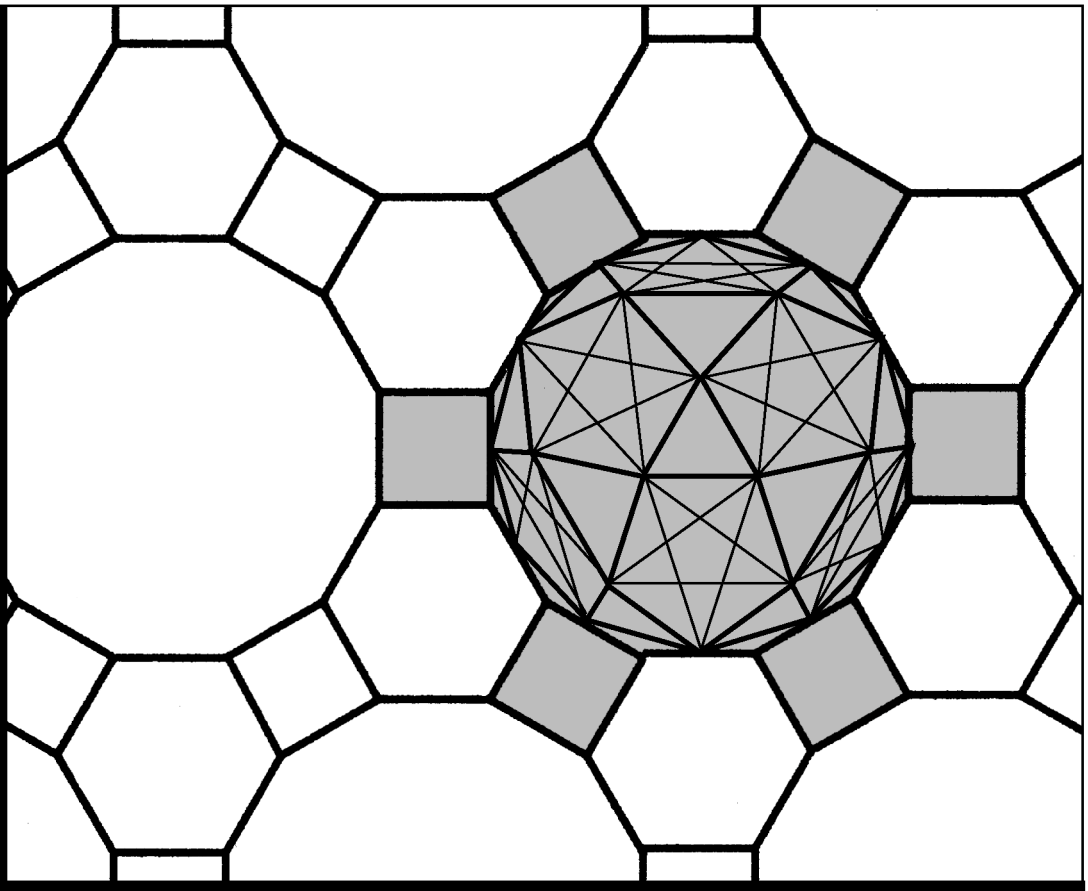
Workbook 36



Descriptive Statistics

HELM: Helping Engineers Learn Mathematics

http://helm.lboro.ac.uk

About the HELM Project

HELM (Helping Engineers Learn Mathematics) materials were the outcome of a three-year curriculum development project undertaken by a consortium of five English universities led by Loughborough University, funded by the Higher Education Funding Council for England under the Fund for the Development of Teaching and Learning for the period October 2002 – September 2005, with additional transferability funding October 2005 – September 2006.

HELM aims to enhance the mathematical education of engineering undergraduates through flexible learning resources, mainly these Workbooks.

HELM learning resources were produced primarily by teams of writers at six universities: Hull, Loughborough, Manchester, Newcastle, Reading, Sunderland.

HELM gratefully acknowledges the valuable support of colleagues at the following universities and colleges involved in the critical reading, trialling, enhancement and revision of the learning materials:
Aston, Bournemouth & Poole College, Cambridge, City, Glamorgan, Glasgow, Glasgow Caledonian, Glenrothes Institute of Applied Technology, Harper Adams, Hertfordshire, Leicester, Liverpool, London Metropolitan, Moray College, Northumbria, Nottingham, Nottingham Trent, Oxford Brookes, Plymouth, Portsmouth, Queens Belfast, Robert Gordon, Royal Forest of Dean College, Salford, Sligo Institute of Technology, Southampton, Southampton Institute, Surrey, Teesside, Ulster, University of Wales Institute Cardiff, West Kingsway College (London), West Notts College.

HELM Contacts:
Post: HELM, Mathematics Education Centre, Loughborough University, Loughborough, LE11 3TU.
Email: helm@lboro.ac.uk *Web:* http://helm.lboro.ac.uk

HELM Workbooks List

1	Basic Algebra	26	Functions of a Complex Variable
2	Basic Functions	27	Multiple Integration
3	Equations, Inequalities & Partial Fractions	28	Differential Vector Calculus
4	Trigonometry	29	Integral Vector Calculus
5	Functions and Modelling	30	Introduction to Numerical Methods
6	Exponential and Logarithmic Functions	31	Numerical Methods of Approximation
7	Matrices	32	Numerical Initial Value Problems
8	Matrix Solution of Equations	33	Numerical Boundary Value Problems
9	Vectors	34	Modelling Motion
10	Complex Numbers	35	Sets and Probability
11	Differentiation	36	Descriptive Statistics
12	Applications of Differentiation	37	Discrete Probability Distributions
13	Integration	38	Continuous Probability Distributions
14	Applications of Integration 1	39	The Normal Distribution
15	Applications of Integration 2	40	Sampling Distributions and Estimation
16	Sequences and Series	41	Hypothesis Testing
17	Conics and Polar Coordinates	42	Goodness of Fit and Contingency Tables
18	Functions of Several Variables	43	Regression and Correlation
19	Differential Equations	44	Analysis of Variance
20	Laplace Transforms	45	Non-parametric Statistics
21	z-Transforms	46	Reliability and Quality Control
22	Eigenvalues and Eigenvectors	47	Mathematics and Physics Miscellany
23	Fourier Series	48	Engineering Case Study
24	Fourier Transforms	49	Student's Guide
25	Partial Differential Equations	50	Tutor's Guide

Production of this 2015 edition, containing corrections and minor revisions of the 2008 edition, was funded by the **sigma** Network.

sigma Σ
network for excellence in mathematics and statistics support

Contents 36

Descriptive Statistics

Learning outcomes

In the first Section of this Workbook you will learn how to describe data sets and represent them numerically using, for example, means and variances. In the second Section you will learn how to explore data sets and arrive at conclusions, which will be essential if you are to apply statistics meaningfully to real situations.

Describing Data

36.1

Introduction

Statistics is a scientific method of data analysis applied throughout business, engineering and all of the social and physical sciences. Engineers have to experiment, analyse data and reach defensible conclusions about the outcomes of their experiments to determine how products behave when tested under real conditions. Work done on new products and processes may involve decisions that have to be made which can have a major economic impact on companies and their employees. Throughout industry, production and distribution processes must be organised and monitored to ensure maximum efficiency and reliability. One important branch of applied statistics is quality control. Quality control is an essential part of any production process which aims to ensure that high quality products are made, surely a principle aim of any practical engineer.

This Workbook is intended to give you an introduction to the subject and to enable you to understand in reasonable depth the meaning and interpretation of numerical and diagrammatic statements involving data. This first Section concentrates on the basic tabular and diagrammatic techniques for displaying data and the calculation of elementary statistics representing location and spread.

Prerequisites

Before starting this Section you should ...

- understand the ideas of sets and subsets (HELM 35.1)

Learning Outcomes

On completion you should be able to ...

- explain why statistics is important for engineers.
- explain what is meant by the term descriptive statistics
- calculate means, medians, modes and standard deviations
- draw a variety of statistical diagrams

1. Introduction to statistics

Many students taking degree courses involving the sciences and technology have to study statistics. This Workbook will enable you to understand the meaning and interpretation of numerical and diagrammatic statements involving data.

Many different definitions of statistics have been given. These include the following.

- "The collection, organisation, analysis and interpretation of numerical data"
- "The methodological basis of the natural sciences"
- "The study of uncertainty"

Why is this of any interest?
Consider the following two activities which we will describe as *experiments*. In statistics we use the word "experiment" to refer to an activity which involves recording outcomes, even when this does not take place in a laboratory.

1. We measure the breaking strength of ten 15-cm specimens of nylon cord.
2. We count the number of vehicles turning right at a road junction in a 30-minute period.

The first of these involves *measurement*. The second involves *counting*.
Now consider repeating the experiments. The results will most likely not be the same (cord specimen strengths will vary if measured very accurately; the numbers of vehicles turning right will probably differ). So the result of any one experiment will not be exactly representative of all samples of cord or all 30-minute periods. In this sense, no result is "true". Even if the result happened to be "exactly representative" we would have no way of knowing this.
These problems are almost universal features of investigating the world around us. We have:

1. Measurement error.
2. Differences between individuals. We (usually) can not observe all individuals and it is often more efficient not to do so anyway.
3. Differences between samples of material or time and we can not observe all possible samples.
4. Other variation which affects the results in an unpredictable way.

So what can we do?
Statistics deals with uncertainty. Its aims include the following:

1. To keep uncertainty to a minimum.
2. To quantify the remaining uncertainty.
3. To make precise statements in situations involving uncertainty.

These aims apply in all sorts of situations.

Variables

A *variable* is a quantity which can take different values, perhaps in different experiments. For example this could be the breaking strength of a nylon cord or the number of right-turning vehicles in a 30-minute period.
Variables may be:

Quantitative: i.e. numerical, e.g. breaking strengths of cords, numbers of vehicles.

Qualitative: usually categorical, e.g. make of a car, voting intention of a voter, cause of failure in a machine.

Quantitative variables may be

Continuous: i.e. measured on a continuous scale. E.g. height of a student, breaking strength of a cord.

Discrete: i.e. only certain separate values, e.g. integers, are possible. E.g. number of right-turning vehicles, goal difference of a football team.

When we do an experiment we observe a value taken by a variable. This value is called an *observation* (e.g. the breaking strength of a particular cord specimen). We may make several observations and collectively these are known as *data* (e.g. the breaking strengths of ten particular cord specimens).
Variables are often denoted by upper case letters such as X. Observations are often denoted by the corresponding lower case letters such as x.

Population and sample

Population: Suppose, for example, we are interested in individuals who belong to a certain group (e.g. students in the first year of UK Engineering degree courses or all TV sets made at a particular factory during October 2015). We call the group a *population*.

Sample: Often it is not possible to observe every member of a population and it is usually not sensible to try. Instead we observe only some members of the population. The members which we observe are called a *sample*. For example we might observe a sample of 15 cm long nylon cord specimens. Clearly it would not be sensible to try to test to destruction all possible 15 cm long nylon cord specimens. The nylon cord example illustrates the point that, in some cases, there are so many potential members of a population that the size of the population may be treated as infinite. We could always take more sample specimens and the result for one specimen does not affect the properties of any future specimens which we might take.

A large part of statistics is concerned with studying what we can learn about a population by observing a sample.

2. Introduction to descriptive statistics

In later Workbooks we will consider formal methods for making *inferences* about a population based on observation of a sample. Before doing this we will consider simple methods for describing and sumarising the observations in a sample, both numerically and graphically. These descriptive methods can be an important first step before we go on to apply more formal analyses. In fact errors can often be avoided by considering suitably chosen graphs before we go any further. In addition, clear presentation of numerical and graphical summaries of data is an important feature of reporting results. A *statistic* is a numerical summary of a sample of data. Later in this Workbook we will look at a number of statistics. First we will look at tabulating data and graphical representations.

3. Frequency tables

Rather than just present a list of the data it is useful to tabulate them in a way which is more informative. Often this is done by making a *frequency table*. Usually nowadays we use computer software to do such tasks but it is useful to understand the process involved and the meanings of various terms.

Categorical data

Suppose that we observe vehicles at a crossroads for thirty minutes. Consider just the vehicles arriving at the junction from the South. Such vehicles can carry straight on to go North, turn left to go West or turn right to go East. Suppose that we see 147 of these vehicles go North, 85 go West and 43 go East.

Frequency: The numbers 147, 85 and 43 are the *frequencies* of these three directions. In general, for categorical (i.e. qualitative) data, the frequency for a particular value of the variable is the number of times that value is observed in the sample.

Total frequency: The *total frequency* is the total number of observations. In this case it is $147 + 85 + 43 = 275$.

Relative frequency: The *relative frequency* is the proportion of the observations which fall in a particular category. It is the frequency divided by the total frequency. In the example the relative frequencies are $147/275 = 0.535$, $85/275 = 0.309$ and $43/275 = 0.156$. The total of the relative frequencies is 1.

Here is a frequency table for the road junction data.

Direction	Frequency	Relative frequency
North	147	0.535
West	85	0.309
East	43	0.156
Total	275	1.000

Note that, in some cases, it is necessary to include an "Other" category for unusual observations.

Discrete data

A frequency table for discrete data is very similar to one for categorical data except that we put the values of the variable in increasing order.
For example, the numbers of vehicles passing a point on a road in each of 100 intervals of length one minute are recorded. These data are summarised in the frequency table below.

Number of Vehicles x	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
0	0	0	0.00	0.00
1	0	0	0.00	0.00
2	0	0	0.00	0.00
3	4	4	0.04	0.04
4	3	7	0.03	0.07
5	5	12	0.05	0.12
6	8	20	0.08	0.20
7	10	30	0.10	0.30
8	19	49	0.19	0.49
9	12	61	0.12	0.61
10	13	74	0.13	0.74
11	7	81	0.07	0.81
12	4	85	0.04	0.85
13	11	96	0.11	0.96
14	0	96	0.00	0.96
15	1	97	0.01	0.97
16	2	99	0.02	0.99
17	0	99	0.00	0.99
18	1	100	0.01	1.00
$x > 18$	0	100	0.00	1.00
Total	100		1.00	

We have also included the *cumulative frequency* and the *cumulative relative frequency*.

Cumulative frequency: The *cumulative frequency* for a value x_i of a discrete variable X is the total frequency for values of X where $X \leq x_i$.

Cumulative relative frequency: The *cumulative relative frequency* is the cumulative frequency divided by the total frequency.

Notice that we have also included a category "$x > 18$" since there is not a definite upper limit to the number of vehicles which we might observe.

Continuous data

When we have continuous data it does not make sense to count how many times each value occurs. This is because, at least if we measured to a great enough precision, no two observations would be exactly the same. Therefore we need to form the observations into groups, called *classes*. We do this by dividing the range of possible values of the variable into a set of non-overlapping intervals called *class intervals*.

The following data are the heights (to the nearest tenth of a centimetre) of 30 students studying engineering statistics.

150.2	167.2	176.2
160.1	151.8	166.3
162.3	167.4	178.3
181.2	175.7	161.1
179.3	168.9	164.8
165.0	177.1	183.2
172.1	180.2	168.2
173.8	164.3	176.8
184.2	170.9	172.2
168.5	169.8	176.7

Notice first of all that all of the numbers lie in the range 150 cm. - 185 cm. This suggests that we try to organize the data into classes as shown below. This first attempt has deliberately taken *easy* class intervals which give a reasonable number of classes and span the numerical range covered by the data.

Class	Class Interval
1	150 - 155
2	155 - 160
3	160 - 165
4	165 - 170
5	170 - 175
6	175 - 180
7	180 - 185

The number of classes which we use will depend on the number of observations and the purpose of the frequency table.

Our definition of the classes above is ambiguous. For example, to which class should the number 165 be allocated? Should it be in the class interval 160-165 or in the class interval 165-170?

Rather than adopt an arbitrary convention such as always placing boundary values in the higher (or lower) class we usually define the class boundaries in such a way that such difficulties do not occur. Therefore we define the classes in a way which resolves this ambiguity. For example we can define the second class as $155 \leq x < 160$, where x is the value of the observation. Alternatively we can simply write the class interval as $[155, 160)$ where the square bracket indicates that the boundary value is included and the round bracket indicates that it is not. Our classes are now as follows.

Class	Class Interval
1	$[150,\ 155)$
2	$[155,\ 160)$
3	$[160,\ 165)$
4	$[165,\ 170)$
5	$[170,\ 175)$
6	$[175,\ 180)$
7	$[180,\ 185)$

Notice that the need to consider observations which lie on a class boundary only arises because, in reality, we only record values to a limited number of decimal places. Another method which is sometimes suggested is to define the class boundaries to one more decimal place than the observations.

When working by hand, the data can be organised into a frequency table using a *tally count*. To do a tally count, as you work through the data set, you make a mark, called a *tally mark*, next to the class to which an observation belongs and lightly mark or cross off the observation. When you have finished, the total number of tally marks should be equal to the number of observations. This process gives the tally marks and the corresponding frequencies as shown below.

Class Interval (cm)	Tally	Frequency
[150, 155)	11	2
[155, 160)		0
[160, 165)	11111	5
[165, 170)	11111111	8
[170, 175)	1111	4
[175, 180)	1111111	7
[180, 185)	1111	4

It is now easier to see some of the information contained in the original data set. For example, we now know that there are no observations in the class interval [155, 160) and that the class interval [165, 170) contains the most entries.

Understanding the information contained in the original table is now rather easier but, as we shall see, diagrams make the situation easier to visualise.

4. Graphical representations

Bar charts

Bar charts can be used to represent the frequencies for categorical or discrete data. In a bar chart, we draw a bar for each value of the variable. The length of the bar is proportional to the frequency for that value. Bars can be drawn vertically or horizontally.

Absenteeism can be a problem for some engineering firms. The following discrete data represent the number of days off taken by 50 employees of a small engineering company.

6	4	4	5	0	4	3	6	1	3
8	3	6	1	0	6	11	5	10	8
2	4	6	6	6	6	5	13	11	6
4	8	4	7	7	6	8	3	3	6
3	2	3	6	2	2	3	2	4	0

In order to construct a bar chart we can proceed as follows.

1. Construct a frequency table for the data.

2. Draw a diagram with the range (0 - 13) on one axis and the number of days corresponding to each value (number of days off) on the other. The length of each bar is proportional to the frequency (that is proportional to the number of staff taking that number of days off).

The results appear as follows:

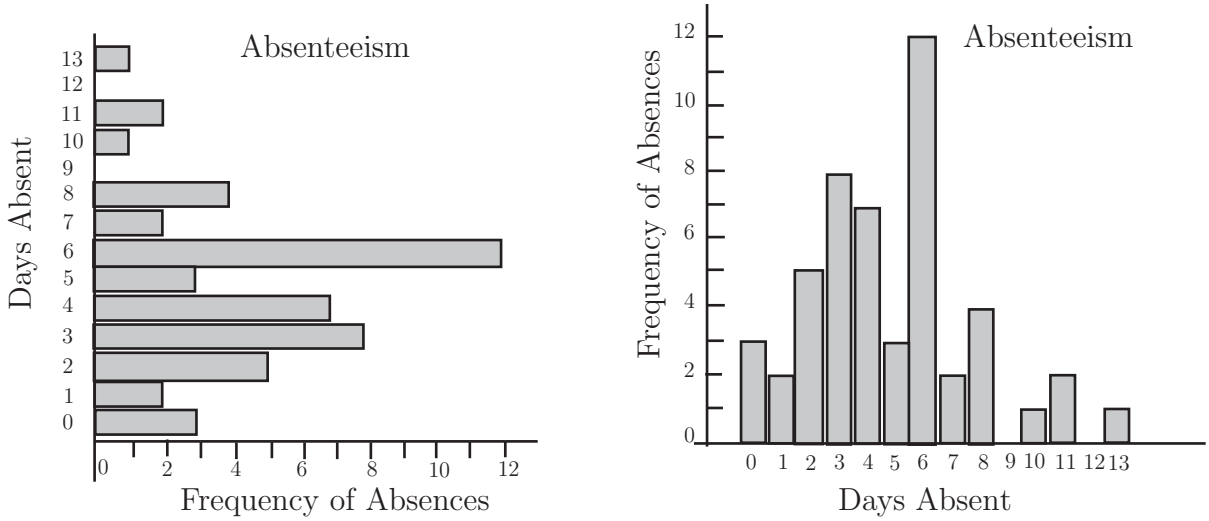


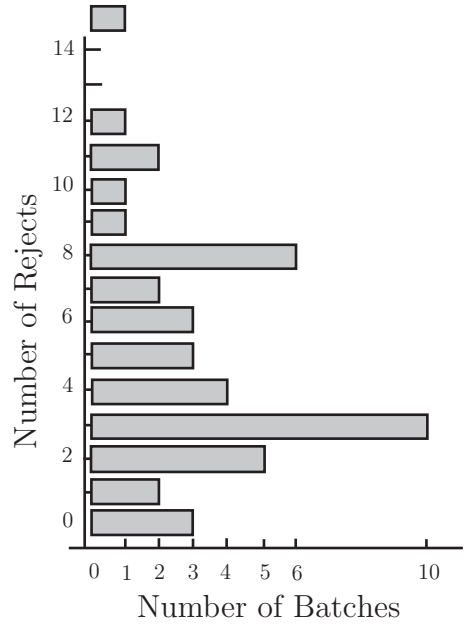
Figure 1

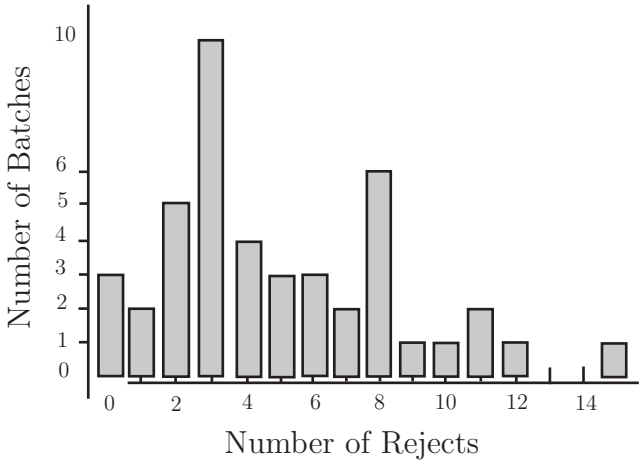
The following data give the number of rejects in fifty batches of engine components delivered to a motor manufacturer. Draw two bar charts representing the data, one with the bars vertical and one with the bars horizontal. Draw one chart manually and one using a suitable computer package.

2	3	5	6	8	1	2	0	3	4
8	3	6	1	0	6	11	5	10	8
3	5	9	12	3	8	5	11	15	3
4	8	4	7	7	6	8	3	3	6
3	2	3	6	2	2	3	2	4	0

Your solution

Answer





Pie charts

Pie charts are often seen in magazines and newspapers. A pie chart is simply a circular diagram where the circle is divided into sectors and the angles of the sectors are proportional to the quantity represented. The quantities may be the frequencies of values of a categorical variable in a sample. Since the total area of the circle is fixed, pie charts are considered to be useful for representing proportions of a total.

The following data represent the time spent weekly on a variety of activities by the full-time employees of an engineering company.

Hours spent on:	Males	Females
Travel to and from work	10.5	8.4
Paid activities in employment	47.0	37.0
Personal sport and leisure activities	8.2	3.6
Personal development	5.6	6.4
Family activities	8.4	18.2
Sleep	56.0	56.0
Other	32.3	38.4

To construct a pie chart showing how the male employees spend their time we proceed as follows. Note that the total number of hours spent is 168 (7×24).

1. Express the time spent on any given activity as a proportion of the total time spent;

2. Multiply the number obtained by 360 thus converting the proportion to an angle, in degrees;

3. Draw a chart consisting of (in this case) 6 sectors having the angles given in the table below subtended at the centre of the circle.

Hours spent on:	Males	Proportion of Time	Sector Angle
Travel to and from work	10.5	$\frac{10.5}{168}$	$\frac{10.5}{168} \times 360 = 22.5$
Paid activities in employment	47.0	$\frac{47}{168}$	$\frac{47}{168} \times 360 = 100.7$
Personal sport and leisure activities	8.2	$\frac{8.2}{168}$	$\frac{8.2}{168} \times 360 = 17.6$
Personal development	5.6	$\frac{5.6}{168}$	$\frac{5.6}{168} \times 360 = 12.0$
Family activities	8.4	$\frac{8.4}{168}$	$\frac{8.4}{168} \times 360 = 18.0$
Sleep	56.0	$\frac{56}{168}$	$\frac{56}{168} \times 360 = 120.0$
Other	32.3	$\frac{32.3}{168}$	$\frac{32.3}{168} \times 360 = 69.2$

The pie chart obtained is illustrated below.

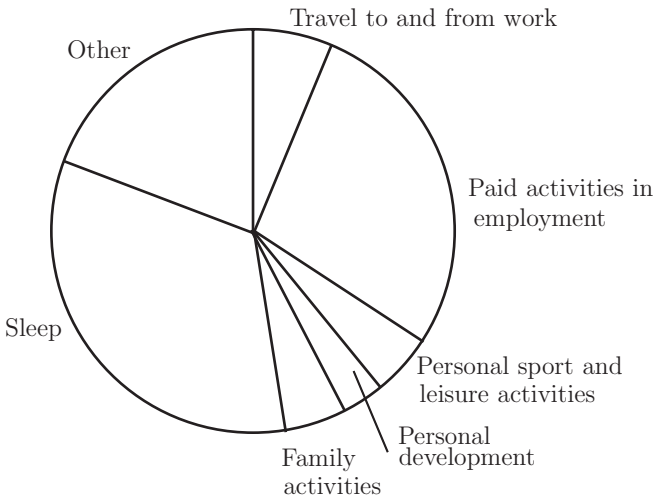


Figure 2

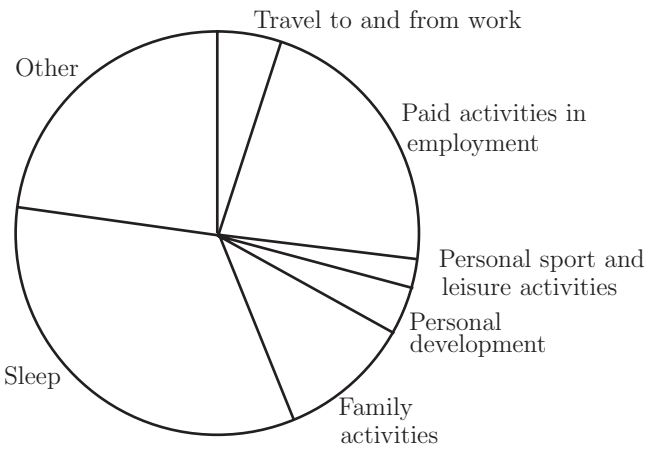
Task

Construct a pie chart for the female employees of the company and use both it and the pie chart in Figure 3 to comment on any differences between male and female employees that are illustrated.

Your solution

Answer

Hours spent on:	Females	Proportion of Time	Sector Angle
Travel to and from work	8.4	$\frac{8.4}{168}$	$\frac{8.4}{168} \times 360 = 18.0$
Paid activities in employment	37.0	$\frac{37}{168}$	$\frac{37}{168} \times 360 = 79.3$
Personal sport and leisure activities	3.6	$\frac{3.6}{168}$	$\frac{3.6}{168} \times 360 = 7.7$
Personal development	6.4	$\frac{6.4}{168}$	$\frac{6.4}{168} \times 360 = 13.7$
Family activities	18.2	$\frac{18.2}{168}$	$\frac{18.2}{168} \times 360 = 39.0$
Sleep	56.0	$\frac{56}{168}$	$\frac{56}{168} \times 360 = 120.0$
Other	38.4	$\frac{38.4}{168}$	$\frac{38.4}{168} \times 360 = 82.3$



Comments: Proportionally less time spent travelling, more on family activities etc.

Histograms

We use a *histogram* to represent the *frequency distribution* of a sample of continuous data. In a histogram, like a bar chart, we draw an element of the diagram to represent each frequency, in this case the frequency for each class. However, because a continuous variable is measured on a continuous scale, we do not have gaps like the gaps between the bars in a bar chart. Instead we draw *columns*, or *blocks*. The base of a column is the class interval on the x-axis.

Strictly speaking, the *areas* of the blocks forming the histogram represent the frequencies since this gives the histogram the necessary flexibility to deal with frequency tables whose class intervals are not of equal width. However, when the class intervals are of equal width, the heights of the columns are proportional to the frequencies.

Sometimes the approximate shape of the distribution of data is indicated by a **frequency polygon** which is formed by joining the mid-points of the tops of the blocks forming the histogram with straight lines.
A histogram, with a frequency polygon, is shown in Figure 3.

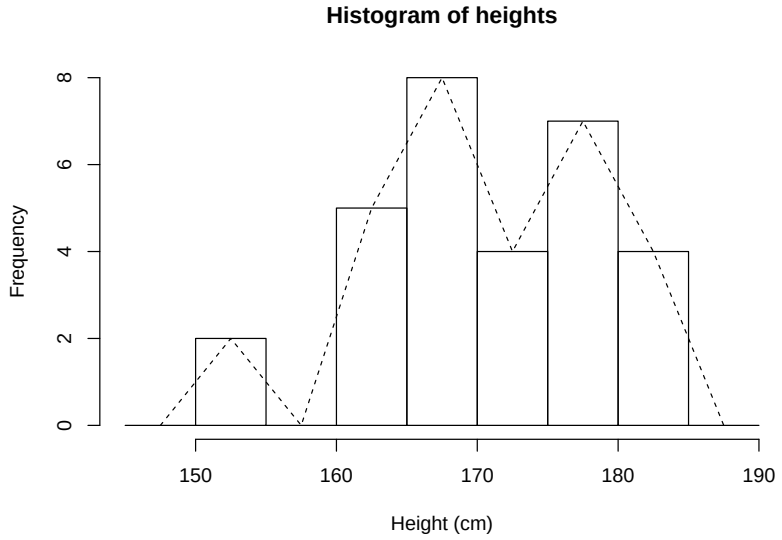


Figure 3

The following data are the heights (to the nearest tenth of a centimetre) of a second sample of 30 students studying engineering statistics.

Organise the data into classes using class intervals $[145,\ 150)$, $[150,\ 155), \ldots, [185,\ 190)$ and construct a frequency table of the data. Use your table to represent the data diagrammatically using a histogram. Hint:- All the data values lie in the range 145-190.

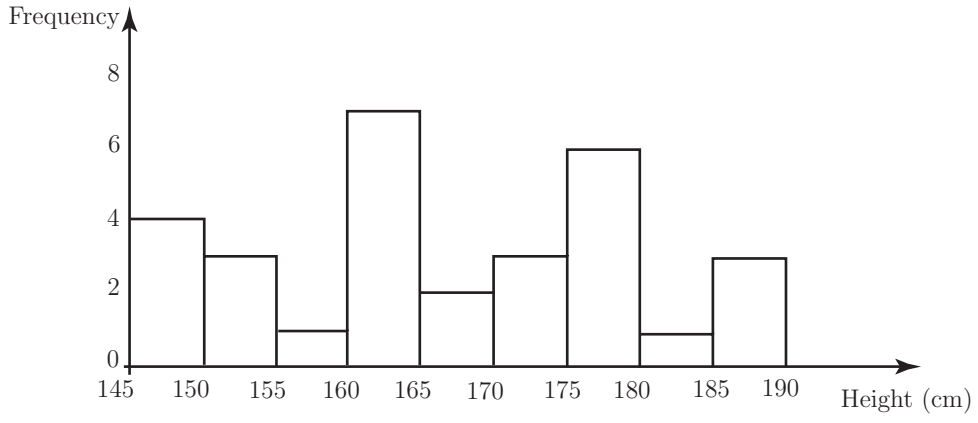
155.3	177.3	146.2	163.1	161.8	146.3	167.9	165.4	172.3	188.2
178.8	151.1	189.4	164.9	174.8	160.2	187.1	163.2	147.1	182.2
178.2	172.8	164.4	177.8	154.6	154.9	176.3	148.5	161.8	178.4

Your solution

Answer

Class Interval (cm)	Tally	Frequency
[145, 150)	1111	4
[150, 155)	111	3
[155, 160)	1	1
[160, 165)	11111 11	7
[165, 170)	11	2
[170, 175)	111	3
[175, 180)	11111 1	6
[180, 185)	1	1
[185, 190)	111	3

The histogram is shown below.



5. Location and spread

The statistical properties of the observations in a sample are often described as a *frequency distribution*, referring to the fact that some values, in the case of discrete or categorical variables, or some ranges of values, in the case of continuous variables, may have greater frequencies than others. Very often the distribution of observations in a sample is summarized using two statistics, one of which measures the *location* or central value of the distribution and one which measures the *dispersion* or spread of the values about this central value. As we shall see later, it is often important to convey information on the *shape* of a distribution as well. However, in many cases in practice, the shape of the distribution, as represented, for example, by a histogram, is reasonably symmetrical and roughly follows the" bell-shape" illustrated below. In such cases the location-and-spread summary may suffice.

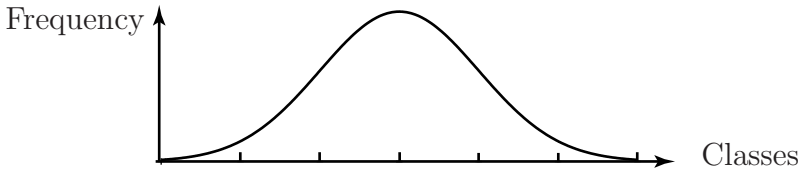


Figure 4

Notation

It is important to remember the distinction between a *population* and a *sample*. In later Workbooks we will look at how sample statistics can be used to give us information about summaries of a population. For now, however, we are concentrating on the sample statistics. To help keep the distinction clear, we usually use Greek letters, such as μ or σ, to represent population quantities, the values of which are, of course, usually unknown, and the equivalent Roman letters, such as m or s, to represent sample statistics.

Measures of location

There are three widely used measures of location, these are:

- The Mean, the arithmetic average of the data;
- The Median, the central value of the data;
- The Mode, the most frequently occurring value in the data set.

Mean

This section of the booklet will show you how to calculate the mean. Strictly, we should say that this is the *arithmetic mean*. There are other kinds of means as well. This is by far the most common, though, and so we often just say "mean".

Key Point 1

If we take a set of numbers $x_1, x_2, \ldots, x_n$, its arithmetic mean is defined as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This is often shortened to:

$$\bar{x} = \frac{1}{n}\sum x$$

In words, this formula says

sum the values of x and divide by the number of numbers you have summed.

In the case of discrete data, we have a frequency for each possible value of the variable and we can express the calculation in terms of these frequencies. Suppose that the possible values are $x_{(1)}, x_{(2)}, \ldots, x_{(J)}$ (where we could have $J \to \infty$) and that these values have observed frequencies $f_1, f_2, \ldots, f_J$. Then it is easily seen that

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{J} f_j x_{(j)}.$$

A common case is when the possible values are the nonnegative integers, $0, 1, 2, \ldots$. Then

$$\bar{x} = \frac{1}{n}\sum_{k=0}^{\infty} k f_k = \frac{\sum k f_k}{\sum f_k}.$$

Example

The numbers of telephone calls made on a particular day by 500 subscribers to a mobile telephone network gave the following frequency table, from which we can calculate the sample mean.

Number of Calls x	Frequency f	fx
0	13	0
1	30	30
2	85	170
3	98	294
4	96	384
5	81	405
6	42	252
7	32	224
8	17	136
9	4	36
10	1	10
11	1	11
	$\sum f = 500$	$\sum fx = 1952$

The sample mean is given by $\bar{x} = \dfrac{\sum fx}{\sum f} = \dfrac{1952}{500} = 3.904$

The following data set gives the number of thread-breaks in each of 50 one-hour period in a spinning machine. Use these data to form a frequency table and calculate the mean of the data.

0	0	1	1	0	0	1	0	2	2
3	0	0	3	2	1	1	0	3	2
0	1	0	1	3	1	2	0	1	0
0	0	1	3	2	0	0	1	1	2
0	1	0	1	0	0	2	1	0	1

Your solution

Answer

Observation x	Frequency f	fx
0	21	0
1	16	16
2	8	16
3	5	15
Total	50	47

Mean $= 47/50 = 0.94$

Median and Quartiles

Some statistics are derived from quantitative data which are placed in **rank order**. Ranking data simply means that the data are placed in order from the smallest to the largest. The smallest observation thus gets rank 1, the second smallest gets rank 2 and so on. Such statistics are called *order statistics*. The median is an order statistic which is used as a measure of location. It divides a

sample of data into two equally sized groups, one containing the smaller observations and the other containing the larger observations. Two other order statistics which we will describe here together with the median are the *Lower Quartile* and the *Upper Quartile*. The lower quartile, median and upper quartile are also called the first, second and third quartiles respectively. They divide a sample of data into four portions according to the sizes of the observations, each portion containing the same number of observations. As you will see the *Ogive* or *Cumulative Frequency Curve* enables us to find these statistics for large data sets. Definitions for the three sample quartiles are given below. A number of slightly different definitions are used for the lower and upper quartile. One version is given here. The different definitions will usually lead to only slightly different results.

The Median; this is the central value of a sample where the numbers of larger observations and smaller observations are equal.

The Lower Quartile; this is the value which divides the observations into a lower 25% and an upper 75%.

The Upper Quartile; this is the value which divides the observations into a lower 75% and an upper 25%.

We calculate the median and the lower and upper quartiles as follows. Let the number of observations be n. We sort the observations into increasing order and allocate ranks. If there are ties then we just use integer ranks as usual. For example, the data 3.1, 4.6, 5.3, 5.3, 7.8, 9.5 would get ranks 1, 2, 3, 4, 5 so that the observations with ranks 3 and 4 are both 5.3.

- For the **median**, calculate $r_2 = \dfrac{n+1}{2}$. If n is odd r_2 is an integer and the sample median is the observation with rank r_2. If n is even then r_2 is not an integer and we average the values on either side. That is we average the observations with ranks $\dfrac{n}{2}$ and $\dfrac{n+2}{2}$. For example, if $n = 16$, then $r_2 = 8.5$ and we average the values with ranks 8 and 9.

- For the **lower quartile**, calculate $r_1 = \dfrac{n+2}{4}$. If r_1 is an integer, the sample lower quartile is the observation with rank r_1. If r_1 is not an integer then we average the values on either side. For example, if $n = 16$, then $r_1 = 4.5$ and we average the values with ranks 4 and 5.

- For the **upper quartile**, calculate $r_3 = \dfrac{3n+2}{4}$. If r_3 is an integer, the sample lower quartile is the observation with rank r_3. If r_3 is not an integer then we average the values on either side. For example, if $n = 16$, then $r_3 = 12.5$ and we average the values with ranks 12 and 13.

Note that $r_2 = \dfrac{2n+2}{4}$ so the three r-values are

$$r_1 = \frac{n+2}{4},\ r_2 = \frac{2n+2}{4} \text{ and } r_3 = \frac{3n+2}{4}.$$

For the simple data set 1.2, 3.0, 2.5, 5.1, 3.5, 4.1, 3.1, 2.4 the process is illustrated by placing the members of the data set in rank order:

1.2, 2.4, 2.5, 3.0, 3.1, 3.5, 4.1, 5.1

Here $n = 8$ so

- $r_1 = 2.5$ and the lower quartile is $\dfrac{2.4 + 2.5}{2} = 2.45$.

- $r_2 = 4.5$ and the median is $\dfrac{3.0 + 3.1}{2} = 3.05$
- $r_3 = 6.5$ and the upper quartile is $\dfrac{3.5 + 4.1}{2} = 3.8.$

Find the median, lower quartile and upper quartile for the data set:

0.7, 1.2, 2.4, 2.5, 3.0, 3.1, 3.5, 4.1, 5.0

Here
$n = 9$ so $r_2 = 5$ and the median is the observation with rank 5.
Median $= 3.0$

$r_1 = 2.75$ and the lower quartile is the average of the observations with ranks 2 and 3: $\dfrac{1.2 + 2.4}{2} = 1.8.$

$r_3 = 7.25$ and the upper quartile is the average of the observations with ranks 7 and 8: $\dfrac{3.5 + 4.1}{2} = 3.8.$

In the case of larger samples the quantities can be approximated by using a **cumulative frequency curve** or **ogive**.

The cumulative frequency distribution for the heights of the 30 students given earlier (on page 7) is shown below. Notice that here, the class intervals are defined in such a way that the frequencies accumulate (hence the term *cumulative frequency*) as the table is built up.

Height	Cumulative Frequency	Cumulative Relative Frequency
less than 150	0	0.000
less than 155	2	0.067
less than 160	2	0.067
less than 165	7	0.200
less than 170	15	0.467
less than 175	19	0.633
less than 180	26	0.867
less than 185	30	1.000

To plot the ogive or cumulative frequency curve, we plot the heights on the horizontal axis and the cumulative frequencies or cumulative relative frequencies on the vertical axis. The corresponding ogive is shown in Figure 5.

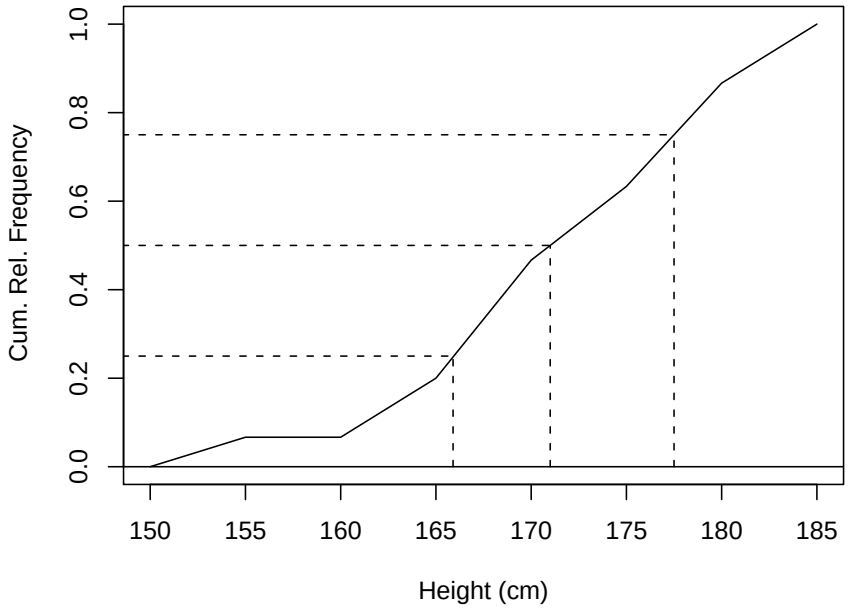


Figure 5

The three statistics defined above can be read from the diagram as indicated. They are the values where the cumulative relative frequency is 0.25, 0.50 and 0.75. With $n = 30$, this corresponds to cumulative frequencies of 7.5, 15.0 and 22.5. For the data set giving the heights of the 30 students, the three statistics, as obtained using the ogive, are as follows.

1. **The Lower Quartile** is 165.9.
2. **The Median** is 171.0.
3. **The Upper Quartile** is 177.5.

The Mode

For discrete data, the mode is the value with the greatest frequency.

Example

In a quality control system, batches of twenty mass-produced items are examined and the number which fall outside an acceptable range in each batch is counted. The numbers of out-of-range items in fifty batches were as follows.

Number out of range (x)	0	1	2	3	4	5	6
Frequency (f)	9	9	13	9	8	1	1

The mode here is 2 since this value has the greatest frequency.

Note that the mean for these data is 2.1 and the median is 2.

In the case of continuous data, we do not look for a most frequent value. Strictly speaking, no two observations should be exactly the same. Instead we form a frequency table and identify the *modal class*, the class with the greatest frequency.

Comparison

The mean, median and mode all have their different uses. If the histogram of a sample of data is more-or-less symmetric then the median and the mean will be very similar. If it is also *unimodal*, that is there is only one maximum in the histogram and, since the histogram is symmetric, this is in the middle, then the mode will also be very similar. The mode will, of course be the location of the maximum in the histogram. If a frequency distribution is not symmetric (we say it is *skewed* in this case) then the three values can be quite different. For example, consider the distribution of incomes among the citizens of a country. Typically, the very large incomes of a small proportion of people increase the mean so that the mean is greater than the incomes of most people. Therefore the median is less than the mean but might be a more relevant measure for most people. The mode is smaller still because of the large numbers of people with relatively small incomes.

Measures of spread

As well as summarising the *location* of the distribution of values in a sample using, for example, the mean, we usually also need to describe how close to this central value or far away from it the individual observations tend to be. We need to summarise the *spread* of the data.
The two data sets below have the same mean of 7 but clearly have different spreads about the mean.

Data set A: 5, 6, 7, 8, 9
Data set B: 1, 2, 7, 12, 13

There are several ways in which one can measure the spread of a distribution about a mean, for example

- the **range** - the difference between the greatest and least values;
- the **inter-quartile range** or **IQR** - the difference between the upper and lower quartiles;

Each of these measures has advantages and problems associated with it.

Measure of Spread	**Advantages**	**Disadvantages**
Range	Easy to calculate	Depends on two extreme values and does not take into account any intermediate values
Inter-Quartile Range	Is not as susceptible to the influence of extreme values.	Measures only the central 50% of a distribution.

Variance and standard deviation

By far the most common measures of the spread of a sample are the sample **standard deviation** and the sample **variance**. The standard deviation is just the square root of the variance so we will describe them together.

The idea of the variance is that it is the average *squared* deviation of an observation from the sample mean. The deviations are squared so that the negative deviations, of observations less than the mean, do not cancel out the positive deviations, of observations greater than the mean.

Suppose that we have observations $x_1, \ldots, x_n$. Then the *sample variance* is defined as follows.

Key Point 2

Sample Variance

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

where s^2 is the sample variance and $\bar{x}$ is the mean of the data in the sample of size n.

Notice that we divide by $n-1$ rather than n. This is because we are using the deviations $x_i - \bar{x}$ from the sample mean, $\bar{x}$, and these necessarily sum to zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0.$$

Therefore, if we knew the values of $(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_{n-1} - \bar{x})$, we would be able to work out the value of $(x_n - \bar{x})$. We say that there are only $n-1$ **degrees of freedom**. As we shall see in Workbook 40, it also turns out that, when we want to use the sample variance to inform us about the value of the *population variance*, σ^2, dividing by $n-1$ instead of n is advantageous.

The sample **standard deviation** is simply the square root of the sample variance. It is therefore the root mean square deviation of the observations from the sample mean.

Key Point 3

Sample Standard Deviation

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

The standard deviation is often preferred to the variance as a summary because the standard deviation is in the same units as the data. For example, if we have a sample of weights, in kg, then the standard deviation is also in kg, whereas the variance is in kg^2.
Consider the two data sets A and B given below. In each case the sample size is $n = 10$ and the sample mean is $\bar{x} = 7$ but the spreads are different.

DATA SET *A*

x	$x - \bar{x}$	$(x - \bar{x})^2$
4	-3	9
5	-2	4
5	-2	4
6	-1	1
7	0	0
7	0	0
8	1	1
9	2	4
9	2	4
10	3	9
		$\sum(x - \bar{x})^2 = 36$

DATA SET *B*

x	$x - \bar{x}$	$(x - \bar{x})^2$
1	-6	36
3	-4	16
3	-4	16
5	-2	4
7	0	0
7	0	0
9	2	4
11	4	16
11	4	16
13	6	36
		$\sum(x - \bar{x})^2 = 144$

For Data Set A, the sample variance is $36/9 = 4$ and the sample standard deviation is $\sqrt{4} = 2$.
For Data Set B, the sample variance is $144/9 = 16$ and the sample standard deviation is $\sqrt{16} = 4$.
In Data Set B the deviations from the sample mean are exactly twice the size of those in Data Set A and consequently, the sample standard deviation is also twice as big.

Calculate the sample variance and sample standard deviation of the data set: 2, 4, 4, 5, 6, 6, 7, 7, 9, 10

Your solution

Answer

Data x	$x-\bar{x}$	$(x-\bar{x})^2$
2	-4	16
4	-2	4
4	-2	4
5	-1	1
6	0	0
6	0	0
7	1	1
7	1	1
9	3	9
10	4	16
$\sum x = 60$	$\sum(x-\bar{x}) = 0$	$\sum(x-\bar{x})^2 = 52$

sample size $n = 10$

sample mean $\bar{x} = 60/10 = 6$

sample variance $s^2 = 52/9 = 5.7778$

sample standard deviation $s = \sqrt{52/9} = 2.4037$

Calculations

To calculate either the sample variance or the sample standard deviation, we need $\sum(x_i - \bar{x})^2$. We can calculate this by subtracting the sample mean, $\bar{x}$, from each observation and squaring the result then summing these. However we have the following result, using Key Point 1:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\end{aligned}$$

It might be more convenient to use the form

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

or

$$\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} x_i\right]^2$$

both of which are equal to $\sum(x_i - \bar{x})^2$.

In particular, for discrete data, where each value $x_{(j)}$ of the variable has a frequency f_j in the sample, we can replace $\sum x_i$ with $\sum f_j x_{(j)}$ and replace $\sum x_i^2$ with $\sum f_j x_{(j)}^2$. For example, for the road traffic data given on page 6, we have

Number of Vehicles x	Frequency f	fx	fx^2
3	4	12	36
4	3	12	48
5	5	25	125
6	8	48	288
7	10	70	490
8	19	152	1216
9	12	108	972
10	13	130	1300
11	7	77	847
12	4	48	576
13	11	143	1859
14	0	0	0
15	1	15	225
16	2	32	512
17	0	0	0
18	1	18	324
Total	100	890	8818

Therefore the sample mean is

$$\bar{x} = \frac{890}{100} = 8.9$$

and

$$\sum (x - \bar{x})^2 = 8818 - 100 \times 8.9^2 = 897.0$$

or

$$\sum (x - \bar{x})^2 = 8818 - \frac{1}{100} \times 890^2 = 897.0.$$

So the sample variance is

$$s^2 = \frac{897}{99} = 9.0606$$

and the sample standard deviation is

$$s = \sqrt{9.0606} = 3.010.$$

Further comments

The sample mean and sample standard deviation are often useful to provide a concise summary of an observed frequency distribution. This is particularly the case when the frequency distribution is *symmetric* and *unimodal*. A *unimodal* frequency distribution is one where the histogram, at least approximately, has a shape with a single peak. If the frequencies decrease in the same way on each side of the peak, we say that the distribution is *symmetric*. The regions on either side of the distribution where the frequencies die out are called the left and right *tails* of the distribution. In a symmetric distribution, the left and right tails are like mirror images of each other.

Symmetric unimodal distributions are common but there are exceptions. The distributions of some variables tend to be asymmetric, with one tail longer than the other. For example lifetime distributions, that is the distributions of the times to failure for pieces of equipment, tend to be asymmetric, with the frequencies decreasing more slowly to the right of the mode. This is called *positive skew*, which means that the longer tail is on the right. In such cases it may be more useful to use the median and lower and upper quartiles to describe the distribution. Figure 6 shows a histogram of a distribution with a positive skew.

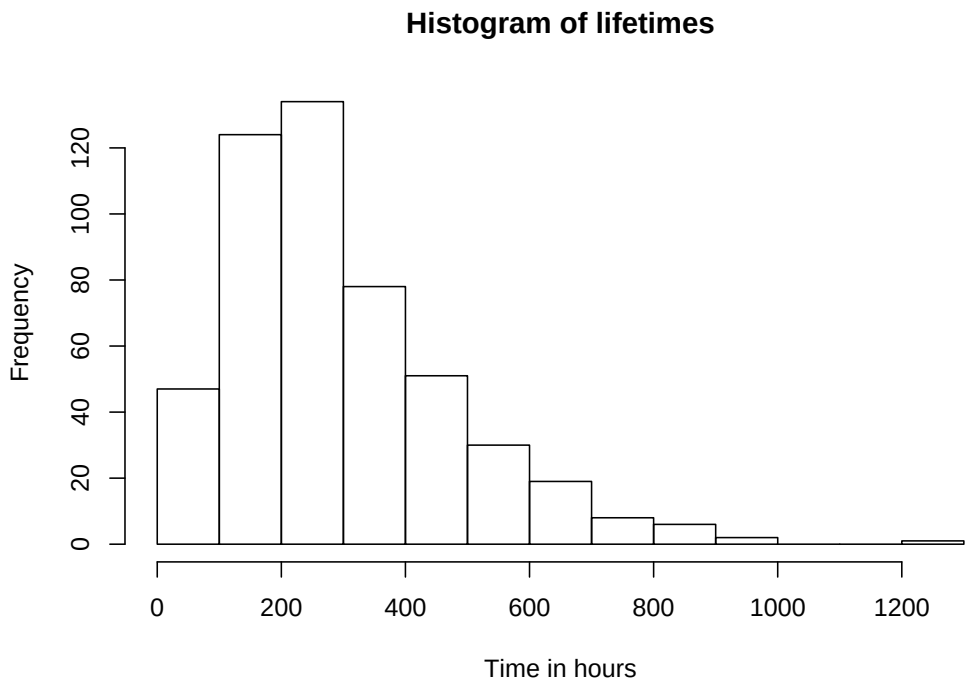


Figure 6

For the following data set of student heights, calculate the mean, variance and standard deviation of the data.

155.3	177.3	146.2	163.1	161.8	146.3	167.9	165.4	172.3
188.2	178.8	151.1	189.4	164.9	174.8	160.2	187.1	163.2
147.1	182.2	178.2	172.8	164.4	177.8	154.6	154.9	176.3
148.5	161.8	178.4						

Your solution

$$\sum x = 5010.3, \qquad \sum x^2 = 841616.65.$$

Hence the sample mean is

$$\bar{x} = \frac{5010.3}{30} = 167.01$$

We have

$$\sum (x - \bar{x})^2 = 841616.65 - 30 \times 167.01^2 = 4846.447$$

so the sample variance is

$$s^2 = \frac{4846.447}{29} = 167.119$$

and the sample standard deviation is

$$s = \sqrt{s^2} = 12.927$$

Exercises

1. Find (a) the mean and standard deviation, (b) the median and inter-quartile range, of the following data set which represents the numbers of days between consecutive failures in a machine (sorted into increasing order):

   $$3,\ 10,\ 11,\ 12,\ 13,\ 13,\ 18,\ 25,\ 30,\ 43,\ 49,\ 52,\ 52,\ 57,\ 67,\ 75$$

   Would you say that either summary set is preferable to the other?

   If the number 75 is replaced by the number 150 so that the data set becomes

   $$3,\ 10,\ 11,\ 12,\ 13,\ 13,\ 18,\ 25,\ 30,\ 43,\ 49,\ 52,\ 52,\ 57,\ 67,\ 150$$

   calculate the same statistics again and comment on which set you would use to summarise the data.

2. The following data give the abrasion losses, in g per hour, from fifty samples in an experiment investigating the resistance to abrasion of synthetic rubber.

165	133	89	174	75	109	277	106	145	93	154	95	75	114	192
525	580	115	116	193	211	117	198	139	274	117	158	219	126	181
194	158	343	85	159	69	176	272	57	169	204	145	248	38	89
146	157	201	96	156										

   (a) Organise the data into a frequency table using the class intervals

   $$[0,\ 100),\ \ [100,\ 200),\ \ [200,\ 300),\ \ [300,\ 400),\ \ [400,\ 500),\ \ [500,\ 600)$$

   and draw a histogram representing the data.

   Comment on the histogram.

   (b) Calculate the sample mean and sample standard deviation.

3. A lecturer gives a science test to two classes and calculates the results as follows:

   Class A - average mark 36% Class B - average mark 40%

   The lecturer reports to her Head of Department that the average mark over the two classes **must** be 38%. The Head of Department disagrees, who is right?

   Do you need any additional information, if so what, to make a decision as to who is right?

Answers

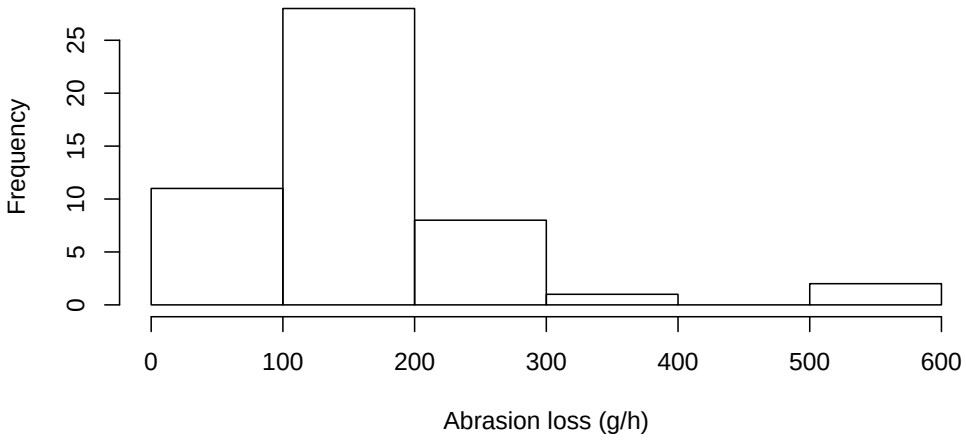
1. Mean = 33.125, standard deviation = 23.131, median=27.5, lower quartile = 12.5, upper quartile = 52 so the inter-quartile range is 39.5. The data are somewhat asymmetric and the mean is a little larger than the median. The median and IQR might represent the bulk of the data more effectively in this case.

   For the second set of data the median and quartiles are unchanged but the mean becomes 37.813 and the standard deviation becomes 36.130. The mean and standard deviation have been markedly increased. Here the median and inter-quartile range are preferable to the mean and standard deviation - they represent the bulk of the data much more realistically.

2. (a) Frequency table:

Abrasion loss x	Frequency	Abrasion loss x	Frequency
$0 \leq x < 100$	11	$300 \leq x < 400$	1
$100 \leq x < 200$	28	$400 \leq x < 500$	0
$200 \leq x < 300$	8	$500 \leq x < 600$	2

   Histogram:



   The distribution is rather asymmetric, with a longer tail to the right.

   (b) The sample mean is 168.54 g/h.
   The sample standard deviation is 100.70 g/h.

3. The Head of Department is right. The lecturer is only correct if both classes have the same number of students. Example: if class A has 20 students and class B has 60 students, the average mark will be: $(20 \times 36 + 60 \times 40)/(20 + 60) = 39\%$.

Exploring Data

36.2

Introduction

Techniques for exploring data to enable valid conclusions to be drawn are described in this Section.

The diagrammatic methods of stem-and-leaf and box-and-whisker are given prominence.

You will also learn how to summarize data using sets of statistics which have meaning in cases where a data set is not symmetrical. You should note that statistics such as the mean and variance are of limited use in such situations. Finally, you will encounter outliers. These are values which lie outside the main body of the data set and which may be treated separately as exceptional cases.

Prerequisites

Before starting this Section you should ...

- understand the ideas of sets and subsets (HELM 35.1)

Learning Outcomes

On completion you should be able to ...

- undertake Exploratory Data Analysis (EDA)
- construct stem-and-leaf diagrams and box-and-whisker plots
- explain the significance of outliers, skewness, gaps and multiple peaks

1. Exploratory data analysis

Introduction

The title 'Exploratory Data Analysis' (EDA) is usually taken to mean the activity by which a data set is explored and organized in order that information it contains is made clear. The usefulness of the summary statistics used in this branch of statistics does not depend on the frequency distribution having a standard shape. The techniques used in EDA were first developed by the statistician John Tukey and for details of EDA which are beyond this open learning booklet, you are referred to the text Exploratory Data Analysis, by J.W. Tukey, Addison-Wesley, 1977. Tukey's techniques have been used in innumerable papers and books since that date.

In many branches of statistics, it is often assumed that frequency distributions will, at least approximately, follow a standard shape known as a *normal distribution*. When a distribution does not have this shape this is sometimes called a *departure from normality*. Statistical methods which do not strongly depend on the assumption of normality are said to be *resistant to departures from normality*. We will see more about the normal distribution in Workbook 39.

The basics of EDA

The basic principles followed in EDA are:

- To measure the location and spread of a distribution we use statistics which are resistant to departures from normality;
- To summarise shape location and spread we use several statistics rather than just two;
- Visual displays as well as numerical displays are used to summarise information obtained about shape, location and spread.

You can see these principles illustrated below. Often, the location and spread of a distribution are measured by calculating its mean and standard deviation. The problem with these statistics is that they are sensitive to the influence of extreme values. For example, the data set

1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6

has mean $\bar{x} = 3.5$ and standard deviation $s = 1.45$. These values are quite acceptable since the distribution is symmetrical about its mean of 3.5. The symmetry is easily seen simply by inspecting the data although the bar chart below might make the symmetry more obvious (Figure 7).

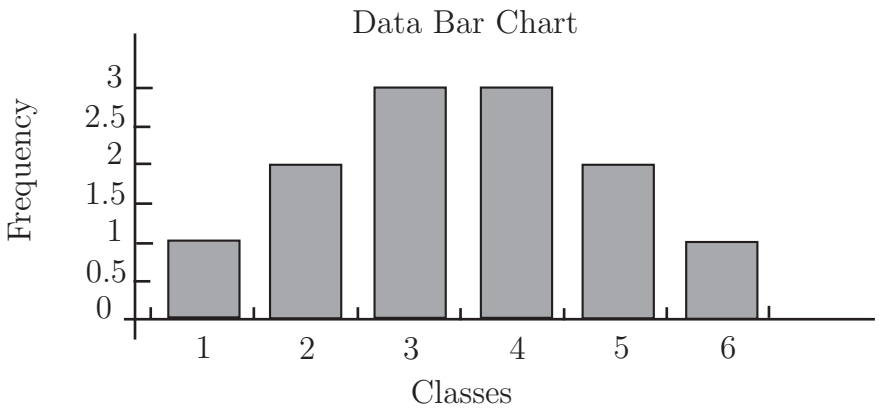


Figure 7

The shape of the distribution may also be shown by the **stem-and-leaf** diagram below. Notice that the *stem* consists of the numbers 1 to 6 and the *leaves* are just the members of each class.

1	1
2	2 2
3	3 3 3
4	4 4 4
5	5 5
6	6

Figure 8

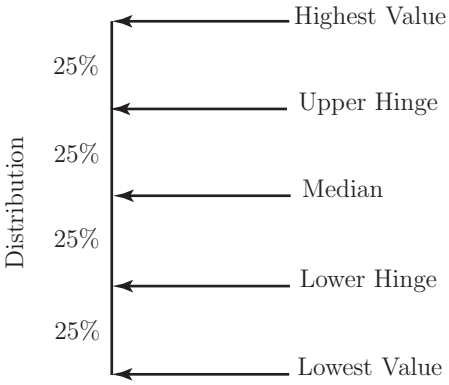
You will study the stem-and-leaf diagram in more detail later in this Workbook.

The effects of changes in extreme values are easily illustrated by looking at what happens if we take the last number to be 60 instead of 6. This destroys the symmetry of the distribution and gives mean $\bar{x} = 8$ and standard deviation $s = 16.42$. Clearly, these values do not describe the distribution very well at all. A mean which is higher than 92% of the members of the distribution can hardly be described as representative!

The simplest and most common examples of **resistant statistics** are those based on the idea of rank order - we simply order a distribution starting at the smallest value and ending at the largest value so the smallest observation gets rank 1, the second smallest gets rank 2 and so on.

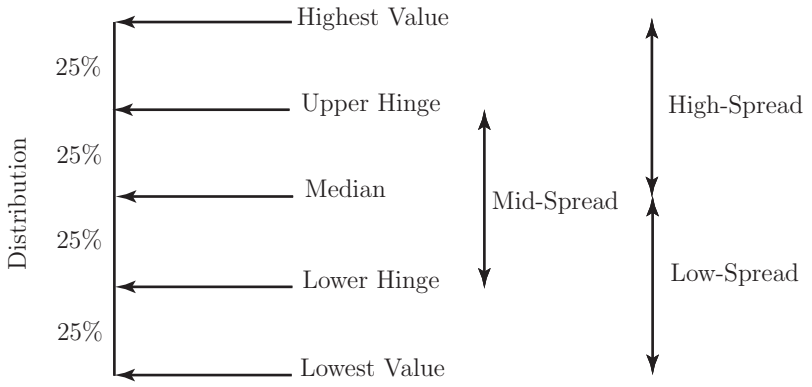
Key Point 4

The five essential statistics based on rank order are illustrated in the diagram below:



Key Point 5

Using the values in Key Point 4 other statistics which represent the shape or spread of the distribution may be defined. These statistics are known as the Mid-Spread, High-Spread and Low-Spread and their definition is indicated in the diagram below.



Elementary EDA recommends the use of a **five-number summary** consisting of:

1. the lowest value;
2. the lower hinge;
3. the median;
4. the upper hinge;
5. the highest value.

to summarize a distribution. The lower and upper *hinges* are the lower and upper sample quartiles. Thus the mid-spread is the inter-quartile range (IQR). You will find that the five-number summary, especially when used in conjunction with the three spreads shown in Key Point 5 above gives an adequate representation of a non-symmetrical distribution.

Notice that:

- the spreads shown in Key Point 5 are easily calculated once the five-number summary is known;
- the median and the hinges are unaffected by changes in extreme values.

Find the five number summary and the mid-spread, high-spread and low-spread for the distribution given below.

1 9 17 2 9 17 3 10 18 3 11 19 4 12 19
5 12 20 6 13 21 6 13 22 7 14 23 8 16 27

Your solution

Answer

1 Lowest Value = 1
2
3
3
4
5
6
6 Lower Hinge = 6 Low-Spread = 11
7
8
9
9
10
11
12 Median = 12 Mid-Spread = 12
12
13
13
14
16
17
17
18 Upper Hinge = 18 High-Spread = 15
19
19
20
21
22
23
27 Highest Value = 27

The stem-and-leaf diagram

You have already seen a basic stem-and-leaf diagram and you know that it shows the shape of a distribution well. Here you will learn how to handle larger amounts of data to form stem-and-leaf diagrams. As you will see, one set of data can give rise to more than one stem-and-leaf diagram which highlight different aspects of the data. Look at the data set below:

11 9 6 27 17 2 19 12 8 17 3 10 23 6 18
13 11 22 13 19 4 12 23 34 19 15 7 40 16 20

We obtain the stem-and-leaf diagram shown below if we define the "stem unit" to be 10. So the numbers to the left of the vertical bar are the numbers of "10s" in the observations and the remainders are shown by the "leaves" to the right of the bar. Thus we can read the data, in increasing order, from the stem-and-leaf diagram. They are 2, 3, 4, 6, 6, 7, 8, 9, 10, 11, 11, 12, ...

0	2 3 4 6 6 7 8 9
1	0 1 1 2 2 3 3 5 6 7 7 8 9 9 9
2	0 2 3 3 7
3	4
4	0

Notice that the skewed nature of the data stands out immediately. The following features are also clear:

- the 10s class has the highest number of members;
- the modal (most frequently occurring) value is 19;
- the 30s and 40s tie for the least number of members (one each).

This is not new information, we could have written these fact down after properly inspecting the original raw data. The advantage of the stem-and-leaf diagram is that it enables these facts to be expressed in a clear and obvious way. As a further illustrative example, look at the data in the table below which we will use to draw two stem-and-leaf diagrams.

Data Set A

9.5	11.9	20.0	33.4	40.1	50.0	12.7	21.0	33.6	40.6
50.0	15.5	26.4	35.4	41.1	50.0	17.7	37.9	41.3	50.0
41.9	50.4	43.0	43.3	43.6	43.7	43.8	44.7	44.9	45.0
45.1	45.2	45.3	45.5	46.1	46.5	46.6	47.1	48.0	48.2
48.5	48.4	48.6	48.7	48.8	48.9	49.4	49.5	49.6	49.8

Drawing a stem-and-leaf diagram

We can start by looking at the data as they are displayed by a stem-and-leaf diagram. Here we will use two-digit leaves with the first digit representing units and the second digit representing tenths. The tens are represented by the stems, that is the numbers to the left of the bar.

0 | 95
1 | 19, 27, 55, 77
2 | 00, 10, 64
3 | 34, 36, 54, 79
4 | 01, 06, 11, 13, 19, 30, 33, 36, 37, 38, 47, 49, 50, 51, 52, 53, 55, 61, 65, 66, 71, 80, 82, 84, 85, 86, 87, 88, 89, 94, 95, 96, 98
5 | 00, 00, 00, 00, 04

Notice that all we have really done is order the data from the lowest value to the highest value reading from top to bottom. This particular display has over half of its members in one class - the 4-class.

It may be informative to split the classes and look more closely at the data.
This can be done by:

1. rounding the raw data to the nearest whole number;
2. splitting each class according to the rule

second digit 0 - 4 ..... * second digit 5 - 9 ..... •

The rounded data now appear as follows

Data Set B

10	12	20	33	40	50	13	21	34	41
50	16	26	35	41	50	18	38	41	50
42	50	43	43	44	44	44	45	45	45
45	45	45	46	46	47	47	47	48	48
49	48	49	49	49	49	49	50	50	50

The stem and leaf diagram now becomes

0*	
0•	
1*	0 2 3
1•	6 8
2*	0 1
2•	6
3*	3 4
3•	5 8
4*	0 1 1 1 2 3 3 4 4 4
4•	5 5 5 5 5 5 6 6 7 7 7 8 8 8 9 9 9 9 9 9
5*	0 0 0 0 0 0 0 0

Essentially, the classes have been split according to the usual rule for rounding decimals. This process can make certain information contained in the data a little more obvious than the previous stem and leaf diagram. For example:

- the values in the 3-class are evenly distributed between both halves of the class in the sense that each half has two members;
- the 4-class is split in the ratio 2:1 in favour of the upper half of the class;
- the values in the 5-class are all in the lower half of the class.

You should have realised that:

- this is not *new* information - the new display has merely highlighted certain aspects of the raw data;
- some of the conclusions may have been affected by the rounding process.

Looking at the original stem and leaf diagram of Data Set A (page 35), it is easy to produce a five-number summary of the data.

The summary is:

1. The lowest value, this is 9.5;
2. The lower hinge, this is 40.1 (the observation with rank 13);
3. The median, this is 45.05 (the average of the observations with ranks 25 and 26);
4. The upper hinge, this is 48.6 (the observation with rank 38);
5. The highest value, this is 50.4.

The corresponding spreads are:

1. The low-spread, this is 45.05 - 9.50 = 35.55;
2. The mid-spread, this is 48.60-40.10=8.50;
3. The high-spread, this is 50.40 - 45.05 = 5.35.

Notice that the spreads indicate a considerable deviation from normality.
For an ideal normal distribution, we would expect:

- The distances between the median and hinges to be equal
- The high-spread and low-spread to be equal
- The distances between the hinges and the extremes to be equal

as shown in the following diagram.

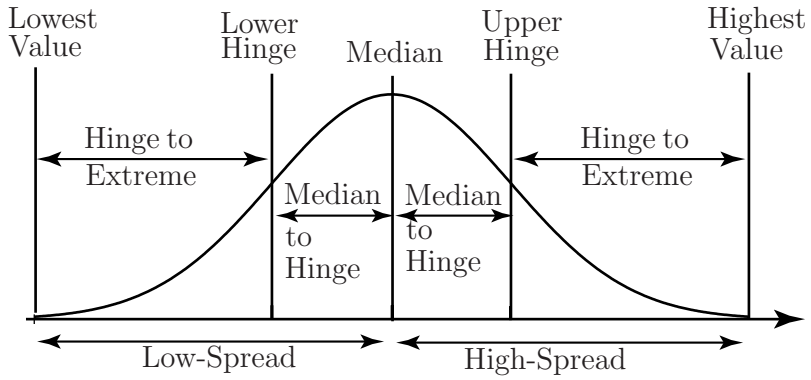


Figure 9

Using the rounded data given above as Data Set B (page 36), find the five number summary. Use your summary to check the data for normality and comment on any deviations from normality that you find.

Your solution

Answer

Data

10 Lowest Value = 10
12
13
16
18
20
21
26
33
34
35
38
40 Lower Hinge = 40 Low-Spread = 35
41 Hinge to
41 Extreme = 30
41
42
43
43
44
44
44
45
45
45 Median = 45 Median to
45 Lower Hinge = 5
45
45 Median to
46 Upper Hinge = 4
46
47
47
47
48
48
48
49
49 Upper Hinge = 49 High-Spread = 5
49
49
49
49
50 Hinge to
50 Extreme = 1
50
50
50
50
50
50 Highest Value = 50

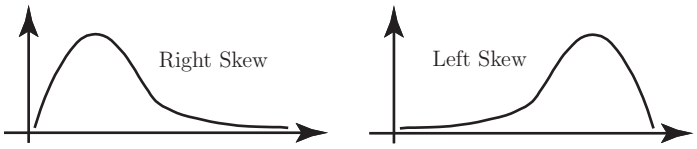
Comparing values as indicated by the diagram on page 37 gives the following results:

Low-Spread = 35	High-Spread = 5
Lower Hinge to Extreme = 30	Upper Hinge to Extreme = 1
Median to Lower Hinge = 5	Median to Upper Hinge = 4

While there are no hard-and-fast rules for comparing figures such as those obtained here, some authors suggest that the figures should be within 10% of each other before normality can be assumed. This is clearly not the case here. We conclude that the distribution of data being investigated is not symmetrical. In fact the figures above suggest that the distribution is skewed to the left, a fact supported by the stem-and-leaf diagram of the same data to be found above. [Note: skewness is defined on page 46.]

Answer
continued . . .

Remember that the term 'skewness' refers to the location of the 'tail' of a distribution.



The box-and-whisker diagram

In order to summarise a data set visually we can use a **box and whisker** plot as well as a stem-and-leaf diagram. A box-and-whisker diagram of the original (unrounded) Data Set A is shown below and the procedure necessary for drawing a plot is discussed.

You should note that there are several similar methods recommended by different authors for drawing box-and-whisker plots and so the methods recommended in statistical texts may vary a little from those given below.

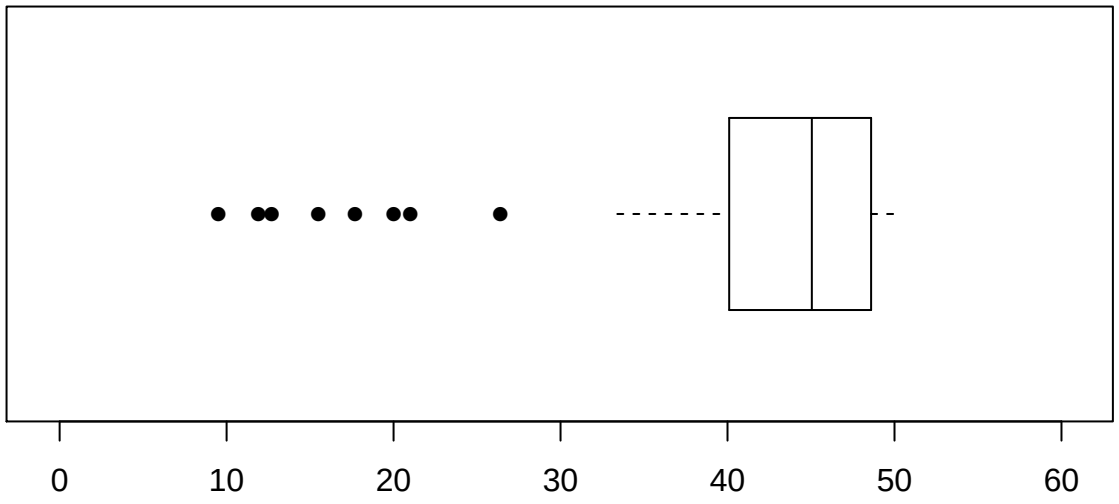


Figure 10

The diagram is constructed as follows, referring to the summary statistics on page 37:

1. **The Box**

   (a) The left-hand vertical is placed at the lower hinge (40.1) ;

   (b) The right-hand vertical is placed at the upper hinge (48.6);

   (c) The vertical in the box is placed at the median (45.05).

2. **The Whiskers**

   Notice that the mid-spread of the data (the difference between the hinges) is 8.5.

   (a) Find the greatest value which is within 1.5 mid-spread (12.75) of the upper hinge (48.6). Here $48.6 + 12.75 = 61.35$ so the greatest value is 50.4.

   (b) Find the least value which is within 1.5 mid-spread (12.75) of the lower hinge (40.1). Here $40.1 - 12.75 = 27.35$ so the least value is 33.4.

   Connect the greatest and least values to the box by means of dashed lines.

3. **The Outlying Values**

Mark as large dots any values which are **more** than 1.5 mid-spreads from the hinges. In this case 1.5 mid-spreads give a value of 12.75 and so we mark dots which represent values which are higher than $48.6 + 12.75 = 61.35$ and values which are lower than $40.1 - 12.75 = 27.35$. In this example there are no values greater than 61.35, but there are 8 values which are less than 27.35. Notice that half of the data values lie in the box and that the tails show up well in the diagram. The diagram shows the left-skew (skewness refers to the tail) present in the data.

2. Outliers

Outliers are values which are well outside the range covered by the bulk of a data set. There is no standard precise definition but some simple criteria do exist which may be used to detect outliers and accept or reject outliers. The eight values shown as large dots above illustrate the concept of outliers. Outliers can be extremely important since they may be (for example) erroneous data or they may point the way to further investigations of a data set.

Example: Particulate air pollution with a particle size up to 10μg is known as "PM-10". Such particles, particularly of soot, are emitted, for example, by Diesel engines. Data on PM-10 concentration are collected by a monitoring instrument. The box-and-whisker plot in Figure 11 summarises 100 daily observations.

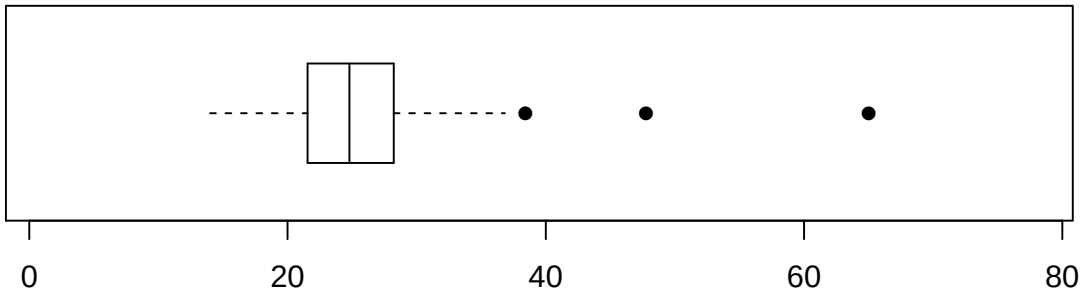


Figure 11. 100 daily PM-10 concentrations in μg/m^3

Three potential outliers are identified. The largest of these, in particular, merits investigation. Is the observation correct? Sometimes automatic monitors can malfunction. Did something happen that day, such as a large fire nearby, to explain the unusual observation?

Place the items in the data set below in rank order and use your rank ordering to find the five number summary of the data.

155.3	177.3	146.2	163.1	161.8	146.3	167.9	165.4	172.3	188.2
178.8	151.1	189.4	164.9	174.8	160.2	187.1	163.2	147.1	182.2
178.2	172.8	164.4	177.8	154.6	154.9	176.3	148.5	161.8	178.4

Construct a box-and-whisker diagram representing the data.
Does the box-and-whisker diagram tell you that the data set that you are working with is symmetrical? Record the reasons for your comments.

Your solution

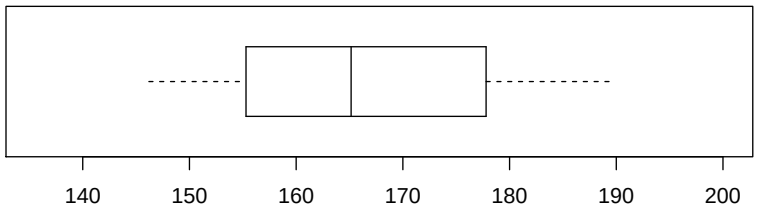
Work the solution on a separate piece of paper. Record the main stages in the calculation and your conclusions here.

Answer

Data

146.2	Lowest Value = 146.2	
146.3		
147.1		
148.5		
151.1		
154.6		
154.9		Low-Spread = 18.95
155.3	Lower Hinge = 155.3	
160.2		
161.8		
161.8		
163.1		
163.2		
164.4		
164.9	Median = 165.15	Mid-Spread = 22.5
165.4		
167.9		
172.3		
172.8		
174.8		
176.3		
177.3		
177.8	Upper Hinge = 177.8	
178.2		High-Spread = 24.25
178.4		
178.8		
182.2		
187.1		
188.2		
189.4	Highest Value = 189.4	

The Box-and-Whisker plot is:



The plot indicates that the distribution is not quite symmetrical, for example you would expect the median value to appear midway between the hinges for a symmetrical distribution.

Criteria for rejecting outliers

As you already know, outliers may be taken to be observations which lie well outside the range of most of a sample. They are important for several reasons:

1. they can have misleading effect on statistics such as the mean and standard deviation;

2. their occurrence may be due to incorrect observation, measurement or recording. In this case it is often possible to correct the data;

3. their presence can induce a false skewness in a data set;

4. they may actually be members of a population not under consideration. For example, data on road traffic speeds collected at a point in a highway may be intended to provide information on the speeds of motor vehicles but the data are likely to include some observations for bicycles and other slower-moving types of traffic.

Simple criteria exist which facilitate the detection of outliers. These criteria should be used with some caution and never automatically used simply to reject an outlier. You should always ask why such a value occurred in the first place and work to answer such a question sensibly before considering rejection. Two criteria for the detection of outliers are given below. Criterion 1 may be applied to data sets that are known to have the shape known as a *normal distribution*. This distribution will be discussed in Workbook 39. Many variables tend to give distributions with this shape, at least approximately. Criterion 2 uses the five-number summary discussed above and may be applied to any data sets.

Criterion 1

We know that, for variables where the distribution has a "normal" shape, we can expect only about 1 in 1000 observations to lie more than 3.3 standard deviations away from the mean. So we could treat any value further than 3.3 standard deviations from the mean as an outlier. This choice essentially implies that a value has less than 1 in a 1000 of chance of occurring naturally outside the range defined by 3.3 standard deviations from the mean. Thus an observation x would be regarded as an outlier if

$$\left|\frac{x-\bar{x}}{s}\right| > 3.3$$

Note that the property that 0.1% of observations are more than 3.3 standard deviations from the mean really refers to the *population* mean and standard deviation. Here we have used $\bar{x}$ and s the sample mean and standard deviation. However this will be a reasonable approximation in reasonably large samples, say $n > 30$.

Criterion 2

We have previously identified as outliers observations which are more than 1.5 mid-spreads (or IQRs) from the hinges (or quartiles). We might also regard as *extreme outliers* any observations which are more than 3 mid-spreads from the hinges.
While all values classified as outliers should be investigated, this is particularly true of those classified as extreme outliers.

Transformations. Note that distributions which are not symmetric can sometimes be made more symmetric by applying a suitable transformation, such as log or square root. This can also cause apparent outliers no longer to appear outlying.

Manufacturing processes generally result in a certain amount of wasted material. For reasons of cost, companies need to keep such wastage to a minimum. The following data were gathered over a five-week period by a manufacturing company whose production lines run seven days per week. The numbers given represent the percentage wastage of the amount of material used in the manufacturing process.

Daily Losses (%)

17	6	8	17	23	18	10	15	17	4
17	18	15	19	11	15	22	12	15	16
11	18	17	17	13	15	9	21	17	16
14	13	15	11	12					

1. Find the mean and standard deviation of the percentage losses of material over the two week period.

2. Assuming that the losses are roughly normally distributed, apply an appropriate criterion to decide whether any of the losses are smaller or larger than might be expected by chance.

Your solution

Answer

1. We will treat any value further than 3.3 standard deviations from the mean as an outlier (criterion 1). With x_0 as the potential outlier we need to calculate the quantity $\left|\dfrac{x_0-\bar{x}}{s}\right|$. This is known as a *standardized score*. We then accept x_0 as a member of the distribution if $\left|\dfrac{x_0-\bar{x}}{s}\right| \leq 3.3$. Otherwise we reject x_0 as an outlier.

   Calculation gives:

   $\bar{x} = 14.69 \qquad s = 4.22$ and the values of the standardized score are

0.55	-2.06	-1.58	0.55	1.97	0.79	-1.11	0.07	0.55	-2.53
0.55	0.79	0.07	1.02	-0.87	0.07	1.73	-0.64	0.07	0.31
-0.87	0.79	0.55	0.55	-0.40	0.07	-1.35	1.50	0.55	0.31
-0.16	-0.40	0.07	-0.87	-0.64					

2. The calculation shows that all values of $\left|\dfrac{x_0-\bar{x}}{s}\right| \leq 3.3$ and so we conclude that the daily losses are all within the range indicated by chance variation.

3. Skewness and multiple peaks

When exploring a data set, as well as looking for any outliers, it is a good idea to look at the shape of the distribution. Features of the shape which might affect how we draw conclusions from the data include whether the distribution is symmetric or *skewed* and whether the distribution has more than one peak or *mode*.

Skewness

If a skewed distribution is represented purely by two numbers, say the mean and standard deviation, then the representation will be inadequate. Remember that the term 'skewness' refers to the location of the 'tail' of a distribution.

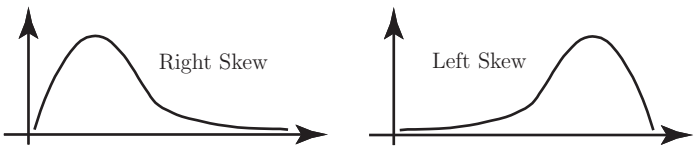


Figure 12

As an example, the data set below gives the current required to burn out a component under test.

9.5	11.9	20.0	33.4	40.1	50.0	12.7	21.0	33.6	40.6
50.0	15.5	26.4	35.4	41.1	50.0	17.7	37.9	41.3	50.0
41.9	50.4	43.0	43.3	43.6	43.7	43.8	44.7	44.9	45.0
45.1	45.2	45.3	46.1	46.5	46.6	47.1	48.0	48.2	45.3
48.5	48.4	48.6	48.7	48.8	48.9	49.4	49.5	49.6	49.8

The data were obtained by measuring the current in mA applied to an electronic component under conditions of destructive testing, giving the following values for the mean, standard deviation, median and mid-spread:

$\bar{x} = 40.72 \quad s = 11.49 \quad$ median $= 45.05 \quad$ and $\quad$ mid-spread $= 9.55$

The median is a more useful summary than the mean here because it has a direct interpretation as the current which would burn out 50% of components of this type. Similarly the upper quartile is the current which would burn out 75%. The relationship of the mean and standard deviation to these is not clear as the distribution is not symmetric.

Multimodal distributions

A distribution with more than one peak is called a **multimodal** distribution. A distribution with exactly two peaks is called a **bimodal** distribution. Such distributions can be very difficult to summarise. The stem-and-leaf plot shown below and the box-and-whisker plot in Figure 13 summarise 200 time intervals in days between failures in aircraft air-conditioning systems. In the stem-and-leaf plot, the stem units are 100s and the leaf units are 10s. So, for example, the last three entries are 480, 490, 540.

```
0 | 000000011111111111111111111222222222233333333333444
0 | 55567778
1 | 0011112444
1 | 555556666666777777888888899999
2 | 000000011111111222222222222233333344444
2 | 555555566677777888899999
3 | 000001122233444
3 | 556677889
4 | 00111111344
4 | 89
5 | 4
```

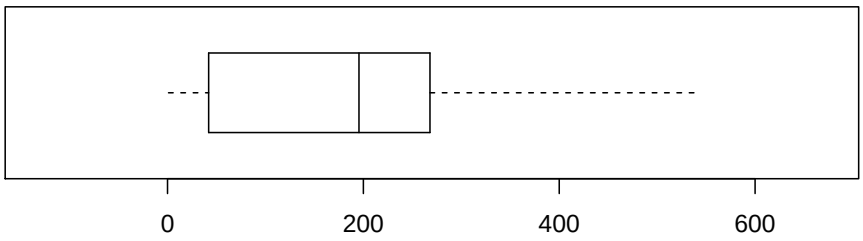


Figure 13. Time intervals in days between air-conditioning failures.

The stem-and-leaf plot reveals relatively few failures of 50-140 days. It looks as though there are two common kinds of failure interval, short intervals of around 0-50 days and long intervals from around 150 days upwards. Taking short intervals as less than 50 days, the distribution of the longer

intervals (over 50 days) has **right** skew. (In fact, closer inspection of the shorter intervals shows that the distribution in this group also has right skew).
Recall that the term skewness refers to the **tail** of a distribution.
The usual summary statistics that you might be tempted to calculate are:

$$\bar{x} = 184.0 \quad \text{and} \quad s = 128.3 \quad \text{or} \quad \text{median} = 195.5 \quad \text{and} \quad \text{mid-spread} = 226$$

In this case, neither set of statistics is very informative since neither set indicates the bimodality or the skewness. Without visual representation, a single peaked (i.e. unimodal) distribution tends to be assumed and this is, of course, not the case here.

The stem-and-leaf plot is more informative than the box-and-whisker plot since it shows the bimodality.
In practice we would work with the two constituent distributions and attempt to relate the results in a practical way.

Final comments on data representations

1. You should not rely on summary statistics such as the mean and standard deviation or median and mid-spread alone to represent a data set. Remember that if a distribution has outliers, gaps, skewness or multiple peaks, then shape is probably more important than location and spread.

2. The shape of a distribution is better shown visually than numerically. Remember that a stem-and-leaf diagram retains the data and arranges the data in rank order and that a box-and-whisker plot emphasises the detail contained in the tails of a distribution.

Exercises

1. The following data give the lifetimes in hours of 50 electric lamps.

1337	1437	1214	1300	1124	1065	1470	1488	1103	978
1177	1289	1045	947	969	1339	1594	812	1277	1032
1167	974	1131	974	1727	1378	1385	1330	1672	1604
1493	1521	1235	1682	1136	1229	803	1166	1494	1733
978	1110	1055	1438	1436	1424	766	1283	829	1652

   (a) Represent the data using a stem-and-leaf diagram with two-digit leaves.

   (b) Calculate the mean lifetime from these data.

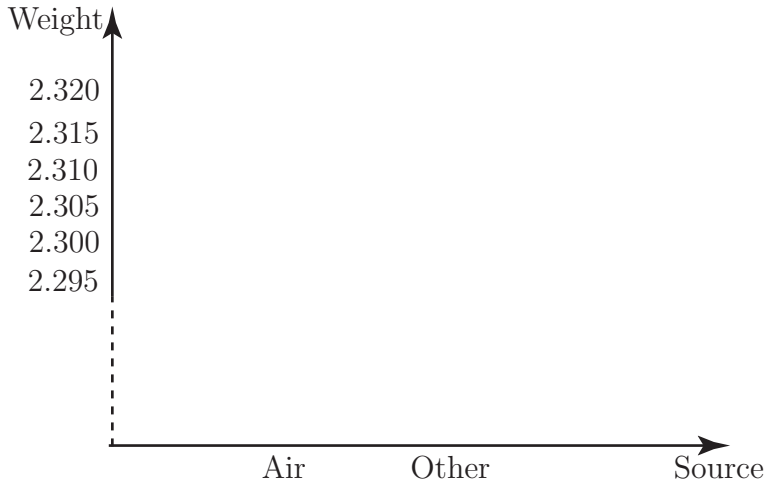
   (c) Does the mean lifetime give a good indication of the expected lifetime of a lamp?

2. During the winter of 1893/94 Lord Rayleigh conducted an investigation into the density of nitrogen gas taken from various sources. He had previously found discrepancies between the density of nitrogen obtained by chemical decomposition and nitrogen obtained by removing oxygen from air. Lord Rayleigh's investigations led to the discovery of argon. The raw data obtained during his investigations are given below.

Date	Source	Weight	Date	Source	Weight
29/11/93	NO	2.30143	26/12/93	N_2O	2.29889
05/12/93	NO	2.29816	28/12/93	N_2O	2.29940
06/12/93	NO	2.30182	09/01/94	NH_4NO_2	2.29849
08/12/93	NO	2.29890	13/01/94	NH_4NO_2	2.29889
12/12/93	Air	2.31017	29/01/94	Air	2.31024
14/12/93	Air	2.30986	30/01/94	Air	2.31030
19/12/93	Air	2.31010	01/02/94	Air	2.31028
22/12/93	Air	2.31001			

   (a) Organise the data into a frequency table using the classes [2.29-2.30), [2.30-2.31), [2.31-2.32). Draw the histogram representing the data and comment on any unusual features that you may see.

   (b) Classify the data according to the two sources 'Air' and 'Other'. Order each data set and hence find the median, the hinges and the mid-spreads for each data set. Plot box-and-whisker diagrams for the data on a diagram similar to the one shown below.



Comment on any unusual features that you see. What do the box-and-whisker plots tell you about the nitrogen obtained from the two sources?

3. Answer the following questions:

   (a) Is the variance measured in the same units as the mean?

   (b) Is the mean measured in the same units as the median?

   (c) Is the standard deviation measured in the same units as the mode?

   (d) Is the mode measured in the same units as the mid-spread?

   (e) Is the high-spread measured in the same units as the low-spread?

   (f) Is the mid-spread measured in the same units as the hinges?

Answers

1. (a) Stem and leaf diagram (stem – hundreds, 2 digit leaves – tens and units).

7	66
8	03,12,29
9	47,69,74,74,78,78
10	32,45,55,65
11	03,10,24,31,36,66,67,77
12	14,29,35,77,83,89
13	00,30,37,39,78,85
14	24,36,37,38,70,88,93,94
15	21,94
16	04,52,72,82
17	27,33

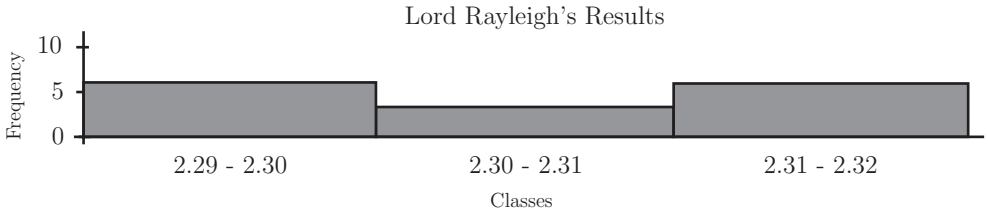
   (b) The sum of the lifetimes is $\sum x = 62802$. So the mean is

$$\frac{62802}{50} = 1256.04.$$

   (c) Yes. The mean lifetime gives a reasonable indication of what can be expected since the distribution is fairly symmetrical. However it does not, of course, give any indication of the spread.

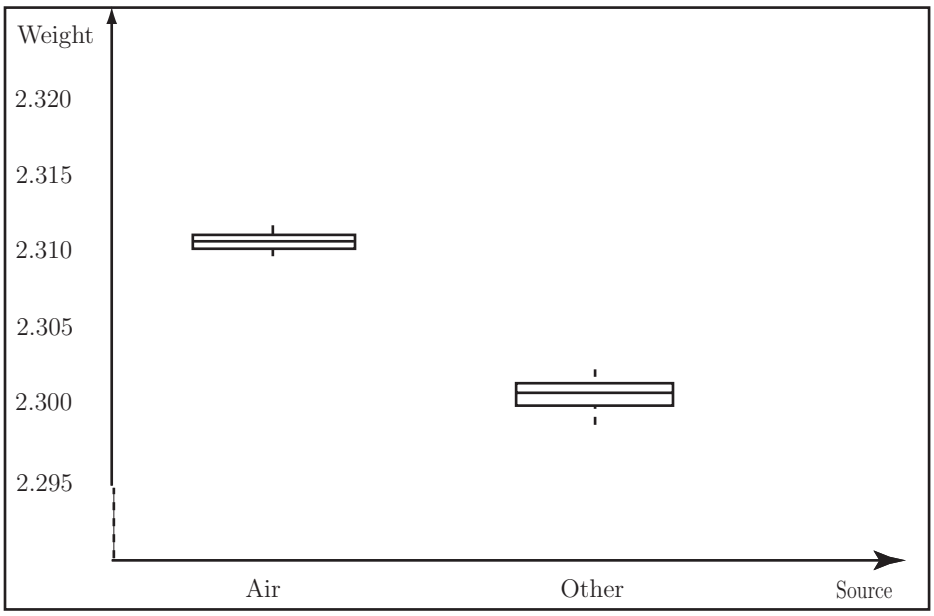
Answers

2. (a)



The lowest class is obtained entirely from non-air sources, the highest class is obtained entirely from air.

(b)



Comment. Box-and-whisker plot tells us that some other element is present in Air which is responsible for the additional weight. This *additional* element subsequently proved to be the inert gas argon.

3. (a) No (b) Yes (c) Yes (d) Yes (e) Yes (f) Yes

Index for Workbook 36

Workbook 36

HELM®

HELM: Helping Engineers Learn Mathematics

http://helm.lboro.ac.uk