated value at the chosen significance level (and w.r.t. the number of degrees of freedom).

d.f. ($n - 1$)	Significance level		
	5%	1%	0.1%
1	12.706	63.657	636.619
	\vdots	\vdots	\vdots
5	2.571	4.032	6.859
	\vdots	\vdots	\vdots
9	2.262	3.250	4.781
10	2.228	3.169	4.587
	\vdots	\vdots	\vdots
20	2.086	2.845	3.850
	\vdots	\vdots	\vdots

Example: For a given sample mean $\bar{X} = 0.785$, sample standard deviation $s = 0.2251$, and $n = 11$, we calculate the 99%CI as follows:

$$99\% \text{ CI}(\text{mean}) = \bar{X} \pm (t_{(n-1)}(1\%) \times \overline{\text{SE}})$$

For the standard error $\overline{\text{SE}} = \frac{s}{\sqrt{n}} = \frac{0.2251}{\sqrt{11}} = 0.0678$ and 10 degrees of freedom, we obtain from Table 5.1, $t_{10}(1\%) = 3.169$. We therefore have

$$99\% \text{ CI}(\text{mean}) = 0.785 \pm (3.169 \times 0.0678) = [0.57, 1] .$$

The 99% confidence interval is therefore $[0.57, 1]$.

Note: One must be careful interpreting the meaning of the confidence limits of a statistic. When we set the lower/upper limits $\pm(t_{(n-1)}(1\%) \times \overline{\text{SE}})$ to a statistic, we imply that the probability of this interval covering the mean is 0.99 or, what is the same, we argue that on average, 99 out of 100, confidence intervals similarly obtained would cover the mean. Note that this is different from saying that there is a probability of 0.99 that the true mean is contained within any particular observed confidence limits.

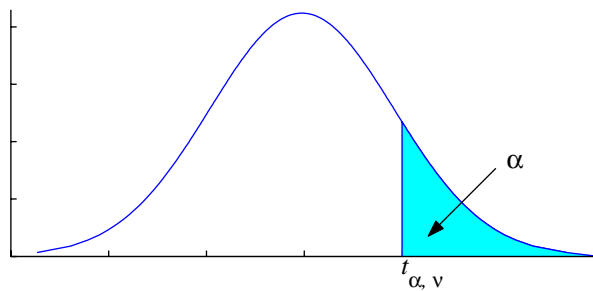


Figure 5.2: Illustration of the values listed in Table 5.2.

Note: Statistical tables published in books differ. For example, in [1], the same information has to be extracted from a table listing the percentage points of the t -distribution for one tail only. In this case, the 5% significance level corresponds to the 100α percentage point and is found in the column for $\alpha = 0.025$. Similar the 1% and 0.1% significance levels are found in the columns for $\alpha = 0.005$ and $\alpha = 0.0005$, respectively. Table 5.2 shows an extract. See also Figure 5.2.

PRACTICE. Try the following problems.

1. Using Table 5.1, we wish to compare two samples, both of which have a sample mean equal to 4.7 and sample variance 0.0507.
 - (a) For a sample of 11 observations, estimate the standard error and calculate the 95% confidence limits on the mean.
 - (b) For $n = 21$, estimate the standard error and calculate the 95% and 99% confidence limits for the mean. What is the effect of an increased sample size?
2. Using the following random sample, construct the 95% confidence interval for the sample mean.

Table 5.2: Percentage points of the t distribution [1]. The table gives the value of $t_{\alpha,\nu}$ - the 100% percentage point of the t distribution for ν degrees of freedom as shown in Figure 5.2. The tabulation is for one tail only, i.e., for positive values of t . For $|t|$ the column headings for α must be doubled.

$\alpha \rightarrow$	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
\vdots							\vdots
$\nu = 9$	1.383	1.833	2.262	2.821	3.250	4.297	4.781
$\nu = 10$	1.372	1.812	2.228	2.764	3.169	4.144	4.587
\vdots							\vdots
$\nu = 19$	1.328	1.729	2.093	2.539	2.861	3.579	3.883
\vdots							\vdots

49	83	58	65	68	60	76	86	74	53
71	74	65	72	64	42	62	62	58	82
78	64	55	87	56	50	71	58	57	75
58	86	64	56	45	73	54	86	70	73

[Solution: 66.0 ± 3.8]

Remark: Comparing two samples by comparing their mean and standard variation, it is important to state the confidence interval, (especially if the sample sizes varied). In graphical representations this is often shown using **error bars**.

6 Testing for Differences: The t-Test

With the material of the previous sections we have now available some of the tools that are necessary for the most frequent application of statistics in biology: testing a hypothesis related to a sample of data. The purpose of statistical **hypothesis testing** is to establish *significance tests* helping us in decision making and quantifying the uncertainty in this process. For example, taking two separate samples, we wish to compare the average values and test whether they are different. From the previous section, we know that sample means itself vary and a numerical difference between two sample means does not necessarily mean that this corresponds to a difference in the population means. The difference between two sample means may happen by chance. Figure 6.1, illustrates the probability of getting a sample mean that is one standard error \overline{SE} greater or smaller than the expected value μ .

In the following we consider three tests for different scenarios: testing the difference between a sample and an expected value, testing the difference between two samples from the same population, and testing the difference between two samples from two populations. Since the test we consider here involve inferences about population parameters, they are also referred to as **parametric tests**. The t -tests are valid for relatively small samples ($n < 30$).

Hypothesis testing

Parametric tests

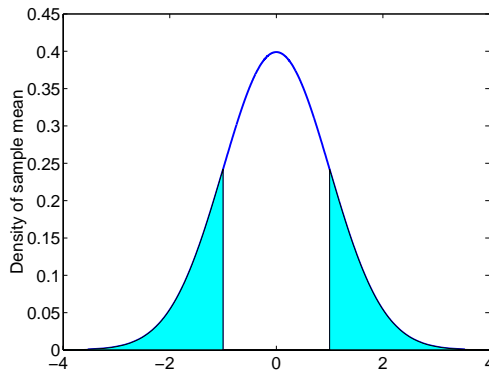


Figure 6.1: The distribution of sample means. The shaded area corresponds to the probability of getting a sample mean that is one standard error \overline{SE} greater or smaller than the expected value μ .

The one-sample t-test

Given some **expected value**, which may be assumed, we here test whether a sample taken from a population is different to the expected value. The one-sample t -test determines how many standard errors the sample mean is away from the expected value: The further the sample mean is away, the less likely it is that the mean and expected value are the same.

Before using a test, statisticians define a **null-hypothesis**, H_0 , stating the opposite of what you are testing. If you are testing for a difference, the null-hypothesis states that there is no difference. For the one-sample t -test, the null-hypothesis is that the mean of the population is *not* different from the expected value. The test will give us a probability to either accept or reject the null-hypothesis.

Null-hypothesis H_0

The next step is to calculate the test statistic t , which defines the number of standard errors the sample mean is away from the expected value, the latter of which is denoted by the letter E :

$$t = \frac{\text{sample mean} - \text{expected value}}{\text{standard error of mean}} = \frac{\bar{X} - E}{\overline{SE}} . \quad (6.1)$$

Once the t statistic is calculated, we can compare its absolute value, written $|t|$, with the critical value of the t statistic for $(n - 1)$ degrees of freedom, at the 5% level, i.e., $t_{(n-1)}(5\%)$, obtained from a statistical table, such as Table 5.1. The decision is made as follows:

- ✘ If $|t|$ is greater than the critical value, the test concludes that the mean is significantly different from the expected value: you must reject the null hypothesis.
- ✓ If $|t|$ is less than the critical value, the mean is not significantly different from the expected value and there is therefore no evidence to reject the null hypothesis.

Using a statistical software package, you will also be able to obtain the probability P that the absolute value of t would be this high or greater if the null hypothesis were true. Note that the smaller the value of $|t|$, the greater the value of P . This probability P is called the **significance**

probability. In many textbooks this is also referred to as the P -value or “achieved significance level”. In general, the P -value is the probability of observing the given sample result under the assumption that the null-hypothesis is true. Using some statistical programme, you can make your decisions depending on the significance probability instead of using the table:

- ✗ If $P < 0.5$, the null hypothesis should be rejected.
- ✓ If $P \geq 0.5$, there is no evidence to reject the null hypothesis.

Finally, you can calculate the 95% confidence limits for the difference by using the following equation:

$$95\% \text{ CI(difference)} = \bar{X} - E \pm (t_{(n-1)}(5\%) \times \overline{\text{SE}}) . \quad (6.2)$$

Example: Ten independent observations are taken from a normal distribution with mean μ . The sample variance is $s^2 = 40$ and the sample mean is $\bar{X} = 16$. We wish to test whether the sample mean is significantly different from the expected value. The hypotheses are therefore

- Null hypothesis: $\mu = 20$
- Alternative hypothesis: $\mu \neq 20$.

The standard error of the sample mean is

$$\overline{\text{SE}} = \frac{s}{\sqrt{n}} = \frac{6.325}{3.162} = 2 .$$

The t -statistic is

$$t = \frac{16 - 20}{2} = -2 .$$

In other words, the sample mean is -2 standard errors away from the expected value. From Table 5.2, for a 1% significance level we look at the column for $\alpha = 0.005$ and row $\nu = 9$, to obtain $t_{\alpha,\nu} = t_{0.005,9}(1\%) = 3.25$. Since $|t|$ is much less than the critical value, we have no evidence to reject the null hypothesis and conclude that the sample mean is not significantly different from the expected value.

The 99% confidence interval for the difference is calculated as

$$99\% \text{ CI(difference)} = 16 - 20 \pm (3.25 \cdot 2) = -4 \pm 6.5 = [-10.5, 2.5] .$$

In other words, 99% of all observed differences would be expected to lie in this interval. Note that the sample mean obtained here does not fall into this range and that this conclusion is quite independent of the decision to accept or reject the null hypothesis.

Example: The mean time taken for a plant to die when treated with a solution is known to be 12 days. When the solution dose was twice as much in a sample of ten plants, the survival times were 10, 10.7, 12.4, 12.2, 9.8, 9.9, 10.4, 10.8, 10.1, 11.2.

If the survival times are following a Normal distribution, test whether the results suggest that the increased solution dosage does lead to a decreased survival time. If we denote the mean survival time for double the dose with μ , then the null hypothesis H_0 is $\mu = 12$ and the alternative hypothesis H_1 is $\mu < 12$ (Note the difference to the previous example!):

$$\begin{aligned} H_0 : \mu &= 12 \\ H_1 : \mu &< 12 . \end{aligned}$$

Choosing a 1% significance level, we are going to look for critical region in one tail of the t -distribution (because of H_1). The 1% significance level means that we are looking for an area in the left tail of the t -distribution which has probability 0.01 (a one-sided hypothesis test). The sample mean

$$\bar{X} = \frac{1}{10} \cdot \sum_{i=1}^{10} x_i = 10.75$$

The sample variance

$$s^2 = \frac{1}{9} \sum_{i=1}^{10} x_i^2 - 10\bar{X}^2 = \frac{1163.39 - 1155.625}{9} = 0.863 \quad \text{and} \quad s = 0.93 .$$

The value of the t -statistic will be

$$t = \frac{10.75 - 12}{0.93/\sqrt{10}} = \frac{-1.25}{0.29} = -4.31 .$$

From Table 5.2, for $\nu = 9$, (the table provides values for positive ts), the critical region in the left tail of the t -distribution is for values of t smaller than -2.821 . Since the value $t = -4.31$ is much further to the left in the critical region, H_0 is rejected. In other words, doubling the dose reduces the survival time of the plants.

Note: The choice of the null hypothesis should be made *before* the data are being analysed as otherwise one might introduce a bias into the analysis. We speak of a **one-sided hypothesis** if we test for a statistic being greater or smaller than a value (e.g. $\mu > 0.3$) and a hypothesis is called **two-sided** if we test whether the statistic is different to a value (e.g. $\mu \neq 0.3$).

Remark: You should read the following definitions carefully and try to remember them. The **P-value** is the probability of the observed data (or data showing a considerable departure from the null hypothesis) when the null hypothesis is true. The P -value is *not* the probability of the null hypothesis nor is it the probability that the data have arisen by chance. The **significance level** is the level of probability at which it is agreed that the null hypothesis will be rejected. Conventionally this value is set to 0.05. A **significance test** is then a statistical procedure that when applied to a set of data results in a P -value relative to some hypothesis.

PRACTICE. Using following data set [7],

4.5	5.2	4.9	4.3	4.6	4.8	4.6	4.9
4.5	5.0	4.8	4.6	4.6	4.7	4.5	4.7

1. estimate the population mean and variance.
2. decide whether the sample mean is significantly different from a population with a mean value of 4.5.
3. use a statistical software package, such as MS Excel to calculate the *P*-value (significance probability).
4. use Table 5.1 to obtain the critical value of *t* at the 5% significance level.
5. calculate the 95% confidence limits for the difference.

The paired t-test

With the paired *t*-test we compare the means from *two* samples obtained from what we consider to be a *single* population. For example, you may take two samples at different times from the same culture (colony, or Petri dish). Other typical experiments for which this test is used include “before/after” or “treated/untreated” descriptions of the experiment. Let X_A and X_B denote the two samples, d is the difference, $X_A - X_B$, of the two samples, and \bar{d} is the average, $1/n \sum d$, of the differences. As with the one-sample *t*-test, the steps to follow are:

Step 1: The null-hypothesis is that the mean difference, \bar{d} is *not* different from zero.

Step 2: The test statistic *t* is the number of standard errors the difference is away from zero:

$$t = \frac{\text{mean difference}}{\text{standard error of difference}} = \frac{\bar{d}}{\overline{\text{SE}}_d}$$

where

$$\overline{\text{SE}}_d = \frac{s_d}{\sqrt{n}} .$$

Step 3: Calculate the significance probability *P* that the absolute value of the test statistic would be equal or greater than *t* if the null hypothesis were true. Using a statistical table, compare the value $|t|$ calculated above with the critical value of the *t* statistic for $(n - 1)$ degrees of freedom and at the 5% level, i.e., $t_{(n-1)}(5\%)$. The bigger the value of $|t|$, the smaller the value of *P*.

Step 4: Hypothesis testing:

- ✕ If $|t|$ is greater than the critical value, the null hypothesis is rejected: The mean difference is significantly different from zero.

- ✓ If $|t|$ is less than the critical value, then there is no evidence to reject the null hypothesis.

Using a statistical software package,

- ✗ If $P < \alpha = 0.05$, reject the null hypothesis.
- ✓ If $P \geq \alpha = 0.05$, there is no evidence to reject the null hypothesis, the mean difference is not significantly different from zero.

Step 5: Calculate the 95% confidence limits for the mean difference as

$$95\% \text{ CI(difference)} = \bar{d} \pm (t_{(n-1)}(5\%) \times \overline{\text{SE}}_d) .$$

Since the decision whether to accept or reject a hypothesis is made on the basis of data that are randomly selected, an incorrect decision is possible. If we reject the null hypothesis H_0 when it is true, this is called a **Type I error**. Similarly, if we accept H_0 when it is false, we commit a **Type II error**. By choosing α (usually 1% or 5%) we fix the Type I error to some acceptable low level. If the P -value is *less than* the chosen Type I error, the null hypothesis is rejected.

Type I error
Type II error

The two-sample t-test

The purpose of this test is to decide whether the means of two samples obtained from *two* populations are different from each other. We assume that both **samples are independent** of each other. For example, this test does not apply to samples taken from the same culture.

Both sample means will have a distribution associated with it, and as illustrated in Figure 6.2, the test effectively tests the overlap between the distributions of the *two* sample means. Here we consider only the case, when it is reasonable to assume that the two **populations have the same variance**. (Most software packages will have available tests for populations with different variances.)

Step 1: The null-hypothesis is that the mean of the differences *is not* different from zero. In other words, the two groups A and B from which the samples were obtained have the same mean.

Step 2: The test statistic t is given by the following formula:

$$t = \frac{\text{mean difference}}{\text{standard error of difference}} = \frac{\bar{X}_A - \bar{X}_B}{\overline{\text{SE}}_d}$$

The standard error of the difference $\overline{\text{SE}}_d$ is more difficult to calculate because this would involve comparing each member of the first population with each member of the second. Assuming that the variance of both populations is the same, we can estimate $\overline{\text{SE}}_d$ using the following equation:

$$\overline{\text{SE}}_d = \sqrt{(\overline{\text{SE}}_A)^2 + (\overline{\text{SE}}_B)^2} ,$$

where $\overline{\text{SE}}_A$ and $\overline{\text{SE}}_B$ are the standard errors of the two populations.

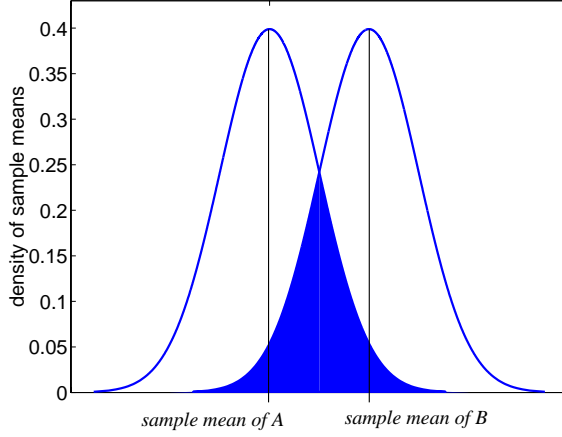


Figure 6.2: In the two-sample t -test, we wish to decide whether the means of two samples, obtained from two populations are different. In other words, we wish to quantify the overlap between the distributions of the sample means.

Step 3: Calculate the significance probability P that the absolute value of the test statistic would be equal to or greater than t if the null hypothesis were true. There are $n_A + n_B - 2$ degrees of freedom, where n_A and n_B are the sample sizes of groups A and B .

Step 4: Using a statistical software package,

- ✗ If $P < 0.05$, reject the null hypothesis, the sample means are significantly different from each other.
- ✓ If $P \geq 0.05$, there is no evidence to reject the null hypothesis, the two sample means are not significantly different from each other.

Step 5: The 95% confidence interval for the mean difference is given by

$$95\% \text{ CI(difference)} = \bar{X}_A - \bar{X}_B \pm (t_{(n_A+n_B-2)}(5\%) \times \overline{\text{SE}}_d) .$$

Example: We obtain two independent samples X_A , X_B and we wish to calculate the 95% confidence interval for the difference of the two group means:

$$X_A = \{64, 66, 89, 77\}, \quad X_B = \{56, 71, 53\} .$$

We calculate $\bar{X}_A = 296/4 = 74$ and $\bar{X}_B = 180/3 = 60$, $s_A = 11.5181$ and $s_B = 9.643$; $\overline{\text{SE}}_A = 5.7591$, $\overline{\text{SE}}_B = 5.5678$; $\overline{\text{SE}}_d = 8.01$. Thus,

$$95\% \text{ CI(difference)} = 74 - 60 \pm (2.571 \cdot 8.01) = 14 \pm 21 .$$

Example: In an experiment we are comparing an organism for which the cells were generated by two independent methods (A and B). At a certain stage of the development the length is measured. The data are summarised in Table 6.1. If the lengths are following a Normal distribution, we wish to test whether they are significantly different for the two groups. The hypotheses are:

$$\begin{aligned} H_0 : \mu_A &= \mu_B \\ H_1 : \mu_A &\neq \mu_B . \end{aligned}$$

Table 6.1: Experimental data set.

Origin	length (mm)	\pm	s (mm)
A	87.04	\pm	7.15
B	77.77	\pm	4.70

We have $\bar{A} = 87.04$, $\bar{B} = 77.77$, $s_A = 7.15$, $s_B = 4.70$ and thus

$$\overline{SE}_A^2 = \frac{s_A^2}{n_A} = 2.56, \quad \overline{SE}_B^2 = \frac{s_B^2}{n_B} = 1.10, \quad \overline{SE}_d = \sqrt{\overline{SE}_A^2 + \overline{SE}_B^2} = 1.91$$

and therefore

$$t = \frac{\bar{A} - \bar{B}}{\overline{SE}_d} = 4.84 .$$

Choosing a 1% significance level, we find that for $\nu = 38$ most tables will not list the desired values. We can however interpolate (from tables like Table 5.2 [1]) such that if values are given for say $\nu = 30$ and $\nu = 40$, this gives us a range for the t -statistic to lie between approximately -2.72 and 2.72 for H_0 to be accepted. Since 4.84 is considerably larger than 2.72 , the null hypothesis is rejected.

The calculation of the t -statistic for the two-sample t -test can be done in different ways and textbooks will sometimes provide the following description of Step 2: If the null hypothesis is correct, the following t -statistic has a Student's t -distribution with $\nu = n_A + n_B - 2$ degrees of freedom:

$$t = \frac{\bar{X}_A - \bar{X}_B}{s_p \cdot \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \quad \text{where} \quad s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

Apart from the different calculation there is no change, i.e., using a statistical table for the t -distribution, we would check whether the calculated t -statistic falls into the critical region.

Equation using s_p is based on the idea, that since we assume that both samples have the same variance, we can ‘pool’ them:

$$s_p^2 = \frac{\sum (n_A - 1)^2 + \sum (n_B - 1)^2}{(n_A - 1) + (n_B - 1)} .$$

Since

$$s_A^2 = \frac{\sum (X_A - \bar{X}_A)^2}{n_A - 1} \quad \text{and} \quad s_B^2 = \frac{\sum (X_B - \bar{X}_B)^2}{n_B - 1} ,$$

we have

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \quad \text{or} \quad s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} .$$

As an exercise you should compare the two different strategies and compare the difference.

Note: If you want to compare the means of more than two groups, you cannot use the t -test. For this case, a more sophisticated test, called **ANOVA** test (analysis of variance) is available. We have also not considered experiments associated with distributions others than the Normal distribution. Statistical t -tests are valid for small sample sizes (say n less than 30). For larger samples z -tests should be used.

Remark: One usually doesn't learn about the origins of a mathematical concept although in statistics it is often rather interesting to know how the various, often alternative, techniques have developed. The following story about the t distribution and test is frequently told. William Sealy Gosset (1876 - 1937) studied chemistry at Oxford University and later worked for the Guinness brewery. Investigating the relationship between the final product and the raw materials, he developed the statistics we have discussed here. The company did not allow publications of this nature and he choose a pseudo name 'student'. Many textbooks will still refer to the distribution as the 'students t distribution. Why he choose to call it a ' t '-test when he was working with beer is unknown...

7 Categorical Data: The Chi-Square Test

The chi-squared (χ^2) test is used to determine whether there are differences between real and expected frequencies in one set of categories, or associations between two sets of categories. Also, in previous sections we have assumed that a particular type of distribution is appropriate for the the data. We then estimated parameters of this distribution and tested hypotheses about parameters.

Categorical data are data that are not numbers but measurements assigned to **categories**. Examples of **character states** are the colour of objects, conditions like dead/alive or healthy/diseased. Data with equal character states form categories. Categorical data is quantified by the frequency with which each category was observed. Similar to the t -tests, we can as the following questions:

Character states

- ▷ Are observed frequencies in a single group different from expected values?
- ▷ Are observed frequencies in two or more groups different from each other?

To answer these questions we have two tests available: the χ^2 test for differences and the χ^2 test for association.

Chi square test for differences

The purpose of this test is to decide whether observed frequencies are different from expected values. The χ^2 statistic calculated here is a measure of this difference. The null hypothesis is that the frequencies of the different categories in the population are equal to the expected frequencies. Critical values or percentage points of the χ^2 distribution can be found in tables of the same nature as Tables 5.1, 5.2.

The χ^2 statistic is calculated by the expression

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

where O is the observed frequency and E is the expected frequency for each character state (category). The larger the difference between the frequencies, the larger the value of χ^2 and the less likely it is that observed and expected frequencies are different just by chance. Different samples will give different observed frequencies and hence different values for χ^2 . Thus χ^2 has a probability distribution which is *illustrated* in Figure 7.1. (Actually, there is a small difference between the expression (7.2) and the distribution in Figure 7.1, but since I referred to Figure 7.1 as an illustration this may not be a problem.).

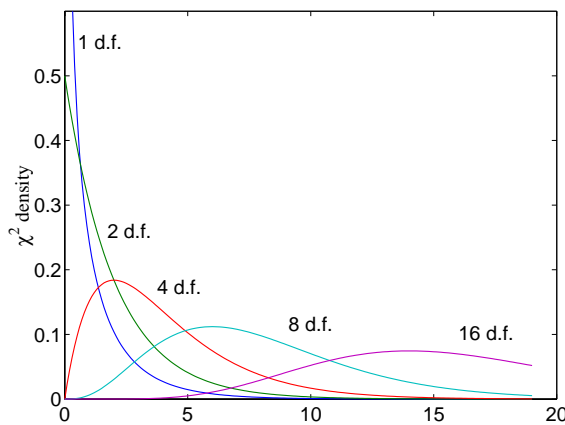


Figure 7.1: Chi square, χ^2 distribution with different degrees of freedom. In practise one would not use such graph but tables to obtain values required for calculations.

The probability P of obtaining χ^2 values equal or greater than the observed values if the null hypothesis were true can be obtained from tables which list the critical value that χ^2 must exceed at $(N - 1)$ degrees of freedom, where N is the number of groups, for the probability to be less than 5%.

Note: The distribution of χ^2 is depends on the number of degrees of freedom - the bigger the sample you take, the more likely you will be to detect any differences. Note that therefore the two tests we introduce here are only valid if all expected values are larger than 5.

Chi square test for association

With this test we wish to decide whether the character frequencies of two ore more groups are different from each other. In other words, we test whether character states are associated in some way. The test investigates whether the distribution is different from what it would be if the character states were distributed randomly among the population.

The null hypothesis is that there is no difference between the frequencies of the groups, hence no association between the character states. Before we can calculate the χ^2 statistic we must calculate the expected

values for each character state if there had been no association. To do this we arrange the data in a **contingency table**:

	Character a	Character b	Total
Group A	frequency	frequency	
Group B	frequency	frequency	
Total			

The expected number E (frequency) if there had been no association between the character states in the two groups is given by

$$E = \frac{\text{column total} \times \text{row total}}{\text{grand total}} \quad (7.1)$$

The grand total is the sum of the two row totals. The significance probability is obtained from a statistical table as the critical value that χ^2 must exceed at $(R - 1) \times (C - 1)$ degrees of freedom, where R is the number of rows in the table above and C is the number of columns, for the probability to be less than 5%. If χ^2 is greater than the critical value, the null hypothesis is rejected - there is a significant associations between the characters.

PRACTICE. *Through experiments on two groups we found that in group A, out of 30 objects, 18 had character state a and 12 had character state b, while of the 60 objects in group B, 12 had character state a and for 48 objects we observed character state b. Test whether character state a is significantly different in the groups. Using a software package such as Minitab, MS Excel, Matlab or Mathematica,*

1. *formulate the null hypothesis.*
2. *calculate the test statistic.*
3. *determine the P-value.*
4. *decide whether to reject the null hypothesis.*

Chi-square test for goodness of fit

The χ^2 test can also be used to determine how well theoretical distributions (such as the Normal distribution) fit empirical distributions (i.e., those obtained from sample data). As in previous sections, a measure for the **goodness of fit** of the model can be established with the following statistic:

$$\chi^2 = \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i}, \quad (7.2)$$

where m denotes the number of different outcomes. Significantly large values of χ^2 suggest a lack of fit. We are now going to see how the chi-square statistic can be used to test whether a frequency histogram fits the normal distribution.

Fitting a normal curve to the data of Table 2.1, we first calculate standard units for the class boundaries, $z = (x - \bar{X})/s$. The areas under the normal curve can be obtained from tables (e.g. Table 4.1). From this we find the areas under the normal curve between successive values of z as shown in column 5 of Table 7.1. These are obtained by subtracting the successive areas in column 4 when the corresponding z 's have the same sign, and adding them when the z 's have opposite sign. Multiplying the entries in column 5 (rel. freq.) by the total frequency ($n = 100$) gives us the expected frequencies, as shown in column 6. To determine the goodness of fit, we calculate

$$\begin{aligned}\chi^2 = & \frac{(5 - 4.13)^2}{4.13} + \frac{(18 - 20.68)^2}{20.68} + \frac{(42 - 38.92)^2}{38.92} \\ & + \frac{(27 - 27.71)^2}{27.71} + \frac{(8 - 7.43)^2}{7.43} = 0.059\end{aligned}$$

Since the number of parameters used in estimating the expected frequencies is 2, we have $\nu = 5 - 1 - 2 = 2$ degrees of freedom. From a table we find $\chi^2_{.95} = 5.99$. Thus we can conclude that the fit of the data is good.

Table 7.1: Fitting a normal curve to the data in Table 4.1 and testing the fit of the frequency histogram in Figure 2.2 to the normal distribution [5].

class limits	class boundaries	z for class limits	area under normal curve from 0 to z	area for each class	expected frequency	observed frequency
60-62	59.5	-2.72	0.4967	0.0413	4.13, or 4	5
63-65	62.5	-1.70	0.4554	0.2068	20.68, or 21	18
66-68	65.5	-0.67	0.2486	0.3892	38.92, or 39	42
69-71	68.5	0.36	0.1406	0.2771	27.71, or 28	27
72-74	71.5	1.39	0.4177	0.0743	7.43, or 7	8
	74.5	2.41	0.4920			

Note: In this section, we introduced only the most basic concepts for categorical data. Books in the reference list (Section 13, page 39) will provide more details on the rationale behind the tests and will help you in selecting an appropriate test for a problem at hand. Another important issue, which we haven't dealt with, is the design of experiments.

8 Finding Associations: Correlation

... have a look at the references given in Section 13.

9 Modelling Relationships: Linear Regression

... have a look at the references given in Section 13.

10 More Exercises

Exercises for Section 2

The following sample consists of 12 temperature measurements taken every two hours: $-2, -3, -3, -2, 0, 4, 5, 6, 6, 6, 3, 1$. Calculate

1. The temperature average of the day, i.e., the sample mean \bar{X} . Do three calculations:
 - (a) Using equation (3.1).
 - (b) Using equation (3.2).
 - (c) Using the statistical function of your calculator.
2. The sample variance σ_n^2 , the sample standard deviation σ_n , and the unbiased estimate of the population variance s^2 :
 - (a) Using equations (3.3) and (3.4).
 - (b) Using equation (3.5).
 - (c) Using the statistical functions of your calculator.

50,	35,	19,	27,	44,	70,	60,	28,	61,	41,	50,	56
61,	52,	62,	66,	70,	52,	81,	43,	63,	52,	71,	51
60,	35,	49,	57,	44,	30,	60,	28,	61,	44,	55,	36
51,	62,	42,	66,	70,	42,	61,	43,	63,	52,	71,	51
50,	75,	44,	65,	44,	70,	60,	67,	65,	44,	55,	57

Table 10.1: Exam results for 60 students.

Table 10 lists the exam results for 60 students. For the given data set,

1. Calculate the range of the scores.
2. Construct the tally chart for the following score intervals

score	tallies
0 – 9	
10 – 19	
20 – 29	
30 – 39	
40 – 49	
50 – 59	
60 – 69	
70 – 79	
80 – 89	
90 – 100	

3. Determine the relative frequency distribution for the intervals specified above.
4. Visualise the relative frequency distribution with a relative frequency histogram.
5. Calculate and draw the cumulative frequency distribution.

Exercises for Section 3

For the data in Table 10,

1. Calculate the mean, the median, the mode and the standard deviation using your calculator or a software package.
2. Mark the calculated statistics in the relative frequency histogram calculated previously.
3. The n^{th} percentile is the score that leaves $n\%$ of the data to the left. Calculate the 10^{th} , 30^{th} , 60^{th} , and 90^{th} percentiles. Hint: Sort the data from the smallest to the largest value. Mark the percentiles in the relative frequency histogram.

Exercises for Section 4

1. For the distribution of the scores of Table 10 answer the following questions,
 - (a) is the distribution unimodal?
 - (b) is the distribution symmetric about the mean?
 - (c) calculate the percentage of observations falling between $\bar{X} + 0.674s$
 - (d) calculate the percentage of observations falling between $\bar{X} + 1.96s$
 - (e) calculate the percentage of observations falling between $\bar{X} + 2.576s$
2. Do you think that the scores of Table 10 are “normally distributed” (follow a normal or Gaussian distribution)?
3. Are the scores in Table 10.2 “more” normally distributed than those in Table 10?

53,	63,	52,	63,	61,	69,	60,	53,	56,	59,	62,	61
55,	58,	60,	60,	51,	59,	59,	65,	61,	59,	67,	71
60,	59,	69,	55,	44,	60,	59,	57,	58,	69,	56,	74
60,	57,	60,	60,	54,	46,	54,	59,	66,	63,	54,	64
58,	61,	68,	61,	52,	58,	62,	63,	66,	73,	57,	63

Table 10.2: Data set.

Exercises for Section 5

You read in a scientific report that the average age of death for women in your country is 73.2 years. To find out whether the average age of death for men is the same as that of women, a small sample of 25 death certificates shows an average age of 58.4 years and a sample standard deviation of 15 years.

1. Using a significance level of 0.01, choose an appropriate hypothesis test and determine whether the null hypothesis (there is no difference between mean and woman) should be accepted.
2. Use the 99% confidence interval for the men's average age of death to reach the same conclusion.

Exercises for Section 6

1. A researcher believes that the average weight in a group of people is 120 pounds. To test this belief, you determine the weight of 7 people with the following results (in pounds): 121, 125, 118, 130, 117, 123, and 120.
 - (a) Estimate the population mean and variance.
 - (b) Decide whether the sample mean is significantly different from a population with a mean value of 120.
 - (c) Obtain the critical value of t at the 5% significance level.
 - (d) Calculate the 95% confidence limits for the difference.
2. Imagine you want to test whether or not six minutes is enough time for the heart to recover the pulse rate after two minutes of exercise. For a period of one week the pulse rate is measured from one person, every day, before exercise and six minutes after the exercise, obtaining the data summarised in Table 10.3. Do these data indicate that the heart rate after the exercise is higher than before the exercise? Use a 1% level of significance.

Test	1	2	3	4	5	6
Before	69	72	75	73	70	74
After	85	79	83	84	87	78

Table 10.3: Data set.

3. We want to compare the efficiency of the two pieces of equipment, referred to as A and B . In Table 10.4 the first row shows the numerical values obtained for the efficiency measure for meter A and the second row show the results for meter B .
 - (a) Calculate the mean and the standard deviation for each group.
 - (b) Calculate the 95% confidence interval for the difference of the two group means
 - (c) What can you say about the efficiency of the two meters?

A	18	15	18	16	17	15	14	14	14	15
B	24	27	27	25	31	35	24	19	28	23

Table 10.4: Data set.

11 Solutions to Exercises in Section 10

Solutions to Section 10, related to Section 5

- $\overline{SE} = \frac{s}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$
 $t = \frac{\bar{X} - E}{\overline{SE}} = \frac{68.4 - 73.2}{3} = -1.6, |t| = 1.6$
 $t_{(24)}(1\%)$ from the table is 2.797, so we do not reject the null hypothesis.
- $99\%CI = \bar{X} \pm (t_{(24)}(1\%) \times \overline{SE})$
 $99\%CI = 68.4 \pm (2.797 \times 3)$
 $99\%CI = 68.4 \pm 8.3910$
The women's average age of death is included in the range.

Solutions to Section 10, related to Section 6

- Sample mean $\bar{X} = \frac{854}{7} = 122$
Estimated population variance $s^2 = 20$
Estimated population standard deviation $s = 4.472$
 $\overline{SE} = \frac{s}{\sqrt{n}} = \frac{4.472}{\sqrt{7}} = 1.690$
 $t = \frac{\bar{X} - E}{\overline{SE}} = \frac{122 - 120}{1.690} = 1.832, |t| = 1.832$
 $t_{(6)}(5\%)$ from the table is 2.447, so we do not reject the null hypothesis.
 $99\%CI = \bar{X} - E \pm (t_{(6)}(5\%) \times \overline{SE})$
 $99\%CI = 122 - 120 \pm (2.447 \times 1.690)$
 $99\%CI = 2 \pm 4.1352$
- $d = X_A - X_B$
 $\bar{d} = \frac{1}{n} \sum d = 10.5$
 $\overline{SE}_d = \frac{s_d}{\sqrt{n}} = 2.110$
 $t = \frac{\bar{d}}{\overline{SE}_d} = \frac{10.5}{2.110} = 4.977, |t| = 4.977$
 $t_{(5)}(1\%)$ from the table is 4.032, so we reject the null hypothesis.
- $\overline{SE}_A = \frac{S_A}{\sqrt{n}} = \frac{1.5776}{\sqrt{10}} = 0.4989,$
 $\overline{SE}_B = \frac{S_B}{\sqrt{n}} = \frac{4.4485}{\sqrt{10}} = 1.4067,$
 $\overline{SE}_d = \sqrt{\overline{SE}_A^2 + \overline{SE}_B^2} = 1.4926$
 $95\%CI = \bar{X}_A - \bar{X}_B \pm (t_{(18)}(5\%) \times \overline{SE}_d)$
 $95\%CI = 15.6 - 26.3 \pm (2.101 \times 1.4926)$
 $95\%CI = -10.70 \pm 3.136$
 $t = \frac{\bar{X}_A - \bar{X}_B}{\overline{SE}_d} = \frac{15.6 - 26.3}{1.4926} = -7.1687, |t| = 7.1687$
 $t_{(18)}(5\%)$ from the table is 2.101 so the means are different.

12 Symbols and Notation

The following symbols are used in the notes. They are listed in the order of their appearance.

A, B	samples or groups,
$F(x)$	cumulative distribution function.
X	sample.
$x_i \in X$	element, observation or measurement in sample.
$\{ \}$	list, set.
n	sample size (number of elements in sample).
\bar{X}	sample mean.
μ	population mean.
σ_n^2	sample variance.
$s^2 = \sigma_{n-1}^2$	unbiased estimator of population variance.
σ^2	population variance.
$\sigma = \sqrt{\sigma^2}$	population standard deviation.
$\sum_{i=1}^n$	sum of elements with index $i = 1, 2, \dots, n$.
f_i	frequency of observation (of x_i or in class i).
m	number of distinct value of x in a sample.
$p(x)$	probability density function.
$P(\cdot)$	probability.
$[a, b]$	interval ranging from a to b .
z	standard unit.
$\Phi(z)$	cum. distr. fcn. of standardized normal variable z .
SE	standard error.
\overline{SE}	estimated standard error.
CI	confidence interval.
d.f.	degrees of freedom. See also ν .
$t_{(n-1)}(\cdot)$	critical value of t distribution.
t	t statistic.
α	100% percentage point (of t distribution).
ν	degree of freedom.
E	expected value, expected frequency.
$ \cdot $	absolute value.
d	difference.
\bar{d}	average of differences.
P	P -value.
H_0, H_1	null hypothesis, alternative hypothesis.
\overline{SE}_d	standard error of mean difference.
χ^2	chi-square distribution and test.
O	observed frequency.
N	number of groups.
R, C	number of rows and columns in contingency table.

13 Further Reading

There are many statistics books available and I strongly recommend to buy at least one. The small book by Rowntree [3] is the most basic and yet very good introduction for those who are worried about the mathematics. It shows that ideas come first, and equations should only follow as a means towards an end. Since there are so many different techniques and concepts in statistics, it is very important to ask the ‘right’ question in order to identify a suitable technique, test or tool. Freedman’s [4] is a very good book in this respect too but much more comprehensive. [2] has many exercises and is at an introductory level. The Schaum Outline [5] is one of a series of well known introductory texts to mathematics. Wonnacott and Wonnacott [6] is a hardcover text and is likely to last a little longer. Although the area from which examples are taken shouldn’t matter too much for an introductory book, here the focus is on examples from business and economics. It has exercises and is suitable as a basic reference. Previous books contained general examples from science, engineering, economics etc, while [7] is a well written basic introduction for biologists - good value for money too. The recent book by Quinn and Keough [9] is a comprehensive treatment that is written for biologists. It is a textbook that is also a good reference. In addition to the basic statistics, it covers multivariate data analysis (clustering, principal component analysis,...) and regression techniques. Finally, Sokal and Rohlf [8] wrote one of the most comprehensive statistics book aimed at biologists. It is a very good but also advanced reference book.

Bibliography

- [1] Murdoch, J. and Barnes, J.A. (1974): *Statistical Tables for Science, Engineering, Management and Business Studies*. 2nd ed. Macmillan, London.
- [2] Upton, G. and Cook, I. (2001): *Introducing Statistics*. Oxford University Press. 470 pages. ISBN 9 780199 148011. £22
- [3] Rowntree, D. (1981): *Statistics without Tears*. Penguin. 199 pages. ISBN 0 14 013632 0. £7
- [4] Freedman, D. and Pisani, R. and Purves, R. (1998): *Statistics*. W.W. Norton & Company. 700 pages. ISBN 0 393 97121 X. £19
- [5] Spiegel, M.R. and Stephens, L.J. (1998): *Statistics*. Schaum's Outlines. 538 pages. ISBN 0 07 060281 6. £12
- [6] Wonnacott, T.H. and Wonnacott, R.J. (1990): *Introductory Statistics*. John Wiley & Sons. 711 pages. ISBN 0 471 61518 8. £16
- [7] Ennos, R. (2000): *Statistical and Data Handling Skills in Biology*. Prentice Hall. 132 pages. ISBN 0 582 31278 7. £10
- [8] Sokal, R.R. and Rohlf, F.J. (1995): *Biometry*. W.H. Freeman and Company. 887 pages. ISBN 0 7167 2411 1. £38
- [9] Quinn, G.P. and Keough, M.J. (2002): *Experimental design and Data Analysis for Biologists*. Cambridge University Press. 537 pages. ISBN 0 521 00976 6. £30