Statistics

1. Introduction

1.1. Chapter overview

This chapter introduces the idea of 'statistics', looking at the different ways in which you can gather information (or **data**) and analyse it.

You may wish to gather information concerning the salaries of people who work at your firm or on the performance of a fund manager. Either way the method of presentation of the data is critical to how easy it is for other people to interpret. This chapter introduces different ways in which data can be presented.

If you gather data it can be important to identify the central point around which the data seems to be clustered. This is known as a measure of **central tendency.** The most important measure of central tendency is the **arithmetic average**, which is a simple average of all the data you have collected.

Once you have a central tendency figure you will want to know how closely clustered to that central tendency your actual data points are, i.e. you will want a measure of **dispersion**. The most important measure of dispersion you will learn is **standard deviation**. This chapter will show you the formula for standard deviation and how to calculate it on a scientific calculator. You will see that the bigger the standard deviation figure the bigger the level of dispersion around the arithmetic mean. In other words, the bigger the standard deviation, the more spread out the data will be.

Standard deviation is probably the single most important calculation for statisticians as it is very useful in analysing the variability of returns to an investment.

1.2. Learning outcomes

On completion of this module you will:

Types and sources of data

- 7.1.1 Distinguish between primary and secondary sources of data
  - 7.1.2 Identify examples of primary and secondary data
  - 7.1.5 Distinguish between continuous and discrete data
  - 7.1.6 Define categorical data and explain how it can be converted to ordinal data

Populations and samples

- 7.1.3 Distinguish between a population and a sample
- 7.1.4 Explain the key sampling methods

Presentation of data

- 7.1.7 Interpret a frequency and relative frequency distribution
  - 7.1.8 Explain the use of the following in the presentation of data: pie chart, bar chart, histogram, scatter plots, graphs

Summary statistics

- 7.2.1 Define, explain and calculate the following measures of central tendency for both raw data and interval data: arithmetic mean, geometric mean, median, mode
- 7.2.4 Define, explain and calculate the following measures of dispersion for both raw data and interval data: standard deviation (population and sample), variance, range, quartiles and percentiles, inter-quartile range
- 7.2.2 Distinguish between symmetric and skewed data
  - 7.2.3 Explain the relationship between the mean, median and mode for symmetric and skewed data
  - 7.2.5 Explain the notion of probability distributions and identify the properties of the normal distribution
  - 15.1.2. Explain the implications of assuming returns are normally distributed
  - 7.2.6 Explain the notion of statistical significance in the context of investment decisions

Correlation and bivariate linear regression

- 7.3.3 Explain and interpret the correlation coefficient in the context of linear regression
- 7.3.5 Define the concept of autocorrelation and describe the impact of extreme events on correlation
- 15.1.10 Calculate correlation coefficients from standard deviation/covariance of two investments
- 7.3.1 Explain the least-squares regression technique in deriving a line of best fit
  - 7.3.2 Calculate and interpret a forecast value for the dependent variable given the intercepts and gradients of a regression line equation
  - 7.3.4 Explain the shortfalls in the application of linear regression to forecasting, including why correlation does not imply causation, and the pitfalls of data-mining

2. Types and sources of data

2.1. Types and sources of data

Background

Investment decisions require analysis and interpretation of a wide variety of information or 'data'.

Fund managers and investment analysts, for instance, require data on such things as company performance, industry life cycle and other macroeconomic factors influencing the investment decision, e.g. interest rates, exchange rates etc.

This module begins by describing the different types of data and how they are collected. The module then moves on to examine the ways in which the collected data can be presented and summarised in a useful and informative way.

Primary data

Primary data is collected with a **particular purpose in mind**. For example, an advertising agency researching consumer attitudes to various brands of chocolate.

Primary data refers to data that an investigator has collected themselves. The investigator therefore knows the conditions under which the data was collected and is aware of any limitations it may contain.

Secondary data

Secondary data is collected by government agencies and other international bodies which have been formed specifically to gather and distribute economic and social data in a convenient form. The Office for National Statistics (ONS), for example, collects economic data on inflation and employment.

Users of secondary data do not have a full understanding of the background and circumstances under which the data was initially collected. Consequently, users of secondary data may be unaware of any limitations it may contain.

Other sources of secondary data could be:

- Bank of England
- HM Treasury
- Credit rating agencies, such as Fitch, Moody and S&P

Discrete data

Discrete data refers to data where the units of measurement cannot be split up. For example, if the data refers to the number of people using a particular tube station each day, then the recorded figures might be 824 **or** 825 people, but never 824 ½.

Data can be put into groups or categories, for example, the answers to a question could be coded 1 for yes, 2 for no and 3 for maybe. This process would separate the responses to form categorical data.

Sometimes categorical data may be ranked or ordered according to a set criteria, e.g. a first or second class degree. It is the order of these numbers that matters. This is known as an **ordinal data**.

It will not generally be possible to directly apply descriptive statistics to such categorical data as the actual number itself is arbitrary.

Continuous data

Continuous data is where the units have a constant scale and all points between the units have meaning. For example, the distance travelled by a person to work can be expressed as 5 miles, 5.1 miles, 5.12 miles and so on to an unlimited number of decimal places.

The level of accuracy in recording continuous data depends on the precision of the measuring device itself.

3. Populations and samples

3.1. Populations

A population is the **entire** set of items which have the desired characteristics under investigation. For example, if the TV viewing habits of males under 40 years of age was under investigation, then the population refers to **all** males under 40 years of age.

A population will give a complete set of data, but will be very difficult and time consuming to collect.

3.2. Samples

Using samples

Sometimes it is impractical, if not impossible, to examine every member of the population under investigation. Instead, only a part, or sample, of the population is tested. A sample is a sub-set of items taken from the population with the characteristics under investigation. Samples of around 1,000 are considered suitable to reflect the population of the UK.

Selecting a sample can be done either on a random or non-random basis.

Random samples

A random sample is a sample selected in such a way that every item of the population has an **equal** chance of being selected.

Non-random samples

The alternative to random sampling is to employ a non-random (or **non-probability**) method of selection.

An example of non-random selection is **quota** sampling, often used in market research. Such a quota is usually categorised into different types of individual items, e.g. professional or manual workers, with 'sub-quotas' for each type.

For example, quota sampling might involve interviewing all people the investigator meets in a city centre up to a given number, (i.e. the quota). Quota sampling might involve choosing to interview 520 women and 480 men in order to reflect the gender split of the UK. This additional condition set on the sample is called **stratified sampling**, and is designed to reduce sampling error. It does this by selecting a sample that represents the population.

Another form of non-random sampling is systematic sampling. This is where researchers select the Nth record of a population. For example, if analysing how far your employees travel to work on average, we may ask every fifth person on an alphabetical list of employees.

Other sampling methods

Convenience sampling – choosing the sample that is easiest to collect information from. Choosing people in your local town to represent the UK, for example.

Judgement sampling – making a judgement of the sample that would best represent the sample. Choosing people who live in Swindon to represent the UK, for example.

Snowball sampling – This is typically used when the subjects of the data are rare. It relies on referrals from initial subjects.

4. Presentation of data

4.1. Frequency distribution tables

A frequency distribution table is one of the more straightforward methods of data presentation. Such a table involves categorising the number of times something has occurred in each category under investigation.

For example, the data of 129 peoples' salaries is illustrated on a frequency distribution table below. As illustrated, the salary range is grouped into subsets in the left hand column and the number of people is shown on the right.

Table 1. Frequency distribution table (salary example)

Salaries £ 000's pa (nearest £)	No. of people
4.999 or less	2
5 to 9.999	3
10 to 14.999	6
15 to 19.999	14
20 to 24.999	29
25 to 29.999	38
30 to 34.999	18
35 to 39.999	9
40 to 44.999	6
45 to 49.999	3
50 and above	1
	Total 129

As a result of grouping, it is possible to detect patterns in the data. For instance, it is clear from the above table that the majority of people earn between £15,000 - £35,000 pa.

4.2. Relative frequency distribution

A relative frequency distribution table allows us to easily see the size of the category frequency in comparison with the total. Each frequency is calculated as percentage of the whole, for example, to calculate to calculate the relative frequency of a salary of £4,999 or less, you would divide 2 by 129 giving 1.55%.

If we add this to the previous table, it looks as follows.

Table 2. Relative frequency distribution example

Salaries £ 000's pa (nearest £)	No. of people	Relative Frequency
4.999 or less	2	1.55
5 to 9.999	3	2.33
10 to 14.999	6	4.65
15 to 19.999	14	10.85
20 to 24.999	29	22.48
25 to 29.999	38	29.46
30 to 34.999	18	13.95
35 to 39.999	9	6.98
40 to 44.999	6	4.65
45 to 49.999	3	2.32
50 and above	1	0.78
	Total 129	Total 100

4.3. Cumulative frequency distribution

A cumulative frequency distribution table identifies the number of times something has occurred **up to** the category under investigation, i.e. it shows the proportion of a sample (or population) taking a value less than or equal to a given number. The table below adds the cumulative frequency to the above data, in percentage terms.

Table 3. Cumulative frequency distribution example

Salaries £ 000's pa (nearest £)	No. of people	Relative Frequency (%)	Cumulative Frequency
4.999 or less	2	1.55	1.55
5 to 9.999	3	2.33	3.88
10 to 14.999	6	4.65	8.53
15 to 19.999	14	10.85	19.38
20 to 24.999	29	22.48	41.86
25 to 29.999	38	29.46	71.32
30 to 34.999	18	13.95	85.27
35 to 39.999	9	6.98	92.25
40 to 44.999	6	4.65	96.90
45 to 49.999	3	2.32	99.22
50 and above	1	0.78	100.00
	Total 129	Total 100	Total 100

The above table clearly identifies that 71% of the population earns £29,999 or less.

4.4. Visual presentation of discrete data

Introduction

Visual techniques are also used to present data in a user-friendly manner.

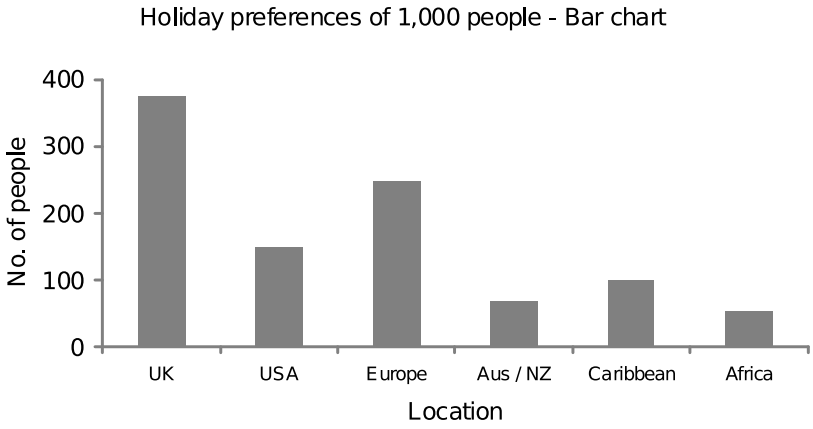
To illustrate, suppose 1,000 people were asked where they spent their annual holiday with the following results:

- UK: 375 people
- USA: 150 people
- Europe: 250 people
- Australia: 70 people
- Caribbean: 100 people
- Africa: 55 people

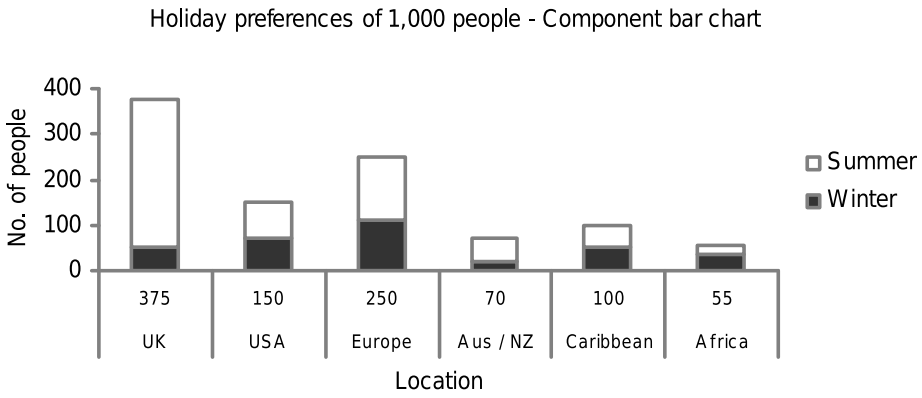
There are two main methods of visually presenting this (discrete data): bar charts and pie charts.

Bar charts

Bar charts display information as 'bars' (or columns) representing each class of data (in this case, holiday location) according to their frequency of occurrence. The **height** of the bar represents the frequency of occurrence.



More information could be added by stipulating whether the holiday was a winter or summer holiday. This would create a **component bar chart**.

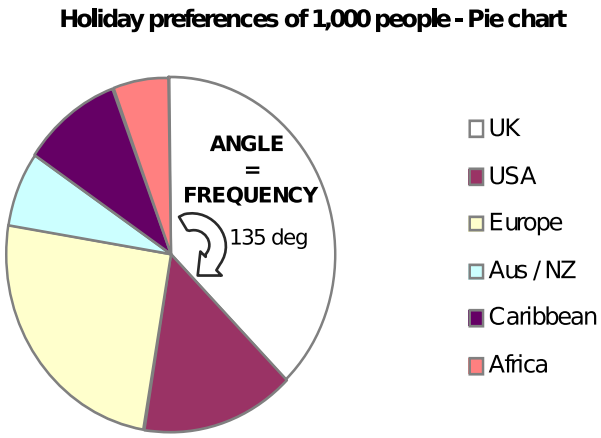


Pie charts

A pie chart is a circle that has been divided into different 'slices', each representing a different category of data.

The area of each slice is proportional to the frequency of occurrence. The area of the slice is determined by its **angle** on the pie chart.

The key to understanding pie charts is to remember that there are 360 degrees in a circle! In the example above, 375 people, or 37.5%, took their holiday in the UK: 37.5% of 360 degrees equals 135 degrees.



The above pie chart shows that most people in the population spent their holiday in the UK, i.e. the greater the relative frequency of occurrence, the bigger the angle of the slice.

4.5. Visual presentation of continuous data

Introduction

As previously mentioned, continuous data can take **any** value. The only limitation is the degree of precision of the measuring equipment. There are four main methods of visually presenting continuous data. These are:

- Histograms

- Time series graphs
- Semi-log graphs
- Scatter diagrams – these will be covered later in the chapter

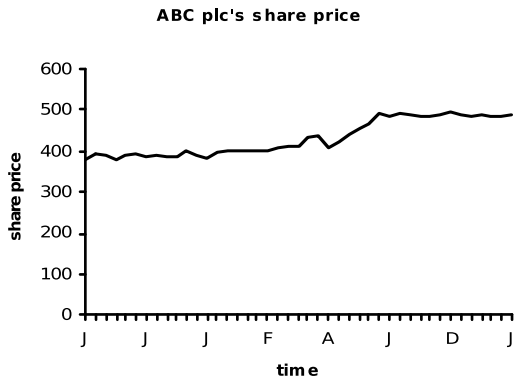
Histograms

Histograms may look similar to bar charts, but it is the **area** (not the height) of the bar that represents the frequency of occurrence.

As illustrated above, a histogram groups continuous data into appropriate intervals and represents the frequencies by the area of the bars.

Time series graphs

Time series graphs are very common in the financial world. The graph displays data **over time** e.g. a share price:



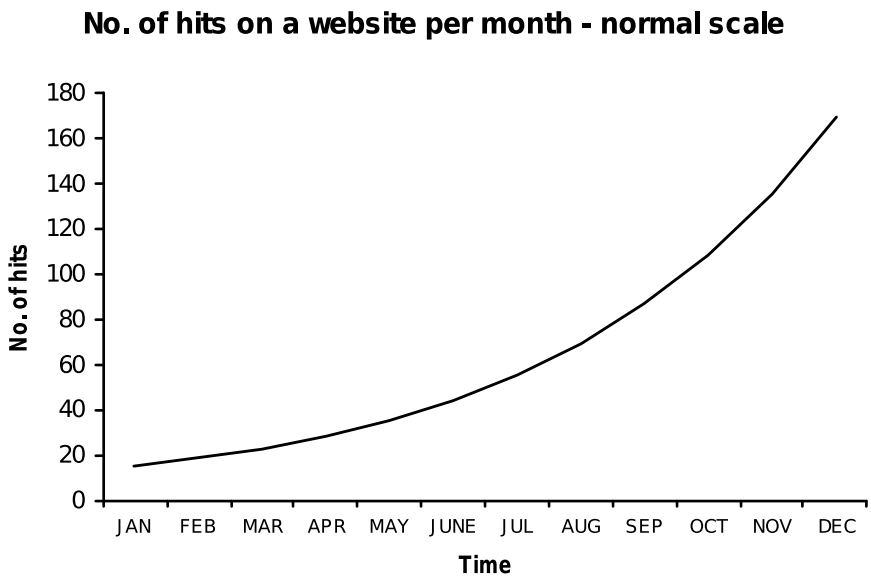
A time series graph displays the path of a variable (in this case, a share price) in chronological order.

Log (semi-log) graphs

A (semi-) log graph is used to illustrate the **rate of change** of a variable.

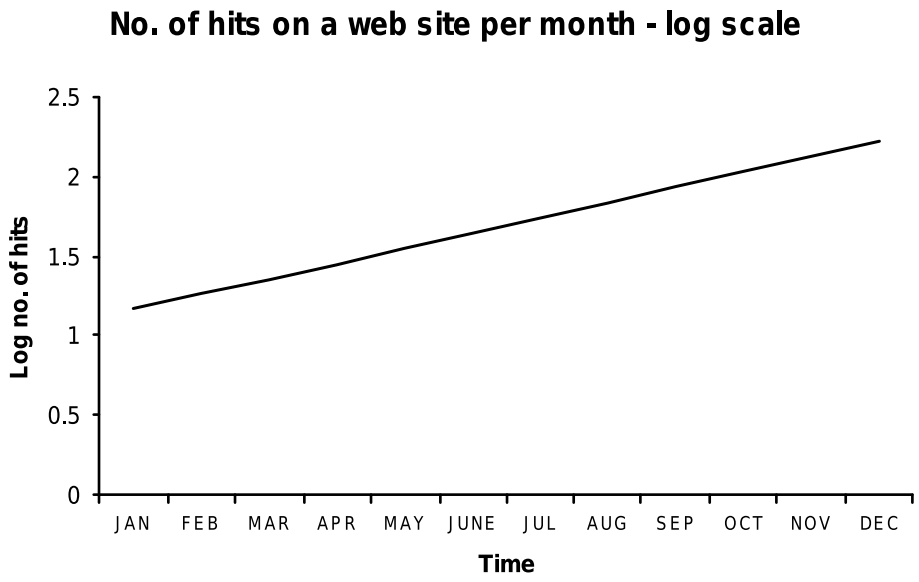
For example, consider the launching of a new website; as time goes by, the site will (hopefully) receive more and more 'hits'.

If the number of people hitting the website over a year was monitored, the results might look something like the graph below:



The graph illustrates that the number of people visiting the site is **accelerating** over time, i.e. there is exponential growth.

A log graph is constructed in order to determine the **rate** of acceleration over time.



The gradient of the slope is 1 in 4 or 25%. In other words the hits are increasing at a steady rate of 25% per month.

The constant 1 in 4 gradient of the log graph shows that not only are the number of website hits increasing, but that they increasing at a **constant rate** of 25% each month.

Where the rate of growth is not constant, the gradient of the line on the graph will vary: steepening to show increased growth and flattening to show the growth slowing.

5. Summary statistics

5.1. Introduction

Summary statistics, also known as 'descriptive statistics', summarise two key features about a set of data:

- The ‘typical’ value contained within the data set, i.e. the measure of **central tendency**
- How widely spread-out the set of data is, i.e. the measure of **dispersion**

These statistics are used to compare two (or more) sets of populations and/or samples.

The aim of descriptive statistics is to efficiently summarise large quantities of data before making comparisons between different samples and/or populations.

There are three pairs of measures for central tendency and dispersion. These are:

Table 4. Central tendency and dispersion measurement summary table

Central tendency (typical values)	Measures of dispersion
Mean	Standard deviation
Mode	Range
Median	Inter-quartile range

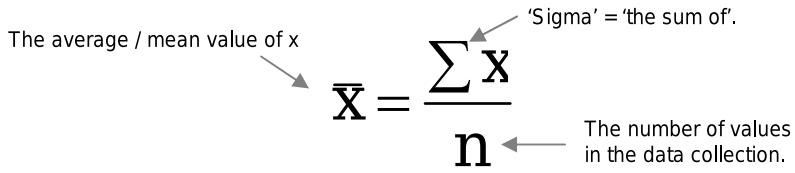
The most appropriate pair of measures for a given set of data depends on the features of the data itself.

5.2. Mean and standard deviation

The mean

The mean - also known as the 'simple arithmetic mean', is calculated as the **average** of a set of values.

The formula for calculating the mean is:



For example, should five funds each achieve returns of 10%, 12%, 12%, 15% and 18% respectfully, then the mean return is calculated as:

(10% + 12% + 12% + 15% + 18%) / 5 = 13.4%.

Standard deviation

The standard deviation measures the level of distribution, i.e. dispersion, around the mean of a set of data.

The formula for the standard deviation is:

Standard deviation (σ) formula

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

1. The sum of all the differences between each value and the mean.

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

2. The 'square' removes any negative values e.g. $-2^2 = 4$

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

3. The number of values in the data set.

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

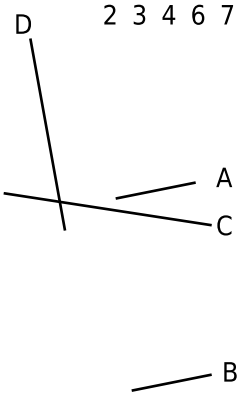
4. The square root. This counterbalances the squaring done in step 2.

A far quicker method of calculation is to use a calculator or computer. The following example demonstrates how to use the Casio FX83 in order to calculate the mean and standard deviation of a data set.

Using the fx-83GT to calculate the mean and standard deviation.

Q. Calculate the mean and standard deviation of the following data:

2 3 4 6 7 9 9 10 11 14 15 16 19 23 27



1) Press the SHIFT (C) followed by MODE (A) and then scroll down using the REPLAY key and then press 3 and then 1 to turn on the frequency function.

2) Press the MODE (A) key followed by 2 and then 1 to enter STAT mode. The frequency table will be displayed on the screen.

3) Next, enter each value using the number keys followed by the '=' button (B). Your key strokes will be:

[2] [=] [3] [=] [4] [=] [6] [=] etc…… [2] [7] [=]

After inputting the last number which should be in the 15th row of the table it is important to press the AC key. (Don't worry the data will not be lost!) You can return to the frequency table at any time by pressing SHIFT (C) 1 2.

4) To calculate the mean press the 'SHIFT' button (C) followed by key number 1 (STAT) then key number 4 followed by 2 and then =

[SHIFT] [1] [4] [2] [=]

The answer is:

$\bar{x}$ 11.66666667

4) For the standard deviation press [SHIFT] [1] [4] [3] [=] giving you:

$X\sigma n$ 7.077350414

5) If you do not get the right answers first time it is probably because you have made some inputting errors. You can return to the frequency table to edit by pressing SHIFT 1 2 and then using REPLAY to scroll through the data. To clear the table and start again press:

[MODE] [2] [1]

The previous example calculated the mean to be 11.67 and the standard deviation to be 7.07. This can be expressed as 11.67 +/- 7.07.

Grouped data

Often, data is grouped together.

Calculator point – Have you turned on the frequency so the calculator will accept grouped data?

- SHIFT, SET UP, V, 3:STAT, 1:ON

Consider the following example:

Table 5. Grouped data example

The level of return achieved by 18 different fund managers is represented below:	
Return achieved / % pa	No. of fund managers
0-5%	5
5-10%	3
10-15%	4
15-20%	6
	18

In this case, the procedure to calculate the mean and standard deviation is shown below:

- Put the calculator into standard deviation mode by pressing **MODE**, 2, 1
- Enter the data. Note that when facing grouped data, it is common practice to choose the mid-point in each group of data e.g. 2.5%, 7.5%, 12.5% and 17.5% and enter these under X in the first column. Enter the frequency in the second column by using the replay key to navigate to the correct cell
- Calculate the mean value by: SHIFT, 1, 4, 2, =
- Calculate the standard deviation: SHIFT, 1, 4, 3 =

As demonstrated above, the mean fund return equals 10.55% and the standard deviation equals 6.04%. In other words, 10.55% +/- 6.04%.

Approximately 68.26% of observations in the distribution will be within 1 standard deviation either side of the mean, i.e. in the previous example, approximately 2/3 (68.26%) of all funds have a return between 4.51% (10.55% - 6.04%) and 16.59% (10.55% + 6.04%).

Approximately 95.5% of all observations will be within 2 standard deviations either side of the mean.

Approximately 99.75% of all observations will be within 3 standard deviations either side of the mean.

Sample standard deviation

The sample standard deviation is used as a measure of dispersion for small samples of data. The limitations of small data sets include the fact that they may not be a good representative of the population as a whole, i.e. 'sampling error' may arise. The accuracy of the standard deviation is therefore put into question.

To calculate the sample standard deviation, a slight adjustment to the standard deviation formula is made; the number (‘n’) of values is reduced by one.

The overall effect of this adjustment is to cause the sample standard deviation to be **greater** than the standard deviation of a set of values.

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

The only difference

The procedure for calculating the sample standard deviation using the FX83 is the same as it is for an ordinary standard deviation (known, sometimes, as a population standard deviation) until the final stage where

SHIFT	1	4	3	=	gives the population SD
SHIFT	1	4	4	=	gives the SAMPLE SD

- Calculate the standard deviation for a sample: SHIFT, 1, 4, 4 =

Variance

The variance is the name given to the square of the standard deviation, i.e. standard deviation raised to the power of 2.

For example, if the standard deviation was 3 the variance would be $3^2=9$

Variance is used as a statistical measure of dispertion and we will meet it again when it is used in the calculation of the correlation coefficient and beta.

Note, the 'sample variance' is the name given to the square of the sample standard deviation.

5.3. Mode and range

The mode

The mode is the most frequently occurring number in a set of data.

The range

The range is calculated as the difference between the highest and lowest values in a set of data.

7 3 14 6 10 9 16 19 2 4 15 11 9 23 27

Mode = 9 (there are two of them and only one of everything else)

Range = 27 – 2
= 25

Problems with the mode and range

As a measure of central tendency, the most obvious problem with the mode is that a set of data may not contain a mode at all. Alternatively, there may be more than one mode in a data set, i.e. 'bi-modal' (two modes) or 'tri-modal' (three modes).

The main problem with using the range as a measure of dispersion is that it is distorted by extreme values.

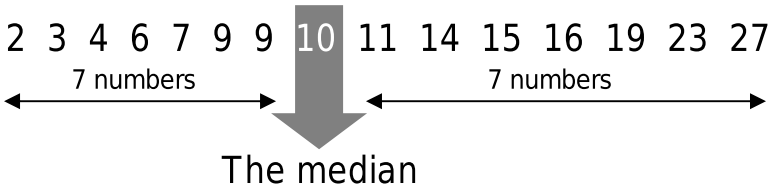
For example, if the (extreme) value of 1,002 is added to the previous set of data, the range increases from 25 to 1,000 - even though most of the numbers are 'clustered' at the lower end of the scale.

5.4. Median and inter-quartile range

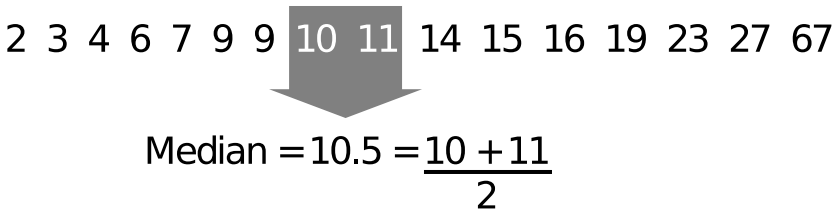
The median

The median is the value of the **middle** item in a set of data arranged in chronological order.

For example:

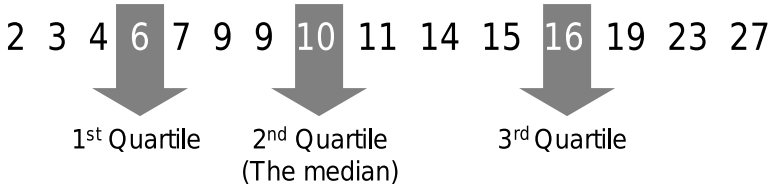


If the data set has an even number of values, then the median is equal to the average of the two middle items:

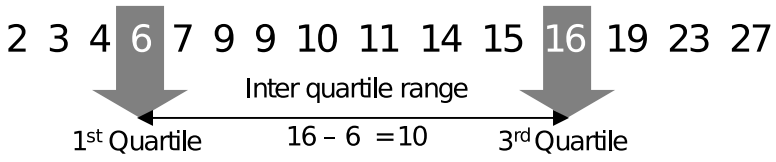


The inter-quartile range

The median is also known as the 'second quartile'. The middle item between the start of a series of numbers and the median is known as the 'first quartile'. The middle item between the median and the end of a series of numbers is known as the 'third quartile'.



The inter-quartile range is the third quartile minus the first quartile:



The inter-quartile range, therefore, is the 'spread' of the middle 50% of items in a data set. As such, it is **not** distorted by extreme values.

5.5. Distributions

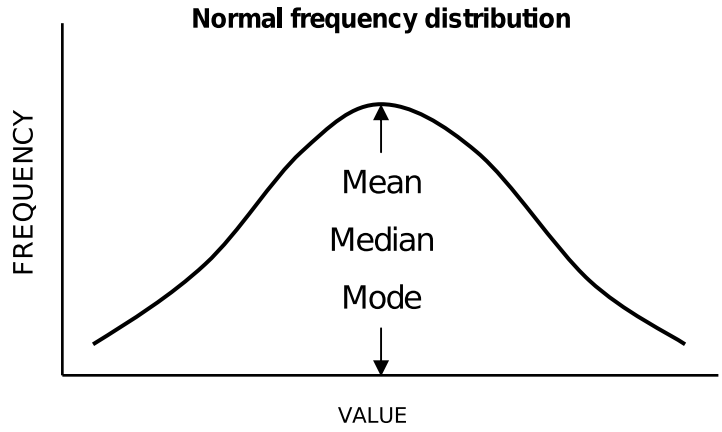
Normal (symmetrical) distributions

Should a large number of histograms be drawn from a wide range of data, a familiar pattern emerges. This pattern results in a high column in the centre of the histogram, with decreasing columns spread symmetrically on either side.

If the class intervals are small enough, the resultant frequency distribution curve will look like a cross-section of a bell, i.e. a 'bell-shaped' curve.

The bell-shaped curve is called the 'normal curve of distributions' (note: the standard deviation forms part of the normal curve for distributions. This is the reason why the standard deviation is so important in statistical analysis).

A normal frequency distribution curve is one where the mean, median and mode all have the same value:



It cannot be assumed, however, that distributions of returns will be normal. When extreme events occur more frequently than is predicted by the normal distribution it is referred to as the distribution having ‘fat tails’.

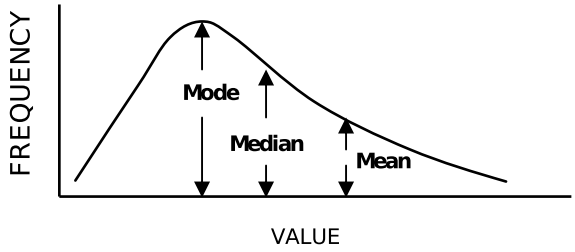
Skewed distributions

Data does not always conform to a normal and expected pattern. In such cases, the frequency distribution curves will be 'skewed'.

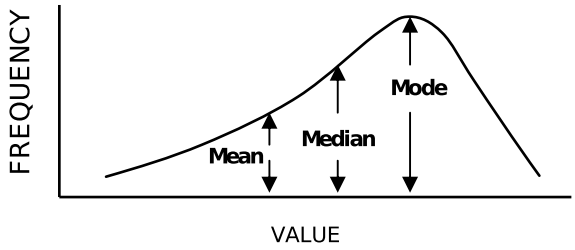
If the peak of the curve lies to the left of centre, it is said to be **positively** skewed.

If the peak of the curve lies to the right of centre, it is said to be **negatively** skewed.

Positively skewed frequency distribution



Negatively skewed frequency distribution



Note that with positively skewed distributions, the mode, median and mean are in **reverse** alphabetical order.

Statistical significance of normal distributions

Normal distribution can be useful when formulating investment decisions. Let's assume we want to buy shares in a company. In order to estimate what kind of return we can expect to earn in the future, we can analyse the historic returns by plotting their frequency. Assuming the result is close to a normal distribution and stock prices behave in the future, as they behaved in the past, we can predict the likelihood that we will earn the mean historic or expected return in the future.

The quality of this prediction - or its statistical significance - depends on two things:

- **The number of observations** (i.e. years of returns observed) underlying our analysis. The more observations, the better. A researcher can be more confident that a sample is representative of the true population if it is larger and has a smaller dispersion
- **The standard deviation** of the historic returns. The smaller the standard deviation, the less variable the returns. Statistically, 68% of all future returns will be within one standard deviation around the mean, 95% within two and 99.7% within three standard deviations around the mean or expected return
    - For example, if standard deviation is 0.5%, for example, we can be 99.7% sure that the returns will be within 1.5% (three standard deviations) of the mean.

The need for caution

As data can be (and often is) skewed, the mean and standard deviation can be a flawed in predicting returns. Equity markets are typically negatively skewed, describing the tendency for the market participants to put more weight on bad news than on good news. This must be considered when using standard deviation as a measure of risk and potential return.

5.6. Geometric mean

The geometric mean measures the **average rate of change** over a given period.

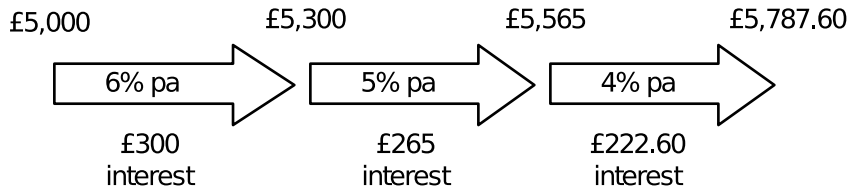
It is particularly useful when looking at compound changes, such as changes in a share price or changes in portfolio returns.

The geometric mean is defined as the nth root of the product of n numbers.

For instance, consider a deposit of £5,000 over three consecutive years at rates of 6% pa, 5% pa, and 4% pa respectively. The geometric mean rate of return is calculated as:

The answer is $\sqrt[3]{(1.06 \times 1.05 \times 1.04)} - 1 = 0.0499649968253$ (or 4.9968253%)

This means that the deposit's average rate of return over the three year period is equal to 4.9968% per year. The proof of this is shown below:

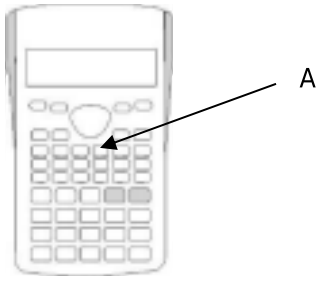


£5,000 x 1.06 x 1.05 x 1.04 = £5,787.60

This gives exactly the same answer as:

$5{,}000\ (1.049968253)^3 = £5{,}787.60$

The following illustration demonstrates how to calculate the geometric mean using the FX83 calculator.



Calculate the average rate of return from 6%, 5% and 4%.

Step 1. Multiply the price relatives together (this will calculate the amount of times bigger or smaller the £5,000 will be after each period of return). So, for example, the price relative of the 6% is 1.06; the £5,000 will be 1.06 times as big as before:

$$1.06 \times 1.05 \times 1.04 = 1.15752$$

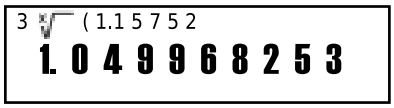
Step 2. There are three rates to average here, so take the 3rd root of this number. To do this on the FX83 calculator follow the key sequences below:

Button A above

[3] [SHIFT] [X] [1] [▪] [1] [5] [7] [5] [2] [=]

Activates the $\sqrt{\ }$ function

The display should look like this:



Step 3. The last thing is to subtract 1 from the answer giving

[-] [1] [=]

The answer of 0.049968253 is 4.9968253% which is the geometric mean.

In other words £5,000 for 3 years (with interest re-invested) at 4.9968 pct p.a. would give us a total of £5,787.60.

6. Correlation and bivariate linear regression

6.1. Correlation

The correlation coefficient measures the **strength** of the relationship between two variables, for example, the strength of the relationship between two share prices.

It is a useful tool when analysing the association between the two variables in a scattergram.

Positive correlation

Positive correlation describes a relationship where an increase in one variable is associated with an increase in another, e.g. frequency of advertising and sales.

Negative correlation

Negative correlation describes a relationship where an increase in one variable is associated with a decrease in another, e.g. sales of umbrellas and sun-tan lotion.

Perfect correlation

Perfect correlation describes a relationship where changes in one variable are reflected by a proportional change in another.

Perfect correlation only exists when all the points on a scattergram lie on the line of best fit.

Autocorrelation

Autocorrelation is the assessment of the correlation of an asset with itself, but staged over deferred time periods. For example, the correlation of returns of share A in 2012 and 2013 with the returns of share A in 2013 and 2014. This correlation can then be used to predict future behaviour of the asset. However, this does lead to a the risk of the asset being underestimated.

The shorter the time lag between the data sets, the stronger the correlation is likely to be.

6.2. Correlation and diversification

Achieving diversification

Diversification, and an associated risk reduction of a portfolio of securities, is achieved by combining securities which are **not** perfectly positively correlated.

For example, a fund manager choosing two securities with perfectly negative correlation of returns achieves a risk-free portfolio.

Risk reduction through diversification is therefore achieved by combining assets with a low (or negative) correlation of returns.

The lower the correlation of returns, the greater the fund's diversification and the lower the risk associated with an expected level of return.

The only instance when no diversification benefits are achieved is when there is a perfect positive correlation of returns.

Correlation does not necessarily imply causation

It should be noted that whilst correlation tells us that historically two variables have moved in similar patterns, this does not necessarily mean that one variable is causing the other to change or vice versa. It could be that both variables are being affected by a third factor or even that the pattern is a coincidence.

Data mining is the use of large subsets of information in order to discover relationships. The technique can be a valuable way to identify previously hidden causation but the large scale indiscriminate processing of data means that many correlations will be purely coincidental.

Extreme events and correlation

Correlation analysis and creating diversified portfolios is undoubtedly good practice. However, it must be noted that in times of extreme market conditions, these established relationships do break down and many assets become strongly positively correlated.

We see this in times of economic uncertainty and crisis, and in times of great optimism.

6.3. Calculating correlation coefficient

Correlation coefficient is calculated by dividing the covariance with the product of the two standard deviations. By virtue of its calculation, correlation will always be between +1 and -1.

$$\text{Correlation coefficient, } (\rho \text{ or } r) = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

Where:

$\text{Cov}(x,y) = \text{Covariance of } x \text{ and } y$

$\sigma_x = \text{Standard deviation of } x$

$\sigma_y = \text{Standard deviation of } y$

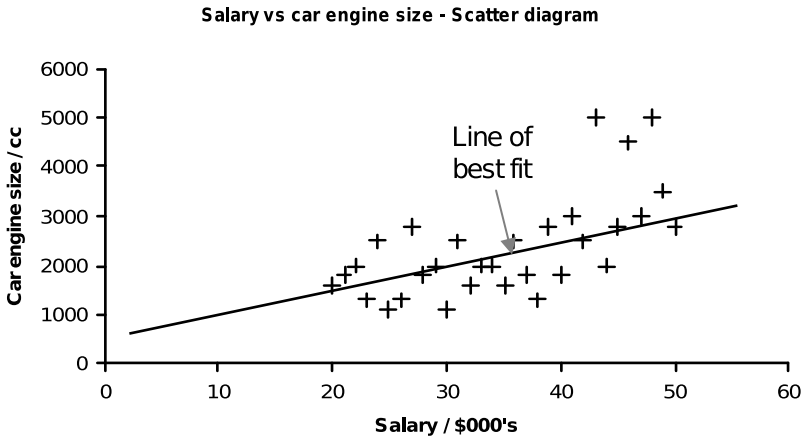
Covariance

The covariance (cov) is a statistical measure of the relationship between two variables, e.g. two share prices.

If the variables tend to move in the same direction, the covariance is positive. If the variables tend to move in the opposite directions, their covariance is negative. If the two variables are independent of each other, the covariance is zero.

6.4. Scattergrams (scatter diagrams)

Scattergrams are used to determine whether there is a **relationship** (correlation) between two variables. In this example, the scattergram illustrates the relationship between salary income and car engine size.



Note how the **dependent** variable (in this case, car engine size) lies along the y-axis and the **independent** variable (i.e. salary income) lies on the x-axis.

The purpose of a scattergram is to demonstrate whether there is any pattern among the plotted points.

6.5. Bivariate linear regression

The value of a scattergram is enhanced by adding a 'line of best fit'. This is the line that best fits the pattern of points.

The line's objective is to minimise the total divergence of the points from the line. The approach minimises the sum of the squares of the distances, and has the effect of giving extra importance to observation. This is often referred to as the 'least squares' method. The line of best fit is calculated by a mathematical process called 'linear regression'.

In the above scattergram, it is clear that there is some pattern, as the points tend to rise from left to right. When there are more data points close to the line of best fit it indicates a stronger correlation.

The equation for the line of best fit can be described as:

$y = a + bx$

Where:

- y is the dependent variable (car engine size in the example)
- x is the independent variable (salary in the example)
- a and b are coefficients of the equation

This equation can be used to either estimate Y outside the original range of values (extrapolation) or to fill in values within the sample range (interpolation).

Forecasting using correlation

Introduction

How useful is linear regression and the line of best fit in providing us with a useful insight into the future relationship between two variables?

The reliability of the tool used to forecast is partially dependent on the quality of data we have collected and the variability of that data. The more data collected on the relationship between the two variables, the better quality line we will get. The less the data deviates from the line of best fit, the more robust any predictions will be.

Regardless of the quality of data, linear regression is, by definition, a linear assumption based on a best fit (or more accurately a least bad fit).

Interpolation

Forecasting is not just looking to extrapolate information from outside the range of data collected, but also to interpolate information from between the values collected. Interest rates are often assumed to work on a linear basis. If the three-month rate is 4% and the six-month rate is 6%, can we then assume a linear relationship and say that the three-month rate beginning in three months' time will be 8%?

The logic being that the six-month rate should be the average of the two three-month rates, i.e. (4% + 8%) / 2 = 6%.

Extrapolation

If the range of data we collected looked at car engine sizes of between 750cc and 2500cc only, is it possible to extrapolate a predicted salary of someone that drives a 3000cc car? There is no guarantee that the linear relationship would hold beyond the range of values collected.

7. Statistics: summary

7.1. Key concepts

Types and sources of data

- 7.1.1 Primary and secondary sources of data
  - 7.1.2 Examples of primary and secondary data
  - 7.1.5 Continuous and discrete data
  - 7.1.6 Categorical data and explain how it can be converted to ordinal data

Populations and samples

- 7.1.3 A population and a sample
- 7.1.4 The key sampling methods

Presentation of data

- 7.1.8 A frequency and relative frequency distribution
- 7.1.8 The use of the following in the presentation of data: pie chart, bar chart, histogram, scatter plots, graphs

Summary statistics

- 7.2.1 The following measures of central tendency for both raw data and interval data: arithmetic mean, geometric mean, median, mode
- 7.2.4 The following measures of dispersion for both raw data and interval data: standard deviation (population and sample), variance, range, quartiles and percentiles, inter-quartile range
- 7.2.2 Symmetric and skewed data
- 7.2.3 The relationship between the mean, median and mode for symmetric and skewed data
- 7.2.5 The notion of probability distributions and identify the properties of the normal distribution
- 7.2.6 The notion of statistical significance in the context of investment decisions

Correlation and bivariate linear regression

- 7.3.3 The correlation coefficient in the context of linear regression
- 7.3.5 The concept of autocorrelation and describe the impact of extreme events on correlation
- 15.1.10 Calculate correlation coefficients from standard deviation/covariance of two investments
- 7.3.1 The least-squares regression technique in deriving a line of best fit
- 7.3.2 The forecast value for the dependent variable given the intercepts and gradients of a regression line equation

- 7.3.4 The shortfalls in the application of linear regression to forecasting, including why correlation does not imply causation, and the pitfalls of data-mining

Now you have finished this chapter you should attempt the chapter questions.