QUANTITATIVE METHODS FOR BUSINESS AND MANAGEMENT

QCF Level 5 Unit

Contents

Chapter	*Title*	*Page*
	Introduction to the Study Manual	**v**
	Unit Specification (Syllabus)	**vii**
	Coverage of the Syllabus by the Manual	**xi**
	Formulae and Tables Provided with the Examination Paper	**xiii**
1	**Data and Data Collection**	**1**
	Introduction	2
	Measurement Scales and Types of Data	3
	Collecting Primary Data	5
	Collecting Secondary Data	10
2	**Sampling Procedures**	**13**
	Introduction	14
	Statistical Inference	15
	Sampling	16
	Sampling Methods	18
	Choice of Sampling Method	23
3	**Tabulating and Graphing Frequency Distributions**	**25**
	Introduction	26
	Frequency Distributions	27
	Class Limits and Class Intervals	29
	Cumulative and Relative Frequency Distributions	31
	Ways of Presenting Frequency Distributions	34
	Presenting Cumulative Frequency Distributions	40
4	**Measures of Location**	**43**
	Introduction	44
	Use of Measures of Location	45
	Means	46
	Median	51
	Quantiles	54
	Mode	57
	Choice of Measure	58
	Appendix: Functions, Equations and Graphs	60

Chapter	*Title*	*Page*
5	**Measures of Dispersion**	**67**
	Introduction	68
	Range	69
	Quartile Deviation	70
	Standard Deviation and Variance	72
	Coefficient of Variation	75
	Skewness	76
6	**Index Numbers**	**79**
	Introduction	80
	Simple (Unweighted) Index Numbers	80
	Weighted index Numbers (Laspeyres and Paasche Indices)	83
	Fisher's Ideal Index	85
	Formulae	86
	Quantity or Volume Index Numbers	87
	Changing the Index Base Year	90
	Index Numbers in Practice	91
7	**Correlation**	**99**
	Introduction	100
	Scatter Diagrams	100
	The Correlation Coefficient	104
	Rank Correlation	108
8	**Linear Regression**	**113**
	Introduction	114
	Regression Lines	115
	Use of Regression	119
	Connection Between Correlation and Regression	119
	Multiple Regression	120
9	**Time Series Analysis**	**121**
	Introduction	122
	Structure of a Time Series	122
	Calculation of Component Factors for the Additive Model	126
	Multiplicative Model	135
	Forecasting	139
	The Z Chart	141
10	**Probability**	**143**
	Introduction	145
	Two Laws of Probability	146
	Permutations	149
	Combinations	152
	Conditional Probability	154
	Sample Space	155
	Venn Diagrams	157

Chapter	*Title*	*Page*
11	**Binomial and Poisson Distributions**	**173**
	Introduction	174
	The Binomial Distribution	175
	Applications of the Binomial Distribution	182
	Mean and Standard Deviation of the Binomial Distribution	184
	The Poisson Distribution	184
	Application of the Poisson Distribution	186
	Poisson Approximation to a Binomial Distribution	188
	Application of Binomial and Poisson Distributions – Control Charts	191
	Appendix: The Binomial Expansion	199
12	**The Normal Distribution**	**201**
	Introduction	202
	The Normal Distribution	202
	Use of the Standard Normal Table	206
	General Normal Probabilities	208
	Use of Theoretical Distributions	210
	Appendix: Areas in the Right-hand Tail of the Normal Distribution	214
13	**Significance Testing**	**215**
	Introduction	216
	Introduction to Sampling Theory	217
	Confidence Intervals	219
	Hypothesis Tests	221
	Significance Levels	228
	Small Sample Tests	229
14	**Chi-squared Tests**	**235**
	Introduction	236
	Chi-squared as a Test of Independence	236
	Chi-squared as a Test of Goodness of Fit	241
	Appendix: Area in the Right Tail of a Chi-squared (X^2) Distribution	245
15	**Decision-making**	**247**
	Introduction	248
	Decision-making Under Certainty	248
	Definitions	249
	Decision-making Under Uncertainty	250
	Decision-making Under Risk	252
	Complex Decisions	255
16	**Applying Mathematical Relationships to Economic and Business Problems**	**261**
	Using Linear Equations to Represent Demand and Supply Functions	262
	The Effects of a Sales Tax	267
	Breakeven Analysis	269
	Breakeven Charts	271
	The Algebraic Representation of Breakeven Analysis	275

Introduction to the Study Manual

Welcome to this study manual for Quantitative Methods for Business And Management.

The manual has been specially written to assist you in your studies for this QCF Level 5 Unit and is designed to meet the learning outcomes listed in the unit specification. As such, it provides thorough coverage of each subject area and guides you through the various topics which you will need to understand. However, it is not intended to "stand alone" as the only source of information in studying the unit, and we set out below some guidance on additional resources which you should use to help in preparing for the examination.

The syllabus from the unit specification is set out on the following pages. This has been approved at level 4 within the UK's Qualifications and Credit Framework. You should read this syllabus carefully so that you are aware of the key elements of the unit – the learning outcomes and the assessment criteria. The indicative content provides more detail to define the scope of the unit.

Following the unit specification is a breakdown of how the manual covers each of the learning outcomes and assessment criteria.

After the specification and breakdown of the coverage of the syllabus, we also set out the additional material which will be supplied with the examination paper for this unit. This is provided here for reference only, to help you understand the scope of the specification, and you will find the various formulae and rules given there fully explained later in the manual.

The main study material then follows in the form of a number of chapters as shown in the contents. Each of these chapters is concerned with one topic area and takes you through all the key elements of that area, step by step. You should work carefully through each chapter in turn, tackling any questions or activities as they occur, and ensuring that you fully understand everything that has been covered before moving on to the next chapter. You will also find it very helpful to use the additional resources (see below) to develop your understanding of each topic area when you have completed the chapter.

Additional resources

- ABE website – www.abeuk.com. You should ensure that you refer to the Members Area of the website from time to time for advice and guidance on studying and on preparing for the examination. We shall be publishing articles which provide general guidance to all students and, where appropriate, also give specific information about particular units, including recommended reading and updates to the chapters themselves.
- Additional reading – It is important you do not rely solely on this manual to gain the information needed for the examination in this unit. You should, therefore, study some other books to help develop your understanding of the topics under consideration. The main books recommended to support this manual are listed on the ABE website and details of other additional reading may also be published there from time to time.
- Newspapers – You should get into the habit of reading the business section of a good quality newspaper on a regular basis to ensure that you keep up to date with any developments which may be relevant to the subjects in this unit.
- Your college tutor – If you are studying through a college, you should use your tutors to help with any areas of the syllabus with which you are having difficulty. That is what they are there for! Do not be afraid to approach your tutor for this unit to seek clarification on any issue as they will want you to succeed!
- Your own personal experience – The ABE examinations are not just about learning lots of facts, concepts and ideas from the study manual and other books. They are also about how these are applied in the real world and you should always think how the

topics under consideration relate to your own work and to the situation at your own workplace and others with which you are familiar. Using your own experience in this way should help to develop your understanding by appreciating the practical application and significance of what you read, and make your studies relevant to your personal development at work. It should also provide you with examples which can be used in your examination answers.

And finally …

We hope you enjoy your studies and find them useful not just for preparing for the examination, but also in understanding the modern world of business and in developing in your own job. We wish you every success in your studies and in the examination for this unit.

Published by:
The Association of Business Executives
5th Floor, CI Tower
St Georges Square
New Malden
Surrey KT3 4TE
United Kingdom

Unit Specification (Syllabus)

The following syllabus – learning objectives, assessment criteria and indicative content – for this Level 5 unit has been approved by the Qualifications and Credit Framework.

Unit Title: Quantitative Methods for Business and Management

Guided Learning Hours: 160

Level: Level 5

Number of Credits: 18

Learning Outcome 1

The learner will: Understand different types of numerical data and different data collection processes, and be able to present data effectively for users in business and management.

Assessment Criteria *The learner can:*	**Indicative Content**
1.1 Explain the main sources and types of data and distinguish between alternative sampling methods and measurement scales.	1.1.1 Explain the main sources and types of data (including primary and secondary data, discrete and continuous data, quantitative and categorical data). 1.1.2 Compare and contrast alternative sampling methods and explain the main features of surveys, questionnaire design and the concept of sampling error and bias. 1.1.3 Distinguish between alternative measurement scales (nominal, ordinal, interval and ratio scales).
1.2 Construct appropriate tables and charts, and calculate and interpret a set of descriptive statistics.	1.2.1 Construct appropriate tables and charts, including frequency and cumulative frequency distributions and their graphical representations. 1.2.2 Calculate and interpret measures of location, dispersion, relative dispersion and skewness for ungrouped and grouped data.
1.3 Compute and interpret index numbers.	1.3.1 Compute unweighted and weighted index numbers and understand their applications. 1.3.2 Change the base period of an index number series.

Learning Outcome 2

The learner will: Understand the basic concepts of probability and probability distributions, and their applications in business and management.

Assessment Criteria *The learner can:*	**Indicative Content**
2.1 Demonstrate an understanding of the basic rules of probability and probability distributions, and apply them to compute probabilities.	2.1.1 Demonstrate an understanding of the basic rules of probability. 2.1.2 Explain the conditions under which the binomial and Poisson distributions may be used and apply them to compute probabilities.

	2.1.3 Explain the characteristics of the normal distribution and apply it to compute probabilities.
2.2 Explain and discuss the importance of sampling theory and the central limit theorem and related concepts.	2.2.1 Explain and discuss the importance of sampling theory and the sampling distribution of the mean. 2.2.2 Discuss the importance of the central limit theorem. 2.2.3 Define the 'standard error of the mean'.
2.3 Construct and interpret confidence intervals and conduct hypothesis tests.	2.3.1 Construct and interpret confidence intervals, using the normal or t distribution, as appropriate, and calculate the sample size required to estimate population values to within given limits. 2.3.2 Conduct hypothesis tests of a single mean, a single proportion, the difference between two means and the difference between two proportions. 2.3.3 Conduct chi-squared tests of goodness-of-fit and independence and interpret the results.

Learning Outcome 3

The learner will: Understand how to apply statistical methods to investigate inter-relationships between, and patterns in, business variables.

Assessment Criteria *The learner can:*	**Indicative Content**
3.1 Construct scatter diagrams and calculate and interpret correlation coefficients between business variables.	3.1.1 Construct scatter diagrams to illustrate linear association between two variables and comment on the shape of the graph. 3.1.2 Calculate and interpret Pearson's coefficient of correlation and Spearman's 'rank' correlation coefficient and distinguish between correlation and causality.
3.2 Estimate regression coefficients and make predictions.	3.2.1 Estimate the regression line for a two-variable model and interpret the results from simple and multiple regression models. 3.2.2 Use an estimated regression equation to make predictions and comment on their likely accuracy.
3.3 Explain the variations in time-series data, estimate the trend and seasonal factors in a time series and make business forecasts.	3.3.1 Distinguish between the various components of a time series (trend, cyclical variation, seasonal variation and random variation). 3.3.2 Estimate a trend by applying the method of moving averages and simple linear regression. 3.3.3 Apply the additive and multiplicative models to estimate seasonal factors. 3.3.4 Use estimates of the trend and seasonal factors to forecast future values (and comment on their likely accuracy) and to compute seasonally-adjusted data

Learning Outcome 4

The learner will: Understand how statistics and mathematics can be applied in the solution of economic and business problems.

Assessment Criteria *The learner can:*	**Indicative Content**
4.1 Construct probability trees and decision trees and compute and interpret EMVs (Expected Monetary Values) as an aid to business decision-making under conditions of uncertainty.	4.1.1 Explain and calculate expected monetary values and construct probability trees. 4.1.2 Construct decision trees and show how they can be used as an aid to business decision-making in the face of uncertainty. 4.1.3 Discuss the limitations of EMV analysis in business decision-making.
4.2 Construct demand and supply functions to determine equilibrium prices and quantities, and analyse the effects of changes in the market.	4.2.1 Use algebraic and graphical representations of demand and supply functions to determine the equilibrium price and quantity in a competitive market. 4.2.2 Analyse the effects of changes in the market (e.g. the imposition of a sales tax) on the equilibrium price and quantity.
4.3 Apply, and explain the limitations of, break-even analysis to determine firms' output decisions, and analyse the effects of cost and revenue changes.	4.3.1 Apply break-even analysis to determine the output decisions of firms and to analyse the effects of changes in the cost and revenue functions. 4.3.2 Discuss the importance and explain the limitations of simple break-even analysis.

Coverage of the Syllabus by the Manual

Learning Outcomes *The learner will:*	***Assessment Criteria*** *The learner can:*	***Manual*** *Chapter*
1. Understand different types of numerical data and different data collection processes, and be able to present data effectively for users in business and management.	1.1 Explain the main sources and types of data and distinguish between alternative sampling methods and measurement scales	*Chaps 1 & 2*
	1.2 Construct appropriate tables and charts, and calculate and interpret a set of descriptive statistics	*Chaps 3 – 5*
	1.3 Compute and interpret index numbers	*Chap 6*
2. Understand the basic concepts of probability and probability distributions, and their applications in business and management.	2.1 Demonstrate an understanding of the basic rules of probability and probability distributions, and apply them to compute probabilities	*Chaps 10 – 12*
	2.2 Explain and discuss the importance of sampling theory and the central limit theorem and related concepts	*Chap 13*
	2.3 Construct and interpret confidence intervals and conduct hypothesis tests	*Chaps 13 & 14*
3. Understand how to apply statistical methods to investigate inter-relationships between, and patterns in, business variables.	3.1 Construct scatter diagrams and calculate and interpret correlation coefficients between business variables	*Chap 7*
	3.2 Estimate regression coefficients and make predictions	*Chap 8*
	3.3 Explain the variations in time-series data, estimate the trend and seasonal factors in a time series and make business forecasts	*Chap 9*
4. Understand how statistics and mathematics can be applied in the solution of economic and business problems.	4.1 Construct probability trees and decision trees and compute and interpret EMVs (Expected Monetary Values) as an aid to business decision making under conditions of uncertainty	*Chap 15*
	4.2 Construct demand and supply functions to determine equilibrium prices and quantities and analyse the effects of changes in the market	*Chap 16*
	4.3 Apply (and explain the limitations of) break-even analysis to determine firms' output decisions and analyse the effects of cost and revenue changes	*Chap 16*

Formulae and Tables Provided with the Examination Paper

FORMULAE

Mean of ungrouped data:

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

Geometric mean of ungrouped data:

$$GM = \sqrt[n]{\Pi x}$$

where: Π = "the product of …"

Mean of grouped data:

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

Median of grouped data:

$$\text{median} = L + \left(\frac{\frac{n}{2} - F}{f} \right) i$$

where: L = lower boundary of the median class
F = cumulative frequency up to the median class
f = frequency of the median class
i = width of the median class.

Mode of grouped data:

$$\text{mode} = L + \left(\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \right) i$$

where: L = lower boundary of the modal class
f_m = frequency of the modal class
f_{m-1} = frequency of the pre-modal class
f_{m+1} = frequency of the postmodal class
i = width of the modal class.

Standard deviation of ungrouped data:

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

Standard deviation of grouped data:

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

Coefficient of skewness:

$$\text{Sk} = \frac{3\left(\bar{x} - \tilde{x}\right)}{s}$$

where: $\tilde{x}$ = median

s = standard deviation

Regression:

$$\hat{y} = a + bx$$

$$b = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - \left(\Sigma x\right)^2}$$

$$a = \bar{y} - b\bar{x}$$

Pearson correlation:

$$R = \frac{n\sum xy - \sum x\sum y}{\sqrt{[n\sum x^2 - \left(\sum x\right)^2][n\sum y^2 - \left(\sum y\right)^2]}}$$

$$R = b\frac{\sigma_x}{\sigma_y}$$

Spearman's rank correlation:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

Laspeyres price index:

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

Paasche price index:

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Binomial distribution:

$$P(x) = {}_nC_x p^x q^{n-x}$$

Poisson distribution:

$$P(x) = \frac{e^{-\mu}\mu^x}{x!}$$

Standard normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

Confidence interval for a mean:

$$\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

Confidence interval for a proportion:

$$\pi = p \pm z\sqrt{\frac{pq}{n}}$$

Test statistic for a single mean:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Test statistic for a difference between means:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Test statistic for a single proportion:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$$

Test statistic for a difference between proportions:

$$z = \frac{p_1 - p_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

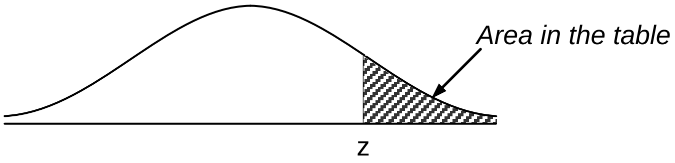
where: $\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$\hat{q} = 1 - \hat{p}$

Chi-squared test statistic:

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

Areas in the Right-Hand Tail of the Normal Distribution



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1496	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1132	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
3.0	.00135									
3.1	.00097									
3.2	.00069									
3.3	.00048									
3.4	.00034									
3.5	.00023									
3.6	.00016									
3.7	.00011									
3.8	.00007									
3.9	.00005									
4.0	.00003									

Chi-Squared Critical Values

df	p value											
	0.25	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.59	6.74	7.78	9.49	11.14	11.67	13.23	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.33	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.53	14.45	15.03	16.81	13.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.63	21.67	23.59	25.46	27.83	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.29	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.93	16.98	18.90	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.40	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	53.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.70
80	88.13	90.41	93.11	96.58	101.90	106.60	108.10	112.30	116.30	120.10	124.80	128.30
100	109.10	111.70	114.70	118.50	124.30	129.60	131.10	135.80	140.20	144.30	149.40	153.20

Chapter 1

Data and Data Collection

Contents ***Page***

A.	**Introduction**	**2**
	The Role of Quantitative Methods in Business and Management	2
	Statistics	2
B.	**Measurement Scales and Types of Data**	**3**
	Measurement Scales	3
	Variables and Data	4
C.	**Collecting Primary Data**	**5**
	Interviews	5
	Advantages of Interviewing	6
	Disadvantages of Interviewing	6
	Self-Completion Questionnaires	7
	Advantages of Self-Completion Questionnaires	8
	Disadvantages of Self-Completion Questionnaires	9
	Non-response Bias and Sampling Error	9
	Personal Observation	9
D.	**Collecting Secondary Data**	**10**
	Scanning Published Data	10
	Internal Data Sources	10
	External Data Sources	11
	ONS Publications	12
	Annual Business Inquiry	12

A. INTRODUCTION

The Role of Quantitative Methods in Business and Management

Quantitative methods play an important role both in business research and in the practical solution of business problems. Managers have to take decisions on a wide range of issues, such as:

- how much to produce
- what prices to charge
- how many staff to employ
- whether to invest in new capital equipment
- whether to fund a new marketing initiative
- whether to introduce a new range of products
- whether to employ an innovative method of production.

In all of these cases, it is clearly highly desirable to be able to compute the likely effects of the decisions on the company's costs, revenues and, most importantly, profits. Similarly, it is important in business research to be able to use data from samples to estimate parameters relating to the population as a whole (for example, to predict the effect of introducing a new product on sales throughout the UK from a survey conducted in a few selected regions). These sorts of business problems require the application of statistical methods such as:

- time-series analysis and forecasting
- correlation and regression analysis
- estimation and significance testing
- decision-making under conditions of risk and uncertainty
- break-even analysis.

These methods in turn require an understanding of a range of summary statistics and concepts of probability. These topics therefore form the backbone of this course.

Statistics

Most of the quantitative methods mentioned above come under the general heading of statistics. The term "statistics" of course is often used to refer simply to a set of data – so, for example, we can refer to a country's unemployment statistics (which might be presented in a table or chart showing the country's unemployment rates each year for the last few years, and might be broken down by gender, age, region and/or industrial sector, etc.). However, we can also use the term "Statistics" (preferably with a capital letter) to refer to the academic discipline concerned with the *collection, description, analysis and interpretation of numerical data*. As such, the subject of Statistics may be divided into two main categories:

(a) Descriptive Statistics

This is mainly concerned with collecting and summarising data, and presenting the results in appropriate tables and charts. For example, companies collect and summarise their financial data in tables (and occasionally charts) in their annual reports, but there is no attempt to go "beyond the data".

(b) Statistical Inference

This is concerned with analysing data and then interpreting the results (attempting to go "beyond the data"). The main way in which this is done is by collecting data from a sample and then using the sample results to infer conclusions about the population. For example, prior to general elections in the UK and many other countries, statisticians conduct opinion polls in which samples of potential voters are asked which political party they intend to vote for. The sample proportions are then used to predict the voting intentions of the entire population.

Of course, before any descriptive statistics can be calculated or any statistical inferences made, appropriate data has to be collected. We will start the course, therefore, by seeing how we collect data. This chapter looks at the various types of data, the main sources of data and some of the numerous methods available to collect data.

B. MEASUREMENT SCALES AND TYPES OF DATA

Measurement Scales

Quantitative methods use quantitative data which consists of measurements of various kinds. Quantitative data may be measured in one of four measurement scales, and it is important to be aware of the measurement scale that applies to your data before commencing any data description or analysis. The four *measurement scale*s are:

(a) Nominal Scale

The nominal scale uses numbers simply to identify members of a group or category. For example, in a questionnaire, respondents may be asked whether they are male or female and the responses may be given number codes (say 0 for males and 1 for females). Similarly, companies may be asked to indicate their ownership form and again the responses may be given number codes (say 1 for public limited companies, 2 for private limited companies, 3 for mutual organizations, etc.). In these cases, the numbers simply indicate the group to which the respondents belong and have no further arithmetic meaning.

(b) Ordinal Scale

The ordinal scale uses numbers to rank responses according to some criterion, but has no unit of measurement. In this scale, numbers are used to represent "more than" or "less than" measurements, such as preferences or rankings. For example, it is common in questionnaires to ask respondents to indicate how much they agree with a given statement and their responses can be given number codes (say 1 for "Disagree Strongly", 2 for "Disagree", 3 for "Neutral", 4 for "Agree" and 5 for "Agree Strongly"). This time, in addition to indicating to which category a respondent belongs, the numbers measure the degree of agreement with the statement and tell us whether one respondent agrees more or less than another respondent. However, since the ordinal scale has no units of measurement, we *cannot* say that the difference between 1 and 2 (i.e. between disagreeing strongly and just disagreeing) is the same as the difference between 4 and 5 (i.e. between agreeing and agreeing strongly).

(c) Interval Scale

The interval scale has a constant unit of measurement, but an arbitrary zero point. Good examples of interval scales are the Fahrenheit and Celsius temperature scales. As these scales have different zero points (i.e. 0 degrees F is not the same as 0 degrees C), it is not possible to form meaningful ratios. For example, although we can say that 30 degrees C (86 degrees F) is hotter than 15 degrees C (59 degrees F), we *cannot* say that it is twice as hot (as it clearly isn't in the Fahrenheit scale).

(d) Ratio Scale

The ratio scale has a constant unit of measurement and an absolute zero point. So this is the scale used to measure values, lengths, weights and other characteristics where there are well-defined units of measurement and where there is an absolute zero where none of the characteristic is present. For example, in values measured in pounds, we know (all too well) that a zero balance means no money. We can also say that £30 is twice as much as £15, and this would be true whatever currency were used as the unit of measurement. Other examples of ratio scale measurements include the average petrol consumption of a car, the number of votes cast at an election, the percentage return on an investment, the profitability of a company, and many others.

The measurement scale used gives us one way of distinguishing between different types of data. For example, a set of data may be described as being "nominal scale", "ordinal scale", "interval scale" or "ratio scale" data. More often, a simpler distinction is made between *categorical* data (which includes all data measured using nominal or ordinal scales) and *quantifiable* data (which includes all data measured using interval or ratio scales).

Variables and Data

Any characteristic on which observations can be made is called a *variable* or *variate*. For example, height is a variable because observations taken are of the heights of a number of people. Variables, and therefore the data which observations of them produce, can be categorised in various ways:

(a) Quantitative and Qualitative Variables

Variables may be either quantitative or qualitative. Quantitative variables, to which we shall restrict discussion here, are those for which observations are numerical in nature. Qualitative variables have non-numeric observations, such as colour of hair, although of course each possible non-numeric value may be associated with a numeric frequency.

(b) Continuous and Discrete Variables

Variables may be either continuous or discrete. A *continuous variable* may take *any value* between two stated limits (which may possibly be minus and plus infinity). Height, for example, is a continuous variable, because a person's height may (with appropriately accurate equipment) be measured to any minute fraction of a millimetre. A *discrete variable* however can take only *certain values* occurring at intervals between stated limits. For most (but not all) discrete variables, these intervals are the set of integers (whole numbers).

For example, if the variable is the number of children per family, then the only possible values are 0, 1, 2, ... etc., because it is impossible to have other than a whole number of children. However in Britain shoe sizes are stated in half-units, and so here we have an example of a discrete variable which can take the values 1, 1½, 2, 2½, etc.

You may possibly see the difference between continuous and discrete variables stated as "continuous variables are measured, whereas discrete variables are counted". While this is possibly true in the vast majority of cases, you should not simply state this if asked to give a definition of the two types of variables.

(c) Primary and Secondary Data

If data is collected for a *specific* purpose then it is known as *primary data*. For example, the information collected direct from householders' television sets through a microcomputer link-up to a mainframe computer owned by a television company is used to decide the most popular television programmes and is thus primary data. The Census of Population, which is taken every ten years, is another good example of primary data because it is collected specifically to calculate facts and figures in relation to the people living in the UK**.**

Secondary data is data which has been collected for some purpose *other* than that for which it is being used. For example, if a company has to keep records of when employees are sick and you use this information to tabulate the number of days employees had flu in a given month, then this information would be classified as secondary data.

Most of the data used in compiling business statistics is secondary data because the source is the accounting, costing, sales and other records compiled by companies for administration purposes. Secondary data must be used with *great care*; as the data was collected for another purpose, and you must make sure that it provides the information that you require. To do this you must look at the sources of the information, find out how it was collected and the exact definition and method of compilation of any tables produced.

(d) Cross-Section and Time-Series Data

Data collected from a sample of units (e.g. individuals, firms or government departments) for a single time period is called *cross-section data*. For example, the test scores obtained by 20 management trainees in a company in 2007 would represent a sample of cross-section data. On the other hand, data collected for a single unit (e.g. a single individual, firm or government department) at multiple time periods are called *time-series data*. For example, annual data on the UK inflation rate from 1985–2007 would represent a sample of time-series data. Sometimes it is possible to collect cross-section over two or more time periods – the resulting data set is called a *panel data* or *longitudinal data* set.

C. COLLECTING PRIMARY DATA

There are three main methods of collecting primary data: by interviews, by self-completion questionnaires or by personal observations. These three methods are discussed below.

Interviews

Interviewing is a common method of collecting information in which interviewers question people on the subject of the survey. Interviews can be face-to-face or conducted by telephone. Face-to-face interviews are relatively expensive, but offer the opportunity for the interviewer to explain questions and to probe more deeply into any answers given. Interviews by telephone are less personal but can be useful if time is short.

Interviews may be structured, semi-structured or unstructured:

(a) Structured Interviews

In a structured interview, the interviewer usually has a well-defined set of prepared questions (i.e. a questionnaire) in which most of the questions are "closed" (i.e. each question has a predetermined set of options for the response, such as a box to be ticked). The design of such questionnaires is essentially the same as that discussed below under the heading *Self-Completion Questionnaires*. Structured interviewing is useful if the information being sought is part of a clearly-defined business research project (such as market research), and if the aim of the survey is to collect numerical data suitable for statistical analysis.

(b) Semi-Structured Interviews

In a semi-structured interview, the interviewer has a set of prepared questions, but is happy to explore other relevant issues raised by the interviewee.

(c) Unstructured Interviews

In unstructured interviews, the interviewer does not have a set of prepared questions and the emphasis is often on finding out the interviewee's point of view on the subject of the survey. Unstructured interviews are more commonly used in qualitative (rather than quantitative) research, though they can also be useful as *pilot studies*, designed to help a researcher formulate a research problem.

Advantages of Interviewing

There are many advantages of using interviewers in order to collect information:

(a) The major one is that a large amount of data can be collected relatively *quickly and cheaply*. If you have selected the respondents properly and trained the interviewers thoroughly, then there should be few problems with the collection of the data.

(b) This method has the added advantage of being very *versatile* since a good interviewer can adapt the interview to the needs of the respondent. If, for example, an aggressive person is being interviewed, then the interviewer can adopt a conciliatory attitude to the respondent; if the respondent is nervous or hesitant, the interviewer can be encouraging and persuasive.

The interviewer is also in a position to explain any question, although the amount of explanation should be defined during training. Similarly, if the answers given to the question are not clear, then the interviewer can ask the respondent to elaborate on them. When this is necessary the interviewer must be *very careful* not to lead the respondent into altering rather than clarifying the original answers. The technique for dealing with this problem must be tackled at the training stage.

(c) This face-to-face technique will usually produce a high response rate. The response rate is determined by the proportion of interviews that are successful. A successful interview is one that produces a questionnaire with every question answered clearly. If most respondents interviewed have answered the questions in this way, then a high response rate has been achieved. A low response rate is when a large number of questionnaires are incomplete or contain useless answers.

(d) Another advantage of this method of collecting data is that with a well-designed questionnaire it is possible to ask a large number of short questions in one interview. This naturally means that the cost per question is lower than in any other method.

Disadvantages of Interviewing

Probably the biggest disadvantage of this method of collecting data is that the use of a large number of interviewers leads to a *loss of direct control* by the planners of the survey. Mistakes in selecting interviewers and any inadequacy of the training programme may not be recognised until the interpretative stage of the survey is reached. This highlights the need to train interviewers correctly.

It is particularly important to ensure that all interviewers ask questions in a similar way. It is possible that an inexperienced interviewer, just by changing the tone of voice used, may give a different emphasis to a question than was originally intended. This problem will sometimes become evident if unusual results occur when the information collected is interpreted.

In spite of these difficulties, this method of data collection is widely used as questions can be answered cheaply and quickly and, given the correct approach, this technique can achieve high response rates.

Self-Completion Questionnaires

Self-completion questionnaires are completed by the respondents with no help from an interviewer. They may be administered by post, by email or by directing potential respondents to a website.

The design of a questionnaire will reflect the way in which it is to be used. Many problems can be avoided by careful design, particularly where the information on the questionnaire has to be transferred to analysis sheets or entered into a computer. For example, if all the responses are aligned down one side of the sheet it is a great deal easier to read them off than if they are scattered around the sheet.

Overall a questionnaire form should not look too overpowering: good layout can improve response considerably. Equally questionnaires should be kept as short as possible (unless there is a legal compulsion to fill it in, as with many government surveys), as a multi-page questionnaire will probably be put on one side and either forgotten or returned late.

The above discussion only touches on a few of the considerations in designing a questionnaire; hopefully it will make you think about what is involved. Professional help is a good idea when designing a questionnaire.

The general principle to keep in mind when designing a set of questions is that, if a question can be misread, it will be. Questions must always be tested on someone who was not involved in setting them, and preferably on a small sample of the people they will be sent to. Testing a new questionnaire on a small sample of potential respondents is sometimes referred to as a *pilot study*.

The principles to observe when designing a questionnaire are:

(a) Keep it as short as possible, consistent with getting the right results.

(b) Explain the purpose of the investigation so as to encourage people to give answers.

(c) Individual questions should be as short and simple as possible.

(d) If possible, only short and definite answers like "Yes", "No" or a number of some sort should be called for.

(e) Questions should be capable of only one interpretation, and leading questions should be avoided.

(f) Where possible, use the "alternative answer" system in which the respondent has to choose between several specified answers.

(g) The questions should be asked in a logical sequence.

(h) The respondent should be assured that the answers will be treated confidentially and not be used to his or her detriment.

(i) No calculations should be required of the respondent.

You should always apply these principles when designing a questionnaire, and you should understand them well enough to be able to remember them all if you are asked for them in an examination question. They are *principles* and not rigid rules – often you have to break some of them in order to get the right information. Governments often break these principles because they can make the completion of the questionnaire compulsory by law, but other investigators must follow the rules as far as practicable in order to make the questionnaire as easy and simple to complete as possible – otherwise they will receive no replies.

If the questionnaire is to be used for a *structured interview*, then the task of collecting the information will be entrusted to a team of interviewers. These interviewers must be trained in the use of the questionnaire and advised how to present it so that *maximum* cooperation is obtained from the respondent. This training is very important and must be carefully thought out. The interviewers must be carefully selected so that they will be suitable for the type of interview envisaged. The type of interviewer and the method of approach must be varied according to the type of respondent selected, e.g. the same technique should not be used for interviewing students and senior bank staff.

What follows is an example of a simple questionnaire:

1. Please tick your sex.	Male	□
	Female	□
2. Which age bracket do you fall in?	Under 25 yrs	□
	25 yrs – under 45 yrs	□
	45 yrs – under 65 yrs	□
	Over 65 yrs	□
3. Which subjects do you enjoy studying most? *You may tick more than one box*.	Maths	□
	Languages	□
	Arts	□
	Sciences	□
	Don't enjoy studying	□
4. Which style of education do you prefer?	Full-time	□
	Part-time/Day release	□
	Evening classes	□
	Correspondence courses	□
	Self-tuition	□
	Other	□
	No preference	□
5. How do you feel at this stage of the course?	Very confident	□
	Confident	□
	Not sure	□
	Unconfident	□
	Very unconfident	□

Your assistance in this matter will help our researchers a great deal. Thank you for your cooperation.

Advantages of Self-Completion Questionnaires

This technique has a number of advantages, the major one being its *cheapness*. As there are no interviewers, the only direct cost is that of the postage. This means that the questionnaires can be distributed to a *wider range* of respondents at a cheaper rate, and this may increase the response rate.

This type of data collection allows the respondents *plenty of time* to consider their answers. Compare this with the interviewing technique where the interviewer requires an immediate response.

The final advantage is the *elimination of interviewer bias*, as even some of the best-trained interviewers will tend to put their own slant on any interview. In some cases, if the interviewer is biased or inadequately trained, this can lead to serious distortion of the results.

Disadvantages of Self-Completion Questionnaires

The major disadvantage of this method of data collection is the inability of the planners to control the number of responses: some respondents will not bother to reply, and others will feel that they are not qualified to reply. For example, if questionnaires about fast motor cars were sent to a cross section of the population, then only those people who owned a fast motor car might return the questionnaire. People without fast cars might think the questionnaire did not apply to them and consequently would not send it back. Therefore, as the percentage of people returning the questionnaire is very low, the *response rate is low*.

This situation can be improved either by sending out a very large number of questionnaires (so that even though the actual response rate is low, the number responding is high enough for the purpose of the survey) or by offering some form of incentive (such as a lottery prize) for the return of the form. Both of these methods would involve an increase in cost which would counteract the greatest advantage of this method, that of cheapness.

The problem introduced by the first method (sending out a very large number of questionnaires) is that even though the number of responses is sufficient, they do not represent the views of a typical cross section of the people first approached. For example, very few replies would be received from those not owning fast motor cars, so that any deductions drawn from the data about the targeted cross section of the population would be biased. So, you can see that you have very little control over the response rate with this method of collection. As there are no interviewers, you have to rely on the *quality* of the questionnaire to encourage the respondents to cooperate.

This means that great care has to be taken with the design of the questionnaire. In particular it is extremely important that the wording of the questions is very simple, and any question that could be interpreted in more than one way should be left out or reworded. The required answers should be a simple yes/no or at the most a figure or date. You should not ask questions that require answers expressing an attitude or opinion while using this particular technique.

Finally, it is important to remember that this type of data collection *takes much longer* to complete than the other methods described. Experience shows that about 15 per cent of the questionnaires sent out will be returned within a week, but the next 10 per cent (bringing the response rate up to a typical 25 per cent), may take anything up to a month before they come back.

Non-response Bias and Sampling Error

The results obtained from a questionnaire survey may be biased (and therefore not representative of the relevant population) if those who fail to respond to the questionnaire differ in any important and relevant ways from those who do respond. For example, if the residents of a town are questioned about the desirability of a new bypass, the people most likely to respond may be those who are currently most affected by traffic congestion and who tend to favour the construction of the bypass. This type of bias is called *non-response bias.* If a sample fails to be representative of the population just by chance, then it is said to exhibit *sampling error.*

Personal Observation

This method is used when it is possible to *observe directly* the information that you wish to collect. For example, data for traffic surveys is collected in this way: observers stand by the roadside and count and classify the vehicles passing in a given time. Increasingly, computers and automated equipment are replacing human observers in this method of data collection as they are considerably cheaper and often more reliable. There are numerous examples of this. For instance, most traffic information is now collected by sensors in rubber tubes laid across the road and linked to a small computer placed alongside the road.

The main advantage of this method of data collection is that the data is observed directly instead of being obtained from other sources. However, when observers are used, you must allow for human error and personal bias in the results. Although this type of bias or error is easy to define, it is sometimes extremely difficult to recognise and even harder to measure the degree to which it affects the results. Personal bias can be more of a problem when only part of the data available is being observed.

This problem will be covered in greater detail in a later chapter which deals with sampling.

Provided proper and accurate instructions are given to the observers in their training, this type of bias can be reduced to a minimum.

Personal observation means that the data collected is usually limited by the resources available. It can often be expensive, especially where large amounts of data are required. For these reasons this method is not widely used. However, the increasing use of the computer is reducing both the amount of bias and the cost.

D. COLLECTING SECONDARY DATA

It is important to consult published sources before deciding to go out and collect your own data, to see if all or part of the information you require is already available. Published sources provide valuable access to secondary data for use in business and management research. We will now describe where to look for business data. You will often find useful information from several sources, both within an organisation and outside.

Scanning Published Data

When you examine published data from whatever source, it is helpful to adopt the following procedure:

(a) Overview the whole publication

Flip through the pages so that you get a feel for the document. See if it contains tables only, or if it uses graphs and tables to describe the various statistics.

(b) Look at the Contents pages

A study of the contents pages will show you exactly what the document contains and give you a good idea of the amount of detail. It will also show you which variables are described in the tables and charts.

(c) Read the Introduction

This will give a general indication of the origin of the statistics in the document. It may also describe how the survey which collected the information was carried out.

(d) Look at part of the Document in Detail

Take a small section and study that in depth. This will give you an appreciation of just what information is contained and in what format. It will also get you used to studying documents and make you appreciate that most tables, graphs or diagrams include some form of notes to help explain the data.

Internal Data Sources

All types of organisation will collect and keep data which is therefore internal to the organisation. More often than not it applies to the organisation where you work, but you should not think of it as meaning just that type of organisation. It is important when looking for some particular types of data to look internally because:

- It will be *cheaper* if the data can be obtained from an internal source as it will save the expense of some form of survey.
- Readily available information can be used much more *quickly* especially if it has been computerised and can be easily accessed.
- When the information is available from within your own organisation, it can be *understood* much more easily as supporting documentation is likely to be readily available.

Overall there are several advantages from using internal data, although there is a tendency when using this type of data to make do with something that is *nearly* right.

Companies' annual reports provide a particularly useful set of data for financial and business research.

External Data Sources

The sources of statistical information can be conveniently classified as:

- central and local government sources together with EU publications
- private sources.

The data produced by these sources can be distinguished as:

- Data collected *specifically* for statistical purposes – e.g. the population census.
- Data arising as a by-product of other functions – e.g. unemployment figures.

This latter distinction is well worth noting because it sometimes helps to indicate the degree of reliability of the data. Do not forget, of course, that very often the statistician has to be his or her own source of information; then he or she must use the techniques of primary data collection which we have already discussed.

The main producer of statistics in the UK is central government, and for this purpose an organisation has been set up called the Office for National Statistics (ONS). The ONS exists primarily to service the needs of central government. However, much of the information it produces is eminently suitable for use by the business community as well, and indeed central government is increasingly becoming aware of the need to gear its publications so that they can be used by the business sector.

Local government also produces a wealth of information, but because of its localised nature it is not often found on the shelves of all libraries or made available outside the area to which it applies. Another information source is the European Union (EU), and data is increasingly becoming available both in printed form and online. Similarly, the United Nations publications and websites are available, which cover world-wide statistics in subjects such as population and trade.

Companies also provide useful financial data in their annual reports and accounts, most of which are now available online, through one of the financial databases, such as Datastream.

ONS Publications

The principal statistics provided by the ONS can be found on the ONS website (www.statistics.gov.uk) and in various publications.

A summary of some of the major ONS publications is given below.

(a) General

Publication	*Description*
Annual Abstract of Statistics	Main economic and social statistics for the UK
Monthly Digest of Statistics	Main economic and social statistics for the UK
Regional Trends (annual)	Main economic and social statistics for regions of the UK

(b) National Income and Expenditure

Publication	*Description*
UK National Accounts (Blue Book) (annual)	National account statistics
Economic Trends (monthly)	Primary statistics on the current economic situation

(c) Other

Publication	*Description*
Financial Statistics (monthly)	UK monetary and financial statistics
Social Trends	Social conditions statistics
UK Balance of Payments	Balance of payments and trade statistics

Annual Business Inquiry

The annual survey of production in the UK, called the *Annual Business Inquiry*, collects employment and financial information covering about two-thirds of the UK economy. It includes manufacturing, construction, wholesale and retail trades, catering, property, services, agriculture, forestry and fishing. The results are used to compile the UK input-output tables in the National Accounts, to re-base the Index of Production, and more generally to provide (through the ONS website) a wealth of information about business activity in the UK.

Chapter 2

Sampling Procedures

Contents ***Page***

A.	**Introduction**	**14**
	Illustrative Example	14
	Definitions	14
	Reasons for Sampling	14
B.	**Statistical Inference**	**15**
C.	**Sampling**	**16**
	Procedure for Selecting the Sample	16
	Sample Size	16
	Bias	17
D.	**Sampling Methods**	**18**
	Probability and Non-Probability Sampling	18
	Simple Random Sampling	18
	Systematic Sampling	19
	Stratified Sampling	19
	Multistage Sampling	20
	Cluster Sampling	21
	Quota Sampling	22
	Judgemental Sampling	22
	Snowball Sampling	22
	Convenience Sampling	22
E.	**Choice of Sampling Method**	**23**

A. INTRODUCTION

A considerable portion of modern statistical theory revolves around the use of samples, and many of the practical applications of this theory are possible only if samples are collected.

Illustrative Example

Preceding a general election, the public is told by the media that the various political parties enjoy the support of certain percentages of the electorate. These results cannot be obtained by asking every voter in the country for his or her political views, as this is clearly impracticable because of cost and time. Instead, some of the voters are asked for their views, and these, after a certain amount of statistical analysis, are published as the probable views of the whole electorate (opinion polls). In other words, the results of a survey of a minority have been *extended* to apply to the majority.

Definitions

The previous example illustrates the principles of sampling, and we must now define some of the terms involved.

- *Population – a* population is the set of all the individuals or objects which have a given characteristic, e.g. the set of all persons eligible to vote in a given country.
- *Sample* – a sample is a subset of a population, e.g. the voters selected for questioning about their views.
- *Sampling* – sampling is the process of taking a sample.
- *Sample Survey* – the process of collecting the data from a sample is called a sample survey, e.g. asking the selected voters their political views is a sample survey.
- *Census* – the process of collecting data from a whole population is called a census, e.g. a population census in which data about the entire population of a country is collected. (Note that the ten-yearly population census taken in the UK is one of the few questionnaires that the head of a household is compelled by law to complete.)

Reasons for Sampling

The advantages of using a sample rather than the whole population are varied:

(a) Cost

Surveying a sample will cost much less than surveying a whole population. Remember that the size of the sample will affect the accuracy with which its results represent the population from which it has been drawn. So, you must balance the size of the sample against the level of accuracy you require. This level of accuracy must be determined before you start the survey (the larger the sample, the greater the reliance that you can put on the result).

(b) Control

A sample survey is easier to control than a complete census. This greater control will lead to a higher response rate because it will be possible to interview every member of the sample under similar conditions. A comparatively small number of interviewers will be needed, so standardisation of the interviews will be easier.

(c) Speed

Apart from the lower cost involved in the use of a sample, the time taken to collect the data is much shorter. Indeed, when a census is taken, a sample of the data is often analysed at an early stage in order to get a general indication of the results likely to arise when the census information is fully analysed.

(d) Quality

When only a few interviews are needed, it is easier to devote a greater degree of effort and control per interview than with a larger number of interviews. This will lead to better quality interviews and to a greater proportion of the questions being answered correctly without the necessity of a call-back. (A call-back is when an interviewer has to return to the respondent, if that is possible, in order to clarify the answer to a question.)

(e) Accuracy

The level of accuracy of a survey is assessed from the size of the sample taken. Since the quality of the data obtained from a sample is likely to be good, you can have confidence in this assessment.

(f) Industrial Application

Sampling is not confined to surveys such as opinion polls which involve interviews and postal questionnaires; it is also very important in controlling industrial production. On an assembly line in a factory producing manufactured goods, it is necessary to check the standard of the product continuously. This is done by selecting a sample of the same product every day and testing to establish that each item in the sample meets the manufacturer's specifications.

Sometimes this testing involves destroying the product. For example, in a tyre factory each tyre will be required to have a minimum safe life in terms of distance driven and to withstand a minimum pressure without a blowout. Obviously the whole population of tyres cannot be tested for these qualities. Even when the testing involves nothing more than measuring the length of a bolt or the pitch of a screw, a sample is used because of the saving in time and expense.

B. STATISTICAL INFERENCE

Among the reasons for taking a sample is that the data collected from a sample can be used to infer information about the population from which the sample is taken. This process is known as *statistical inference*. The theory of sampling makes it possible not only to draw statistical inferences and conclusions from sample data, but also to make precise probability statements about the reliability of such inferences and conclusions. Future chapters will enlarge on this subject.

Before we continue we must define some terms which are generally used in statistical inference:

- *Parameter* – a constant measure used to describe a characteristic of a population.
- *Statistic* – a measure calculated from the data set of a sample.
- *Estimate* – the value of a statistic which, according to sampling theory, is considered to be close to the value of the corresponding parameter.
- *Sampling unit* – an item from which information is obtained. It may be a person, an organisation or an inanimate object such as a tyre.
- *Sampling frame* – a list of all the items in a population.

The sampling theory which is necessary in order to make statistical inferences is based on the mathematical theory of probability. We will discuss probability later in the course.

C. SAMPLING

Once you have decided to carry out a sample survey, there are various decisions which must be made before you can start collecting the information. These are:

- procedure for selecting the sample
- size of the sample
- elimination of bias
- method of taking the sample.

We will discuss these in some detail.

Procedure for Selecting the Sample

In selecting a sample you must first define the sampling frame from which the sample is to be drawn. Let us consider a particular survey and discuss how the stages, defined above, may be carried out.

Example:

> Suppose you are the chairperson of Bank A, which is in competition with Banks B, C and D. You want to find out what people think of your bank compared with the other three banks. It is clearly a case for a sample survey, as cost alone would prohibit you from approaching everyone in the country to find out their views. The information required for the survey would involve questions of *opinion*, so an *interviewing* technique is the best method to use.
>
> If you want a cross section of views throughout the country, then the sampling frame could be all the adults in the country. However, if you are interested only in existing customers' views, then the sampling frame would be all the customers of Bank A. In this case a list of all the customers at the various branches can be obtained. In the former case a list of all the adults in the country can be found in the electoral roll, which is a record of all those people eligible to vote.
>
> You must be careful to make sure that the sampling frame represents the population exactly as, if it does not, the sample drawn will not represent a true cross section of the population. For example, if the electoral roll is used as the sampling frame but the population you want comprises all present and prospective customers, then customers under the age of 18 would not be represented, since only those persons of voting age (18 and over) are included in the electoral roll. So, if you decide that the population should include persons old enough to have bank accounts but under 18, the sampling frame must include school rolls (say) as well. Thus you can see that there might well be several sampling frames available, and you have to take great care in matching the sample frame with the scope of the survey. You have to decide whether the effort and cost involved in extending the sampling frame justifies the benefits gained.

Sample Size

Having chosen the sampling frame, you now have to decide on the size of the sample, and this is a very complex problem. The cost of a survey is directly proportional to the sample size, so you need to keep the sample as small as possible. However, the level of accuracy (and hence the degree of confidence that you can place on your deductions) also depends on the sample size and is improved as the size increases. You have to strike a delicate balance between these conflicting requirements.

In addition the method of analysis depends, to some extent, on the sample size. The relationship between the size of the sample and the size of the population from which it is taken does *not* affect the accuracy of the deductions. This problem will be discussed again later, but the theory on which the decision is based is outside the scope of this course. You only need to be aware of the problem and to know the formulae (given later) used to calculate the degree of confidence associated with deductions.

Bias

In Chapter 1, we referred to the possibility of non-response bias. Three other common sources of bias are:

(a) Inadequacy of Sampling Frame

The sampling frame chosen may not cover the whole population, so that some items will not be represented at all in the sample and some will be over-represented or duplicated. This bias can be avoided by a careful statement of the aim of the survey and a check that none of the sampling units has been ignored.

For example, if a survey of unemployment is undertaken by randomly speaking to people in South-East England, a biased result will be obtained. This is because the survey population does not contain people in the rest of England. Thus although the selection process may have been fair and totally random, it will be very biased and non-representative of the whole of England.

(b) Items of Selected Sample not all Available

It is possible that when a sample has been selected, some of the items chosen cannot be located, e.g. some voters on the electoral roll may not have notified a change of address. If the missing items are not replaced or are incorrectly replaced, a bias will be introduced. This bias can be reduced to a minimum by returning to the sampling frame and using the same method to select the replacements as was used to select the original sample.

For example, a survey on sickness at a large industrial company could be done by randomly drawing a sample of 500 personal files. However, having randomly selected 500 employees it may transpire that some personal files are missing (they may be in transit from other departments). This could be easily rectified by returning to the frame and randomly selecting some replacements.

Care must obviously be taken to ensure that the reason why the files are missing is not related to the survey – e.g. if they are out for updating because the person has just resumed work after yet another period of sickness!

(c) Interviewer or Observer Bias

This is often the commonest of all types of bias. All interviewers and observers are given a list of their sampling units. Sometimes to save time and effort they may substitute missing units on the spot without considering how the other units have been chosen. Other sources of bias arise when the interviewers do not follow the questionnaires exactly, allow their own ideas to become evident, or are careless in recording the responses; observers may measure or record their results inaccurately.

This type of bias is difficult to recognise and correct. It can be reduced by careful choice and training of the team, and by close supervision when the survey is taking place. For example, during a high street survey an interviewer is eager to speed up responses. In order to do so she prompts people who hesitate with replies. Although a question reads, "What type of mineral water do you prefer?", she goes on to add, "Most people have said 'lemonade', which seems quite sensible". This would inevitably lead the respondent either to agree or appear not sensible.

Bias can rarely be eliminated completely, but the results of the survey may still be useful provided that the final report states any assumptions made, even if they are not fully justified, e.g. if the sampling frame is not complete.

D. SAMPLING METHODS

Probability and Non-Probability Sampling

The final decision you have to make is about the method to use to select the sample. The choice will depend on the:

- aim of the survey
- type of population involved, and
- time and funds at your disposal.

An important distinction is made between *probability* and *non-probability* sampling. In probability sampling, every item in the population has a known chance of being selected as a sample member. In non-probability sampling, the probability that any item in the population will be selected for the sample cannot be determined.

The methods from which the choice of sampling is usually made are listed below:

Probability sampling:

- simple random sampling
- systematic sampling
- stratified sampling
- multistage sampling
- cluster sampling.

Non-probability sampling:

- quota sampling
- judgemental sampling
- snowball sampling
- convenience sampling.

In the next section we will define, explain and discuss the major advantages and disadvantages of these methods.

Simple Random Sampling

The word *random* has a definite and specific meaning in the statistical theory of sampling. The dictionary definition of random is "haphazard" or "without aim or purpose", but the statistical definition is:

> *a process by which every available item has an equal chance of being chosen.*

So simple random sampling is probability sampling in which every member of the population has an equal probability of being selected.

For example, looking at the bank survey again and given that the sampling frame is everybody over 18 shown on any electoral roll throughout the UK, everyone on the roll is given a unique number from 1 to n, (n being the total number of people in the sampling frame). Each number is now written on a slip of paper and put in a box. If you want a sample of a thousand people you mix up these slips thoroughly and draw out a thousand slips. The numbers on these slips then represent the people to be interviewed. In theory each slip would stand an equal chance of being drawn out and so would have been chosen in a *random* manner. It is fundamental to simple random sampling that every element of the sampling frame stands an *equal* chance of being included in the sample.

This method sounds almost foolproof but there are some practical difficulties. For instance, if there are 52 million people in the sampling frame, another method of drawing a sample in a random fashion is needed – using a computer, for example.

The most convenient method for drawing a sample for a survey is to use a table of random numbers. Such a table is included in your copy of *Mathematical Tables for Students*. These tables are compiled with the use of a computer, and are produced in such a way that each of the digits from 0 to 9 stands an equal chance of appearing in any position in the table. If a sample of a thousand is required, for example, then the first thousand numbers falling within the range 1 to n that are found in the table *form the sample* (where n is the total number in the sampling frame). Many pocket calculators have a built-in program for selecting random numbers.

- *Advantages* - the advantage of this method of selection is that it always produces an unbiased sample.
- *Disadvantages* – its disadvantage is that the sampling units may be difficult or expensive to contact, e.g. in the bank survey sampling units could be drawn in any area from any part of the country.

Systematic Sampling

Systematic sampling (sometimes called *quasi-random sampling***)** is another probability sampling method. It involves the selection of a certain *proportion of the total population*. Drawing a simple random sample as described above can be very time-consuming. The systematic sampling method simplifies the process.

First you decide the size of the sample and then divide it into the population to calculate the proportion of the population you require. For example, in the bank survey you may have decided that a tenth of the population would provide an adequate sample. Then it would be necessary to select every tenth person from the sampling frame. As before, each member of the population will be given a number from 1 to n. The starting number is selected from a table of random numbers by taking the first number in the table between 1 and 9. Say a 2 was chosen, then the 2nd, 12th, 22nd, 32nd ... person would be selected from the sampling frame. This method of sampling is often used as it reduces the amount of time that the sample takes to draw. However, it is not a purely random method of selecting a sample, since once the starting point has been determined, then the items selected for the sample have also been set.

- *Advantages* – the main advantage of this method is the speed with which it can be selected. Also it is sufficiently close to simple random sampling, in most cases, to justify its widespread use.
- *Disadvantages* – it is important to *check*. A major disadvantage occurs if the sampling frame is arranged so that sampling units with a particular characteristic occur at regular intervals, causing over-representation or under-representation of this characteristic in the sample. For example, if you are choosing every tenth house in a street and the first randomly chosen number is 8, the sample consists of numbers 8, 18, 28, 38 and so on. These are all even numbers and therefore are likely to be on the same side of the street. It is possible that the houses on this side may be better, more expensive houses than those on the other side. This would probably mean that the sample was biased towards those households with a high income. A sample chosen by systematic sampling must always be examined for this type of bias.

Stratified Sampling

Before we discuss this method of sampling, we have to define two different types of population:

- *Homogeneous population*: sampling units are all of the same kind and can reasonably be dealt with in one group.
- *Heterogeneous population*: sampling units are different from one another and should be placed in several separate groups.

In the sampling methods already discussed we have assumed that the populations are homogeneous, so that the items chosen in the sample are typical of the whole population. However, in business and social surveys the populations concerned are very often heterogeneous. For example, in the bank survey the bank customers may have interests in different areas of banking activities, or in a social survey the members of the population may come from different social classes and so will hold different opinions on many subjects. If this feature of the population is ignored, the sample chosen will not give a true cross section of the population.

This problem is overcome by using *stratified sampling*, another example of probability sampling. The population is divided into groups or *strata,* according to the characteristics of the different sections of the population, and a simple random sample is taken from each stratum. The sum of these samples is equal to the size of the sample required, and the individual sizes are proportional to the sizes of the strata in the population. An example of this would be the division of the population of London into various socio-economic strata.

- *Advantages* – the advantage of this method is that the results from such a sample will not be distorted or biased by undue emphasis on extreme observations.
- *Disadvantages* – the main disadvantage is the difficulty of defining the strata. This method can also be time-consuming, expensive and complicated to analyse.

Multistage Sampling

This "probability sampling" method consists of a number of stages and is designed to retain the advantage of simple random sampling and at the same time cut down the cost of the sample. The method is best explained by taking the bank survey already discussed as an example, and working through the various stages.

Suppose you have decided that you need a sample of 5,000 adults selected from all the adults in the UK, but that the expense of running the survey with a simple random sample is too high. Then you could proceed as follows:

Stage 1 Use all the administrative counties of the UK as the sampling units and select a simple random sample of size 5 from this sampling frame.

Stage 2 Each county will be divided into local authority areas. Use these as the sampling units for this stage and select a simple random sample of size 10 from each of the 5 counties chosen in stage 1. You now have 50 local authority areas altogether.

Stage 3 Divide each of the selected local authority areas into postal districts and select one of these districts randomly from each area. So you now have 50 randomly selected small regions scattered throughout the country.

Stage 4 Use the electoral rolls or any other appropriate list of all the adults in these districts as the sampling frame and select a simple random sample of 100 adults from each district.

If you check back over the stages you will find that you have a multistage sample of total size 5,000 which is divided equally between 50 centres. The 100 persons at each centre will be easy to locate and can probably be interviewed by one or two interviewers. The subdivisions at each stage can be chosen to fit in conveniently with the particular survey that you are running. For instance, a survey on the health of school children could begin with local education authorities in the first stage and finish with individual schools.

- *Advantages* – the advantages of this method are that at each stage the samples selected are small and interviews are carried out in 50 small areas instead of in 5,000 scattered locations, thus economising on time and cost. There is no need to have a sampling frame to cover the whole country. The sample is effectively a simple random sample.
- *Disadvantages* – the main disadvantages are the danger of introducing interviewer bias and of obtaining different levels of accuracy from different areas. The interviewers must be well chosen and thoroughly trained if these dangers are to be avoided.

Cluster Sampling

We have already considered the problems of cost and time associated with simple random sampling, and cluster sampling is another probability sampling method which may be used to overcome these problems. It is also a useful means of sampling when there is an *inadequate sampling frame* or when it is too expensive to construct the frame. The method consists of dividing the sampling area into a number of small concentrations or *clusters* of sampling units. Some of these clusters are chosen at random, and every unit in the cluster is sampled.

For example, suppose you decided to carry out the bank survey using the list of all the customers as the sampling frame. If you wished to avoid the cost of simple random sampling, you could take each branch of the bank as a cluster of customers. Then you select a number of these clusters randomly, and interview every customer on the books of the branches chosen. As you interview all the customers at the randomly selected branches, the sum of all interviews forms a sample which is representative of the sampling frame, thus fulfilling your major objective of a random sample of the entire population.

A variation of this method is often used in the United States, because of the vast distances involved in that country (often referred to as *area sampling*). With the use of map references, the entire area to be sampled is broken down into smaller areas, and a number of these areas are selected at random. The sample consists of all the sampling units to be found in these selected areas.

- *Advantages* – the major advantages of this method are the reduction in cost and increase of speed in carrying out the survey. The method is especially useful where the size or constitution of the sampling frame is unknown. Nothing needs to be known in advance about the area selected for sampling, as all the units within it are sampled; this is very convenient in countries where electoral registers or similar lists do not exist.
- *Disadvantages* – one disadvantage is that often the units within the sample are homogeneous, i.e. clusters tend to consist of people with the same characteristics. For example, a branch of a bank chosen in a wealthy suburb of a town is likely to consist of customers with high incomes. If all bank branches chosen were in similar suburbs, then the sample would consist of people from one social group and thus the survey results would be biased. This can be overcome to some extent by taking a large number of small clusters rather than a small number of large clusters. Another disadvantage of taking units such as a bank branch for a cluster is that the variation in size of the cluster may be very large, i.e. a very busy branch may distort the results of the survey.

Quota Sampling

In all the methods discussed so far, the result of the sampling process is a list of all those to be interviewed. The interviewers must then contact these sampling units, and this may take a considerable amount of time. It is possible that, in spite of every effort, they may have to record "no contact" on their questionnaire. This may lead to a low response rate and hence the survey result would be biased and a great deal of effort, time and money would have been wasted.

To overcome these problems the method of quota sampling has been developed, in which a sampling frame and a list of sampling units is not necessary This is an example of a *non-probability sampling method*, because it is not possible to determine the probability that any individual member of the population will be included in the sample. The basic difference between this method and those we have already discussed is that the final choice of the sampling units is left to the sampler in person.

The organisers of the survey supply the sampler, usually an interviewer, with the area allocated to him or her and the number and type of sampling units needed. This number, called a *quota*, is usually broken down by social class, age or sex. The interviewers then take to the street and select the units necessary to make up their quota. This sounds simple but in reality selecting the quota can be difficult, especially when it comes to determining certain characteristics like the social class of the chosen person. It requires experience and well-trained interviewers who can establish a good relationship quickly with those people being interviewed.

- *Advantages* – the advantages of this method are that it is probably the cheapest way of collecting data; there is no need for the interviewers to call back on any respondent, they just replace any respondent with another more convenient to locate; it has been found to be very successful in skilled hands.
- *Disadvantages* – the disadvantages are that as the sample is not random, statistically speaking, it is difficult to assess a degree of confidence in the deductions; there is too much reliance on the judgement and integrity of the interviewers and too little control by the organisers.

Judgemental Sampling

Judgemental sampling is a non-probability sampling method in which the researcher uses his or her judgement to choose appropriate members of the population for the sample. Often, the sample members are selected because they are thought to be experts in the field who can provide useful information on the research topic.

Snowball Sampling

Snowball sampling is a non-probability sampling method in which a small sample is first selected (using, for example, either random or judgemental sampling) and then each sample member is asked to pass on a questionnaire to acquaintances. In this way, a much larger sample can be obtained.

Convenience Sampling

Convenience sampling is a non-probability sampling method in which the sample members are selected because of their availability and willingness to participate. For example student researchers, who are collecting primary data for a dissertation, may collect data from their fellow students. Such samples are unlikely to be representative of the entire population.

E. CHOICE OF SAMPLING METHOD

After all the preliminary steps for a survey have been taken, you may feel the need for a trial run before committing your organisation to the expense of a full survey. This trial run is called a *pilot survey* and will be carried out by sampling only a small proportion of the sample which will be used in the final survey. The analysed results of this pilot survey will enable you to pick out the weaknesses in the questionnaire design, the training of the interviewers, the sampling frame and the method of sampling. The expense of a pilot survey is worth incurring if you can correct any planning faults before the full survey begins.

The sampling method is probably the factor which has most effect on the quality of survey results so it needs very *careful thought*. You have to balance the advantages and disadvantages of each method for each survey. When you have defined the aim of the survey, you have to consider the type of population involved, the sampling frame available and the area covered by the population.

If you are to avoid bias there should be some element of randomness in the method you choose. You have to recognise the constraints imposed by the level of accuracy required, the time available and the cost.

If you are asked in an examination to justify the choice of a method, you should list its advantages and disadvantages and explain why the advantages outweigh the disadvantages for the particular survey you are required to carry out.

Chapter 3

Tabulating and Graphing Frequency Distributions

Contents ***Page***

A.	**Introduction**	**26**
	Collection of Raw Data	26
	Form of Raw Data	26
	Arrays	26
B.	**Frequency Distributions**	**27**
	Ungrouped Frequency Distribution	27
	Grouped Frequency Distribution	28
	Construction of a Grouped Frequency Distribution	28
C.	**Class Limits and Class Intervals**	**29**
	Choosing Class Limits	29
	How to Record Observations	29
	Class Intervals	30
	Unequal Class Intervals	30
	Open-ended Classes	30
	Choosing Class Limits and Intervals	31
	Reasons for Choice	31
D.	**Cumulative and Relative Frequency Distributions**	**31**
	Relative Frequency	32
	Cumulative Relative Frequency	33
E.	**Ways of Presenting Frequency Distributions**	**34**
	Histograms	34
	Frequency Polygon	38
	Frequency Curve	39
G.	**Presenting Cumulative Frequency Distributions**	**39**
	Cumulative Frequency Polygon	39
	Percentage Ogive	41

A. INTRODUCTION

Collection of Raw Data

Suppose you were a manager of a company and wished to obtain some information about the heights of the company's employees. It may be that the heights of the employees are currently already on record, or it may be necessary to measure all the employees. Whichever the case, how will the information be recorded? If it is already on record, it will presumably be stored in the files of the personnel/human resources department, and these files are most likely to be kept in alphabetical order, work-number order, or some order associated with place or type of work. The point is that it certainly will *not* be stored in height order, either ascending or descending.

Form of Raw Data

It is therefore most likely that when all the data has been collected it is available for use, but it is not in such a form as to be instantly usable. This is what usually happens when data is collected; it is noted down as and when it is measured or becomes available. If, for example, you were standing by a petrol pump, noting down how many litres of petrol each motorist who used the pump put into his or her car, you would record the data in the order in which it occurred, and not, for example, by alphabetical order of car registration plates.

Suppose your company has obtained the measurements of 80 of its employees' heights and that they are recorded as follows:

Table 3.1: Heights of company employees in cm

173	177	168	173	182	176	179	173
179	163	180	168	188	167	183	187
160	173	174	184	163	188	176	169
175	178	177	162	176	181	188	183
181	170	179	173	170	169	164	176
164	175	164	180	174	165	174	179
183	181	170	177	185	173	171	165
189	181	175	186	166	177	179	169
179	183	182	165	180	171	173	174
172	166	182	186	181	178	178	187

Table 3.1 is simply showing the data in the form in which it was collected; this is known as *raw data*. What does it tell us? The truthful answer must be, not much. A quick glance at the table will confirm that there are no values above 200, and it appears that there are none below 150, but within those limits we do not have much idea about any pattern or spread in the figures. (In fact, all the values are between 160 and 190.) We therefore need to start our analysis by rearranging the data into some sort of order.

Arrays

There is a choice between the two obvious orders for numerical data, namely ascending and descending, and it is customary to put data in *ascending order*. A presentation of data in this form is called an *array*, and is shown in Table 3.2.

It becomes immediately obvious from this table that all the values are between 160 and 190, and also that approximately one half of the observations occur within the *middle third*, between 170 and 180. Thus we have information not only on the *lower and upper limits* of the set of values, but also on their *spread* within those limits.

Table 3.2: Array of heights of company employees in cm

160	166	170	173	176	179	181	184
162	166	171	174	176	179	181	185
163	167	171	174	177	179	181	186
163	168	172	174	177	179	182	186
164	168	173	174	177	179	182	187
164	169	173	175	177	180	182	187
164	169	173	175	178	180	183	188
165	169	173	175	178	180	183	188
165	170	173	176	178	181	183	188
165	170	173	176	179	181	183	189

B. FREQUENCY DISTRIBUTIONS

Ungrouped Frequency Distribution

However, writing out data in this form is a time-consuming task, and so we need to look for some way of presenting it in a more concise form. The name given to the number of times a value occurs is its *frequency*. In our array, some values occur only once, i.e. their frequency is 1, while others occur more than once, and so have a frequency greater than 1.

In an array, we write a value once for every time it occurs. We could therefore shorten the array by writing each value only once, and noting by the side of the value the frequency with which it occurs. This form of presentation is known as an *ungrouped frequency distribution*, because all frequency values are listed and not grouped together in any form (see Table 3.3). By frequency distribution we mean the way in which the frequencies or occurrences are distributed throughout the range of values.

Note that there is no need to include in a frequency distribution those values (for example, 161) which have a frequency of zero.

Table 3.3: Ungrouped frequency distribution of heights of company employees in cm

Height	Frequency	Height	Frequency	Height	Frequency
160	1	171	2	181	5
162	1	172	1	182	3
163	2	173	7	183	4
164	3	174	4	184	1
165	3	175	3	185	1
166	2	176	4	186	2
167	1	177	4	187	2
168	2	178	3	188	3
169	3	179	6	189	1
170	3	180	3		

Total frequency = 80

Grouped Frequency Distribution

However the ungrouped frequency distribution does not enable us to draw any further conclusions about the data, mainly because it is still rather lengthy. What we need is some means of being able to represent the data in summary form. We are able to achieve this by expressing the data as a *grouped frequency distribution*.

In a grouped frequency distribution, certain values are grouped together. The groups are usually referred to as *classes*. We will group together all those heights of 160 cm and upwards but less than 165 cm into the first class; from 165 cm and upwards but less than 170 cm into the second; and so on. Adding together the frequencies of all values in each class gives the following grouped frequency distribution – Table 3.4.

(Always total up the frequencies, as it gives you a good check on the grouping you have carried out.)

Table 3.4: Grouped frequency distribution of heights of company employees in cm

Heights *(cm)*	**Frequency**
160 – under 165	7
165 – under 170	11
170 – under 175	17
175 – under 180	20
180 – under 185	16
185 – under 190	9
Total	80

This table is of a more manageable size and the clustering of a majority of the observations around the middle of the distribution is quite clear. However, as a result of the grouping we no longer know exactly how many employees were of one particular height. In the first class, for example, we know only that seven employees were of a height of 160 cm or more but less than 165 cm. We have no way of telling, just on the information given by this table, exactly where the seven heights come within the class. As a result of our grouping therefore, we have lost some accuracy and some information. This type of trade-off will always result.

Construction of a Grouped Frequency Distribution

We obtained the grouped frequency distribution of employees' heights from the raw data by constructing an array from all of the data, then constructing an ungrouped distribution, and finally a grouped distribution. It is not necessary to go through all these stages – a grouped frequency distribution may be obtained directly from a set of raw data.

A short study of a set of raw data will enable you to determine (not necessarily exactly – approximate values are sufficient at this stage) the highest and lowest observations, and the spread of the data, e.g. are the observations closely packed together; are there a few extreme observations? On this basis you will be able to set up initial classes.

Then go through the raw data item by item, allocating each observation to its appropriate class interval. This is easily done by writing out a list of classes and then using "tally marks" – putting a mark against a particular class each time an observation falls within that class; every fifth mark is put diagonally through the previous four. Thus the marks appear in groups of five.

This makes the final summation simpler, and less liable to error. Thus, for the "160 to under 165 cm" class in the distribution of heights, the tally marks would appear as *~~////~~ //* giving a frequency of 5 + 2 = 7. Similarly, the "165 to under 170 cm" class would appear as giving a frequency of //. Having obtained the frequencies for each class, first check that they do sum to the known total frequency. It is essential that errors are eliminated at this stage.

By looking at the grouped frequency distribution you have constructed, you will be able to see if it can be improved. You may decide that groups at either end of the distribution have such low frequencies that they need to be combined with a neighbouring class; and a look at exactly where the extreme observations lie will help you to make the decision as to whether or not the first and last classes should be open-ended. You may decide that, although your class intervals are correct, the class limits ought to be altered.

If some classes (particularly those near the middle of the distribution) have very high frequencies compared with the others, you may decide to split them, thus producing a larger number of classes, some of which have a smaller interval than they did originally. With practice you will acquire the ability to make such decisions.

C. CLASS LIMITS AND CLASS INTERVALS

Choosing Class Limits

If we divide a set of data into classes, there are clearly going to be values which form dividing lines between the classes. These values are called *class limits*. Class limits must be chosen with *considerable care*, paying attention both to the form of the data and the use to which it is to be put.

Consider our grouped distribution of heights. Why could we not simply state the first two classes as 160–165 cm, and 165–170 cm, rather than 160 to under 165 cm, etc.? The reason is that it is not clear into which class a measurement of *exactly 165 cm* would be put. We could not put it into both, as this would produce double counting, which must be avoided at all costs. Is one possible solution to state the classes in such terms as 160–164 cm, 165–169 cm? It would appear to solve this problem as far as our data is concerned. But what would we do with a value of 164.5 cm? This immediately raises a query regarding the recording of the raw data.

How to Record Observations

The raw data consisted of the heights of 80 people in measured centimetres, and all were exact whole numbers. Could we honestly expect that 80 people would all have heights that were an exact number of centimetres? Quite obviously not. Therefore, some operation must have been performed on the originally measured heights before they were noted down, and the question is, what operation? There are two strong possibilities. One is that each height was rounded to the nearest cm; the other is that only the whole number in centimetres of height were recorded, with any additional fraction being ignored (this procedure is often referred to as *cutting*).

Let us consider what would produce a recorded value of 164 cm under both these procedures.

(a) Rounding

A value of 163.5 cm would be recorded as 164 cm (working on the principle that decimals of 0.5 and above are always rounded up), and so would all values up to and including 164.49999 ... cm.

(b) Cutting

A value of 164 cm would be recorded as 164 cm, and so would all values up to and including 164.99999 ... cm.

Applying the same principles to other values, we can see that if the data had been cut, the class stated as "160 – under 165 cm" would represent exactly that, i.e. all values from exactly 160 cm up to but not including 165 cm. If the data had been rounded however, the class would represent all values from exactly 159.5 cm up to, but not including, 164.5 cm. In other words, a measured value of, say, 164.7 cm would be recorded as a rounded value of 165 cm and would appear in the grouped frequency distribution in the "165 to under 170 cm" class, *not* the "160 to under 165 cm" class. From this you can see that it is advisable always to discover how data has been recorded before you do anything with it.

Thus we can see that both the form in which raw data has been recorded, and whether the variable in question is discrete or continuous (discrete and continuous variables are discussed in Chapter 4), play an important part in determining class limits.

Class Intervals

The width of a class is the difference between its two class limits, and is known as the *class interval*. It is essential that the class interval should be able to be used in calculations, and for this reason we need to make a slight approximation in the case of continuous variables.

The true class limits of the first class in our distribution of heights (if the data has been rounded) are 159.5 cm and 164.4999 ... cm. Therefore the class interval is 4.999 ... cm. However, for calculation purposes we approximate slightly and say that because the lower limit of the first class is 159.5 cm and that of the next class is 164.5 cm, the class interval of the first class is the difference between the two, i.e. 5 cm.

Unequal Class Intervals

You will see that, using this definition, all the classes in our group frequency distribution of heights have the same class interval, that is 5 cm. While this will almost certainly make calculations based on the distribution simpler than might otherwise be the case, it is *not* absolutely necessary to have equal class intervals for all the classes in a distribution. If having equal intervals meant that the majority of observations fell into just a few classes while other classes were virtually empty, then there would be a good reason for using unequal class intervals. Often it is a case of trial and error.

Open-ended Classes

Sometimes it may happen that one or both of the end limits of the distribution (the lower limit of the first class and the upper limit of the last class) are not stated exactly. This technique is used if, for example, there are a few observations which are spread out quite some distance from the main body of the distribution, or (as can happen) the exact values of some extreme observations are not known.

If this had occurred with our distribution of heights, the first and last class could have been stated as *under 165 cm* and *185 cm and over*, respectively. (Note the last class would *not* be stated as over 185 cm, because the value 185 cm would then not be included in any class.) Classes such as this are said to be *open-ended* and, for calculation purposes, are assumed to have class intervals equal to those of the class next to them. This does introduce an approximation but, provided the frequencies in these classes are small, the error involved is small enough to be ignored.

Choosing Class Limits and Intervals

In choosing classes into which to group the set of heights, you have two decisions to make: one related to the position of class limits, the other to the size of class intervals. We chose limits of 160 cm, etc., and a class interval of 5 cm (although, as we have seen, we need not have kept the class interval constant throughout the distribution). These are not the *correct* values, because there *are no such things as correct values in this context*. There are, however, some values which are better than others in any particular example.

Reasons for Choice

Firstly, we noted that the observations taken as a set were quite compact; there were no extreme values widely dispersed from the main body of the distribution. Consequently we did not need to use open-ended classes, or classes with a wider than normal interval, to accommodate such values. We could thus make all the class intervals equal.

Purely for the sake of ease of calculation and tidiness of presentation, we chose a class interval of 5 cm. Why did we not use 10 cm as the class interval? Surely, you may ask, that would make calculation even easier? Yes, it would; but it would also mean that we would have only three or four classes (depending on where we fixed the class limits), and that is not enough.

We have seen that grouping data simplifies it, but it also introduces a considerable amount of approximation. The smaller the number of classes, the wider will be the class intervals, and so the greater the approximation.

The three guidelines that you must consider when choosing class limits and intervals are as follows:

- As far as practicable, have *equal* class intervals, but if the spread of the observations implies that you need to use unequal class intervals and/or open-ended classes, then do so.
- For ease of calculation, try to work with values which are multiples of 5 to 10, but if this would impose unwarranted restriction on your choice in other ways, then ease of calculation should be sacrificed. (Remember, the main consideration is that your grouped distribution should bear a reasonable resemblance to the original data.)
- Try to keep the number of classes between 5 and 15. This will make your distribution simple enough to interpret and work with, but also accurate enough for you to have confidence in the results of your calculations.

D. CUMULATIVE AND RELATIVE FREQUENCY DISTRIBUTIONS

Cumulative Frequency

So far we have discovered how to tabulate a frequency distribution. There is a further way of presenting frequencies and that is by forming *cumulative frequencies*. This technique conveys a considerable degree of information and involves adding up the number of times (frequencies) values less than or equal to a particular value occur.

You will find this easier to understand by working through our example on employees' heights in Table 3.4. We start with the value 0 as there are no employees less than 160 cm in height. There were seven employees with a height between 160 and less than 165 cm. Therefore the total number of employees less than 165 cm in height is seven. Adding the number in the class "165 but under 170 cm", you find that the total number of employees less than 170 cm in height is 18. There are 35 employees who are not as tall as 175 cm, and so on. The cumulative frequencies are shown in Table 3.5.

Table 3.5: Less than cumulative frequencies table of employees' heights

Height *(cm)*	**Frequency**	**Cumulative Frequencies**
Under 165	7	0 + 7 = 7
Under 170	11	7 + 11 = 18
Under 175	17	18 + 17 = 35
Under 180	20	35 + 20 = 55
Under 185	16	55 + 16 = 71
Under 190	9	71 + 9 = 80
	80	

You can see that the simplest way to calculate cumulative frequencies is by adding together the actual frequency in the class to the cumulative frequencies of the previous classes. It can also work in reverse if you want to obtain class frequencies from cumulative frequencies. Work it out for the above example, and you will see how easy it is. You will also notice in the table that the class descriptions have changed slightly, to read "Under 165 cm", etc. This is a true description of what the cumulative frequencies actually represent.

It is possible to switch the descriptions round, so that they read: "More than 160", "More than 165", etc., as shown in the following table. This is known as the *more than cumulative frequency distribution*, as set out in Table 3.6.

Table 3.6: More than cumulative frequency table of employees' heights

Heights *(cm)*	**Cumulative Frequencies**
More than 160	80
More than 165	73
More than 170	62
More than 175	45
More than 180	25
More than 185	9

However, distributions are not usually presented in this way. In future examples we shall deal solely with the less than cumulative frequency distribution.

Relative Frequency

Relative frequencies are the actual class frequencies divided by the total number of observations, i.e.:

$$\text{Relative frequency} = \frac{\text{Actual frequency}}{\text{Total number of observations}}$$

Let us go back to our example of employees' heights. There are 7/80 or 0.0875 employees who are less than 165 cm tall, and 20/80 or 0.25 (one quarter) who are between 175 cm and under 180 cm tall. Table 3.7 shows the relative frequencies.

Table 3.7: Relative frequencies of employees' heights

Heights *(cm)*	**Frequency**	**Relative Frequency**
160 – under 165	7	7/80 = 0.0875 or 8.75%
165 – under 170	11	11/80 = 0.1375 or 13.75%
170 – under 175	17	17/80 = 0.2125 or 21.25%
175 – under 180	20	20/80 = 0.25 or 25.00%
180 – under 185	16	16/80 = 0.2 or 20.00%
185 – under 190	9	9/80 = 0.1125 or 11.25%
	80	

In Table 3.7 we have expressed the fractions also as percentages, something that is extremely useful and that improves a table. You can see at a glance that 20 per cent of all employees measured were more than 180 cm, but less than 185 cm tall. The main advantage of relative frequencies is their ability to describe data better.

Cumulative Relative Frequency

We have seen how to calculate cumulative frequencies. Using the same logic, you can obtain cumulative relative frequencies by adding the relative frequencies in a particular class to that already arrived at for previous classes. See Table 3.8:

Table 3.8: Cumulative relative frequencies of employees' heights

Heights *(cm)*	**Cumulative Relative Frequency**	**Cumulative Percentage**
Under 165	0.0875	8.75
Under 170	0.225	22.5
Under 175	0.4375	43.75
Under 180	0.6875	68.75
Under 185	0.8875	88.75
Under 190	1.0000	100.00

You will notice that, in the above table, an extra column has been added which is labelled "Cumulative Percentage". This column is the cumulative relative frequency converted to a percentage. This makes it easier for conclusions to be drawn from this table. For example, 88.75 per cent of all employees measured were less than 185 cm tall.

E. WAYS OF PRESENTING FREQUENCY DISTRIBUTIONS

We have seen how to tabulate frequency distributions, and we now have to consider ways of bringing these distributions to life by presenting them in such a way that, even though some of the detail may be lost, the main points contained in the data come across to the reader. We shall look at various types of diagram that are commonly used to represent frequency distributions.

Histograms

A histogram can be used to present discrete data, although it is more commonly used to illustrate continuous data. However, first we will look at its use for discrete variables; this will make it easier to follow its use in describing a frequency distribution of a continuous variable.

(a) Discrete Variables

Our data is in Table 3.9:

Table 3.9: Hand-built cars produced in a month

Cars produced	Frequency
1	4
2	5
3	11
4	6
5	3
6	1
7	0
	30

We shall plot vertical rectangles on the x-axis (horizontal axis). The width of these rectangles on the x-axis will depend on the class each represents. In our case, as the values in each class are discrete, i.e. 0, 1, 2, ... 7, the width is ±0.5, with the discrete values being the midpoints.

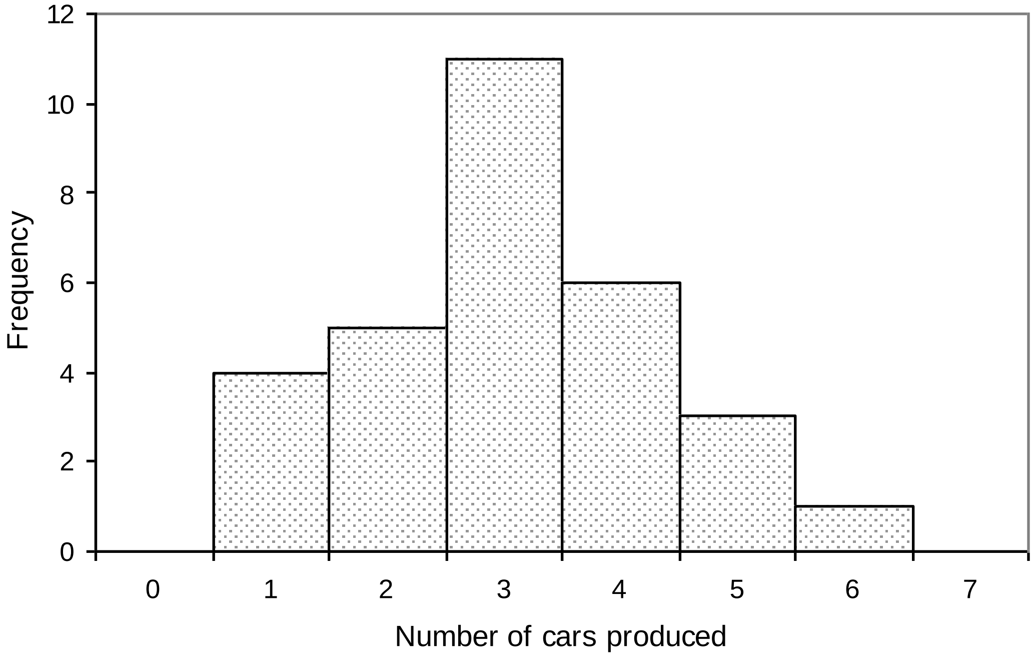
Each measurement or value of the variable falls within a class (as described earlier). Every class has a midpoint and a pair of limits, referred to as *class limits*. Sometimes, as in our example, there will be only one value in a class and the midpoint of that class is that value. A class encompasses all values between the class limits. In the example, the class limits are ±0.5. Therefore, in the class 0.5 to 1.5 all values between these limits would be added into the class. A *midpoint* is the centre of the class and is the value you usually consider that class to represent.

Obviously, in an example which includes only discrete variables, this would apply only to the discrete values. However, it is more relevant to continuous variables, as you will see later.

Each class is plotted on the x-axis. The y-axis measures the frequency with which the observations occurred in each of the x-axis classes. In a histogram, each observation is represented by a *finite amount of space*. The space assigned to each observation is the same in every case. Since each space represents an observation and the dimensions of that finite space *do not vary* within each class, the sum of these spaces constitutes an area analogous to the sum of the frequencies.

Figure 3.1 represents the data on hand-built cars. As you can see, all classes are one car in size, although the class limits range from ±0.5. The midpoints are 0, 1, 2, 3, ... 7.

Figure 3.1: Data on hand-built cars



In Figure 3.1 the frequency in each class is represented by the rectangle. However, it is important to realise that it is *not the height of the rectangle* that represents the frequency, but *the area within the rectangle*. If we now move on to using histograms to plot continuous variables, this point will become clearer.

(b) Continuous Variables

Consider the following data on employees' heights:

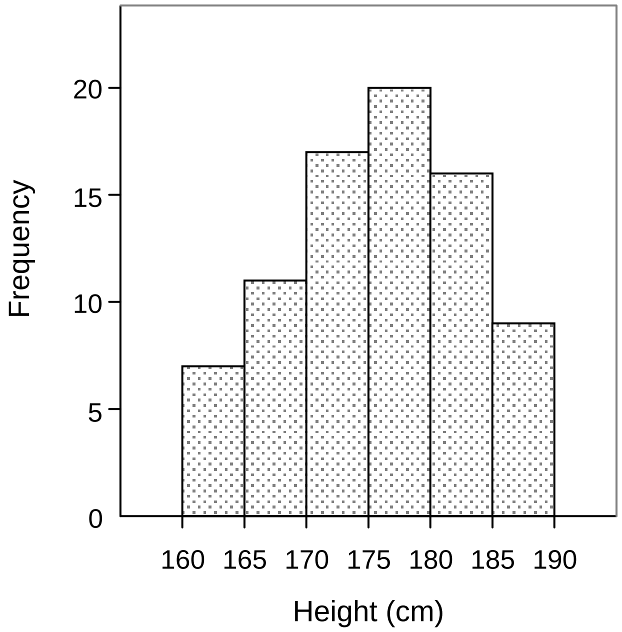
Table 3.10: Employees' heights

Height *(cm)*	**Frequency**
160 but under 165	7
165 but under 170	11
170 but under 175	17
175 but under 180	20
180 but under 185	16
185 but under 190	9
	80

A histogram would show this information in a much more clear and interesting way. If the frequencies involved are very large, it is unrealistic to plot every observation separately, so a scale can be introduced on the y-axis. See Figure 3.2.

The class limits are 160–164.99, 165–169.99, 170–174.99, ... 185–189.99, all having a range of 5 cm. Therefore the midpoints are 162.5, 167.5, 172.5, ... 187.5.

Figure 3.2: Histogram of employees' heights



You can see from the histogram that the most frequent class is 175 to 180 cm.

Let us look at the problem involved in drawing a histogram where the class limits are *unequal*. We will keep the example as before but change the class limits and rework the frequencies, as follows:

Table 3.11: Revised frequency distribution

Height *(cm)*	**Frequency**
160 but under 164	4
164 but under 168	9
168 but under 176	25
176 but under 184	32
184 but under 190	10
	80

First look at the class widths of our new frequency distribution:

Table 3.12: Revised class widths

Height *(cm)*	**Class Widths** *(cm)*
160 but under 164	4
164 but under 168	4
168 but under 176	8
176 but under 184	8
184 but under 190	6

In drawing the histogram, this time one unit in the horizontal scale will represent a class width of 4 cm in height. Therefore the first two classes are one unit wide, the next two classes are two units (i.e. 8 cm ÷ 4 cm), with the final class being 1.5 units (i.e. 6 cm ÷ 4 cm).

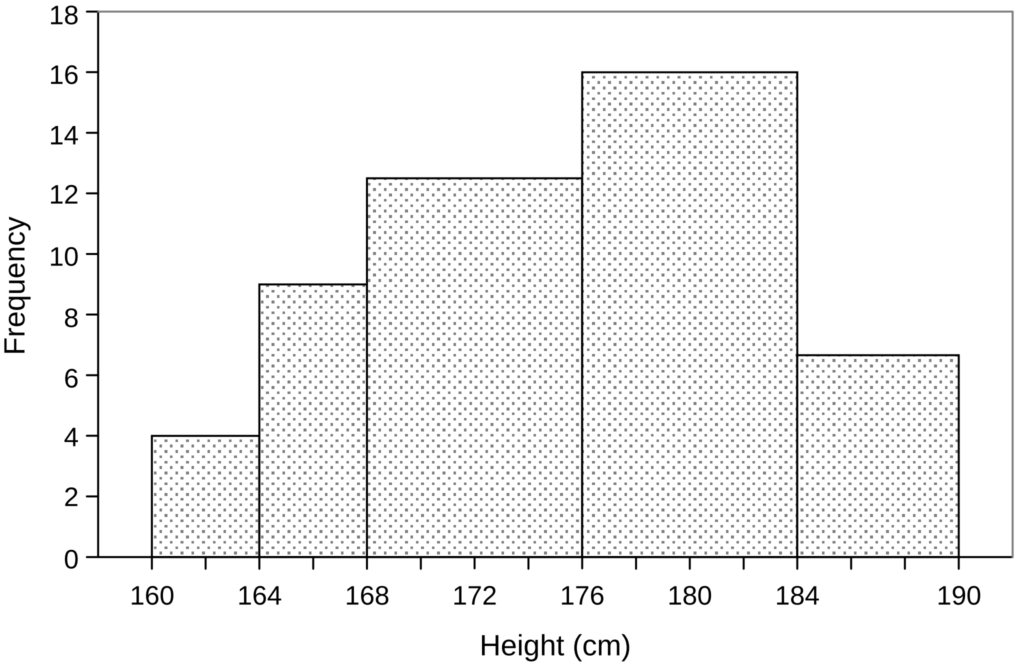
Using the frequencies set out above, the *heights of the rectangles* in each class are the *frequencies in each class divided by the number of units applicable to that class*. These heights work out as follows:

Table 3.13: Heights of histogram rectangles

Heights *(cm)*	**Height of Rectangle**
160 but under 164	$\frac{4}{1} = 4$
164 but under 168	$\frac{9}{1} = 9$
168 but under 176	$\frac{25}{2} = 12.5$
176 but under 184	$\frac{32}{2} = 16$
184 but under 190	$\frac{10}{1.5} = 6.66$

The histogram is shown in Figure 3.3.

Figure 3.3: Histogram of revised frequency distribution



The reduction in the number of classes from six to five has meant some loss of detail in the histogram, compared with that shown in Figure 3.2.

This last example has shown you how to handle grouped frequency data with classes of unequal size. Although this practice is not recommended, it is often used and this type of question may well come up in the examination.

Another practice sometimes used, although again not recommended, is open-ended classes. In our original example, the last class could be referred to as "185 and above". It would thus be necessary to make an intelligent assessment of the likely class width. This guess could be based on an inspection of the known class limits and the shape of the distribution drawn so far.

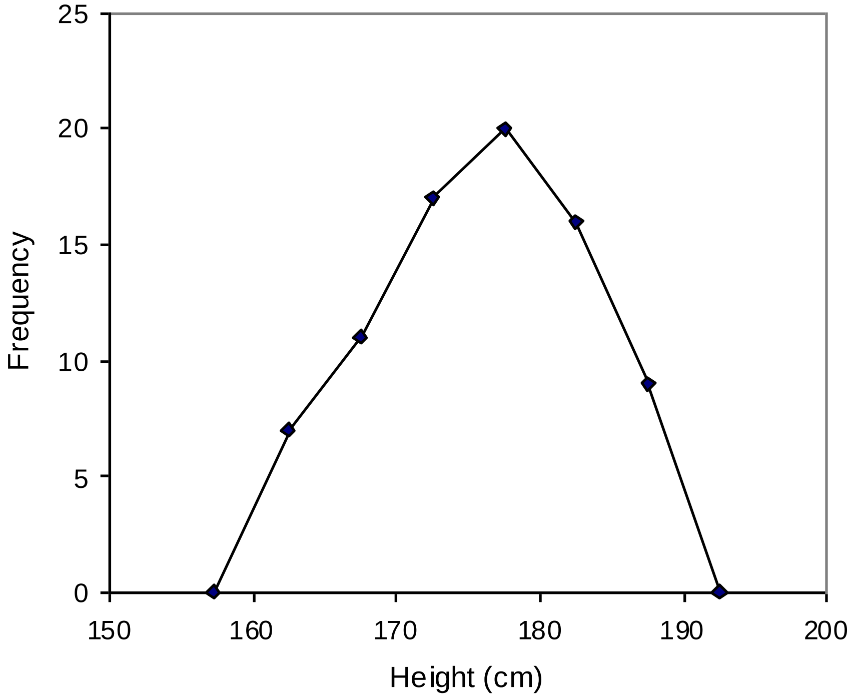
In our example, after the peak is reached at the class "175 but under 180", frequencies diminish. It would therefore be foolish for you to group all observations under the value 185 and end up by drawing a tall, thin rectangle nearly as high as the previous class. It would be realistic to use the same class widths as all previous classes – which is what happened in our frequency distribution. Thus, with open-ended classes, *use your common sense*, tempered with an inspection of the shape of the distribution.

There is one further diagram that can be used to represent frequency distributions of both discrete and continuous variables. This is called a *frequency polygon*.

Frequency Polygon

This is a very quick method of drawing the shape of a frequency distribution. Refer back to Figure 3.2, which shows the histogram drawn for our original data on heights of individuals. By joining together the midpoints at the top of each rectangle, we form a frequency polygon (see Figure 3.4).

Figure 3.4: Frequency polygon



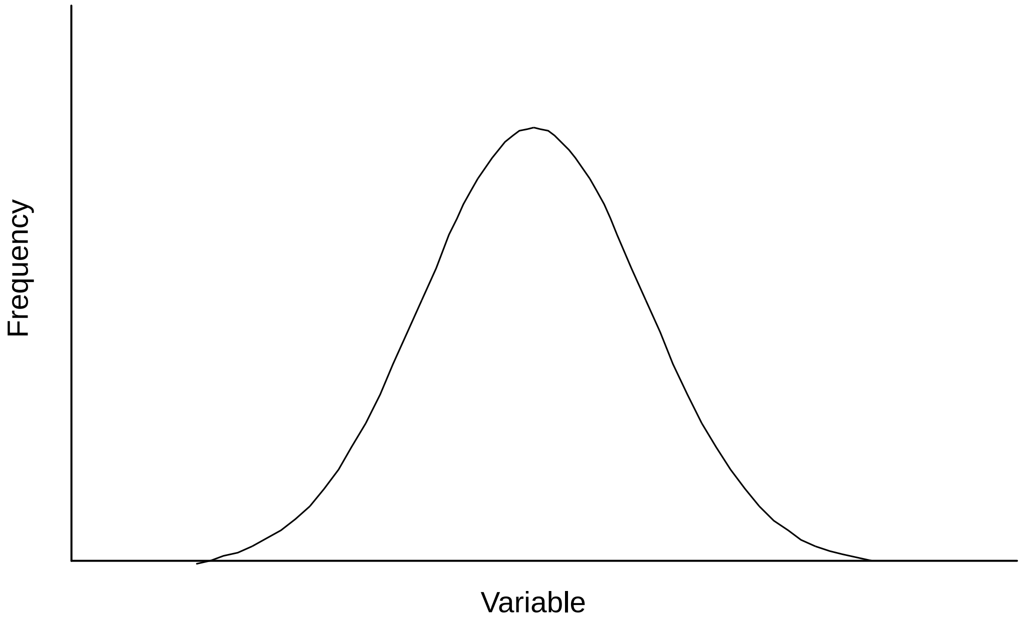
As you can see in Figure 3.4, it is necessary to extend the lines at each end of the histogram to the midpoints of the next highest and lowest classes, which will have a frequency of zero. The lines are extended to the x-axis, so that the area of the polygon will equal that of the histogram it represents. This is *vital*; this principle is extremely important in a number of branches of statistics. It is not necessary, however, to draw the histogram first. The polygon can be drawn just by plotting the frequencies of the classes at their midpoints.

You will often find that an examination question will ask you to describe a frequency polygon and its uses. Be sure you understand exactly what it is and how to draw one.

Frequency Curve

Suppose your raw data consisted of a large number of observations of a continuous variable, which were subsequently formed into a large number of classes. It would of course be possible to construct a histogram using these classes. The larger the number of classes, the narrower would become the rectangles in the histogram. If a line was drawn through the tops of the rectangles as if constructing a frequency polygon, eventually, as the numbers of classes increased, so the straight lines joining together the rectangles would become a smooth curve. It is this curve that is known as a *frequency curve* and is illustrated in Figure 3.5:

Figure 3.5: Frequency curve



The concept behind a frequency curve is fundamental to all statistics.

Almost all distributions that we obtain by one means or another are incomplete. They contain only a *proportion* of what may really be available. Take our example of heights of employees. We were only able to draw a distribution containing only 80 measurements. In real life this may have been 1/10th or 1/100th of the total number of employees available. Therefore, the resulting frequency curve which we drew was only an approximation of the true curve which would be obtained from *all* the employees. As we shall see later, it is a good approximation but nevertheless it is still only part of the whole distribution.

In later chapters, notably on *the normal distribution*, we shall use the idea of a frequency curve to illustrate distributions of variables, even though actual figures may not be presented.

G. PRESENTING CUMULATIVE FREQUENCY DISTRIBUTIONS

Cumulative Frequency Polygon

This type of diagram is used to represent *cumulative frequencies*. As in all diagrams that plot frequency distributions, the frequencies are plotted on the y-axis, with the values of the variables on the x-axis. The difference this time is that the y-axis contains the cumulative frequencies, starting from zero and finishing at the *total* number of frequencies. The cumulative frequencies are plotted not on the midpoints of their respective classes, as for histograms, but at the upper limits of the class. This is called a *less than* cumulative frequency distribution, and is the one most commonly plotted.

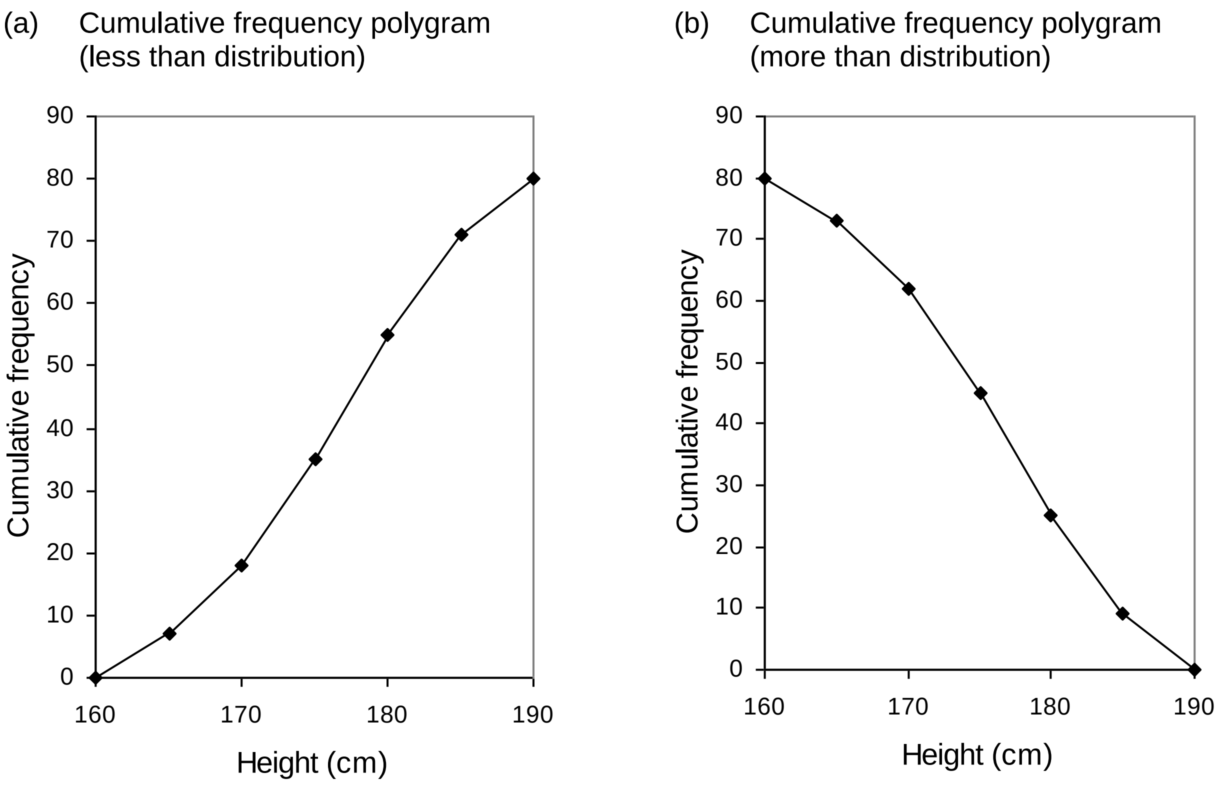
Table 3.14: Less than and more than cumulative frequency distribution

Height	Cumulative Frequency
Less than 160	0
Less than 165	7
Less than 170	18
Less than 175	35
Less than 180	55
Less than 185	71
Less than 190	80

Height	Cumulative Frequency
More than 160	80
More than 165	73
More than 170	62
More than 175	45
More than 180	25
More than 185	9
More than 190	0

If the more than type of cumulative frequency distribution is required, then the upper limits are still used but the diagram starts with the total frequency and finishes with zero. The less than and more than cumulative frequency polygons for our example on employees' heights are shown in Figure 3.6(a) and (b) respectively.

Figure 3.6: "Less than" and "more than" cumulative frequency polygons



(a) Cumulative frequency polygram (less than distribution)

(b) Cumulative frequency polygram (more than distribution)

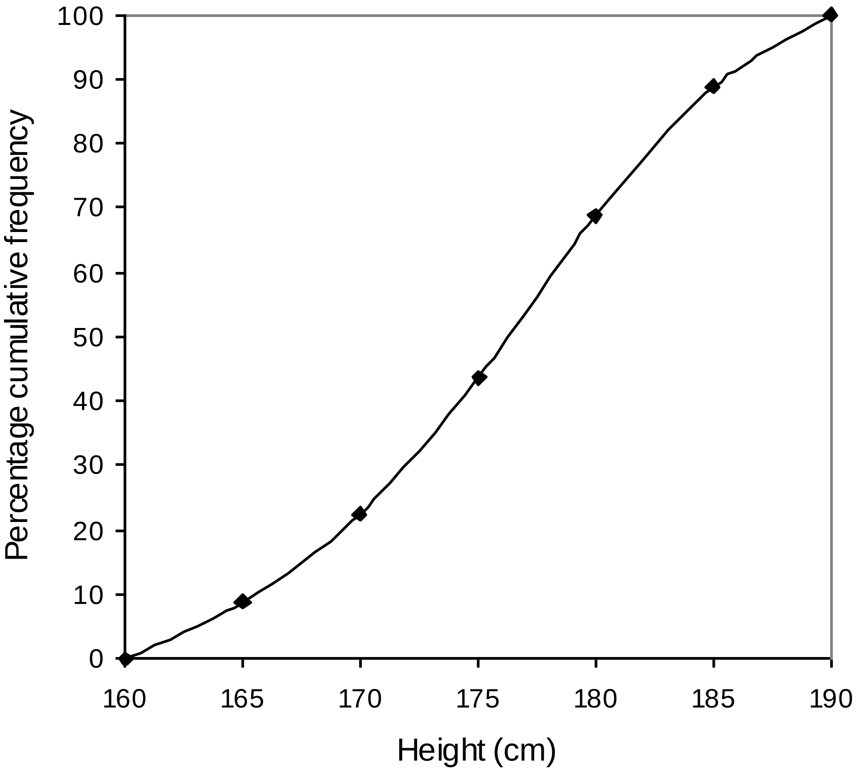
In a cumulative frequency polygon, the cumulative frequencies are joined together by straight lines. In a cumulative frequency curve, a smooth curve joins the points.

This type of diagram is often referred to as an *ogive*. An ogive is basically the name given to a graph or diagram of a cumulative frequency distribution.

Percentage Ogive

If you wish to present the *percentage relative cumulative frequency*, which is often referred to as a percentage cumulative frequency, then a percentage ogive is used. In this diagram the percentage cumulative frequencies are again plotted on the y-axis and the points joined together by a *smooth* curve. Figure 3.7 shows the percentage ogive of the information on employees' heights.

Figure 3.7: Ogive for employees' heights



The diagram plots an upward curve starting at 0 per cent and ending at 100 per cent. This is extremely useful when it comes to comparing several distributions. Say, for example, you have collected two sets of observations on employees' heights – one on women's heights and the other on men's heights. It would be possible to plot the percentage cumulative frequencies for each distribution on the same diagram. Thus the differences, if any, could be detected at a glance.

A percentage ogive could also be used to make statements concerning the characteristics of the observed distribution in percentage terms. For example, 50 per cent of those employees measured were taller than 176 cm.

Chapter 4

Measures of Location

Contents ***Page***

Introduction **44**

A. Use of Measures of Location **45**

B. Means **46**

Arithmetic Mean 46

Using a Calculator 48

Advantages and Disadvantages of the Arithmetic Mean 49

Weighted Mean 49

Geometric Mean 50

Harmonic Mean 50

C. Median **51**

Definition 51

Calculation of the Median 51

Advantages and Disadvantages of the Median 53

D. Quantiles **54**

Definitions 54

Calculation of Quantiles 55

E. Mode **57**

Definition 57

Calculation of Mode 57

Advantages and Disadvantages of the Mode 58

F. Choice of Measure **58**

Appendix: Functions, Equations and Graphs **60**

Basic Rules 60

Functions 61

Equations 62

Graphs 64

INTRODUCTION

In this chapter we shall start dealing with functions and equations, mathematical techniques with which you should be familiar from previous studies. However, should you need a reminder of these techniques, the Appendix to this chapter provides a brief revision of the basics.

In Chapter 3, we constructed frequency distributions from sets of raw data. At the time we noticed four features of these frequency distributions:

- All the available data is used in their construction.
- The result of the tabulation is to rearrange individual measurements according to their size instead of the order in which they were collected.
- The resulting tables can be illustrated graphically in several different ways.
- The type of variable which is measured affects the method of construction.

The two types of variables which you have to learn to recognise are:

(a) *Continuous variables*: these may take all values in a given range. In theory all measurements are continuous variables and in general discussion they are treated as such, e.g. time, length, weight.

(b) *Discrete variables*: may take only specified values, e.g. the number of children in families, shoe sizes, clothes sizes.

In this chapter we will discuss calculated functions of a set of data known as *measures of location* or *measures of central tendency*. These functions describe a set of data by giving the position of its "centre".

We shall be using the *sigma notation* throughout this section. Σ is a Greek letter (pronounced "sigma") and is used to denote the *summation* of a number of terms.

Thus, if we wish to add together four numbers $X_1 + X_2 + X_3 + X_4$, we could say that we wish to add together all the X_i's for $i = 1$ to 4 inclusive. This is written as:

$$\sum_{i=1}^{4} X_i$$

Similarly, $X_3 + X_4 + X_5 + X_6 + X_7$ would be written as:

$$\sum_{i=3}^{7} X_i$$

Where all the values in a set of data denoted by X are to be added, we usually omit the subscripts and superscripts and simply write ΣX. For simplicity, we shall do this throughout the remainder of this course.

A. USE OF MEASURES OF LOCATION

The main measures of location are the:

- mean
- median
- mode.

Let us first consider, in general terms, why we need these measures.

(a) Descriptive Use

The main purpose of a statistical analysis is to review unwieldy sets of data so that they may be understood and used in planning economic and business policies. A measure of location describes one feature of a set of data by a single number.

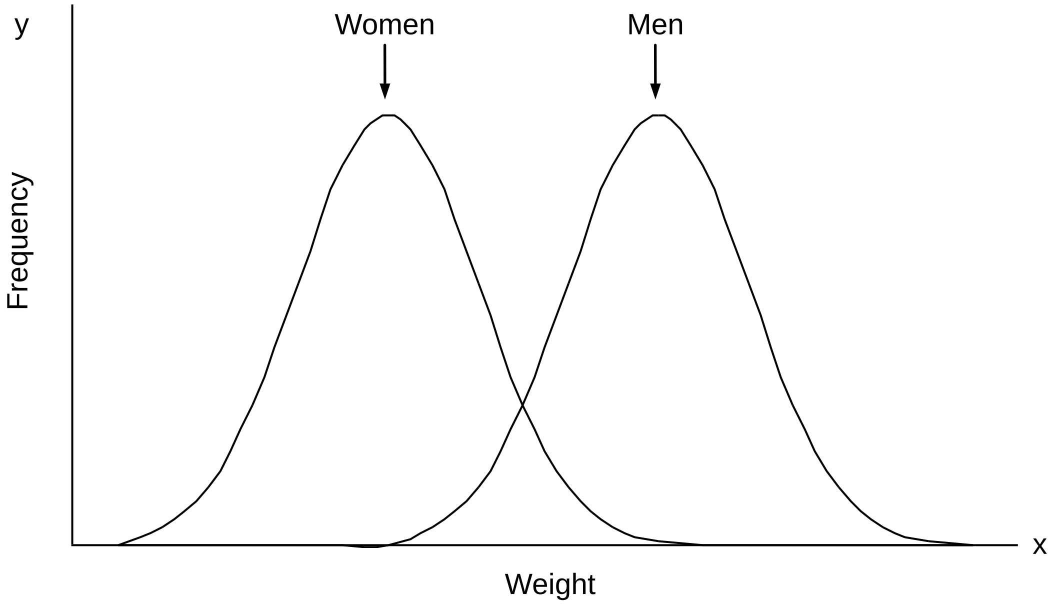
You have to discriminate between the various "centres" as each has its advantages and disadvantages. You must inspect any set of data carefully and choose the "centre" which is best for the problem you have to solve.

(b) Comparison of Distributions

Suppose you wish to compare the distribution of the weights of men and women in a given population. The data has been summarised to give two frequency distributions, and these distributions give the frequency curves shown in Figure 4.1.

The two curves overlap, showing, as you would expect, that some of the women are heavier than some of the men but, in general, the women are lighter than the men. The curves are the same shape and are symmetrical, so by symmetry, without any calculations, you can read off a value of x from each distribution which you could call the "centre". Since every other visible feature of the curves is the same, these two values of x describe their difference.

Figure 4.1 Weight distribution of men and women



B. MEANS

Arithmetic Mean

The arithmetic mean of a set of observations is the *total sum of the observations divided by the number of observations*. This is the most commonly used measure of location and it is often simply referred to as "the mean".

Example 1:

Find the mean monthly rainfall in Town A from the twelve monthly observations given in Table 4.1:

Table 4.1: Monthly rainfall in Town A

Month	**Rainfall** *(inches)*
Jan	5.4
Feb	6.8
Mar	7.2
Apr	6.5
May	5.2
June	4.2
July	2.1
Aug	2.8
Sept	3.9
Oct	4.5
Nov	4.8
Dec	5.3

We will use the sigma notation to work out this problem. We let the variable x denote the rainfall in Town A in each month. Then Σx represents the total rainfall for the given year. Remembering that there are 12 months in the year, we can calculate the mean monthly rainfall in inches (denoted by the symbol $\bar{x}$) as:

$$\bar{x} = \left(\sum x\right) \div 12 = \frac{58.7}{12} = 4.89$$

From this example we can deduce the general formula for finding the mean of a set of n observations:

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

In this formula, remember that:

$\bar{x}$ represents the arithmetic mean of the sample of observations;

x represents the sample of observations;

Σx represents the sum of the sample of observations;

n represents the number of observations.

Example 2:

Table 4.2 is the frequency distribution of the number of days on which 100 employees of a firm were late for work in a given month. Using this data, find the mean number of days on which an employee is late in a month.

Table 4.2: Number of days employees are late in a month

Number of Days Late *(x)*	**Number of Employees** *(f)*	**Number of Days** *(fx)*
1	32	32
2	25	50
3	18	54
4	14	56
5	11	55
Total	100	247

This is a little more difficult than Example 1. Begin by looking closely at the data; it is given in the form of a simple frequency distribution and the variable, x, is discrete and exact. It is almost certain that, in an exam question, you would only be given the first two columns without the symbols x and f. You should construct the full table and include it in your answer; then define the symbols you are using.

Let x = the possible values for the number of days late

f = the frequencies associated with each possible value of x

Then:

Total number of days late $= \sum fx = 247$

Total Number of employees $= \sum f = 100$

and $\bar{x} = \dfrac{\sum fx}{\sum f} = \dfrac{247}{100} = 2.47$

To deduce the general formula for a problem of this type, we let n (see the general formula on the previous page) be the number of values that the variable may take (i.e. Σf). Then:

$$\boxed{\bar{x} = \frac{\sum fx}{\sum f}}$$

Example 3:

You are given a frequency distribution of heights of employees, shown in Table 4.3. As in Example 2, the data is grouped into a number of classes, but the variable, x, is a measurement, so it is continuous and it is quite likely that all the 80 values are different. The frequency distribution only tells us the class boundaries, e.g. we know that 7 observations lie between 160 and 165 cm but we do not know any of their values. To calculate the mean we must have a single value of x which is typical of all the values in each class. We choose the *midpoint* of each class as this typical value and assume that each observation in the class is equal to the midpoint value.

Then the value of $\bar{x}$ is calculated in exactly the same way as in Example 2. As long as we realise that $\bar{x}$ is an approximate value only, this result is accurate enough for statistical analysis.

Table 4.3: Heights of employees in cm

Class Boundaries *(cm)*	**Class Midpoints** *(x)*	**Frequency** *(f)*	*(fx)*
160 – under 165	162.5	7	1,137.5
165 – under 170	167.5	11	1,842.5
170 – under 175	172.5	17	2,932.5
175 – under 180	177.5	20	3,550.0
180 – under 185	182.5	16	2,920.0
185 – under 190	187.5	9	1,687.5
		80	14,070.0

Let x = the midpoints of the classes

f = frequencies associated with each class

Then $\bar{x} = \frac{\sum fx}{\sum f}$

$$= \frac{14{,}070.0}{80} = 175.875$$

The general formula for calculating the mean of a frequency distribution is:

$$\boxed{\bar{x} = \frac{\sum fx}{\sum f}}$$

This formula looks exactly the same as the previous one. It is most important that you recognise the difference between them. It lies in the definition of x as representing the midpoints of the classes instead of the possible values of the variable.

If you are given a frequency distribution with classes of different widths, make sure that you use the correct class midpoints, e.g. if the first two classes above had been combined to give boundaries of 160 – under 170, the class midpoint would have been 165.

Using a Calculator

These three examples cover all the types of problems on arithmetic means that you will have to solve. You will see that if x, f and n are large, the amount of arithmetic required can be extensive. Modern calculators have statistics modes that enable you to input sets of ungrouped and grouped data, and then read off the arithmetic mean directly. It is important that you read your calculator manual carefully and practise calculating arithmetic means with your own calculator prior to the examination.

Advantages and Disadvantages of the Arithmetic Mean

(a) Advantages

(i) It is easy to calculate as the only information you need is the sum of all the observations and the number of observations.

(ii) It is a well known statistic and it is easily manipulated to calculate other useful statistical measures.

(iii) It uses the values of all the observations.

(b) Disadvantages

(i) A few extreme values can cause distortion which makes it unrepresentative of the data set.

(ii) When the data is discrete it may produce a value which appears to be unrealistic, e.g. in Example 2, the mean number of days on which an employee is late is 2.47.

(iii) It cannot be read from a graph.

Weighted Mean

A firm owns six factories at which the basic weekly wages are given in column 2 of Table 4.4. Find the mean basic wage earned by employees of the firm.

Table 4.4: Basic weekly wage at factories

Factory	**Basic Weekly Wage £(x)**	**Number of Employees (w)**	**(wx)**
A	85	50	4,250
B	105	80	8,400
C	64	40	2,560
D	72	35	2,520
E	96	90	8,640
F	112	75	8,400
Total	534	370	34,770

If you have no further information than column 2 then:

$$\bar{x} = \frac{\sum x}{n} = \frac{534}{6} = £89$$

But suppose you also know the number of employees at each factory (column 3), then:

$$\bar{x} = \frac{\text{Total wagebill}}{\text{Number of employees}} = \frac{\sum wx}{\sum w} = \frac{34{,}770}{370} = £93.97$$

This second result, which takes account of the number of employees, is a much more realistic measure of location for the distribution of the basic wage than the straight mean we found first. The second result is called the *weighted mean* of the basic wage, where the weights are the numbers of employees at each factory.

In Example 2 earlier, the mean calculated could be called the weighted mean of the number of days, where the weights are the frequencies.

The advantage of using a weighted mean is that the weights can be chosen to vary the effect of some values of the variable on the measure of location.

Geometric Mean

The geometric mean is seldom used outside of specialist applications. It is appropriate when dealing with a set of data such as that which shows exponential growth (that is where the rate of growth depends on the value of the variable itself), for example population levels arising from indigenous birth rates, or that which follows a geometric progression, such as changes in an index number over time, for example the Retail Price Index.

It is sometimes quite difficult to decide where the use of the geometric mean over the arithmetic mean is the best choice. We will return to the use of geometric means in the next chapter. The geometric mean (GM) is evaluated by taking the n^{th} root of the product of all n observations, that is:

$$\boxed{\text{GM} = \sqrt[n]{x_1 \times x_2 \times x_3 \times \ldots \times x_n} = \sqrt[n]{\Pi x}}$$

where: **π** represents "the product of".

Example:

In 1980 the population of a town is 300,000. In 1990 a new census reveals it has risen to 410,000. Estimate the population in 1985. If we assume that was no net immigration or migration then the birth rate will depend on the size of the population (exponential growth) so the geometric mean is appropriate.

$$\text{Geometric mean } = \sqrt[2]{300{,}000 \times 410{,}000} = 350{,}713$$

(Note that this is appreciably less than the arithmetic mean which is 355,000.)

Harmonic Mean

Another measure of central tendency which is only occasionally used is the harmonic mean. It is most frequently employed for averaging speeds where the *distances* for each section of the journey are equal.

If the speeds are x then:

$$\boxed{\text{harmonic mean} = \frac{N}{\sum \frac{1}{x}}}$$

Example:

An aeroplane travels a distance of 900 miles. If it covers the first third and the last third of the trip at a speed of 250 mph and the middle third at a speed of 300 mph, find the average speed.

$$\text{Average speed } = \frac{3}{\frac{1}{250} + \frac{1}{250} + \frac{1}{300}} = \frac{3}{0.01133} = 264.7$$

Note: if the average speed is to be calculated where the *times* for each section of the journey are the same, the appropriate average is the arithmetic mean. Note also that the distance the aeroplane travels – 900 miles – is irrelevant to the problem.

C. MEDIAN

Definition

If a set of n observations is arranged in order of size then, if n is *odd*, the median is the *value of the middle observation*; if n is *even*, the median is the *value of the arithmetic mean of the two middle observations*.

Note that the same value is obtained whether the set is arranged in ascending or descending order of size, though the ascending order is most commonly used. This arrangement in order of size is often called *ranking*.

The rules for calculating the median are:

(a) If n is odd and M is the value of the median then:

M = the value of the $\left[\frac{n+1}{2}\right]^{th}$ observation.

(b) If n is even, the middle observations are the $\frac{n}{2}^{th}$ and the $\left[\frac{n}{2}+1\right]^{th}$ observations and then:

M = the value of the mean of these two observations.

To show how the median is calculated for data presented in different forms, we will use the data from the three examples used before.

Calculation of the Median

Example 1:

When n is relatively small and all the individual observations are listed, begin by ranking the observations.

Arrange the monthly rainfall observations given in Table 4.1 in ascending order of size:

2.1, 2.8, 3.9, 4.2, 4.5, 4.8, 5.2, 5.3, 5.4, 6.5, 6.8, 7.2

n = 12 (i.e. even), so $\frac{n}{2}^{th}$ observation is the 6th and $\left[\frac{n}{2}+1\right]^{th}$ observation is the 7th.

Therefore:

M = mean of 6th and 7th observations:

$$M = \frac{(4.8+5.2)}{2} = 5.0$$

Example 2:

The data is given in the form of a simple frequency table so the values of the variable have already been arranged in ascending order of size; the smallest value x_1 occurs f_1 times and the general value x_i occurs f_i times.

Table 4.5: Number of days employees are late in a month

Number of Days Late *(x)*	**Number of Employees** *(f)*	**Cumulative Frequency**
1	32	32
2	25	57
3	18	75
4	14	89
5	11	100

The first two columns of Table 4.5 repeat the simple frequency distribution of Table 4.2. The third column is the cumulative frequency of the distribution. As n = 100 the median value is the mean of the 50th and 51st observations. From the table you can see at once that these values are both 2, so the median value is 2.

Example 3:

The data is given in the form of a grouped frequency distribution (see Table 4.6). Since we do not know the actual values of any of the observations, we cannot order them. We can only say that all the observations in the first class are smaller than all the observations in the second class and so on. In order to calculate the median value we assume that the f_i observations in the i^{th} class are equally spaced across the class. This means that the median value found is an approximation only.

Table 4.6: Height of employees in cm

Heights: Class Boundaries ***(cm)***	**Frequency**	**Cumulative Frequency**	**Percentage Cumulative Frequency**
160 – under 165	7	7	8.75
165 – under 170	11	18	22.50
170 – under 175	17	35	43.75
175 – under 180	20	55	68.75
180 – under 185	16	71	88.75
185 – under 190	9	80	100.00
	80		

Since n = 80, the median value of this distribution is the mean of the 40th and 41st observation values, and you can see that both these observations lie in the 4th class. The value of the lower class boundary is 175 cm, the first observation in this class is the 36th, and the width of the class is 5 cm. So the 40th and 41st observations are the 5th and 6th in the class, which contains 20 observations.

Therefore, assuming that the observations are evenly spaced in the class:

$$\text{value of 40}^{\text{th}}\text{ observation } = \left[175 + \frac{5}{20} \times 5\right]\text{cm}$$

$$\text{value of 41}^{\text{st}}\text{ observation } = \left[175 + \frac{6}{20} \times 5\right]\text{cm}$$

$$\text{median } = \frac{1}{2}(176.25 + 176.5)\text{cm} = 176.375 \text{ cm.}$$

The values in the fourth column of Table 4.6 are the percentage cumulative frequencies and are found by expressing each of the cumulative frequencies as a percentage of the total frequencies (80).

Note that if the total frequency is 100, the cumulative and percentage cumulative frequencies are the same. Percentage cumulative frequencies are useful for comparing different sets of data and for calculating quantiles (see later).

Note also that if the sample size is large and the data continuous, it is acceptable to find the median of grouped data by estimating the position of the $(n/2)^{th}$ value. In this case the following formula may be used:

$$\boxed{\text{median } = L + \left(\frac{\frac{n}{2} - F}{f}\right) i}$$

where: L is the lower class boundary of the median class

F is the cumulative frequency up to but not including the median class

f is the frequency of the median class

i is the width of the class interval.

Also, the median can be found graphically by plotting the *ogive*. If the y-axis is the percentage cumulative frequency axis, then the median can be read very easily by finding the value of x corresponding to the 50 % on the y-axis, as shown in Figure 4.2. This gives M = 176 cm, which is very close to the calculated approximate value.

Advantages and Disadvantages of the Median

(a) Advantages

(i) Its value is not distorted by extreme values, open-ended classes or classes of irregular width.

(ii) All the observations are used to order the data even though only the middle one or two observations are used in the calculation.

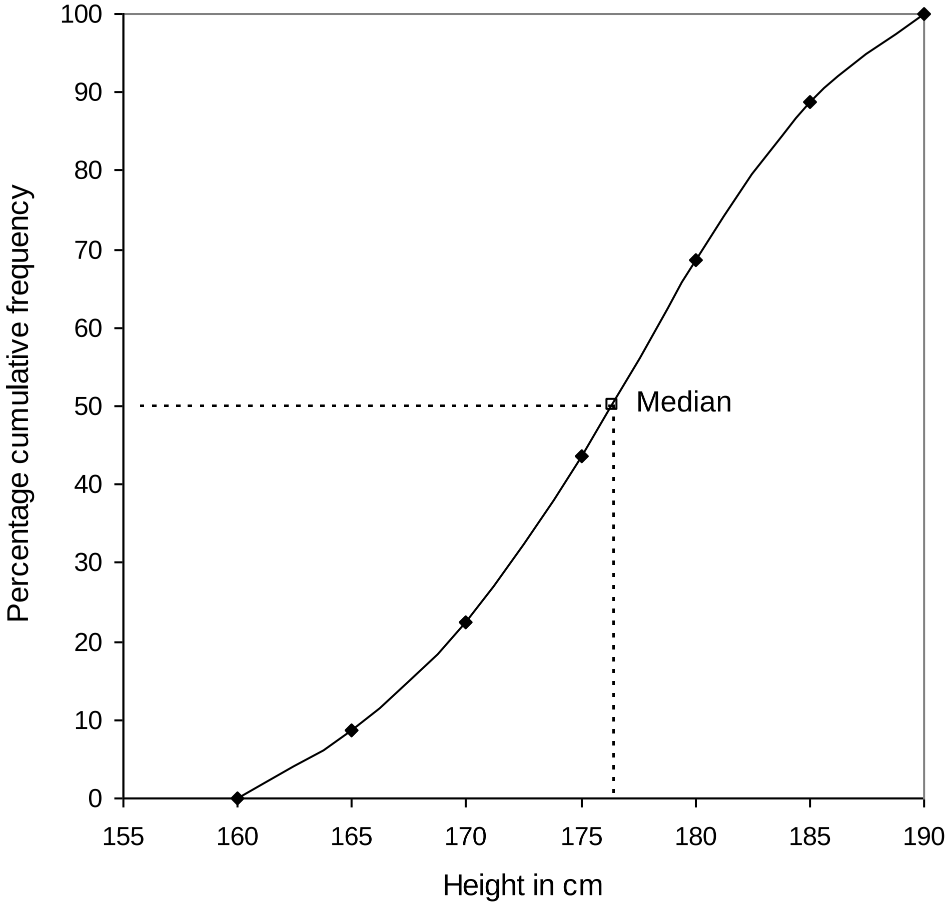
(iii) It can be illustrated graphically in a very simple way.

(b) Disadvantages

(i) In a grouped frequency distribution the value of the median within the median class can only be an estimate, whether it is calculated or read from a graph.

(ii) Although the median is easy to calculate it is difficult to manipulate arithmetically. It is of little use in calculating other statistical measures.

Figure 4.2: Ogive of heights



D. QUANTILES

Definitions

If a set of data is arranged in ascending order of size, quantiles are the values of the observations which divide the number of observations into a given number of *equal parts*.

They cannot really be called measures of central tendency, but they are measures of location in that they give the position of specified observations on the x-axis.

The most commonly used quantiles are:

(a) Quartiles

These are denoted by the symbols Q_1, Q_2 and Q_3 and they divide the observations into four equal parts:

Q_1 has 25% below it and 75% above it.

Q_2 has 50% below it and 50% above, i.e. it is the *median* and is more usually denoted by M.

Q_3 has 75% below it and 25% above.

(b) Deciles

These values divide the observations into 10 equal parts and are denoted by D_1, D_2, ... D_9, e.g. D_1 has 10% below it and 90% above, and D_2 has 20% below it and 80% above.

(c) Percentiles

These values divide the observations into 100 equal parts and are denoted by P_1, P_2, P_3, ... P_{99}, e.g. P_1 has 1% below it and 99% above.

Note that D_5 and P_{50} are both equal to the median (M).

Calculation of Quantiles

Example:

Table 4.7 shows the grouped distribution of the overdraft sizes of 400 bank customers. Find the quartiles, the 4th decile and the 95th percentile of this distribution.

Table 4.7: Size of overdraft of bank customers

Size (£)	Number of Customers	Cumulative Frequency	Percentage Cumulative Frequency
less than 100	82	82	20.5
100 but less than 200	122	204	51.0
200 but less than 300	86	290	72.5
300 but less than 400	54	344	86.0
400 but less than 500	40	384	96.0
500 but less than 600	16	400	100.0
	400		

The values of these quantiles may be approximated by reading from the ogive as shown in Figure 4.3.

Using appropriately amended versions of the formula for the median given previously, the arithmetic calculations are as follows:

The formula for the first quartile (Q_1) may be written as:

$$Q_1 = L + \left(\frac{\frac{n}{4} - F}{f} \right) i$$

where: L is the lower class boundary of the class which contains Q_1

F is the cumulative frequency up to but not including the class which contains Q_1

f is the frequency of the class which contains Q_1 and

i is the width of the class interval.

This gives:

$$Q_1 = 100 + \left(\frac{100 - 82}{122} \right) 100 \; = \; £114.75$$

The formulas and arithmetic calculations for Q_2, Q_3, D_4 and P_{95} are given below, where in each case L, F, f and i refer to the class which contains the quantile.

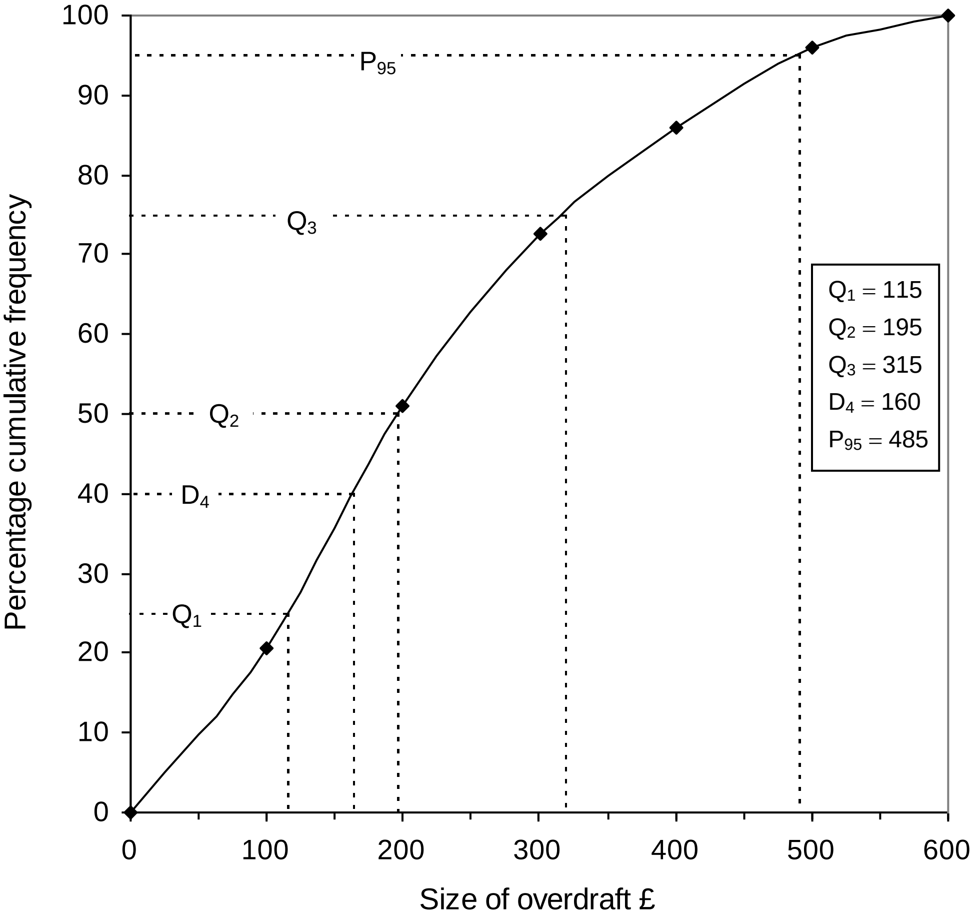
$$Q_2 = L + \left(\frac{\frac{n}{2} - F}{f} \right) i = 100 + \left(\frac{200 - 82}{122} \right) 100 = £196.72$$

$$Q_3 = L + \left(\frac{\frac{3n}{4} - F}{f} \right) i = 300 + \left(\frac{300 - 290}{54} \right) 100 = £318.52$$

$$D_4 = L + \left(\frac{\frac{4n}{10} - F}{f} \right) i = 100 + \left(\frac{160 - 82}{122} \right) 100 = £163.93$$

$$P_{95} = L + \left(\frac{\frac{95n}{100} - F}{f} \right) i = 400 + \left(\frac{380 - 344}{40} \right) 100 = £490$$

Figure 4.3: Ogive of size of overdrafts



You can see from Figure 4.3 that the two methods give approximately the same results.

E. MODE

Definition

If the variable is *discrete*, the mode is that *value of the variable which occurs most frequently*. This value can be found by ordering the observations or inspecting the simple frequency distribution or its histogram.

If the variable is *continuous*, the mode is located in the *class interval with the largest frequency*, and its value must be estimated.

As it is possible for several values of the variable or several class intervals to have the same frequency, a set of data may have several modes.

- A set of observations with one mode is called *unimodal*.
- A set of observations with two modes is called *bimodal*.
- A set of observations with more than two modes is called *multimodal*.

Calculation of Mode

(a) Discrete Variable

Example 1:

The following is an ordered list of the number of complaints received by a telephone supervisor per day over a period of a fortnight:

3, 4, 4, 5, 5, 6, 6, 6, 6, 7, 8, 9, 10, 12

The value which occurs most frequently is 6, therefore:

$$\text{mode} = 6$$

Suppose one 6 is replaced by a 5, then 5 and 6 both occur three times and the data is bimodal, with modal values 5 and 6.

Example 2:

For the simple frequency distribution shown in Table 4.2, the number of days late with the greatest frequency is 1. Therefore:

$$\text{mode} = 1$$

(b) Continuous Variable

There are various methods of estimating the modal value (including a graphical one). A satisfactory result is obtained easily by using the following formula:

$$\boxed{\text{Mode} = L + \left(\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \right) i}$$

where: L = lower boundary of the modal class

i = width of the modal class interval

f_m = the frequency of the modal class

f_{m-1} = the frequency of the pre-modal class

f_{m+1} = the frequency of the post-modal class

Example:

Find the modal value of the height of employees from the data shown in Table 4.3.

The largest frequency is 20, in the fourth class, so this is the modal class, and the value of the mode lies between 175 and 180 cm.

So, using the above formula:

$$\text{Mode} = 175 + \left(\frac{20 - 17}{40 - 17 - 16} \right) 5$$

$$= 177.1 \text{ cm (to one decimal place).}$$

Note: the formula assumes that all the class intervals are the same width. If they are *not*, you must adjust the frequencies to allow for this before you use the formula. If you cannot remember how to do this, look back at the section on frequency distributions and histograms in the previous chapter.

Advantages and Disadvantages of the Mode

(a) Advantages

(i) It is not distorted by extreme values of the observations.

(ii) It is easy to calculate.

(b) Disadvantages

(i) It cannot be used to calculate any further statistic.

(ii) It may have more than one value (although this feature helps to show the shape of the distribution).

F. CHOICE OF MEASURE

We have now covered all the common measures of location. If you have a choice (particularly in an examination question), which measure should you choose? Before you decide, inspect the data carefully and consider the problem you have to solve.

- The arithmetic mean, although it involves the most arithmetic, is the most obvious choice. It is easy to understand, it uses all the data and it is needed in the calculation of a number of other statistics.
- However, the arithmetic mean is not perfect; if it is distorted by extreme values or uneven class intervals, the median or the mode may be better. If a large number of observations are concentrated at one end of the distribution, the median is likely to be the best measure. If there are concentrations of observations at intervals over the distribution, it may be multimodal and so the mode will give a more realistic picture.
- The mode will be a more useful measure than the mean for a manufacturer of dresses for example. It would be more useful to know that the greatest demand is for size 14 rather than that the mean size is 15.12.
- You should *not* choose the geometric or harmonic means as they are difficult to calculate and of little practical interest.
- You use the weighted mean whenever you calculate the arithmetic mean from a frequency distribution.
- Apart from the median, you will use other quantiles only when the question specifically asks for them.

If you are asked to comment on a result, you can often give a better answer if you compare several measures. Remember that all these measures of location replace the details of the raw data by one number, which summarises the details to tell you *one* characteristic of the distribution. To get a good summary of the data you will need other statistics as well.

APPENDIX: FUNCTIONS, EQUATIONS AND GRAPHS

The purpose of mathematics in this course is to provide a theoretical basis for some of the concepts we study in economics and business. This Appendix provides a brief revision of some basic arithmetic and algebra, and their application in functions, equations and graphs, which you will be using as you continue your studies in this course.

Basic Rules

You will already be familiar with the four basic arithmetical operations:

addition	+
subtraction	–
multiplication	×
division	÷

To undertake a series of calculations, we use the *order of operations* rule, which sets out the order in which we need to perform the calculations, as follows:

1. remove brackets
2. divide
3. multiply
4. add
5. subtract.

For example, if we have to calculate:

$$9 \times (5 + 7 - 4) \div 2$$

we perform the following steps.

- First, we calculate the part in brackets:

  $$5 + 7 - 4 = 8$$

- Then we perform the division:

  $$8 \div 2 = 4$$

- Finally, we perform the multiplication:

  $$9 \times 4 = 36$$

For any three numbers, the *distributive law* states:

$$a(b + c) = ab + ac$$

(As you will know, the multiplication sign is often omitted, so that in full this equation would read: $a \times (b + c) = a \times b + a \times c$.)

This can be applied whatever the number of terms inside the brackets. For example, if there are two pairs of brackets, we multiply each term in the first pair of brackets by each term in the second pair of brackets, as follows:

$$(a + b)(c + d) = ac + ad + bc + bd$$

For example:

$$(2 + 5)(7 + 3) = (2 \times 7) + (2 \times 3) + (5 \times 7) + (5 \times 3) = 70$$

There are also some important rules about how the signs of positive and negative numbers change, when we are multiplying or adding them together, which are shown in the diagrams below. For example, multiplying a positive number by a negative number always results in a negative number.

Adding

	+	−
+	+	#
−	#	−

# sign of largest number

Multiplying

	+	−
+	+	−
−	−	+

A mathematical expression containing letters is known as an *algebraic expression*, for example:

$$4(x+5)$$

An algebraic expression takes different values for different values of x – a procedure known as *substitution*. For example:

When, $x=1$

$$4(x+5)=4(1+5)=4\times 6=24$$

The different parts of an algebraic expression that are separated by $+$ or $-$ signs are known as *terms*. If they contain the same combinations of letters, they are known as *like terms*, such as 2x and $-8x$, and they can be added and subtracted together. Terms which are *unlike* cannot be added or subtracted together; examples of these would be 2x, $4x^3$ and 7xy.

Functions

As you will have already seen in your studies of Economics, we often want to analyse how one variable affects another, such as how the price of strawberries affects the amount that consumers will buy. We also noted that these relationships are known as functional relationships, because one variable is dependent upon the other, that is, it is a function of the other.

A function is expressed as:

$$y=f(x)$$

which means that y is a function of x. The *dependent* variable is y and x is the *independent* variable.

A *linear* function is a relationship which when plotted on a graph produces a straight line.

An example of linear function is:

$$y=4+2x$$

As you can see, this is also an equation.

Functional relationships can be expressed mathematically as equations or graphs.

Equations

A mathematical expression which comprises two parts separated by an equals sign ($=$) is known as an equation. As we have just seen, a function can be expressed as an equation. If it is a linear function, the equation is known as a *linear* equation. So

$$y = 4 + 2x$$

is therefore a linear equation.

Expressing functions as equations enables us to apply to them the mathematical techniques which are used to manipulate equations. For example, the equation:

$$8x = 3x + 2y + 5$$

is also a function $y = f(x)$.

We can re-arrange the equation so that x becomes the subject of the equation, as follows:

$$8x - 3x = 2y + 5$$

$$5x = 2y + 5$$

$$x = \frac{2y + 5}{5}$$

$$x = \frac{2}{5}y + 1$$

The solution of an equation is the value of the unknown variable which satisfies the equation.

To solve a linear equation, we carry out the following operations:

1. Simplify the expression by multiplying out the brackets.
2. Move all the unknown variables to one side and all the numerical terms to the other side.
3. Perform the arithmetical operations, leaving an expression of the following form:

   $$ax = k$$

   where: $a = \text{constant};$

   $k = \text{constant};$

   $a \neq 0.$
4. Divide both sides by the value in front of x to obtain the solution, expressed as:

   $$x = \frac{k}{a}$$

Here is an example of a linear equation:

$$2(x - 1) = \frac{4x}{3}$$

To solve the equation, we first multiply out the brackets:

$$2x - 2 = \frac{4x}{3}$$

Then we move the unknown variables to one side and the numerical terms to the other side:

$$4x - 2x = 2 \times 3$$

Next we perform the arithmetical operations:

$$2x = 6$$

Finally, we divide both sides by the value in front of x to obtain the solution:

$$x = \frac{6}{2} = 3$$

If there are two unknown variables that we have to find, we need two equations, which are known as *simultaneous equations*. We can then eliminate one of the unknown variables, producing just one equation with one unknown, which we can solve. Then we substitute the known variable into one of the original equations and solve the equation for the other variable.

A set of simultaneous equations is shown below:

$$4x + 3y = 11$$

$$2x + y = 5$$

Let us proceed to solve these.

First, we eliminate x by making the numbers in front of x the same. In this example, we can multiply the first equation by 2 and the second equation by 4:

$$8x + 6y = 22$$

$$8x + 4y = 20$$

Then we subtract one equation from the other to eliminate x:

$$\begin{array}{rrcr} & 8x + \ 6y & = & 22 \\ -(& 8x + \ 4y & = & 20) \\ \hline & 2y & = & 2 \\ & y & = & 1 \end{array}$$

Finally, we substitute the value of y into one of the equations:

$$4x + 3y = 11$$

$$4x + 3(1) = 11$$

$$4x + 3 = 11$$

$$4x = 8$$

$$x = 2$$

Graphs

We use graphs in economic analysis when we want to depict a functional relationship.

To graph a function, we carry out the following operations:

1. Select a series of values for the independent variable x.
2. Substitute the values of x into the function to find the corresponding values of the dependent variable y. Each pair of values represents a point on a graph, known as *co-ordinates* and expressed as (x, y).
3. Plot the co-ordinates on a graph and join them up. If we are depicting a linear function, we will find that we can join the co-ordinates with a straight line.

Let us consider how to graph the following example of a linear function:

$$y = 3 + 2x$$

First, we select two values for the independent variable (x). These can be any values which are convenient to graph, such as:

$x = 1$

$x = 3$

Then we substitute these values of x into the equation to find corresponding values of the dependent variable y. This operation gives us the following pairs of values:

when $x = 1$, $y = 5$;

when $x = 3$, $y = 9$.

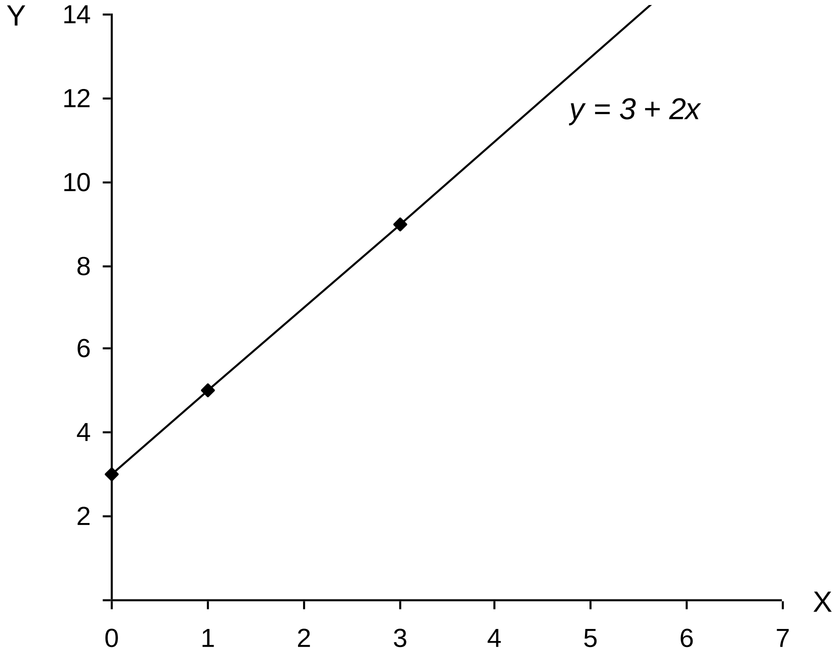
Each pair of values are co-ordinates:

(1, 5)

(3, 9)

Next, we graph these co-ordinates and join them up, as shown in the following figure.

Figure: Graph of y = 3 + 2x



The graph of a linear function is always a straight line. The general form of a linear function is:

$$y = a + bx$$

where: $a =$ the point where the line crosses the vertical axis

$b =$ the slope of the line.

To draw a straight line, we only need two points, (x_1, y_1) and (x_2, y_2).

For any two points, we can obtain the gradient of a straight line by using the following formula:

$$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\text{differencein y coordinates}}{\text{differencein x coordinates}}$$

Chapter 5

Measures of Dispersion

Contents ***Page***

Introduction **68**

A. Range **69**
Definition and Calculation 69
Advantages and Disadvantages 70

B. Quartile Deviation **70**
Definition and Calculation 70
Advantages and Disadvantages 71

C. Standard Deviation and Variance **72**
Definition and Calculation 72
Using a Calculator 74
Interpreting the Standard Deviation 75
Advantages and Disadvantages of the Standard Deviation 75

D. Coefficient of Variation **75**

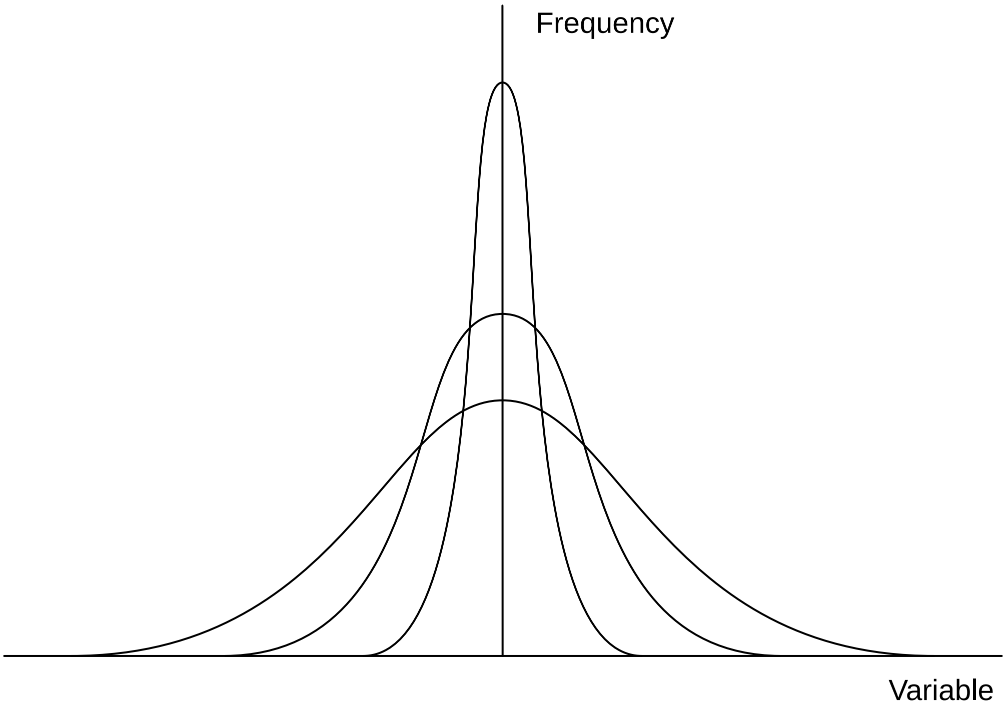
E. Skewness **76**

Summary **78**

INTRODUCTION

In the previous chapter we defined ways of describing distributions by calculating measures of location. However, one value of a measure of location can represent several sets of data. The distributions shown in Figure 5.1 all have the same "centre", but their shapes are quite different. So we need another number to show this difference, and this number is called a *measure of dispersion*. This describes the way in which the observations are *spread* about the "centre", i.e. it is a measure of the variability of the data. A summary of any set of data is not complete unless *both* a measure of location and a measure of dispersion are given.

Figure 5.1: Distributions with different dispersions



If the observations in a set of data are very variable, some of them will be a long way from the "centre", the curve will be wide and flat, and the value of the measure of dispersion will be large. If the observations are not very variable, they will all be close to the "centre", the curve will be narrow and tall, and the measure of dispersion will be small.

You can see from Figure 5.1 that, in these distributions, the part of the curve to the right of the centre is a mirror image of the part to the left. Distributions which have this characteristic are said to be *symmetrical*. You can see without any calculation that if a distribution is symmetrical, the value of its mean, median and mode will be equal. If a distribution does not have this characteristic, it is said to be *skewed* or *asymmetrical*, and the mean, median and mode will all have different values.

There are several different measures of dispersion. The most important of these (which we will describe in this chapter) are the:

- range
- quartile deviation
- standard deviation and variance

Two further measures, the coefficient of variation and the coefficient of skewness, will also be discussed.

A. RANGE

Definition and Calculation

The range of a distribution is the difference between the largest and the smallest values in the set of data.

Example:

Look back to the data given in Table 4.1 in Chapter 4.

The largest monthly rainfall for Town A is 7.2 inches in March, and the smallest is 2.1 inches in July. Therefore:

range = (7.2 – 2.1) inches = 5.1 inches.

Table 5.1 gives the monthly rainfall in another town:

Table 5.1: Monthly rainfall in Town B

Month	Rainfall in Inches
Jan	6.2
Feb	6.6
Mar	10.6
Apr	5.1
May	4.3
June	2.2
July	0.4
Aug	2.8
Sept	3.4
Oct	4.8
Nov	5.4
Dec	6.9

The largest monthly rainfall for Town B is 10.6 inches in March, and the smallest is 0.4 inches in July. Therefore:

range = (10.6 – 0.4) inches = 10.2. inches.

that is, the range for Town B is double that for Town A.

However, if you calculate the mean rainfall for Town B, you find that:

$$\bar{x} = \frac{\sum x}{12} = \frac{58.7}{12} = 4.89 \text{ inches,}$$

which is exactly the same as for Town A.

If only the means are known, you would say that the rainfall distributions for these two towns are identical, but when the range is known you can see that they are different.

- If the data is given in the form of a *simple frequency* distribution, the range is the difference between the *largest and smallest possible values of the variable*.
- If the data is given in the form of a grouped frequency distribution, the range is the difference between the *highest upper class boundary* and the *lowest lower class boundary*.

Advantages and Disadvantages

(a) Advantages

(i) It is easy to understand.

(ii) It is simple to calculate.

(iii) It is a good measure for comparison as it spans the whole distribution.

(b) Disadvantages

(i) It uses only two of the observations and so can be distorted by extreme values.

(ii) It does not indicate any concentrations of the observations.

(iii) It cannot be used in calculating other functions of the observations.

B. QUARTILE DEVIATION

Definition and Calculation

The quartile deviation is half the difference between the third quartile and the first quartile, and for this reason it is often called the *semi-interquartile range*.

$$\boxed{\text{Quartile Deviation} = ½(Q_3 - Q_1)}$$

To calculate the quartile deviation, you must first order the set of observations, so this measure of dispersion is related to the median.

Example 1:

Find the quartile deviation of the monthly rainfall for the two towns whose rainfall is given in Tables 4.1 and 5.1. The ordered amounts of rainfall are:

Town A: 2.1, 2.8, 3.9, 4.2, 4.5, 4.8, 5.2, 5.3, 5.4, 6.5, 6.8, 7.2

Town B: 0.4, 2.2, 2.8, 3.4, 4.3, 4.8, 5.1, 5.4, 6.2, 6.6, 6.9, 10.6

Since n (=12) is divisible by 4, none of the quartiles is equal to an observation. Q_1 is a quarter of the way between the 3rd and 4th observations, Q_2 is halfway between the 6th and 7th observations and Q_3 is three-quarters of the way between the 9th and 10th observations.

For Town A:

$$\begin{aligned} Q_1 &= 3.9 + 0.075 & \qquad Q_3 &= 5.4 + 0.825 \\ &= 3.975 & &= 6.225 \end{aligned}$$

So the quartile deviation = ½(6.225 – 3.975) inches = 1.125 inches.

For Town B:

$$\begin{aligned} Q_1 &= 2.8 + 0.15 & \qquad Q_3 &= 6.2 + 0.3 \\ &= 2.95 & &= 6.5 \end{aligned}$$

So the quartile deviation = ½(6.5 – 2.95) inches = 1.775 inches.

It is interesting to compare the medians.

We have already calculated that for Town A, $M = Q_2 = 5.0$ inches. For Town B:

$M = Q_2 = ½(4.8 + 5.1) = 4.95$ inches.

These two values are very close but not identical, while the quartile deviation of Town B is still much larger than that of Town A.

Example 2:

Find the quartile deviation of the number of days late, from the simple frequency distribution given in Table 4.5 in the previous chapter.

Here $n = 100$, so Q_1 is a quarter of the way between the 25th and 26th observations; both of these are 1 so $Q_1 = 1$.

Q_3 is three-quarters of the way between the 75th and 76th observations; the 75th observation is 3 and the 76th is 4 so $Q_3 = 3.75$.

Therefore:

quartile deviation $= ½(3.75 - 1) = 1.375$ days.

Example 3:

Find the quartile deviation for the size of overdrafts using the quantiles calculated in the previous chapter from Table 4.7.

$Q_1 = £114.75$ $\quad Q_3 = £318.52$

quartile deviation $= £½(318.52 - 114.75) = £101.89$

Note that this method must be used because you are asked to *calculate* the value. Otherwise the graphical method for finding Q_1 and Q_3 would be acceptable, though not quite as accurate.

Advantages and Disadvantages

(a) Advantages

(i) The calculations are simple and quite quick to do.

(ii) It covers the central 50% of the observations and so is not distorted by extreme values.

(iii) It can be illustrated graphically.

(b) Disadvantages

(i) The lower and upper 25% of the observations are not used in the calculation so it may not be representative of all the data.

(ii) Although it is related to the median, there is no direct arithmetic connection between the two.

(iii) It cannot be used to calculate any other functions of the data.

C. STANDARD DEVIATION AND VARIANCE

These two measures of dispersion can be discussed in the same section because the standard deviation is the positive square root of the variance. So even if you are asked to find the standard deviation of a set of data, you will have to find the variance first.

The variance is of great theoretical importance in advanced statistical work, but all you need to know about it is its definition, how to calculate it, and its relationship to the standard deviation.

Since the variance is a function of the squares of the observations, the unit in which it is measured is the square of the unit in which the observations are measured. So its square root, i.e. the standard deviation, is measured in the *same* unit as the observations.

Definition and Calculation

We will define the standard deviation first as you are more likely to be asked for this in your examination than the variance.

The standard deviation is the *positive square root of the mean of the squares of the differences between all the observations and their mean*. You will find this definition quite easy to understand when you see it written in symbols.

Let σ (the small Greek letter "sigma") be the standard deviation. (You will sometimes find "s" or "sd" used instead.)

The formula used to calculate the standard deviation of a set of data is:

$$\boxed{\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}} \qquad \textit{Formula (a)}$$

where: x = values of the observations

$\bar{x}$ = mean of the observations

n = number of observations.

As the standard deviation $= \sqrt{\text{variance}}$

$\sigma^2 = \text{variance}.$

Example 1:

Using the data of Table 4.1, find the standard deviation of the monthly rainfall of Town A. The data is shown again in the first two columns of Table 5.2.

Table 5.2: Monthly rainfall for Town A

Month	**Rainfall in Inches**	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
Jan	5.4	+0.51	0.2601
Feb	6.8	+1.91	3.6481
Mar	7.2	+2.31	5.3361
Apr	6.5	+1.61	2.5921
May	5.2	+0.31	0.0961
June	4.2	–0.69	0.4761
July	2.1	–2.79	7.7841
Aug	2.8	–2.09	4.3681
Sep	3.9	–0.99	0.9801
Oct	4.5	–0.39	0.1521
Nov	4.8	–0.09	0.0081
Dec	5.3	+0.41	0.1681
			25.8692

Table 5.4 shows the calculation of $\sum(x_i - \bar{x})^2$ using $\bar{x} = 4.89$, which we have already calculated and so can assume to be known, and $n = 12$.

Then, substituting in the formula gives:

$$\sigma = \sqrt{\frac{25.8692}{12}} = \sqrt{2.16} = 1.47$$

$$\text{variance} = \sigma^2 = (1.47)^2 = 2.16$$

By expanding the expression $\sum(x_i - \bar{x})^2$ we can rewrite the formula in the following useful way:

$$\boxed{\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}} \qquad \textit{Formula (b)}$$

The choice of the formula to use depends on the information that you already have and the information that you are asked to give. In any calculation you should keep the arithmetic as simple as possible and the rounding errors as small as possible.

- If you already know $\bar{x}$ and it is a small integer, there is no reason why you should not use formula (a).
- If you already know $\bar{x}$ but it is not an integer, as in Example 1, then formula (b) is the best to use.

When you are given the data in the form of a simple or grouped frequency distribution then the alternative formulae for calculating the standard deviation are:

$$\boxed{\sigma = \sqrt{\frac{\Sigma f\left(x - \bar{x}\right)^2}{\Sigma f}}}$$

Formula (c)

OR

$$\boxed{\sigma = \sqrt{\frac{\Sigma f x^2}{\Sigma f} - \bar{x}^2}}$$

Formula (d)

Formula (c), like formula (a), is derived directly from the definition of σ. Formula (d) is obtained by using the sigma notation as in formula (b) and this is the one to use in calculations.

Example 2:

Using the data of Table 4.3, find the standard deviation of the heights of employees.

The first four columns of Table 5.3 are exactly the same as those in Table 4.3. Column five is added so that we have all the sums required in formula (d). So from the table:

$$n = \Sigma f = 80; \Sigma fx = 14{,}070.0; \Sigma fx^2 = 2{,}478{,}750$$

Table 5.3: Heights of employees in cm

Height *(cm)*	**Midpoint** *(x)*	**Frequency** *(f)*	**fx**	**fx^2**
160 – under 165	162.5	7	1,137.5	184,843.75
165 – under 170	167.5	11	1,842.5	308,618.75
170 – under 175	172.5	17	2,932.5	505,856.25
175 – under 180	177.5	20	3,550.0	630,125.00
180 – under 185	182.5	16	2,920.0	532,900.00
185 – under 190	187.5	9	1,687.5	316,406.25
Totals		80	14,070.0	2,478,750.00

Then:

$$\sigma = \sqrt{\frac{2478750}{80} - \left(\frac{14070}{80}\right)^2}$$

$$\sigma = \sqrt{52.36} = 7.24$$

Using a Calculator

As you can see, the arithmetic involved in calculating a standard deviation is extensive even with a small number of classes. Modern calculators have statistics modes that enable you to input sets of ungrouped and grouped data, and then read off the standard deviation directly. It is important that you read your calculator's manual carefully and practise calculating standard deviations with your own calculator prior to the examination.

Interpreting the Standard Deviation

We have seen that the standard deviation is one of the measures used to describe the variability of a distribution. However, the standard deviation is in general not easy to interpret. Clearly larger values mean that the set of data is more spread out around the mean, while smaller values mean that the set of data is less spread out around the mean. If the data set has an approximately symmetrical and bell-shaped distribution (i.e. is approximately normally distributed), like the frequency curve illustrated in Fig 3.5, then we can say that:

- Roughly 68% of the data will lie within ONE standard deviation of the mean.
- Roughly 95% of the data will lie within TWO standard deviations of the mean.
- More than 99% of the data will lie within THREE standard deviations of the mean.

Furthermore, the standard deviation has an additional use which makes it more important than the other measures of dispersion. It may be used as a unit to measure the *distance between any two observations*. For example, in the distribution of employees' heights, the distance in the distribution between a height of 160 cm and a height of 189 cm is 29 cm. But as the standard deviation is 7.24 cm, we can also say that the distance between the two observations is (29 ÷ 7.24), or just over 4 standard deviations.

Advantages and Disadvantages of the Standard Deviation

(a) Advantages

(i) It uses all the observations.

(ii) It is closely related to the most commonly used measure of location, i.e. the mean.

(iii) It is easy to manipulate arithmetically.

(b) Disadvantages

(i) It is rather complicated to define and calculate.

(ii) Its value can be distorted by extreme values.

D. COEFFICIENT OF VARIATION

The standard deviation is an *absolute* measure of dispersion and is expressed in the units in which the observations are measured. The coefficient of variation is a *relative* measure of dispersion, i.e. it is independent of the units in which the standard deviation is expressed.

The coefficient of variation is calculated by expressing the standard deviation as a percentage of the mean:

$$\text{coefficient of variation (CV)} = \frac{\sigma}{\bar{x}} \times 100\% .$$

By using the coefficient of variation you can compare dispersions of various distributions, such as heights measured in centimetres with a distribution of weights measured in kilograms.

Example:

Compare the dispersion of the monthly rainfall for Town A with the dispersion of employees' heights.

Town A: $\sigma = 1.47$ inches $\bar{x} = 4.89$ inches

$$\text{therefore: } CV = \frac{1.47}{4.89} \times 100\% = 30.06\%$$

Employees' heights: $\sigma = 7.24$ cm $\bar{x} = 175.875$ cm

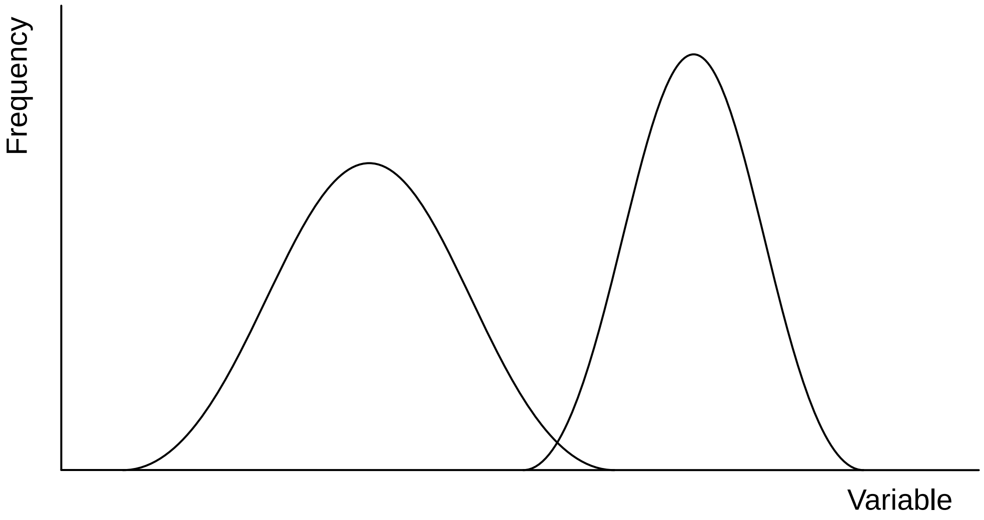
$$\text{therefore: } CV = \frac{7.24}{175.875} \times 100\% = 4.12\%$$

This shows that the rainfall distribution is much more variable than the employees' heights distribution.

E. SKEWNESS

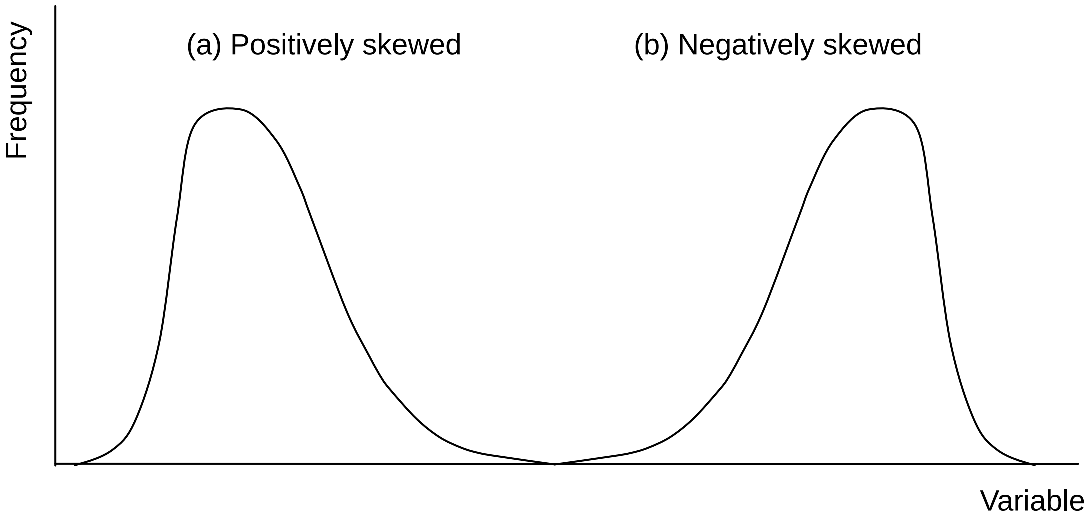
When the items in a distribution are dispersed equally on each side of the mean, we say that the distribution is *symmetrical*. Figure 5.2 shows two symmetrical distributions.

Figure 5.2: Symmetrical distributions



When the items are not symmetrically dispersed on each side of the mean, we say that the distribution is *skewed* or asymmetric. Two skewed distributions are shown in Figure 5.3. A distribution which has a tail drawn out to the right, as in Figure 5.3 (a), is said to be *positively skewed*, while one with a tail to the left, like (b), is *negatively skewed*.

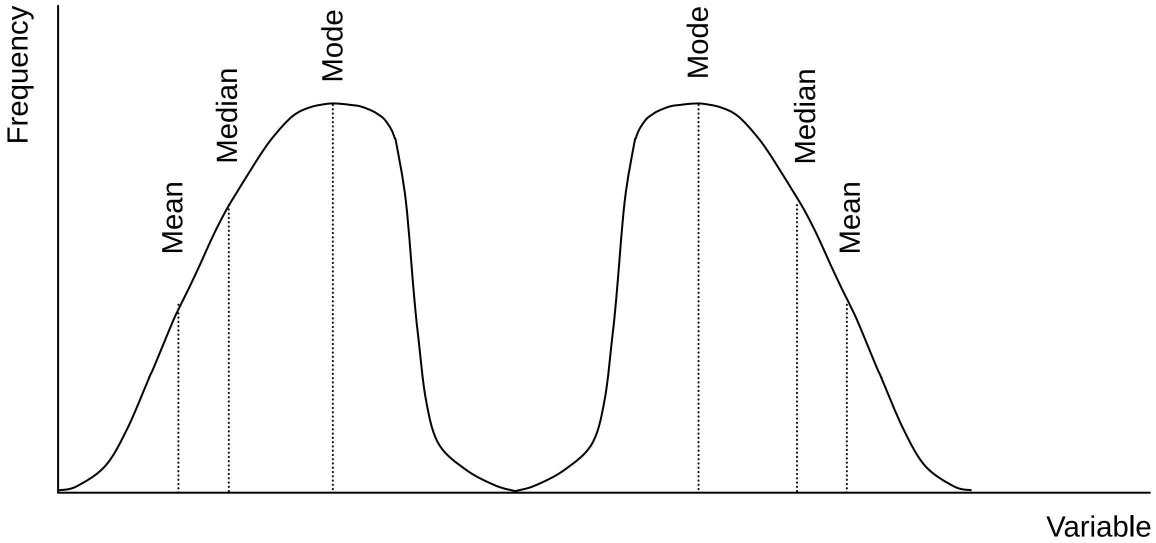
Figure 5.3: Skewed distributions



Two distributions may have the same mean and the same standard deviation but they may be differently skewed. This will be obvious if you look at one of the skewed distributions and then look at the *same one* through from the other side of the paper! What, then, does skewness tell us? It tells us that we are to expect a few unusually high values in a positively skewed distribution or a few unusually low values in a negatively skewed distribution.

If a distribution is symmetrical, the mean, mode and the median all occur at the same point, i.e. right in the middle. But in a skew distribution, the mean and the median lie somewhere along the side with the "tail", although the mode is still at the point where the curve is highest. The more skew the distribution, the greater the distance from the mode to the mean and the median, but these two are always in the same order; working outwards from the mode, the median comes first and then the mean, as in Figure 5.4:

Figure 5.4: Measures of location in skewed distributions



For most distributions, except for those with very long tails, the following relationship holds approximately:

$$\text{mean} - \text{mode} = 3(\text{mean} - \text{median})$$

The more skew the distribution, the more spread out are these three measures of location, and so we can use the amount of this spread to measure the amount of skewness. The most usual way of doing this is to calculate:

$$\text{Pearson's first coefficient of skewness} = \frac{\text{mean - mode}}{\text{standard deviation}}$$

However, the mode is not always easy to find and so we use the equivalent formula:

$$\text{Pearson's second coefficient of skewness} = \frac{3(\text{mean - median})}{\text{standard deviation}}$$

This second formula for the coefficient of skewness (Sk) may be written as:

$$\text{Sk} = \frac{3(\bar{x} - \tilde{x})}{s}$$

where: $\tilde{x}$ = median

s = standard deviation

You are expected to use one of these formulae when an examiner asks for the skewness (or coefficient of skewness) of a distribution. When you do the calculation, remember to get the correct sign (+ or –) when subtracting the mode or median from the mean and then you will get negative answers for negatively skew distributions, and positive answers for positively skew distributions. The value of the coefficient of skewness is between –3 and +3, although values below –1 and above +1 are rare and indicate very skewed distributions.

Examples of variates (variables) with positive skew distributions include size of incomes of a large group of workers, size of households, length of service in an organisation, and age of a workforce. Negative skew distributions occur less frequently. One such example is the age at death for the adult population of the UK.

SUMMARY

The definitions and formulae introduced in this chapter are very important in statistical analysis and interpretation and you are likely to get questions on them in your examination. You should learn all the formal definitions and the formulae thoroughly. Make sure you know when each of the formulae should be used, and that you can distinguish between those formulae which are only a statement of the definition in symbols and those which are used to calculate the measures.

Remember that even if you are allowed to use a pocket calculator, you must show all the steps in a calculation.

Examination questions will often ask you to comment on the results you have obtained. You will be able to make sensible comments if you have studied the use and the advantages and disadvantages of each of the measures.

Chapter 6

Index Numbers

Contents ***Page***

Introduction **80**

A. Simple (Unweighted) Index Numbers **80**
Simple Index 80
Price Relatives 82

B. Weighted index Numbers (Laspeyres and Paasche Indices) **83**
Weighted Aggregative Index Numbers 83
Weighted Price-relative Index Numbers 84

C. Fisher's Ideal Index **85**

D. Formulae **86**

E. Quantity or Volume Index Numbers **87**
Worked Example 87

F. Changing the Index Base Year **90**
Worked Example 90

G. Index Numbers in Practice **91**
Practical Problems with Index Numbers 91
What Items to Include 91
How to Obtain the Data 92
What Weights to Use 92
The Kind of Average to Use 93
What Period to Use as Base 93
Criteria for a Good Index 95
Index Numbers in Use 96

INTRODUCTION

Table 6.1 shows the monthly profits of Firm X for a period of one year. We could plot profits against time (i.e. each month) and draw a graph. However, if we are interested in *changes* in profits rather than in the actual level of profits, we can use one month's figures (say January) as standard and express all the others as percentages of this standard. Because we are dealing with percentages, we use a standard figure of 100.

In Table 6.1, the right-hand column shows January set to the standard figure of 100 and all the other profit values set to percentages of this standard.

Table 6.1: Monthly profits of Firm X

Month	Profit	Profit Based on Jan = 100
Jan	512	100
Feb	520	102
Mar	530	104
Apr	531	104
May	546	107
Jun	549	107
Jul	560	109
Aug	565	110
Sep	568	111
Oct	573	112
Nov	584	114
Dec	585	114

The percentage figures in the right-hand column are called *index numbers* of profits and, in this case, January is known as the *base* month against which all others are compared.

The essentials of an index number then are that it illustrates *changes* by expressing the items in a time series as *percentages* of the same item at a chosen *base* period.

A. SIMPLE (UNWEIGHTED) INDEX NUMBERS

In commercial and economic affairs there are some very important quantities which are too complex to be measured directly; such things as the "level of industrial production" or the "cost of living". These are real enough things, but they are made up of a very large number of component parts, all affecting the main issue in different ways or to different extents. Index numbers are especially suited to dealing with such matters.

You should note that an index number is sometimes called an *index* (plural: *indices*).

Simple Index

Let us first of all take a much over-simplified family food bill from which we want to show the effects of changes in price over a period of time, i.e. a very elementary cost of living index. Suppose we consider four items of food – milk, butter, tea and potatoes – and we know the average weekly consumption of these items in a typical household for Year 1 and Year 10 and the price of each item, as set out in Table 6.2:

Table 6.2: Household consumption of food items

Item	Year 1		Year 10	
	Average Weekly Consumption	Price	Average Weekly Consumption	Price
Milk	8 pints	5p/pt	5 pints	21p/pt
Butter	1 kg	40p/kg	500g	100p/kg
Tea	250g	60p/kg	125g	200p/kg
Potatoes	8 lbs	2p/lb	6 lbs	7p/lb

We somehow want to combine the price per given unit of each of the items so that we have a single index number comparing prices in Year 10 with those in Year 1.

- Because we are combining the four food items into a single "shopping basket", the index is called an *aggregative index*.
- It makes no allowance for the different quantities of item used – butter as compared to tea, for example – and so is called a *simple index*.
- Finally, because we are comparing prices, it is called a *price index*.

All we do in this extremely simple situation is total the prices for each given unit for each year and express that for Year 10 as a percentage of that for Year 1:

Simple aggregative price index (Year 10 compared to Year 1)

$$= \left[\frac{21 + 100 + 200 + 7}{5 + 40 + 60 + 2}\right] \times 100 = \frac{328}{107} \times 100 = 306.5$$

This tells us that prices as a whole were more than three times higher in Year 10 than in Year 1, i.e. prices had increased by 206.5 per cent in that period.

Notes:

(a) We shall work out all the index numbers in this chapter correct to one decimal place. This is precise enough for most comparisons, and particularly in times of rapid inflation when there are large changes in price indices.

(b) There are no units to an index number as we are expressing one price as a percentage of another price. We must remember to have our prices in the same units, in this case pence, when calculating an aggregative index.

(c) Year 1 is the base year in this example. Instead of having to state this every time in words, it is customary to write: Price index for Year 10 $=$ 306.5 (Year 1 $=$ 100).

You may already have some criticisms to make of this simple approach to constructing an index. Firstly, it depends on the units of the commodities given. Suppose for tea we had said that its price was 30p/500g and 100p/500g instead of 60p/kg and 200p/kg, then:

Simple aggregative price index for Year 10 (Year 1 $=$ 100)

$$= \left[\frac{21 + 100 + 100 + 7}{5 + 40 + 30 + 2}\right] \times 100 = \frac{228}{77} \times 100 = 296.1$$

In other words, we get a completely different value for the index, which is obviously unsatisfactory.

Price Relatives

We can get round this problem by using the *ratio of prices* of a given item rather than the actual prices themselves. Thus, the price of a pint of milk in Year 10 as a percentage of its price in Year 1 is:

$$\frac{21}{5} \times 100 = 420.0.$$

This ratio, 420.0, is called the *price relative* for milk in Year 10 (Year 1 $=$ 100).

Similarly, we can work out price relatives for the other items. (Remember, all we are doing is making the current price into a percentage of the base year price.)

Table 6.3: Commodity price relatives

Commodity	Price Relatives in Year 10 (Year 1 = 100)
Milk	$\frac{21}{5} \times 100 = 420.0$
Butter	$\frac{100}{40} \times 100 = 250.0$
Tea	$\frac{200}{60} \times 100 = 333.3$
Potatoes	$\frac{7}{2} \times 100 = 350.0$

From these price relatives we can now construct another index number called the *arithmetic mean of price relatives*, which is calculated as follows:

Arithmetic mean of price relatives for Year 10 (Year 1 $=$ 100)

$$= \frac{420.0 + 250.0 + 333.3 + 350.0}{4} = \frac{1{,}353.3}{4} = 338.3$$

In other words, on this basis prices in general appear to have risen 238 per cent over the given period.

Another advantage of this price-relative type of index number is that the prices of all the commodities do not have to be in the same units, although the prices of *each individual item* must be in the same units. This is a useful feature if you are dealing with results from different countries.

Since price relatives are rates of change, a more accurate average would be obtained by calculating the geometric mean rather than the arithmetic mean. It is possible to compute the *geometric mean of price relatives* as the n^{th} root of the product of the price relatives:

Geometric mean of price relatives for Year 10 (Year 1 $=$ 100)

$$= \sqrt[4]{420 \times 250 \times 333.3 \times 350} = 332.7$$

B. WEIGHTED INDEX NUMBERS (LASPEYRES AND PAASCHE INDICES)

You may think that the mean of price relatives index is still not very satisfactory, in that all items are treated as of equal importance and no account has been taken of the different quantities of the items consumed. For instance, the average family is much more concerned about a 5p increase in the price of a loaf of bread than a 10p increase in the price of a drum of pepper, as far more bread is consumed than pepper.

If you look back at Table 6.2 you will see that we are, in fact, given the average weekly consumption of each item in Year 1 and Year 10. You can see that the consumption pattern, as well as the prices, has changed over the 10-year period. We are interested in calculating an index for *prices*, so we have to be careful not to overemphasise the increase in prices by incorporating the changes in consumption.

Weighted Aggregative Index Numbers

We can adopt either of two approaches:

(a) We can consider the consumption pattern in Year 1 as typical and:

(i) work out the total expenditure on the four items in Year 1; then,

(ii) work out what the total expenditure would have been in Year 10 if the family had consumed the same items at Year 1 levels; and finally,

(iii) express the sum in (ii) as a percentage of (i) to form an index number.

This index is called a *base-weighted aggregative index* and in our example we work as follows:

Year 1 values are (Year 1 consumption × Year 1 prices)

Year 10 values are (Year 1 consumption × Year 10 prices)

In other words, we assume the consumption has not changed, only the prices.

The resulting table of values is:

Table 6.4: Expenditure using Year 1 consumption

Item	**Year 1**	**Year 10**
	Expenditure Using Year 1 Consumption	**Expenditure Using Year 1 Consumption**
Milk	40	168
Butter	40	100
Tea	15	50
Potatoes	16	56
Total	111	374

Base-weighted aggregative index of prices in Year 10 (Year 1 = 100)

$$= \frac{374}{111} \times 100 = 336.9$$

This type of index, where the weights are derived from quantities or values consumed in the base period, is known as a *Laspeyres index*, after the nineteenth-century economist of that name.

The main defect of a Laspeyres index is that the weights become out of date as the pattern of demand changes. A Laspeyres index tends to overstate the change in prices, as it takes no account of the fall in consumption when prices rise.

(b) The alternative method is to regard Year 10 as typical and to work all the figures as before, except that this time assume Year 10 consumption in Year 1.

This index is called the *current-weighted aggregative index*. For our example we have:

Table 6.5: Expenditure using Year 10 consumption

Item	Year 1	Year 10
	Expenditure Using Year 10 Consumption	**Expenditure Using Year 10 Consumption**
Milk	25	105
Butter	20	50
Tea	7.5	25
Potatoes	12	42
Total	64.5	222

Current-weighted aggregative index of prices in Year 10 (Year 1 = 100)

$$= \frac{222}{64.5} \times 100 = 344.2$$

This type of index, where the weights are derived from quantities or values consumed in the current period, is known as a *Paasche index* after the nineteenth-century economist of that name.

The main defect of a Paasche index is that new weights have to be ascertained each time the index is calculated, and this involves time-consuming and expensive survey work. A Paasche index tends to *understate* the changes in prices, as most people tend to buy less of those commodities which have gone up in price.

Weighted Price-relative Index Numbers

We can also form base-weighted or current-weighted price-relative index numbers. As before, we work out the price relatives for each commodity and as we now want to take into account the relative importance of each item in the family budget, we use as weight the actual *expenditure* on each item. The expenditure is used rather than the quantities consumed, to avoid exaggeration of variations arising from the change in consumption pattern rather than the change in price.

(a) Base-weighted Price-relative Index Number (Laspeyres)

Table 6.6: Example of calculation of Laspeyres index

Item	**Price Relative**	**Expenditure in Year 1 (Weight)** *(pence)*	**Price Relative × Weight** *(pence)*
Milk	420.0	40	16,800
Butter	250.0	40	10,000
Tea	333.3	15	5,000
Potatoes	350.0	16	5,600
	Total	111	37,400

Base-weighted price-relative index for Year 10 (Year 1 = 100)

$$= \frac{\sum(\text{price relative} \times \text{weight})}{\sum \text{weights}} = \frac{37{,}400}{111} = 336.9$$

(b) Current-weighted Price-relative Index Number (Paasche)

Table 6.7: Example of calculation of Paasche index

Item	**Price Relative**	**Expenditure in Year 1 (Weight)** *(pence)*	**Price Relative × Weight** *(pence)*
Milk	420.0	105	44,100
Butter	250.0	50	12,500
Tea	333.3	25	8,333
Potatoes	350.0	42	14,700
	Total	222	79,633

Current-weighted price-relative index for Year 10 (Year 1 = 100)

$$= \frac{\sum(\text{price relative} \times \text{weight})}{\sum \text{weights}} = \frac{79{,}633}{222} = 358.7$$

C. FISHER'S IDEAL INDEX

The American economist Irving Fisher made a major study of index numbers in the early 1920s. He proposed a number of tests, which could be applied to decide if an index number was acceptable or not. One of these was called the factor reversal test. This states that if the prices and quantities used in an index number are exchanged, then the product of the two index numbers should be an index of total expenditure.

In practice, none of the commonly used index numbers can satisfy this test. However, an index number designed by Fisher, which he called the "ideal index", does meet the test. It also meets another test of a similar nature called the time reversal test. This index is found using the formula:

$$\text{index number} = \sqrt{\frac{\sum(p_1q_0)}{\sum(p_0q_0)} \times \frac{\sum(p_1q_1)}{\sum(p_0q_1)}} \times 100$$

It is in fact the geometric mean of the Laspeyres and Paasche index numbers (the meaning of p_1, q_0 etc. is given in the next section).

This index has not been widely used, not least because it requires much heavier computation than many other index numbers, and cannot be easily "explained" in the way that an aggregative index number can. Although the computations could now be carried out without trouble using a computer, the ideal index has not found any more favour than it did when it was first suggested.

D. FORMULAE

It will be useful at this stage to summarise our results so far by using formulae. We use the normal notation:

p_o = base year prices

p_1 = current year prices

q_o = base year quantities

q_1 = current year quantities

n = number of commodities considered.

We have the following results:

- Price relative for current year $= \frac{\text{current price}}{\text{base price}} \times 100 = \frac{p_1}{p_0} \times 100$ *(1)*
- Simple aggregative price index $= \frac{\sum \text{current price}}{\sum \text{base price}} \times 100 = \frac{\sum p_1}{\sum p_0} \times 100$ *(2)*
- Arithmetic mean of price relatives $= \frac{\sum \text{price relatives}}{n}$ *(3)*
- Geometric mean of price relatives $= \sqrt[n]{\text{product of price relatives}}$ *(4)*
- Base-weighted aggregative price index (Laspeyres)

  $$= \frac{\sum(\text{Current price} \times \text{Base quantity})}{\sum(\text{Base price} \times \text{Base quantity})} \times 100 = \frac{\sum(p_1q_0)}{\sum(p_0q_0)} \times 100 \qquad (5)$$

- Current-weighted aggregative price index (Paasche)

  $$= \frac{\sum(\text{Current price} \times \text{Current quantity})}{\sum(\text{Base price} \times \text{Current quantity})} \times 100 = \frac{\sum(p_1q_1)}{\sum(p_0q_1)} \times 100 \qquad (6)$$

- Weighted price-relative index $= \frac{\sum(\text{price relatives} \times \text{w eight})}{\sum \text{w eight}}$ *(7)*
- Ideal price index (Fisher's) = $= \sqrt{\frac{\sum(p_1q_0)}{\sum(p_0q_0)} \times \frac{\sum(p_1q_1)}{\sum(p_0q_1)}} \times 100$ *(8)*

In (7), for a base-weighted price-relative index use (Base price × Base quantity) as the weight. For a current-weighted price-relative index, use (Current price × Current quantity) as the weight.

In trying to remember these it is probably simplest to memorise the price-relative, Laspeyres and Paasche formulae, and to deduce the others from their descriptive names.

E. QUANTITY OR VOLUME INDEX NUMBERS

You must not think that we are always concerned with *price* indices. Often we are interested in *volume* or *quantity* indices as, for instance, in the Index of Industrial Production which seeks to measure the changes in volume of output in a whole range of industries over a period of time. We can calculate such quantity index numbers in exactly the same sort of way as we dealt with the price indices, for example:

- Quantity relative of a commodity in current year relative to base year $= \frac{q_1}{q_0} \times 100$
- Base-weighted aggregative quantity index (Laspeyres) $= \frac{\sum(q_1p_0)}{\sum(q_0p_0)} \times 100$
- Base-weighted quantity-relative index (Paasche) $= \frac{\sum\left(\frac{q_1}{q_0} \times 100\right) \times (p_0q_0)}{\sum(p_0q_0)}$

NB: there is no need to memorise these as they are really the same formulae with quantity substituted for price.

Notes:

(a) The *price* of a commodity is now used as the weight for an aggregative quantity index and the *expenditure* on that commodity is used as the weight for a quantity-relative index.

(b) It is usual, if we are considering the situation from a producer's point of view rather than the consumer's, to call the index numbers *volume* indices and $\Sigma(p_0q_0)$, for example, will be the total *value* of production in the base year.

(c) Remember that for any commodity at any one time:

value = price × volume (producer's view)

expenditure = price × quantity (consumer's view).

Worked Example

Table 6.8 shows UK imports of board from Finland. Calculate a base-weighted **price** Laspeyres index for all types of board for Year 3 (Year 1 = 100).

Table 6.8: Imports of board

Type	Year 1		Year 3	
	Quantity (q_o) *(000 tonnes)*	**Value (p_oq_o)** *(£m)*	**Quantity (q_1)** *(000 tonnes)*	**Value (p_1q_1)** *(£m)*
Machine glazed	90	300	180	650
Folding box board	70	250	10	30
Kraft board	180	550	240	650
Woodpulp board	90	250	80	230
Other board	40	100	100	250

As we are asked for a price index, we must first calculate the price per tonne for each type of board using:

$$\text{value} = \text{price} \times \text{quantity, i.e. price} = \frac{\text{value}}{\text{quantity}}$$

Table 6.9: Price per tonne

Year 1	Year 3
Price (p_o) *(£000/tonne)*	**Price (p_1)** *(£000/tonne)*
3.33	3.61
3.57	3.00
3.06	2.71
2.78	2.88
2.50	2.50

We have now to decide whether to use an aggregative index or a price-relative index. We are asked to find a base-weighted index. Interestingly, we should obtain the same answer whichever method we choose. However, there is less calculation involved in this particular example if we choose an aggregative type, so this is the one we shall work first; we will try the other later.

Base-weighted aggregative price index for Year 3 (Year 1 = 100):

$$= \frac{\text{total value at Year 3 prices and Year 1 quantities}}{\text{total value at Year 1 prices and Year 1 quantities}} \times 100$$

We have the Year 1 values in column two of Table 6.8 so we need only sum that column to get the denominator of the expression: £1,450 million.

The numerator is the sum of the product of column one (q_0) in Table 6.8 and column two (p_1) in Table 6.9:

Table 6.10: Value at Year 3 prices, Year 1 quantities

	Value (p_1q_o) at Year 3 prices, *(£m)* **Year 1 quantities**
	324.9
	210.0
	487.8
	259.2
	100.0
Total	1,381.9

Index for Year 3 $= \frac{1{,}381.9}{1{,}450} \times 100 = 95.3$ (to one decimal place).

Therefore, there was an overall *decrease* in prices of 4.7 per cent over the period Year 1 to Year 3.

You can check that using the price-relative method gives the same results. You will need a column for the price relatives and a column for the price relatives weighted with the base-year values:

Table 6.11: Price relatives and price relatives weighted with base year values

$\frac{p_1}{p_0} \times 100$	$\frac{p_1}{p_0}\left(p_0q_0\right) \times 100$
108.4	32,520
84.0	21,000
88.6	48,730
103.6	25,900
100.0	10,000
Total	138,150

Base-weighted price-relative index for Year 3 (Year 1 $=$ 100):

$= \frac{138{,}150}{1{,}450} = 95.3$ (to one decimal place as before).

You will see that this must be so by simplifying the base-weighted price-relative formula. There is not an equivalent rule for current-weighted indices, though.

F. CHANGING THE INDEX BASE YEAR

To convert indices from an earlier to a later base year, divide all the indices by the index for the new base year. This is really a variation on the technique of chain-based indices, except that we relate to one particular year rather than continuing to roll forward. We also multiply by 100 to regain percentage values.

The following indices have a base year of 1995 = 100:

Table 6.12: Indices with base year of 1995

Year	1998	2001	2004	2007
Index	115	126	142	165

We will now convert to base year 1998 = 100 by dividing each index by 115 (1998 index) and multiplying by 100. You will immediately notice that the 1998 index becomes 100 as intended:

Table 6.13: Indices converted to base year 1998

Year	1998	2001	2004	2007
Index	100	$\frac{126}{115} \times 100 = 110$	$\frac{142}{115} \times 100 = 123$	$\frac{165}{115} \times 100 = 143$

Also, the 1995 index becomes $\frac{100}{115} \times 100 = 87$

Worked Example

Table 6.14 shows the average weekly earnings of male workers (aged 21 and over) during the years 1970–78. Also shown is the value of the RPI for these years with 1962 as base period. Determine the real average weekly earnings over the period 1970-78.

Table 6.14: Average weekly earnings of male workers

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978
RPI (1962 = 100)	140.2	153.4	164.3	179.4	208.1	258.5	301.3	349.1	378.0
Earnings (£)	28.05	30.93	35.82	40.92	48.63	59.58	66.97	72.89	83.50

After calculation as above, we obtain:

Table 6.15: Real earnings

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978
RPI (1970 = 100)	100	109.4	117.2	128.0	148.4	184.4	214.9	249.0	269.6
Real Earnings (£)	28.05	28.27	30.56	31.97	32.77	32.31	31.16	29.27	30.97

We thus see that, although from 1970 to 1978 the average weekly earnings had apparently jumped by £55.45, i.e. increased by almost 200 per cent, the real purchasing power had increased by £2.92 or 10 per cent.

With inflation, the public becomes more aware of index numbers, e.g. index-linked pensions, savings, insurance premiums, etc. However, index numbers are of necessity imperfect measures, the values of which can be manipulated by changes in base year or in the weighting system. For pensions, the decision has to be made whether to link them with earnings or with prices, and if with earnings, the earnings of whom: manual workers, all workers, workers in the same industry? With house insurance premiums, is the index used to be based on the estimated market value of the house or on the cost of clearing the site and rebuilding the house? There is increasing discussion on these matters in the press, so do be on the lookout for such articles and relate them to your own knowledge of how index numbers are constructed.

G. INDEX NUMBERS IN PRACTICE

Practical Problems with Index Numbers

In the previous sections you learned about the basic principles involved in the calculation of index numbers. In the rest of this chapter we will deal with the problems which arise when we try to put these principles into practice.

An index number is nothing more than a specialised kind of average, and you must avoid treating it as anything more than that. Index numbers are very useful for indicating changes in the levels of economic activity. But remember that there is no *direct* measure of things such as "cost of living" or "industrial production"; what we have to do is to decide on the things that are relevant and calculate an index. For example, the correct description of the well-known index measuring changes in the cost of living is the Index of Retail Prices.

The main practical points that arise when constructing index numbers are:

- What items to include.
- How to get data relating to the items.
- What weighting factors to use.
- What kind of averaging to use.
- What period to choose as the base period.

We will now deal with each of these in turn.

What Items to Include

A practical index number is, in fact, a particular sort of average of a number of values, such as prices, export values, etc. An average relates to a collection of items, and therefore we have to consider, in planning an index number, what items to use in the average. The main rule to be followed is that the items *must be representative of the topic covered by the index*. In an index of prices intended to show the cost of living for middle-class town families, it would be silly to include the price of a Park Lane flat or the wages of a gamekeeper! Within the limits of this rule, every case must be considered on its own merits.

How many items to include is, of course, part of the problem. As many items as practicable should be included, although the labour of calculation each time the index is worked out should not be excessive. In this connection, the use of computers has enabled some indices to be based on hundreds of items and yet still be quickly and accurately calculated on a daily basis.

How to Obtain the Data

Taking an index of prices as our example, what data do we need to collect? We need data to arrive at:

- The figure to be used as the *price* of each item.
- The figure to be used as the *weight* of each item.

If butter is one of the items in an index of prices, then we know that the price at any time may vary between different places, different shops in the same place and different kinds of butter. If our index is being calculated monthly, then we must remember that the price of the items will also vary throughout the month. Clearly then, the price must be some form of average, and since it is impracticable to average all the prices of all the items in all the shops, some form of *sample survey* must be undertaken. We have covered the principles and techniques of sample surveys, and these must be applied when carrying out price surveys for the calculation of index numbers.

The weights used in calculating a price index are usually based on the relative *quantities* of commodities consumed over a certain period or the relative *values* expended on the commodities. Again, the quantities or values are usually determined by sample surveys; the actual weights are not necessarily the absolute results obtained, but some more convenient numbers which are proportional to the results. Remember – it is the *relative* sizes of the weights, not the actual sizes, that matter. Weights of 2, 7, 10 and 6 will give the same results in a weighted average as weights of 4, 14, 20 and 12, since each is *multiplied* by a constant factor, in this case 2, though it would not do to *add* the same number to each – try it and see!

What Weights to Use

We have to choose between base-period weighting and current-period weighting. A system where the weights are based on the quantities or values of commodities consumed in the base period is known as *base-period weighting*. If the weights are derived from the quantities or values consumed in the current period (for which the index is being calculated), the system is known as *current-period weighting*. As we have seen, the two systems also have names which come from those of two 19th-century economists:

- base-period weighting: Laspeyres index
- current-period weighting: Paasche index.

Theoretical economists argue about the relative merits of these two kinds of index number. There is often not much difference between them, but with higher inflation the differences become larger. The main defect of a Laspeyres index is that the weights become out of date as the pattern of demand changes. The main drawback of a Paasche index is the additional survey and calculation work required each time the index is to be worked out.

Because people tend to buy more of a commodity when it is cheaper and less when it is dearer, prices and weights have an influence on each other. The effect, as you will remember, is that:

- A Laspeyres index tends to overstate the changes in prices because it takes no account of the fall in consumption when prices rise.
- A Paasche index tends to understate the changes in prices as people buy less of those commodities which have gone up most in price.

There are various possible refinements to overcome the difficulty of deciding on current-period or base-period weighting. We can use "typical" period weighting using the quantities consumed in some representative period between the base period and the current period. Alternatively we can use the arithmetic mean of the base-period and current-period quantities as weights, giving the *Marshall-Edgeworth index*. Another possible index, as we have seen, is *Fisher's ideal index* which is the geometric mean of the Laspeyres and Paasche indices.

The Kind of Average to Use

We have mentioned several times that an index number is no more than an average of a special kind. So far, we have taken the word "average" to mean "arithmetic mean". But, there are other kinds of average and, in particular, we could have used the *geometric mean*. At this point you might well go back to the last chapter and revise how to calculate geometric means. Then the following example, which uses data from earlier in the chapter, will be easier to follow.

We are going to calculate a base-period weighted geometric mean of price-relative index number using data as shown in Table 6.16.

Table 6.16: Price relatives

	Year 10 Price Relative *(Yr 1 = 100)*	**Log Price Relative**	**Expenditure Weight (Base Year)** *(Pence)*	**Weight × Log**
Milk	420	2.6232	40	104.928
Butter	250	2.3979	40	95.916
Tea	333.3	2.5229	15	37.8435
Potatoes	350	2.5441	16	40.7056
			111	279.3931

$$\text{log (Index number)} = \frac{279.3931}{111} = 2.5171$$

$$\text{index number} = 328.9$$

Notice that this is *less* than the index number of 336.9 obtained earlier by using the weighted arithmetic mean. This is a general property of a geometric mean – it is always *less* than the arithmetic mean of the same figures. You should note that we have not calculated the geometric mean of the weighted price relatives directly, but instead we worked out a weighted arithmetic mean of the logs of the price relatives and then took the antilog of the answer. This is obviously more trouble to calculate than a weighted arithmetic mean of the price relatives.

An advantage claimed for the geometric mean is that it is less influenced by occasional extreme items than is the arithmetic mean. This is true, but the advantage is slight when large numbers of items are involved. The result is that, in practice, there is not much to choose between the two sorts of average, and the arithmetic mean is most often used. This does not deter examiners from setting questions about geometric means, but do not calculate a geometric-type index unless *specifically* requested. One special point to watch when using a weighted geometric mean is that it is applicable only to a price-relative type index and *not* an aggregative type index.

The one commonly known index that is calculated in this way is the *Financial Times Industrial Ordinary Share Index*. The index is based on the share prices of only 30 large companies. The effect of using a geometric mean is to stop a large fluctuation in the value of the index when only one of these shares shows a sudden rise or fall in price.

What Period to Use as Base

The main point about the base period is that it should, in some way, represent a standard or normal period. It is an economics problem to decide upon a standard period rather than a statistical one, and so we will not dwell too closely on it. However there are a few points about base periods that we do have to deal with.

Firstly, consider how long the base period should be. As with so many questions of this kind, there is no one correct answer. Sometimes you will see an index number the base period of which is one day (e.g. 14th July 2006 = 100), sometimes one month (e.g. July 2006 = 100) and sometimes one year (e.g. 2006 = 100). The general principle behind the decision about length of base period is that the period should be long enough to enable a good average to be taken, without including exceptional times such as booms or depressions. When a long base period is decided upon, it should not exceed a year, and it should preferably be exactly one year so as to include each season once only. We had a similar consideration when dealing with moving averages.

One difficulty about base periods is that, by the very nature of the changing conditions that we are trying to measure, they eventually become out of date and unrepresentative. A base period in which wax candles were in wide use could hardly be representative of life in the age of electric light! There are three customary ways of counteracting (to some extent) this tendency to become out of date:

- Using a Laspeyres index, bring the base period up to date at regular intervals.

  When the base period of an index number is changed, in order to make direct comparisons with historical values, we may need to splice two index series. For example, we may be given:

  Table 6.17: Index series with different bases

	1998	1999	2000	2001	2002	2003	2004	2005
Price index (1990 = 100)	123	135	147	160	182	193	-	-
Price index (2003 = 100)	-	-	-	-	-	100	110	115

  We thus have two separate index series for the same commodities, but after 2003 the base year has been updated to 2003 from 1990. To enable us to compare all the price indices directly, we must recalculate the values for 2004 and 2005 based on 1990.

$$\text{Price index for 2004 (1990 = 100)} = \frac{\text{price index for 2004 (2003 = 100)}}{\text{price index for 2003 (2003 = 100)}} \times \text{price index for 2003 (1990 = 100)}$$

$$= \frac{110}{100} \times 193 = 212.3$$

$$\text{Price index for 2005 (1990 = 100)} = \frac{\text{price index for 2005 (2003 = 100)}}{\text{price index for 2003 (2003 = 100)}} \times \text{price index for 2003 (1990 = 100)}$$

$$= \frac{115}{100} \times 193 = 222.0$$

- Using a Paasche index, the weights are automatically brought up to date each time.
- Use what is called the *chain-base* method. Here the index for the current period is based on the last (i.e. the immediately preceding) period. For example, if we are calculating an index for Year 3, we use Year 2 as the base year; then, when we come to calculate the index for Year 4, we use Year 3 as the base year; and so on. This system has the advantage that it is always up to date and it is easy to introduce new

items or delete old ones gradually without much upset to the reliability of the index. Its disadvantage is that it cannot be used for making comparisons over long periods of time, as we are simply comparing each year with the immediately preceding year.

If we do need to make long-term comparisons when a chain-base index number is in use, then it is necessary to convert the indices from a *chain base* to a *fixed base*. Such conversions are a favourite topic with some examiners. The method of working is shown in the following two examples.

Example 1:

The indices for the years 9, 10, 11, (Year 8 as base = 100) are:

Year 9:	104
Year 10:	104
Year 11:	109

We are required to convert these to a chain-base set of indices. The Year 8 index remains the same at 100; the Year 9 index (based on Year 8) is still 104; the Year 10 index (based on Year 9) is 100 because the two years have the same index; the Year 11 index (based on Year 10) is (109 × 100)/104 = 105.

Example 2:

The following indices were arrived at by a chain-base method. Convert them to Year 7 as a fixed base.

Year 7:	100
Year 8:	106
Year 9:	110
Year 10:	95
Year 11:	100

The Year 7 index remains at 100; the Year 8 index (based on Year 7) remains at 106; the Year 9 index (based on Year 8) is 110 and therefore the Year 9 index will always be 110/100 of the Year 8 index, no matter what base is used. Now, the Year 8 index (based on Year 7) is 106, and so the Year 9 index (based on Year 7) is (110 × 106)/100 = 116.6. Similarly, the Year 10 index will be 95/100 of the Year 9 index, no matter what base is used. So the Year 10 index (based on Year 7) is (95 × 116.6)/100 = 110.8. The Year 11 index (based on Year 10) is 100 and therefore there is no change from Year 10 to Year 11; the Year 11 index (based on Year 7) is consequently the same as the Year 10 index (based on Year 7), namely 110.8.

Criteria for a Good Index

(a) Easy to Understand

An ordinary person can appreciate the influence of the average change in the cost of a fixed bundle of goods on the cost of living or the cost of production, so this type of index is widely used. Thus the Laspeyres and Paasche indices and the means of price relatives are easily understood. However, the Fisher ideal index and others that use geometric means are of more theoretical interest and are more difficult concepts to grasp.

(b) Reliable

Reliability is difficult to define in connection with index numbers as it need not imply a high level of accuracy. The ability to reflect realistically the type of change which is taking place is more important than numerical accuracy. Reliability depends mainly on the weighting used, and so is affected to some extent by the experience of the person constructing the index. In particular, any change in the system of weighting should not generate a major change in the index. Thus a chain-based index is likely to be more reliable than a fixed-base index.

(c) Cost-effective

It is not practicable to spend a considerable amount of time and money in constructing an index that will appeal to only a few specialists. An index must be produced for the minimum cost and appeal to the maximum number of people. In assessing the effectiveness of an index, you must ensure that the delay between the end of the relevant time period and the publication of the index is a minimum, i.e. the initial analysis of any new survey material must be kept to a minimum. In this context the Laspeyres index has the advantage over the Paasche index.

(d) Base Year Up-to-date

Base-weighted indices are the most popular, so the bundle of goods in the base year must be similar to that of the current year. Changes in the bundle take place gradually under normal conditions, so a regular updating of the base year will allow for these changes. For example, indices formed from a family expenditure survey need to be continually updated to take account of changing patterns of family expenditure.

Indices are used to compare conditions for the same bundles in different geographical areas or different bundles in the same area. These comparisons will be valid only if all the indices have the same base year, which should be a recent year. The method of updating an index was mentioned earlier.

Index Numbers in Use

There are several index numbers published and used in the UK. In this section we will look at two well-known ones that are of most interest to the public.

(a) Retail Price Index

This is compiled by the Office for National Statistics, and its purpose is to measure the relative change, every month, of a bundle of goods that represents the expenditure of an average family in that month. For each article or service included in the bundle, the current price is expressed as a relative of the price at the base date. Currently (2008) the base year is 1987.

Each article is weighted with weights derived from an annual government survey called the *Family Spending*. The bundle includes such items as food, housing, fuel and light, household goods, clothing, transport and services. Prices are collected once a month from a number of retail outlets randomly selected throughout the country.

This index is widely used in settling wage claims, but because of the omission of certain items, notably income tax payments, it should not be interpreted as a cost of living index.

(b) Indices of Producer Prices

These indices, produced monthly by the Office for National Statistics, are designed to measure the changes in the output prices of home sales of manufactured goods (the output prices index) and the prices of inputs (i.e. materials and fuels) used by manufacturing industries (the input price index).

These indices are extremely useful sources of market intelligence and very useful to industry.

Chapter 7

Correlation

Contents ***Page***

Introduction **100**

A. Scatter Diagrams **100**
Examples of Correlation 100
Degrees of Correlation 102
Different Types of Correlation 103

B. The Correlation Coefficient **104**
Formula 105
Characteristics of a Correlation Coefficient 106
Significance of the Correlation Coefficient 106
Note on the Computation of R 107

C. Rank Correlation **108**
Relationship between Ranked Variates 108
Ranked Correlation Coefficients 110
Tied Ranks 112

INTRODUCTION

When studying frequency distributions we were always handling only *one variable*, e.g. height or weight. Having learned how to solve problems involving only one variable, we must now discover how to solve problems involving *two variables* at the same time.

For example, if we are trying to assess, for one firm (or a group of firms), whether there is any relationship between revenues and profits, then we are dealing with two variables, i.e. revenues and profits.

Correlation analysis is concerned with measuring whether two variables are associated with each other. If two variables tend to change together in the same direction, they are said to be positively correlated. If they tend to change together in opposite directions, they are said to be negatively correlated. In the example given above, we might expect higher revenues to be associated with higher profits – so we would expect revenues and profits to be positively correlated. On the other hand, higher interest rates may lead to lower profits for many firms – so we would expect interest rates and profits to be negatively correlated. In order to plan for the future, it could be useful in business to know the strength of association between variables such as prices and sales, employee training and productivity, and unit costs and production levels.

However, it is important to emphasise that correlation between two variables does not necessarily mean that changes in one of the variables are causing changes in the other. Indeed, with time-series data, it is possible for spurious correlation to arise. If two variables have strong upward trends, they are likely to be positively correlated even if there is no link between them at all.

A. SCATTER DIAGRAMS

Examples of Correlation

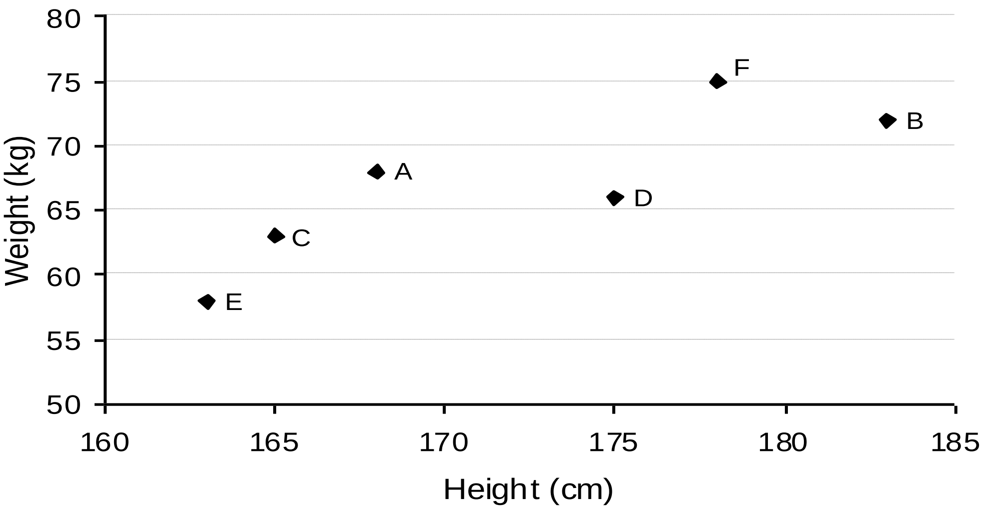
Suppose we have measured the height and weight of 6 men. The results might be as follows:

Table 7.1: Relationship of height and weight

Man	**Height** *(cm)*	**Weight** *(kg)*
A	168	68
B	183	72
C	165	63
D	175	66
E	163	58
F	178	75

A *scatter diagram* or *scattergraph* is the name given to the method of representing these figures graphically. On the diagram, the horizontal scale represents one of the variables (let us say height) while the other (vertical) scale represents the other variable (weight). Each *pair* of measurements is represented by one point on the diagram, as shown in Figure 7.1:

Figure 7.1: Scatter diagram of men's heights and weights



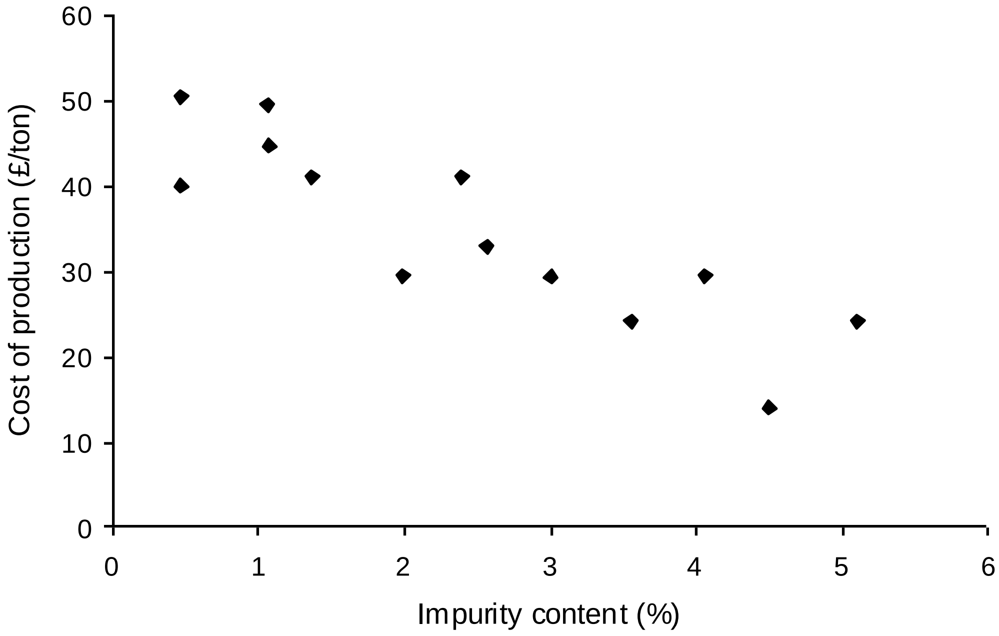
Make sure that you understand how to plot the points on a scatter diagram, noting especially that:

- Each point represents a *pair* of corresponding values.
- The two scales relate to the two variables under discussion.

The term scatter diagram or scattergraph comes from the scattered appearance of the points on the chart.

Examining the scatter diagram of heights and weights, you can see that it shows up the fact that, by and large, tall men are heavier than short men. This shows that some relationship exists between men's heights and weights. We express this in statistical terms by saying that the two variables, height and weight, are *correlated*. Figure 7.2 shows another example of a pair of correlated variables (each point represents one production batch):

Figure 7.2: Cost of production compared with impurity contents



Here you see that, in general, it costs more to produce material with a low impurity content than it does to produce material with a high impurity content. However, you should note that correlation does not necessarily mean an *exact* relationship, for we know that, while tall men are usually heavy, there are exceptions, and it is most unlikely that several men of the same height will have exactly the same weight!

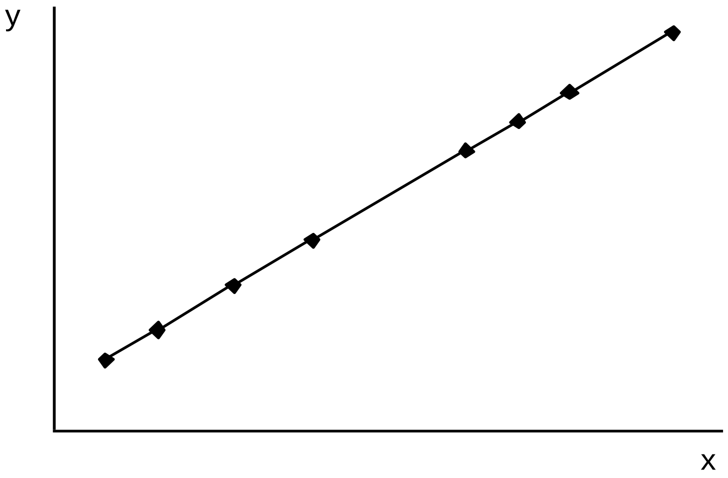
Degrees of Correlation

In order to generalise our discussion, and to avoid having to refer to particular examples such as height and weight or impurity and cost, we will refer to our two variables as x and y. On scatter diagrams, the horizontal scale is always the x scale and the vertical scale is always the y scale. There are three degrees of correlation which may be observed on a scatter diagram. The two variables may be perfectly correlated, uncorrelated or partly correlated.

(a) Perfectly Correlated

If the variables are perfectly correlated the points on the diagram all lie exactly on a straight line (Figure 7.3):

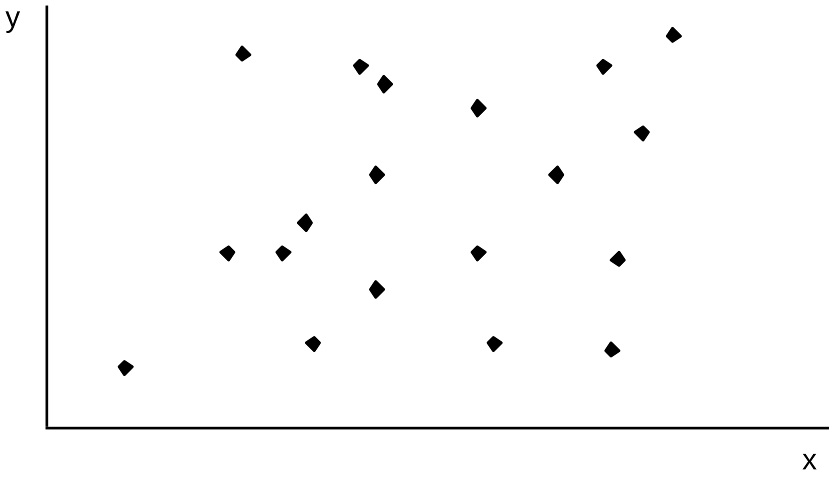
Figure 7.3: Perfect correlation



(b) Uncorrelated

If the points on the diagram appear to be randomly scattered about with no suggestion of any relationship, then the variables are uncorrelated (Figure 7.4):

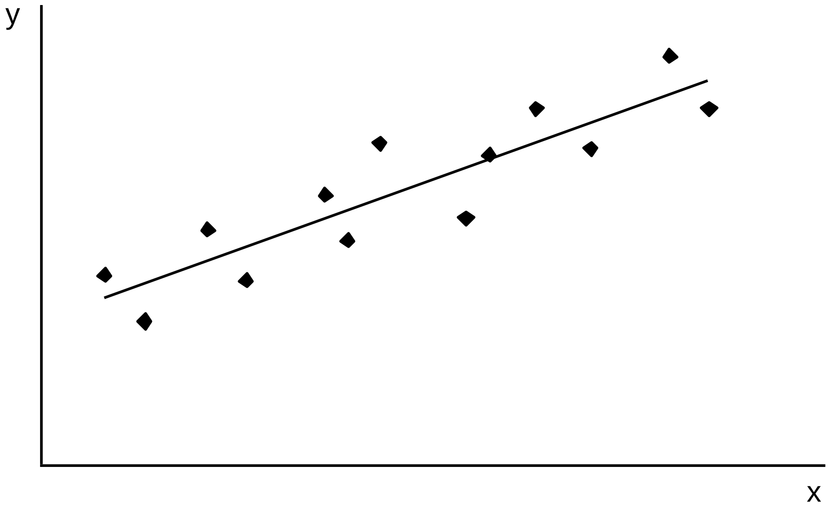
Figure 7.4: Uncorrelated



(c) Partly Correlated

Partly correlated means that the points lie scattered in such a way that, although they do not lie exactly on a straight line, they do display a general tendency to be clustered around such a line (Figure 7.5):

Figure 7.5: Partly correlated

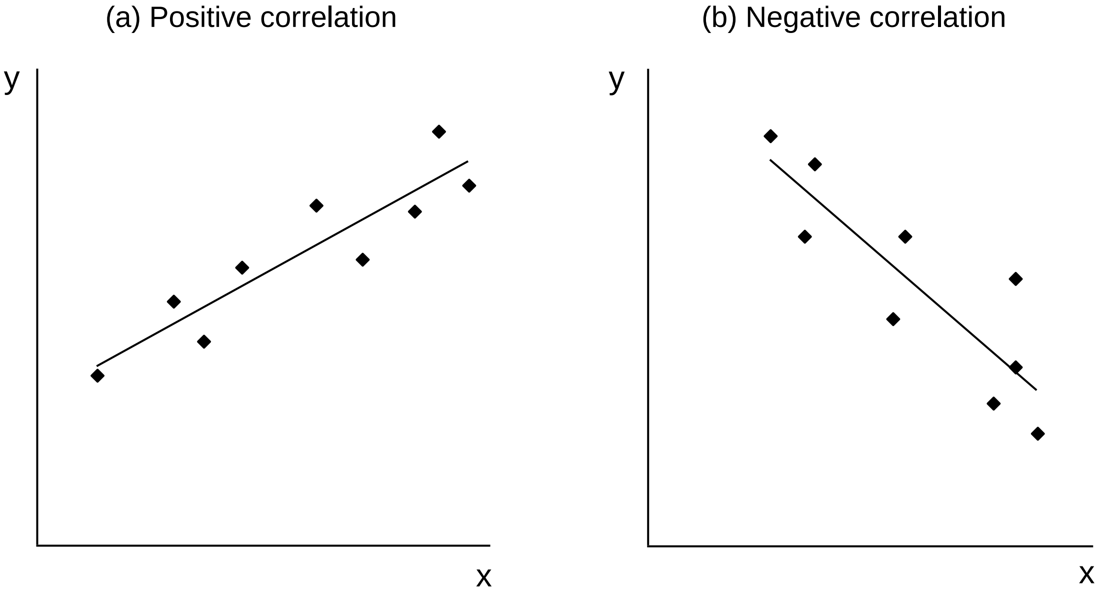


Different Types of Correlation

There is a further distinction between correlations of the height/weight type and those of the impurity/cost type. In the first case, high values of the x variable are associated with high values of the y variable, while low values of x are associated with low values of y. On the scatter diagram in Figure 7.6 (a), the points have the appearance of clustering about a line which slopes *up to the right*. Such correlation is called *positive* or *direct* correlation.

In the other case (like the impurity/cost relationship) high values of the x variable are associated with low values of the y variable and vice versa; on the scatter diagram (Figure 7.6 (b)) the approximate line slopes *down to the right*. This correlation is said to be *negative* or *inverse*.

Figure 7.6: Positive and negative correlation



(a) Linear Correlation

The correlation is said to be linear when the relationship between the two variables is linear. In other words, the scatter of points can be modelled by a straight line. For example, the correlation between car ownership and family income may be linear as car ownership is related in a linear fashion to family income.

(b) Non-linear Correlation

Non-linear correlation is outside the scope of this course, but it is possible that you could be required to define it in an examination question. It occurs when the relationship between the two variables is non-linear. An example is the correlation between the yield of a crop, like carrots, and rainfall. As rainfall increases so does the yield of the crop of carrots, but if rainfall is too large the crop will rot and yield will fall. Therefore, the relationship between carrot production and rainfall is non-linear.

B. THE CORRELATION COEFFICIENT

If the points on a scatter diagram all lie very close to a straight line, then the correlation between the two variables is stronger than it is if the points lie fairly widely scattered away from the line.

To measure the strength (or intensity) of the correlation in a particular case, we calculate a *linear correlation coefficient*, denoted by R. In textbooks and examination papers, you will sometimes find this referred to as *Pearson's product moment coefficient of linear correlation*, after the English statistician who invented it. It is also known as the product-moment correlation coefficient.

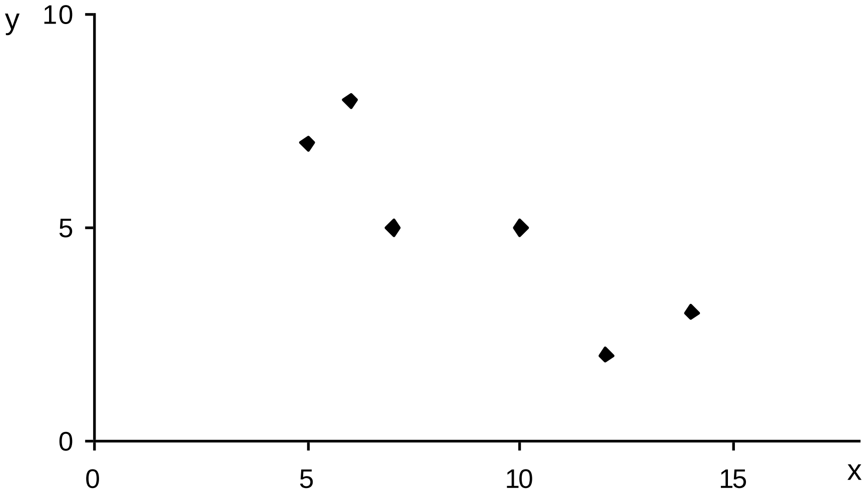
For an illustration of the method used to calculate the correlation coefficient, suppose we are given the following pairs of values of x and y:

Table 7.2: Illustration of correlated data

x	10	14	7	12	5	6
y	5	3	5	2	7	8

We shall plot these on a scatter diagram so that we can make some qualitative assessment of the type of correlation present (Figure 7.7). We see from the scatter diagram that some negative correlation appears to be present:

Figure 7.7: Scatter diagram of given data



Formula

The formula for Pearson's product-moment correlation coefficient (the proof is beyond the scope of this course) is:

$$R = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where n is the number of pairs of readings.

It is a good idea to set out the calculation in tabular form:

Table 7.3: Calculation of R

x	y	x^2	y^2	xy
10	5	100	25	50
14	3	196	9	42
7	5	49	25	35
12	2	144	4	24
5	7	25	49	35
6	8	36	64	48
$\Sigma x = 54$	$\Sigma y = 30$	$\Sigma x^2 = 550$	$\Sigma y^2 = 176$	$\Sigma xy = 234$

$n = 6$,

therefore $R = \dfrac{6 \times 234 - 54 \times 30}{\sqrt{(6 \times 550 - 54^2)\,(6 \times 176 - 30^2)}}$

$$= \frac{1{,}404 - 1{,}620}{\sqrt{(3{,}300 - 2{,}916)(1{,}056 - 900)}}$$

$= \dfrac{-216}{\sqrt{384 \times 156}} = \dfrac{-216}{\sqrt{59{,}904}} = \dfrac{-216}{244.75} = -0.88$ to two decimal places

This result ($R = -0.88$) shows that x and y are negatively correlated.

Note: in Chapter 8, you will learn how to find the equation of the least-squares regression line by calculating values for a and b in the linear equation, y = a + bx. If the value of b has already been calculated, you can find the correlation coefficient between x and y from the following simpler formula:

$$R = b\frac{\sigma_x}{\sigma_y}$$

where σ_x and σ_y are the standard deviations of x and y respectively.

Characteristics of a Correlation Coefficient

We know what the + and – signs of the correlation coefficient tell us: that the relationship is positive (increase of x goes with increase of y) or negative (increase of x goes with decrease of y). But what does the actual numerical value mean? Note the following points (the proofs are again beyond the scope of this course):

(a) The correlation coefficient is always between –1 and +1 inclusive. If you get a numerical value bigger than 1, then you've made a mistake!

(b) A correlation coefficient of –1.0 occurs when there is *perfect negative correlation*, i.e. all the points lie *exactly* on a straight line sloping down from left to right.

(c) A correlation of 0 occurs when there is *no correlation*.

(d) A correlation of +1.0 occurs when there is *perfect positive correlation*, i.e. all the points lie *exactly* on a straight line sloping upwards from left to right.

(e) A correlation of between 0 and + 1.0 indicates that the variables are *partly correlated*. This means that there is a relationship between the variables but that the results have also been affected by other factors.

In our example (R = –0.88), we see that the two variables are quite strongly negatively correlated. If the values of R had been, say, –0.224, we should have said that the variables were only slightly negatively correlated. For the time being, this kind of interpretation is all that you need consider.

Significance of the Correlation Coefficient

Correlation analysis has been applied to data from many business fields and has often proved to be extremely useful. For example, it has helped to locate the rich oil fields in the North Sea, and also helps the stockbroker to select the best shares in which to put clients' money.

Like many other areas of statistical analysis, correlation analysis is usually applied to sample data. Thus the coefficient, like other statistics derived from samples, must be examined to see how far it can be used to make generalised statements about the population from which the samples were drawn. *Significance tests* for the correlation coefficient are possible to make, but they are beyond the scope of this course, although you should be aware that they exist.

We must be wary of accepting a high correlation coefficient without studying what it means. Just because the correlation coefficient says there is some form of association, we should not accept it without some other supporting evidence. We must also be wary of drawing conclusions from data that does not contain many pairs of observations. Since the sample size is used to calculate the coefficient, it will influence the result and, whilst there are no hard and fast rules to apply, it may well be that a correlation of 0.8 from 30 pairs of observations is a more reliable statistic than 0.9 from 6 pairs.

Another useful statistic is R^2 (R squared); this is called the *coefficient of determination* and may be regarded as the percentage of the variation in y directly attributable to the variation in

x. Therefore, if you have a correlation coefficient of 0.8, you can say that approximately 64 per cent (0.8^2) of the variation in y is explained by variations in x. This figure is known as the *explained variation* whilst the balance of 36 per cent is termed the *unexplained variation*. Unless this unexplained variation is small, there may be other causes than the variable x which explain the variation in y, e.g. y may be influenced by other variables or the relationship may be non-linear.

In conclusion then, the coefficient of linear correlation tells you only part of the nature of the relationship between the variables: it shows that such a relationship exists. You have to interpret the coefficient and use it to deduce the form and find the significance of the association between the variables x and y.

Note on the Computation of R

Often the values of x and y are quite large and the arithmetic involved in calculating R becomes tedious. To simplify the arithmetic and hence reduce the likelihood of numerical slips, it is worth noting the following points:

(a) we can take any constant amount off every value of x;

(b) we can take any constant amount off every value of y;

(c) we can divide or multiply every value of x by a constant amount;

(d) we can divide or multiply every value of y by a constant amount,

all without altering the value of R. This also means that the value of R is independent of the units in which x and y are measured.

Let us consider the previous example as an illustration. We shall take 5 off all the x values and 2 off all the y values to demonstrate that the value of R is unaffected. We call the new x and y values, x' (x-dash) and y' respectively:

Table 7.4: Calculation of R

x	**y**	**x′**	**y′**	**$(x')^2$**	**$(y')^2$**	**x′y′**
10	5	5	3	25	9	15
14	3	9	1	81	1	9
7	5	2	3	4	9	6
12	2	7	0	49	0	0
5	7	0	5	0	25	0
6	8	1	6	1	36	6
Totals		24	18	160	80	36

$n = 6$,

therefore: $$R = \frac{n\sum x'y' - \sum x'\sum y'}{\sqrt{[n\sum(x')^2 - (\sum x')^2][n\sum(y')^2 - (\sum y')^2]}}$$

$$= \frac{6 \times 36 - 24 \times 18}{\sqrt{(6 \times 160 - 24^2)\,(6 \times 80 - 18^2)}}$$

$$= \frac{216 - 432}{\sqrt{(960 - 576)(480 - 324)}}$$

$$= \frac{-216}{\sqrt{384 \times 156}}$$

Thus the result is identical and the numbers involved in the calculation are smaller, taken overall.

C. RANK CORRELATION

Sometimes, instead of having actual measurements, we only have a record of the *order* in which items are placed. Examples of such a situation are:

(a) We may arrange a group of people in order of their heights, without actually measuring them. We could call the tallest No. 1, the next tallest No. 2, and so on.

(b) The results of an examination may show only the order of passing, without the actual marks; the highest-marked candidate being No. 1, the next highest being No. 2, and so on.

Data which is thus arranged in order of merit or magnitude is said to be *ranked*.

Relationship between Ranked Variates

Consider, as an example, the case of eight students who have taken the same two examinations, one in Mathematics and one in French. We have not been told the actual marks obtained in the examination, but we have been given the relative position (i.e. the *rank*) of each student in each subject:

Table 7.5: Table of ranked data

Student	**Relative Position**	
	French	***Mathematics***
A	8	6
B	5	5
C	3	4
D	6	7
E	7	8
F	2	1
G	1	3
H	4	2

We see from this table of ranks that student F was top in Mathematics but only second in French. Student G was top of the class in French, student E was bottom of the class (rank 8) in Mathematics, and so on.

A question which naturally arises is: "Is there any relationship between the students' performances in the two subjects?" This question can be put into statistical terms by asking: "Is there any correlation between the students' ranks in Mathematics and their ranks in French?" The answer to the question will fall into one of the following three categories:

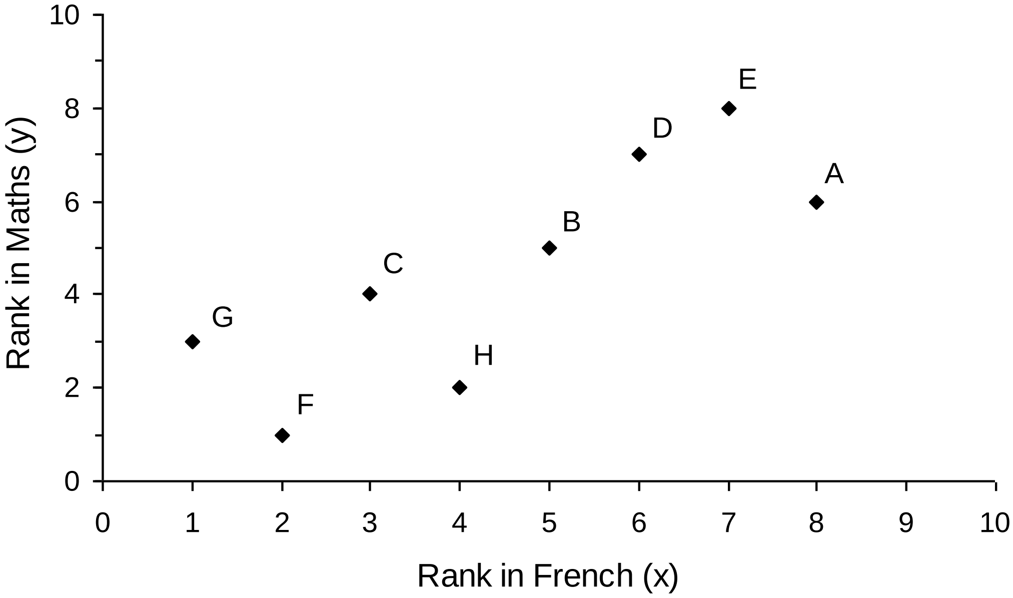
(a) *No correlation*: no connection between performance in the Mathematics examination and performance in the French examination.

(b) *Positive correlation*: students who do well in one of the subjects will, generally speaking, do well in the other.

(c) *Negative correlation*: students who do well in one of the subjects will, generally speaking, do poorly in the other.

We will start our analysis by drawing the scatter diagram as in Figure 7.8. It does not matter which subject we call x and which y.

Figure 7.8: Scatter diagram of students' results



The general impression given by the scatter diagram is that there is positive correlation. To find out how strong this correlation is, we calculate the correlation coefficient:

$$R = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \qquad n = 8$$

Table 7.6: Calculation of R

Student	Rank in French *(x)*	Rank in Maths *(y)*	x^2	y^2	xy
A	8	6	64	36	48
B	5	5	25	25	25
C	3	4	9	16	12
D	6	7	36	49	42
E	7	8	49	64	56
F	2	1	4	1	2
G	1	3	1	9	3
H	4	2	16	4	8
Total	36	36	204	204	196

$$R = \frac{8 \times 196 - (36)^2}{\sqrt{[8 \times 204 - (36)^2][8 \times 204 - (36)^2]}} = \frac{1{,}568 - 1{,}296}{1{,}632 - 1{,}296}$$

$$= \frac{272}{336} = 0.81$$

Ranked Correlation Coefficients

With ranked variates, there are simpler methods of calculating a correlation coefficient.

(a) Spearman's Rank Correlation Coefficient

The formula for Spearman's rank correlation coefficient is given by:

$$R = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

i.e.

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

In this formula, d is the difference between the two ranks for any one item, and n is the number of items involved. In the above example, n = 8. You can follow the calculation of R in the following table:

Table 7.7: Calculation of Spearman's coefficient

Student	**Rank in:**		**d**	**d^2**
	Maths	***French***		
A	6	8	–2	4
B	5	5	0	0
C	4	3	1	1
D	7	6	1	1
E	8	7	1	1
F	1	2	–1	1
G	3	1	2	4
H	2	4	–2	4
Total		*(Check)*	0	16

$$R = 1 - \frac{6 \times 16}{8^3 - 8} = 1 - \frac{96}{512 - 8} = 1 - \frac{96}{504} = 1 - \frac{12}{63}$$

$$= 1 - 0.19 = +0.81$$

When there is perfect agreement between the ranks of the two variates, then all the values of d will be 0 and so the rank correlation coefficient will be +1.0. When there is complete disagreement between the ranks, the values of d will be at their maximum and the rank correlation coefficient is –1.0.

(b) Kendall's Rank Correlation Coefficient

This is usually denoted by the Greek letter τ (tau, pronounced "tow", as in "now"). It does not give exactly the same answer as Spearman's method. Its formula is:

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}$$

where, as before, n is the number of pairs of observations. S is referred to as the score of the ranks.

To work out the score, we first arrange the students in order of their French ranks. We then consider for each student in turn whether the differences in French rankings between him or her and students lower down the list have the same signs as the differences in their Mathematics rankings. If the signs are the same, a pair of students is said to be *concordant*. If the signs are different, the pair is *discordant*. The score, S, is ($n_c - n_d$) where n_c is the total number of concordant pairs and n_d is the total number of discordant pairs. It is easiest to set out the calculation in a table:

Table 7.8: Calculation of Kendall's coefficient

Student	**Rank in:**		n_c	n_d	$n_c - n_d$
	French	***Mathematics***			
G	1	3	5	2	3
F	2	1	6	0	6
C	3	4	4	1	3
H	4	2	4	0	4
B	5	5	3	0	3
D	6	7	1	1	0
E	7	8	0	1	–1
A	8	6	0	0	0
Total					18

Compared with Student G, whose French rank is 1, all other French ranks have a higher numerical value. Student G's Maths rank is 3, however, so there are 5 Maths ranks with a higher numerical value and 2 with a lower numerical value. Thus $n_c = 5$ and $n_d = 2$. Similarly, for Student F, all French ranks below him or her in the table have higher numerical values and so do all the Maths ranks so $n_c = 6$ and $n_d = 0$. Similarly n_c and n_d are found for the other students. Each student should be compared only with those *lower down* the table, so that each pair of French and Maths rankings is considered once only.

$$\tau = \frac{18}{\frac{1}{2}\times 8\times 7} = \frac{36}{56} = 0.64 \text{ to two decimal places}$$

This value, being relatively large and positive, again shows a tendency for a high mark in French to be associated with a high mark in Maths, although the agreement is not perfect.

Tied Ranks

Sometimes it is not possible to distinguish between the ranks of two or more items. For example, two students may get the same mark in an examination and so they have the same rank. Or, two or more people in a group may be the same height. In such a case, we give all the equal ones an average rank and then carry on *as if we had given them different ranks.*

You will see what this means by studying the following examples:

(a) First two equal out of eight:

1½	**1½**	3	4	5	6	7	8
Average of 1 & 2							

(b) Three equal out of nine, but not at the ends of the list:

1	2	3	**5**	**5**	**5**	7	8	9
			Average of 4, 5 & 6					

(c) Last two equal out of eight:

1	2	3	4	5	6	**7½**	**7½**
						Average of 7 & 8	

(d) Last four equal out of eleven:

1	2	3	4	5	6	7	**9½**	**9½**	**9½**	**9½**
							Average of 8, 9, 10 & 11			

Strictly speaking, a rank correlation coefficient should not be used in these cases without making some adjustment for tied ranks. But the formula for the adjustments are a little complex and are outside the scope of this course. The best way for you to deal with tied ranks in practice is to calculate the ordinary (Pearson's) correlation coefficient. If, in an examination, you are specifically asked to calculate a rank correlation coefficient when there are tied ranks, then of course you must do so; but you might reasonably add a note to your answer to say that, because of the existence of tied ranks, the calculated coefficient is only an approximation, although probably a good one.

Final note: rank correlation coefficients may be used when the actual observations (and not just their rankings) *are* available. We first work out the rankings for each set of data and then calculate Spearman's or Kendall's coefficient as above. This procedure is appropriate when we require an approximate value for the correlation coefficient. Pearson's method using the *actual* observations is to be preferred in this case, however, so calculate a rank correlation coefficient only if an examination question specifically instructs you to do so.

Chapter 8

Linear Regression

Contents ***Page***

Introduction **114**

A. Regression Lines **115**

Nature of Regression Lines 115

Graphical Method 116

Mathematical Method 117

B. Use of Regression **119**

C. Connection between Correlation and Regression **119**

D. Multiple Regression **120**

INTRODUCTION

We've seen how the correlation coefficient measures the degree of relationship between two variates (variables). With perfect correlation ($R = +1.0$ or $R = -1.0$), the points of the scatter diagram all lie exactly on a straight line. It is sometimes the case that two variates are perfectly related in some way such that the points would lie exactly on a line, but not a *straight* line. In such a case R would not be 1.0. This is a most important point to bear in mind when you have calculated a correlation coefficient; the value may be small, but the reason may be that the correlation exists in some form other than a straight line.

The correlation coefficient tells us the extent to which the two variates are linearly related, but it does not tell us how to find the particular straight line which represents the relationship. The problem of determining which straight line best fits the points of a particular scatter diagram comes under the heading of *linear regression* analysis. Estimating the equation of the best-fitting line is of great practical importance. It enables us to test whether the independent variable (x) really does have an influence on the dependent variable (y), and we may be able to predict values of y from known or assumed values of x. In business and management research, it is important to gain an understanding of the factors that influence a firm's costs, revenues, profits and other key performance indicators, and to be able to predict changes in these variables from knowledge of possible changes in their determinants. Regression analysis therefore can be of great value in business planning and forecasting.

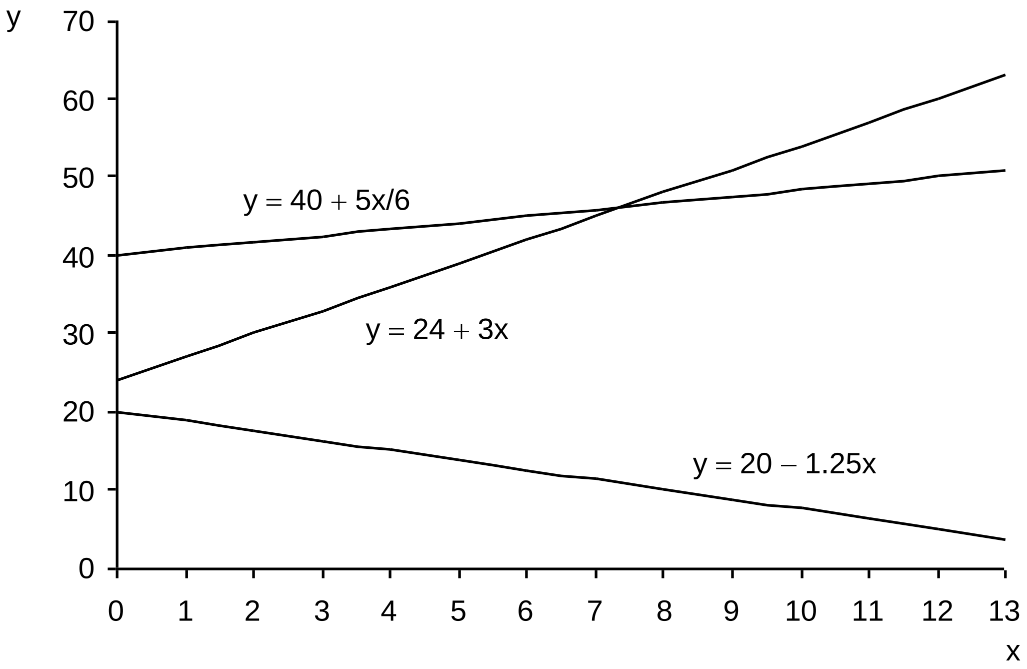
Remember that a straight-line graph can always be used to represent an equation of the form $y = a + bx$. In such an equation, y and x are the variables while a and b are the constants. Figure 8.1 shows a few examples of straight-line graphs for different values of a and b. Note the following important features of these linear graphs:

- The value of a is always the value of y corresponding to $x = 0$.
- The value of b represents the *gradient* or *slope* of the line. It tells us the number of units change in y per unit change in x. Larger values of a mean steeper slopes.
- Negative values of the gradient b mean that the line slopes *downwards* to the right; positive values of the gradient b mean that the line slopes *upwards* to the right.

So long as the equation linking the variables y and x is of the form $y = a + bx$, it is always possible to represent it graphically by a straight line. Likewise, if the graph of the relationship between y and x is a straight line, then it is always possible to express that relationship as an equation of the form $y = a + bx$.

If the graph relating y and x is *not* a straight line, then a more complicated equation would be needed. Conversely, if the equation is *not* of the form $y = a + bx$ (if, for example, it contains terms like x^2 or log x) then its graph would be a curve, not a straight line.

Figure 8.1: Straight line graphs



A. REGRESSION LINES

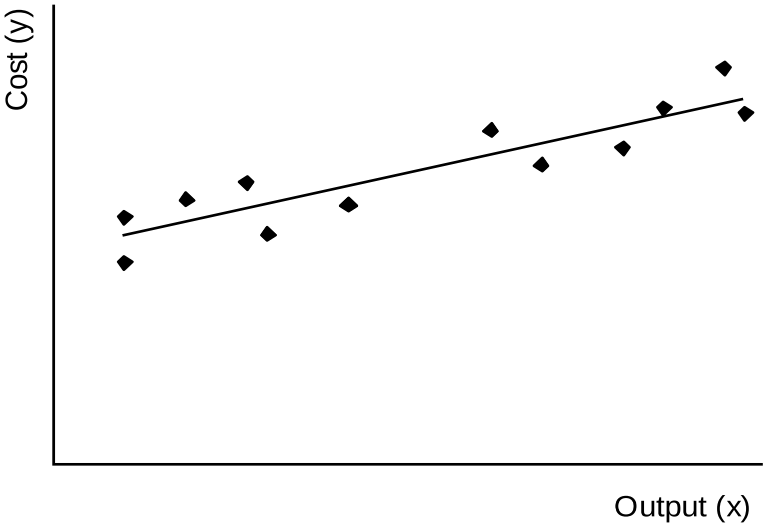
Nature of Regression Lines

When we have a scatter diagram whose points suggest a straight-line relationship (though not an exact one), and a correlation coefficient which supports the suggestion (say, R equal to more than about 0.4 or 0.5), we interpret this by saying that there is a linear relationship between the two variables but there are other factors (including errors of measurement and observation) which operate to give us a scatter of points around the line instead of exactly on it.

In order to determine the relationship between y and x, we need to know *what straight line* to draw through the collection of points on the scatter diagram. It will not go through all the points, but will lie somewhere in the midst of the collection of points and it will slope in the direction suggested by the points. Such a line is called a *regression line*.

In Figure 8.2, x is the monthly output of a factory and y is the total monthly costs of the factory; the scatter diagram is based on last year's records. The line which we draw through the points is obviously the one which we think best fits the situation, and statisticians often refer to regression lines as *lines of best fit*. Our problem is how to draw the best line.

Figure 8.2: Regression line



There are two methods available – a graphical method and a mathematical method.

Graphical Method

It can be proved mathematically (but you don't need to know how!) that the regression line *must* pass through the point representing the arithmetic means of the two variables. The graphical method makes use of this fact, and the procedure is as follows:

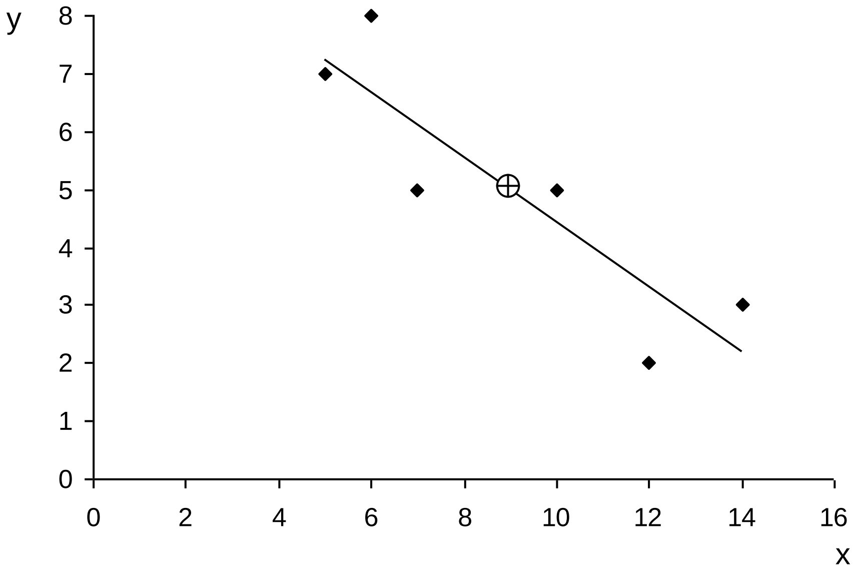
(a) Calculate the means, $\bar{x}$ and $\bar{y}$, of the two variables.

(b) Plot the point corresponding to this pair of values on the scatter diagram.

(c) Using a ruler, draw a straight line through the point you have just plotted and lying, as evenly as you can judge, among the other points on the diagram.

In Figure 8.3 this procedure was followed using the data from the section on the correlation coefficient in the previous chapter. If someone else (you, for example) were to do it, you might well get a line of a slightly different slope, but it would still go through the point of the means (marked ⊕).

Figure 8.3: Graphical method of finding the regression line



Quite obviously, this method is not exact (no graphical methods are) but it is often sufficient for practical purposes. The stronger the correlation, the more reliable this method is, and with perfect correlation there will be little or no error involved.

Mathematical Method

A more exact method of determining the regression line is to find mathematically the values of the constants a and b in the equation $y = a + bx$, and this can be done very easily. This method is called the *least squares* method, as the line we obtain is that which *minimises the sum of the squares of the vertical deviations of the points from the line.* The equation of the least squares line is:

$$\hat{y} = a + bx$$

where the symbol $\hat{y}$ is used to indicate the *predicted* rather than the actual value of y.

The values of the gradient b and the intercept a can be calculated from the following formulas:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$a = \bar{y} - b\bar{x} \quad \text{or} \quad \frac{\sum y - b\sum x}{n}$$

$n = \text{number of pairs of readings.}$

We will now apply these formulae to the example we used when talking about the correlation coefficient. The data is reproduced below:

x	10	14	7	12	5	6
y	5	3	5	2	7	8

If you look back at Chapter 7 you will see that we had the following figures:

$$\Sigma x = 54 \quad \Sigma y = 30 \quad \Sigma x^2 = 550 \quad \Sigma xy = 234 \quad n = 6$$

Therefore $\bar{x} = 9$, and $\bar{y} = 5$

Applying the formulae, we get:

$$b = \frac{6 \times 234 - 54 \times 30}{6 \times 550 - (54)^2} = \frac{-216}{384} = -0.5625$$

$$a = 5 = -(-0.5625)9 = 5 + 5.0625 = 10.0625$$

b and a are termed the *regression coefficients* (and b also represents the gradient, as previously stated).

The equation for the regression line in this case is therefore:

$$\hat{y} = 10.0625 - 0.5625x$$

To draw this line on the scatter diagram, choose two values of x, one towards the left of the diagram and one towards the right. Calculate the y-value for each of these values of x, plot the two points and join them up with a straight line. If you have done the calculations correctly, the line will pass through the ($\bar{x}$, $\bar{y}$) point.

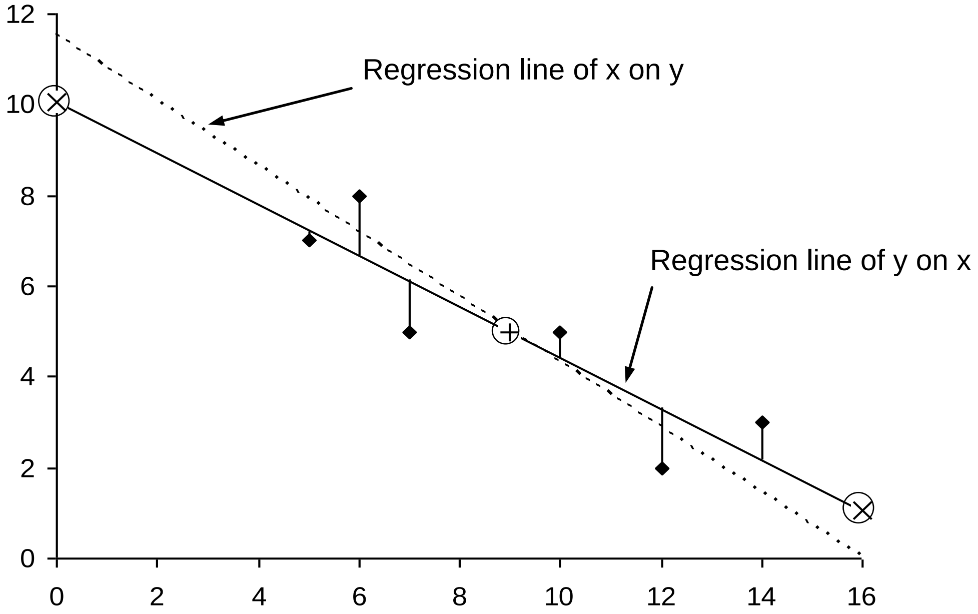
For drawing the regression line, we will choose values of x which are convenient, e.g. x = 0 and x = 16. The corresponding values of y are:

$$\text{for } x = 0, y = 10.0625 - 0 = 10.0625$$

$$\text{for } x = 16, y = 10.0625 - 16\,(0.5625) = 10.0625 - 9.0 = 1.0625$$

These two points are marked in the scatter diagram (⊗) in Figure 8.4, together with the individual points ♦, the regression line (drawn as an unbroken line) and the mean point ⊕.

Figure 8.4: Regression of y on x and x on y



The regression line which we have drawn, and the equation which we have determined, represent the *regression of y on x*. We could, by interchanging x and y, have obtained the *regression of x on y*. This would produce a different line and a different equation. This latter line is shown in Figure 8.4 by a broken line. The question naturally arises: "Which regression line should be used?" The statistician arrives at the answer by some fairly complicated reasoning but, for our purposes, the answer may be summed up as follows:

- Always use the regression of y on x, denoting the variable that is being influenced as y (the *dependent* variable) and the variable that is doing the influencing as x (the *independent* variable). Then use the method described in detail above, putting y on the *vertical* axis and x on the *horizontal* axis.

- If you intend to use the regression line to predict one variable from another, then the variable you want to predict is always treated as y; the other variable is x. For example, if you wish to use the regression line (or its equation) to predict costs from specified outputs, then the outputs will be the x and the costs will be the y. Then y will always be the dependent variable and x the independent variable.

B. USE OF REGRESSION

The main use of a regression line is to calculate values of the dependent variable not observed in the set of data. Take as our example that of employees' heights with a regression equation of:

$$\hat{y} = -345.33 + 2.87x$$

where x is height.

Of the 12 people measured and weighed there was nobody of height 181 cm. Therefore, if we wanted to know the weight of somebody of this height, it would be impossible to read it from the data available. However, by assuming that a linear relationship exists between weight and height it is possible, by using the regression equation, to calculate an estimate of the weight:

$$x = 181$$

$$\hat{y} = 2.87 \times 181 - 345.33 = 174.14 \text{ lb}$$

Therefore the estimated weight of somebody of height 181 cm is 174.14 lb.

Since the value of x (181 cm) lies *within the observed range* of x from the 12 people, we say that we have estimated the value of y by *interpolation*.

However, if we wish to use a regression equation to forecast a result from values which are *outside the range of observations* from which the line is calculated, we have to consider carefully the validity of the estimate obtained. This use of the regression line is called *extrapolation*, and we have to assume that the same linear relationship will exist for observations beyond those from which it has been formulated. For example, say we want to estimate the weight of somebody whose height is 194 cm. This value is outside the range of the 12 people measured but a predicted value of y can still be calculated as:

$$x = 194$$

$$\hat{y} = 2.87 \times 194 - 345.33 = 211.45 \text{ lb}$$

This result seems reasonable, but common sense suggests that values of x much smaller than 160 cm or much larger than 186 cm would be rather improbable.

Sometimes this assumption of the same linear relationship is incorrect, as the factors that influenced the two variables may not remain constant outside the range from which the regression equation is formed, or some extra factor may be introduced.

Consider the relationship between time and the average working wage. If a regression line calculated from data that is collected during years where inflation is very low is used to estimate the wage for years of high inflation, the predicted figure will be much lower than the actual figure, i.e. the change in inflation will change the relationship between the variables. This emphasises that extrapolation gives reliable results only for values *close to the ends of the observed range*.

C. CONNECTION BETWEEN CORRELATION AND REGRESSION

The degree of correlation between two variables is a good guide to the likely accuracy of the estimates made from the regression equation. If the correlation is high then the estimates are likely to be reasonably accurate, and if the correlation is low then the estimates will be poor as the unexplained variation is then high. However, remember that correlation in itself does not necessarily imply *causation*.

You must remember that both the regression equations and the correlation coefficient are calculated from the same data, so both of them must be used with caution when estimates are predicted for values outside the range of the observations, i.e. when values are predicted by extrapolation or the correlation coefficient is assumed to remain constant under these conditions. Also remember that the values calculated for both correlation and regression are influenced by the number of pairs of observations used. So results obtained from a large sample are more reliable than those from a small sample.

D. MULTIPLE REGRESSION

In multiple regression there is one dependent variable, y, but two or more independent variables. You will not be required to calculate a multiple regression equation, but you may be required to write an interpretation of a set of multiple regression results. To show you how to do this, consider a multiple regression model in which there are two independent variables.

Suppose that a business researcher believes that a firm's unit costs will depend on the size of the firm and the wage rate paid to its employees. In this example, y will represent unit costs (the dependent variable) and we will use x to represent firm size and z to represent the wage rate – these are the two independent variables. Thus, the multiple regression equation for this model would be:

$$\hat{y} = a + bx + cz$$

A set of results for this regression of y on x and z may look like this:

$$\hat{y} = 150 - 0.025x + 16.5z$$

$$(2.5) \quad (-4.2) \qquad (1.2) \qquad \textit{(t-values in brackets)}$$

$$R^2 = 0.85$$

The calculated values of a, b and c are called the coefficient estimates. The coefficient estimates indicate how responsive the dependent variable (i.e. unit costs) is to changes in each of the independent variables. In the example, the coefficient of x (firm size) is found to be negative (equal to –0.025), suggesting that a one-unit rise in firm size, other things being equal, will lead to a 0.025 decrease in unit costs. The coefficient of z (wage rate) is positive (equal to 16.5), suggesting that a one-unit increase in the wage rate will tend to increase unit costs by 16.5 units. The calculated value of a (equal to 150) represents an estimate of the firm's fixed costs.

The results also include a set of *t-values* associated with each of the coefficient estimates. We shall be examining t-values in Chapter 13. For now, all you need to know is that the t-values enable us to test whether the coefficient estimates are significantly different from zero. With a sample size greater than 50, a t-value greater than 2 (or less than –2) is required for statistical significance. In the example, the constant term, a, and the coefficient of x (firm size) are significantly different from zero, while the coefficient of z (wage rate) is not significantly different from zero. Thus, we can say that firm size has a significant influence on unit costs, but the wage rate does not.

Finally the results give us an R^2 value of 0.85. As we mentioned in Chapter 7, this is the *coefficient of determination*, and implies that 85 per cent of the variation in costs is being explained by variations in firm size and the wage rate. A value of 0.85 is quite close to 1, suggesting that our multiple regression equation provides quite a good fit to the scatter of points in the sample.

Chapter 9

Time Series Analysis

Contents *Page*

Introduction **122**

A. Structure of a Time Series **122**
Trend 123
Seasonal Variations 124
Cyclical Fluctuations 125
Irregular or Random Fluctuations 125
Summary 125

B. Calculation of Component Factors for the Additive Model **126**
Trend 126
Seasonal Variation 133
Deseasonalised Data and Residual 135

C. Multiplicative Model **135**
Example of a Multiplicative Model 136

D. Forecasting **139**
Assumptions 139
Methods of Forecasting 139

E. The Z Chart **141**

Summary **142**

INTRODUCTION

Businesses and governments use statistical analysis of information collected at regular intervals over extensive periods of time to plan future policies. For example, sales values or unemployment levels recorded at yearly, quarterly or monthly intervals are examined in an attempt to predict their future behaviour. Such sets of values observed at regular intervals over a period of time are called time series.

The analysis of this data is a complex problem as many variable factors may influence the changes. The first step is to plot the observations on a scattergraph, which differs from those scattergraphs we have considered previously as the points are evenly spaced on the time axis in the order in which they are observed, and the time variable is always the independent variable. This scattergraph gives us a good visual guide to the actual changes, but is of very little help in showing the component factors causing these changes or in predicting future movements of the dependent variable.

Statisticians have constructed a number of mathematical models to describe the behaviour of time series, and several of these will be discussed in this chapter.

A. STRUCTURE OF A TIME SERIES

These mathematical models assume that the changes are caused by the variation of four main factors; they differ in the relationship between these factors. It will be easier to understand the theory in detail if we relate it to a simple time series so that we can see the calculations necessary at each stage.

Consider a factory employing a number of people in producing a particular commodity, say thermometers. Naturally, at such a factory during the course of a year some employees will be absent for various reasons. The following table shows the number of days lost through sickness over a five-year period. Each year has been broken down into four quarters of three months. We have assumed that the number of employees at the factory remained constant over the five years.

Table 9.1: Days lost through sickness at a thermometer factory

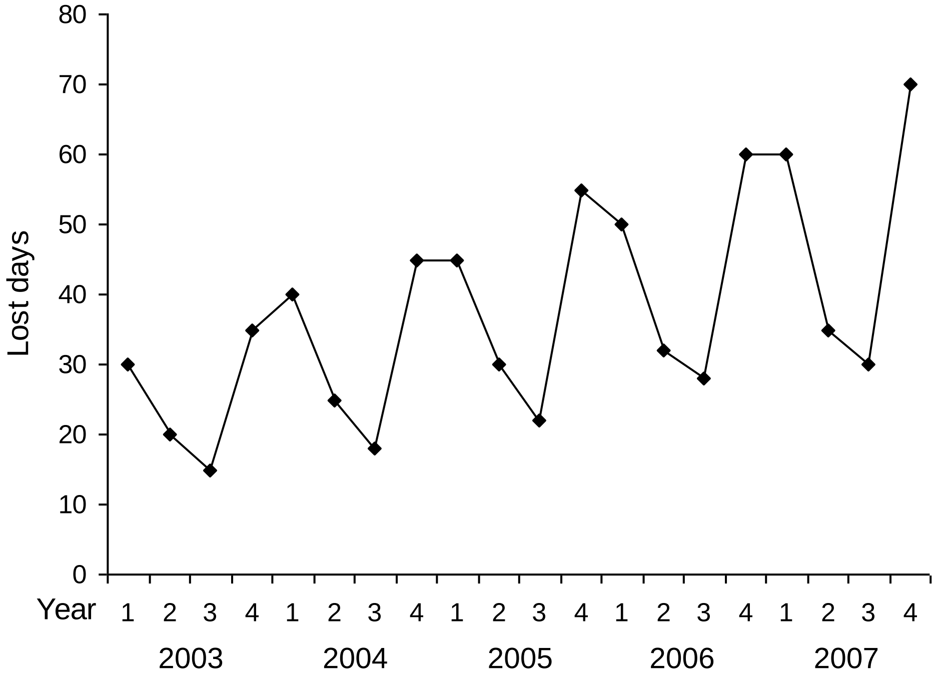
Year	Quarter	Days Lost	Year	Quarter	Days Lost
2003	1	30	2004	1	40
	2	20		2	25
	3	15		3	18
	4	35		4	45
2005	1	45	2006	1	50
	2	30		2	32
	3	22		3	28
	4	55		4	60
2007	1	60			
	2	35			
	3	30			
	4	70			

We will begin by plotting a time-series graph for the data, as shown in Figure 9.1.

Note the following characteristics of a time-series graph:

- It is usual to join the points by straight lines. The only function of these lines is to help your eyes to see the pattern formed by the points.
- Intermediate values of the variables cannot be read from the graph.
- Every time-series graph will look similar to this, but a careful study of the change of pattern over time will suggest which model should be used for analysis.

Figure 9.1: Time series of days lost through sickness



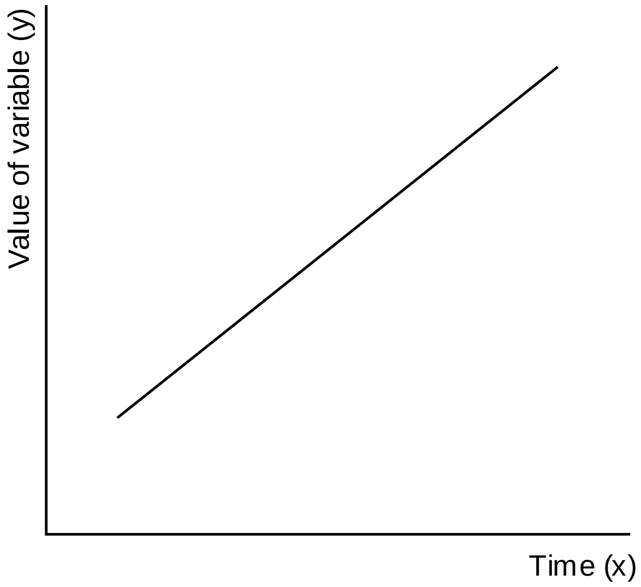
There are four factors that influence the changes in a time series – trend, seasonal variations, cyclical fluctuations, and irregular or random fluctuations. Now we will consider each in turn.

Trend

This is the change in general level over the whole time period and is often referred to as the secular trend. You can see in Figure 9.1 that the trend is definitely upwards, in spite of the obvious fluctuations from one quarter to the next.

A trend can thus be defined as a clear tendency for the time series data to travel in a particular direction in spite of other large and small fluctuations. An example of a linear trend is shown in Figure 9.2. There are numerous instances of a trend, for example the amount of money collected from UK taxpayers is always increasing; therefore any time series describing income from tax would show an upward trend.

Figure 9.2: Example of trend

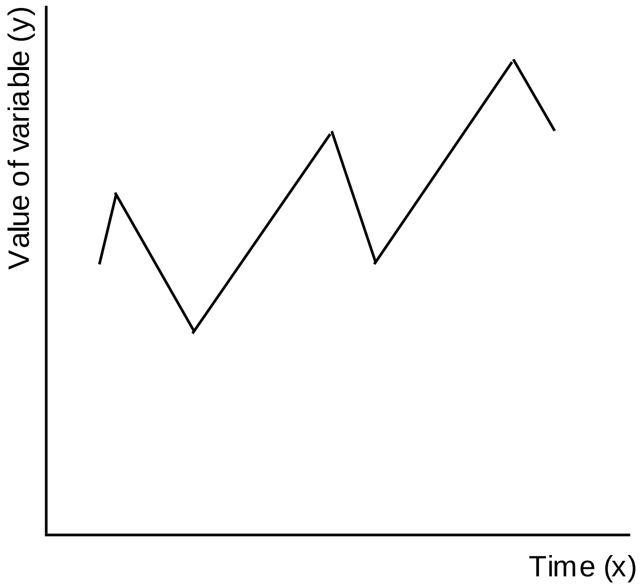


Seasonal Variations

These are variations which are repeated over relatively short periods of time. Those most frequently observed are associated with the seasons of the year, e.g. ice cream sales tend to rise during the summer months and fall during the winter months. You can see in our example of employees' sickness that more people are sick during the winter than in the summer.

If you can establish the variation throughout the year then this seasonal variation is likely to be similar from one year to the next, so that it would be possible to allow for it when estimating values of the variable in other parts of the time series. The usefulness of being able to calculate seasonal variation is obvious as, for example, it allows ice cream manufacturers to alter their production schedules to meet these seasonal changes. Figure 9.3 shows a typical seasonal variation that could apply to the examples above.

Figure 9.3: Seasonal variations

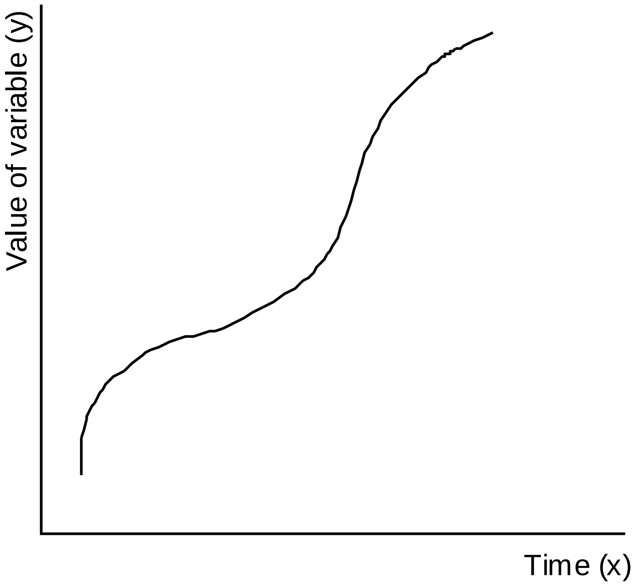


Cyclical Fluctuations

These are long-term but fairly regular variations. They are difficult to observe unless you have access to data over an extensive period of time during which external conditions have remained relatively constant. For example, it is well known in the textile trade that there is a cycle of about three years, during which time demand varies from high to low. This is similar to the phenomena known as the trade cycle which many economists say exists in the trading pattern of most countries, but for which there is no generally accepted explanation.

Figure 9.4 shows how such a cyclical fluctuation would relate to an upward trend. In our example on sickness, a cyclical fluctuation could be caused by, say, a two-year cycle for people suffering from influenza.

Figure 9.4: Cyclical fluctuation



As this type of fluctuation is difficult to determine, it is often considered with the final (fourth) element, and the two together are called the *residual variation*.

Irregular or Random Fluctuations

Careful examination of Figure 9.1 shows that there are other relatively small irregularities which we have not accounted for and that do not seem to have any easily seen pattern. We call these irregular or random fluctuations; they may be due to errors of observation or to some one-off external influence which is difficult to isolate or predict. In our example (Table 1) there may have been a measles epidemic in 2006, but it would be extremely difficult to predict when and if such an epidemic would occur again.

Summary

To sum up, a time series (Y) can be considered as a combination of the following four factors:

- trend (T)
- seasonal variation (S)
- cyclical fluctuation (C)
- irregular fluctuations (I).

It is possible for the relationship between these factors and the time series to be expressed in a number of ways through the use of different mathematical models. We are now going to look in detail at the *additive model* before looking briefly at the multiplicative model.

B. CALCULATION OF COMPONENT FACTORS FOR THE ADDITIVE MODEL

The additive model can be expressed by the equation:

time series = trend + seasonal variation + cyclical fluctuations + random fluctuations

i.e. the value of the time series Y is:

$$Y = T + S + C + I$$

Usually the cyclical and random fluctuations are put together and called the "residual" (R), so:

$$Y = T + S + R$$

Trend

The *most important factor* of a time series is the trend. Before deciding on the method to be used for finding it, we must first decide whether the conditions that have influenced the series have remained stable over time. For example, if you have to consider the production of some commodity and want to establish the trend, you should first decide if there has been any significant change in conditions affecting the level of production, such as a sudden and considerable growth in the national economy. If there has, you must consider breaking the time series into sections over which the conditions have remained stable.

Having decided the time period you will analyse, you can use any one of the following methods to find the trend. The basic idea behind most of these methods is to average out the three other factors of variation so that you are left with the long-term trend.

(a) Graphical Method

Once you have plotted the graph of the time series, it is possible to draw in by eye a line through the points to represent the trend. The result is likely to vary considerably from person to person (unless the plotted points lie very near to a straight line), so it is not a satisfactory method.

(b) Semi-Averages Method

This is a simple method which involves very little arithmetic. The time period is divided into equal parts, and the arithmetic means of the values of the dependent variable in each half are calculated. These means are then plotted at the quarter and three-quarters position of the time series. The line adjoining these two points represents the trend of the series. Note that this line will pass through the overall mean of the values of the dependent variable.

In our example which consists of five years of data, the midpoint of the whole series is mid-way between quarter 2 and quarter 3 of 2005.

For the mean of the first half:

Year	Quarter	Days Lost
2003	1	30
	2	20
	3	15
	4	35
2004	1	40
	2	25
	3	18
	4	45
2005	1	45
	2	30
	Total	303

mean = 30.3

For the mean of the second half:

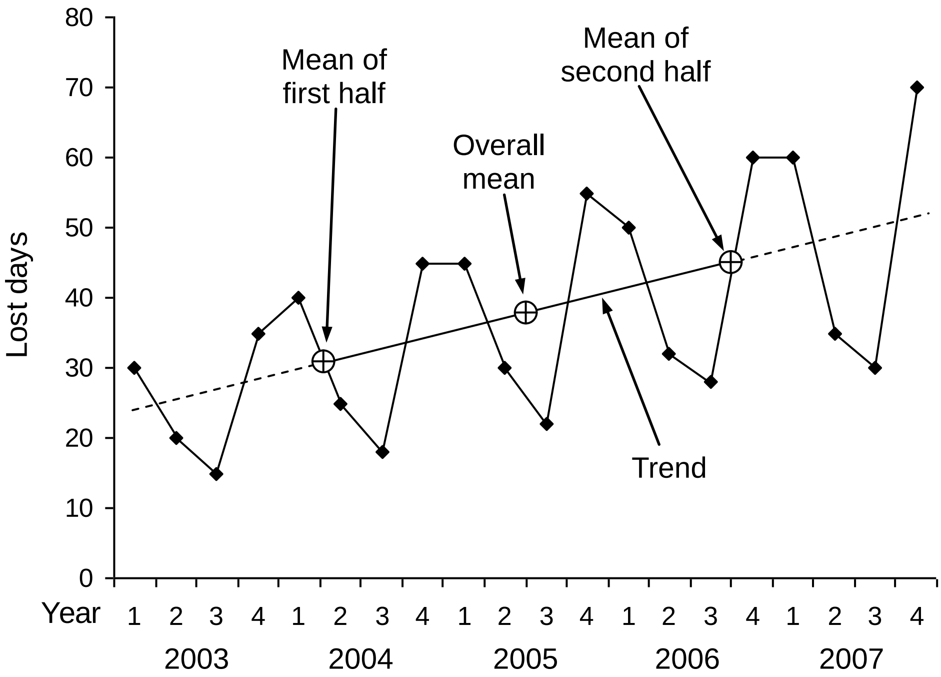
Year	Quarter	Days Lost
2005	3	22
	4	55
2006	1	50
	2	32
	3	28
	4	60
2007	1	60
	2	35
	3	30
	4	70
	Total	442

mean = 44.2

These values are plotted on the time-series graph in Figure 9.5. You will notice that 30.3 days, as it is the mean for the first half, is plotted halfway between quarters 1 and 2 of 2004, and likewise 44.2 days is plotted halfway between quarters 3 and 4 of 2006. The trend line is then drawn between these two points and it can be extrapolated beyond these points as shown by the dotted line.

If there is an odd number of observations in the time series, the middle observation is ignored and the means of the observations on each side of it are calculated.

Figure 9.5: Time series showing the semi-averages method



(c) *Least Squares Method*

The trend line is calculated using the formula for the mathematical method of calculating regression lines in Chapter 8. In fact the trend line is the regression line of y on x where y is the dependent variable and x is the time variable. Since in a time series the observations are always recorded at equally-spaced time intervals, we can represent x by the first n positive integers, where n is the number of observations. (We never calculate the other regression line in time series analysis as it has no significance.) Thus the equation of the trend is:

(1) $y = a + bx$

where

(2) $b = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$

and

(3) $a = \dfrac{\Sigma y - b\Sigma x}{n}$

Using the data given in our earlier example, we set up a table of calculations as follows:

Year	Quarter	x	Days Lost y	x^2	xy
2003	1	1	30	1	30
	2	2	20	4	40
	3	3	15	9	45
	4	4	35	16	140
2004	1	5	40	25	200
	2	6	25	36	150
	3	7	18	49	126
	4	8	45	64	360
2005	1	9	45	81	405
	2	10	30	100	300
	3	11	22	121	242
	4	12	55	144	660
2006	1	13	50	169	650
	2	14	32	196	448
	3	15	28	225	420
	4	16	60	256	960
2007	1	17	60	289	1,020
	2	18	35	324	630
	3	19	30	361	570
	4	20	70	400	1,400
Total		210	745	2,870	8,796

$n = 20$ so therefore:

$$b = \frac{20(8{,}796) - 210(745)}{20(2{,}870) - (210)^2} = \frac{175{,}920 - 156{,}450}{57{,}400 - 44{,}100} = \frac{19{,}470}{13{,}300} = 1.46$$

and

$$a = \frac{745 - 1.46(210)}{20} = \frac{438.4}{20} = 21.92$$

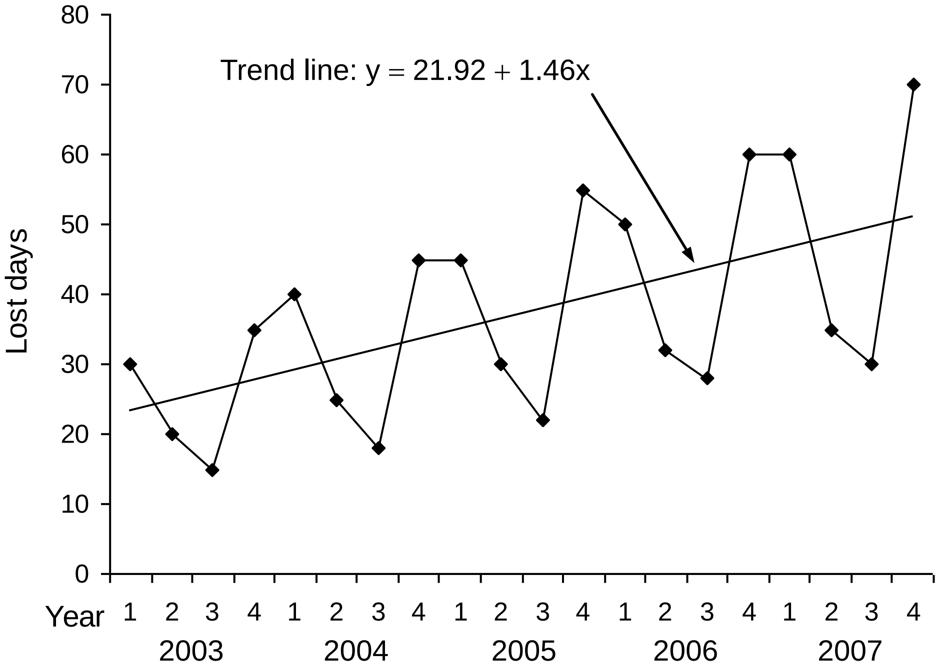
So the equation of the trend line is:

$$y = 21.92 + 1.46x$$

where y is the number of days lost owing to sickness and x is the number given to the quarter required.

We can now draw the line represented by this equation on the time-series graph as shown in Figure 9.6. This method uses all the available information, but it suffers from the same limitations as other regression lines if it is used for prediction by extrapolation.

Figure 9.6: Least squares trend line



(d) Moving Averages Method

So far, the methods we have discussed for finding trends have resulted in a straight line, but the actual trend may be a curve or a series of straight segments. The method of moving averages gives a way of calculating and plotting on the time-series graph a trend point corresponding to each observed point. These points are calculated by averaging a number of consecutive values of the dependent variable, so that variations in individual observations are reduced. The number of consecutive values selected will depend on the length of the short-term or seasonal variation shown on the graph.

The method of calculating a set of moving averages is illustrated by the following simple example. Consider the sequence of seven numbers 6, 4, 5, 1, 9, 5, 6. Now take the number of time periods covered by the fluctuations to be four, as in quarterly figures, so a moving average of order four is needed.

Step 1: Find the average of the first to the fourth numbers.

$$\text{Average } = \frac{6+4+5+1}{4} = 4$$

Step 2: Find the average of the second to the fifth numbers.

$$\text{Average } = \frac{4+5+1+9}{4} = 4.75$$

Step 3: Find the average of the third to the sixth numbers.

$$\text{Average } = \frac{5+1+9+5}{4} = 5$$

Step 4: Find the average of the fourth to seventh numbers.

$$\text{Average } = \frac{1+9+5+6}{4} = 5.25$$

Hence the moving averages of order 4 are 4, 4.75, 5, and 5.25. For monthly data a moving average of order 12 would be needed; for daily data the order would be 7, and so on.

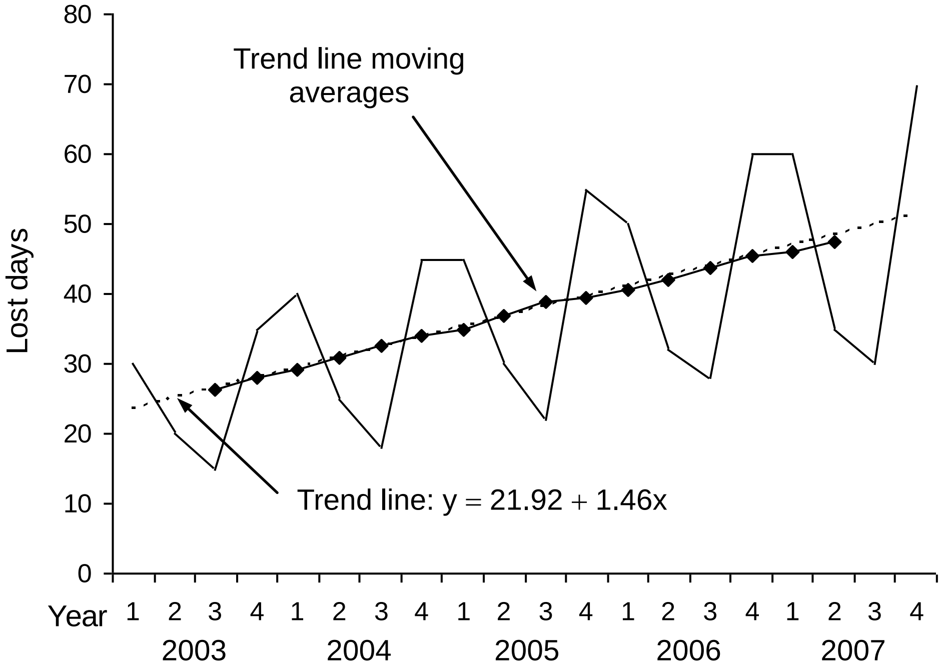
Using the data of the earlier example, we can calculate the trend values and plot them on to Figure 9.6 (as Figure 9.7) so that we can compare the two trend lines. The table of calculations follows:

Year	Quarter	Days Lost	4-Quarter Total	Moving Average	Trend
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
2003	1	30			
	2	20			
			100	25	
	3	15			26.3
			110	27.5	
	4	35			28.1
			115	28.75	
2004	1	40			29.1
			118	29.5	
	2	25			30.8
			128	32.0	
	3	18			32.6
			133	33.25	
	4	45			33.9
			138	34.5	
2005	1	45			35.0
			142	35.5	
	2	30			36.8
			152	38.0	
	3	22			38.6
			157	39.25	
	4	55			39.5
			159	39.75	
2006	1	50			40.5
			165	41.25	
	2	32			41.9
			170	42.5	
	3	28			43.8
			180	45.0	
	4	60			45.4
			183	45.75	
2007	1	60			46.0
			185	46.25	
	2	35			47.5
			195	48.75	
	3	30			
	4	70			

The trend is given correct to one decimal place as this is the greatest accuracy justified by the accuracy of the data. Notice how the table of calculations is set out, with the numbers in columns (4) and (5) placed *midway between* two quarterly readings. This is because we were averaging over an even number of values, so the moving average would have to be plotted in this position on the graph and would not correspond to any particular quarter. Thus it is necessary to add column (6) which gives the mean of successive pairs of moving averages and these numbers are the trend values plotted. (The values in column (6) are often called the *centred moving averages*.)

If we were calculating a moving average with an odd number of values it would not be necessary to carry out this final stage. This is because the moving averages would be centred on an actual observation, and so would be the trend values, e.g. daily observation over a number of weeks or data with a short-term cycle of an odd number of years.

Figure 9.7: Moving average and least squares trend lines



The main advantage of this method is that the trend values take into account the *immediate* changes in external factors which the trend lines, using the previous two methods, are unable to do. However, this method has three disadvantages:

- The trend line cannot be found for the whole of the time series. As you can see from our example, there are no trend values for quarters at the beginning and end of the series.
- Problems can be encountered in deciding the order number, i.e. the period of fluctuation. Unless the seasonal or cyclical movement is definite and clear cut, the moving average method of deriving the trend may yield a rather unsatisfactory line.
- Since the trend is calculated as a simple arithmetic mean it can be unduly influenced by a few extreme values.

Seasonal Variation

As we are assuming at present that the additive model is satisfactory, once we have found the trend by one of the methods described in the previous section we can find the value of the remaining factors for each value of the dependent variable from the equation for the additive model by subtraction, that is:

$$Y = T + S + C + I$$

and so

$$Y - T = S + C + I = S + R$$

($C + I = R$ (the residual) since we cannot usually separate C and I)

Column (5) of the following table shows the value of this difference for all the quarters from 2003 quarter 3 to 2007 quarter 2.

Year	**Quarter**	**Days Lost Y**	**Trend T**	**Y – T**
(1)	(2)	(3)	(4)	(5)
2003	3	15	26.3	-11.3
	4	35	28.1	6.9
2004	1	40	29.1	10.9
	2	25	30.8	-5.8
	3	18	32.6	-14.6
	4	45	33.9	11.1
2005	1	45	35.0	10.0
	2	30	36.8	-6.8
	3	22	38.6	-16.6
	4	55	39.5	15.5
2006	1	50	40.5	9.5
	2	32	41.9	-9.9
	3	28	43.8	-15.8
	4	60	45.4	14.6
2007	1	60	46.0	14.0
	2	35	47.5	-12.5

One of the assumptions we make for the additive model is that the seasonal variations are the same for corresponding quarters in each year. You can see that this is not the case in column (5), except that for each year the first and fourth quarters give a positive result and the second and third a negative one. The variation must be caused by the residual (R), and this factor can be eliminated by calculating the adjusted average for each quarter as shown in the next table:

Year	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr	
2003			–11.3	6.9	
2004	10.9	–5.8	–14.6	11.1	
2005	10.0	–6.8	–16.6	15.5	
2006	9.5	–9.9	–15.8	14.6	
2007	14.0	–12.5			
Total	44.4	–35.0	–58.3	48.1	
Average	11.1	–8.8	–14.6	12.0	(–0.3)
Adjusted Average	11.175	–8.725	–14.525	12.075	

The average fluctuations (11.1, -8.8, -14.6, and 12.0) should add up to zero, but as you can see in the example above, because of rounding errors they do not; therefore a minor adjustment is carried out in the last row. This is done by subtracting a quarter of the total outstanding from each average (in this case ¼ of –0.3, i.e. –0.075).

Therefore the values 11.2, –8.7, –14.5 and 12.1 (all correct to 1 decimal place) are the seasonal fluctuations of the four quarters for the time series of days lost through sickness at the factory.

Deseasonalised Data and Residual

The remaining results that are needed for this analysis are the deseasonalised values (Y – S) and the residuals (Y – S – T). These are shown in columns (4) and (6) of the following table:

Year and Qtr	Days Lost	Seasonal Adjustment	Deseasonalised Data	Trend	Residual
	Y	**S**	**Y – S**	**T**	**R = Y – S – T**
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
2003 3	15	–14.5	29.5	26.3	3.2
4	35	12.1	22.9	28.1	–5.2
2004 1	40	11.2	28.8	29.1	–0.3
2	25	–8.7	33.7	30.7	3.0
3	18	–14.5	32.5	32.6	–0.1
4	45	12.1	32.9	33.9	–1.0
2005 1	45	11.2	33.8	35.0	–1.2
2	30	–8.7	38.7	36.7	2.0
3	22	–14.5	36.5	38.6	–2.1
4	55	12.1	42.9	39.5	3.4
2006 1	50	11.2	38.8	40.5	–1.7
2	32	–8.7	40.7	41.9	–1.2
3	28	–14.5	42.5	43.7	–1.2
4	60	12.1	47.9	45.4	2.5
2007 1	60	11.2	48.8	46.0	2.8
2	35	–8.7	43.7	47.5	–3.8

As you can see, there is no pattern to the residuals but they are fairly small, that is they can be considered as random errors of observation and rounding, though they may contain a systematic cyclic element.

C. MULTIPLICATIVE MODEL

We have seen that the *additive model* is defined as:

$$Y = T + S + C + I$$

By contrast, the *multiplicative model* is defined as:

$$Y = T \times S \times C \times I$$

The difference between these models rests in the *assumptions* made for each. In the case of the additive model, it is assumed that the magnitude of all the factors other than the trend is not affected by the trend, but with the multiplicative model it is assumed that their magnitude is directly proportional to the trend. This means that the effect of a given factor is to change the overall total of the time series by a constant multiple as opposed to a fixed amount. For example, the multiplicative model would assume that sales of umbrellas are 30 per cent above the yearly average in winter, whereas the additive model would assume that sales were 2,000 above the yearly average in winter.

The assumptions made for the *additive model* will be satisfactory as long as the *trend is linear* or alters only slightly during the period of analysis. Under other conditions the multiplicative model is likely to give more reliable results. For example, a company with a turnover of £200,000 per annum is likely to experience considerably greater seasonal fluctuations than one with a turnover of only £10,000 per annum, so that the multiplicative model would be more applicable.

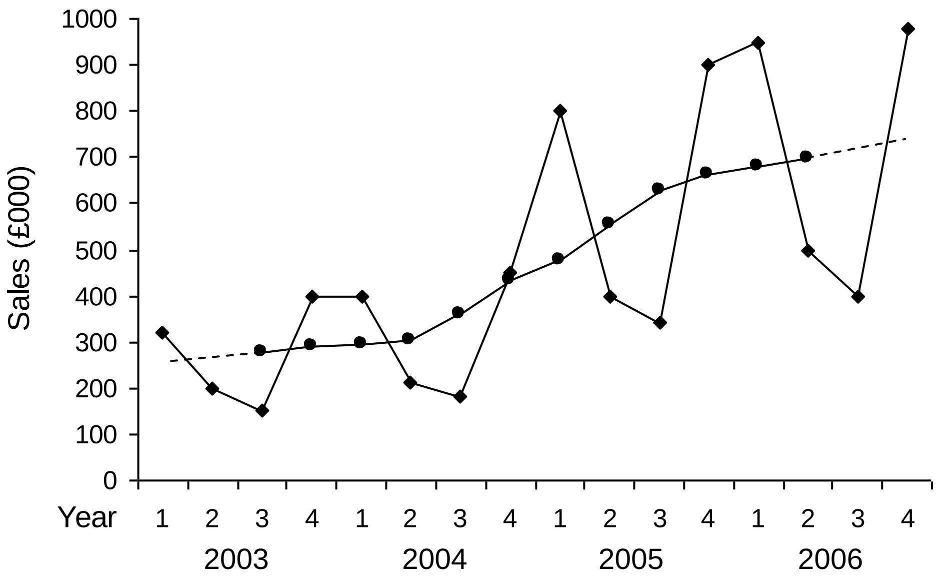
Example of a Multiplicative Model

We will look at the sales of Premium Bonds at a large post office over five years. The data and calculations are shown in the following table and plotted on the graph shown in Figure 9.8. You can see that in this data there is a high growth rate, particularly over the years 2004 and 2005, and it is in this type of time series that the multiplicative model is best used.

Table 9.2: Sales of Premium Bonds

Year and Quarter		**Sales** *(£000)*	**Moving Total**	**Moving Average**	**Trend**	**Sales/ Trend**
2003	1	320				
	2	200				
			1,070	268		
	3	150			278	0.54
			1,150	288		
	4	400			289	1.38
			1,160	290		
2004	1	400			294	1.36
			1,190	298		
	2	210			304	0.69
			1,240	310		
	3	180			360	0.50
			1,640	410		
	4	450			434	1.04
			1,830	458		
2005	1	800			478	1.67
			1,990	498		
	2	400			554	0.72
			2,440	610		
	3	340			629	0.54
			2,590	648		
	4	900			661	1.36
			2,690	673		
2006	1	950			681	1.40
			2,750	688		
	2	500			698	0.72
			2,830	708		
	3	400				
	4	980				

Figure 9.8: Premium Bond sales and trend line



The calculation of the trend can be carried out by any of the methods described previously, but the moving average method is usually preferred. The trend is removed from the series by dividing the sales value by the trend:

Since

$$Y = T \times S \times C \times I$$

then

$$\frac{Y}{T} = S \times C \times I$$

Now we need to remove the cyclical and irregular fluctuations from the time series in order to calculate the seasonal fluctuations. This is achieved in a similar way to the additive model, by calculating the average of the sales/trend ratios and adjusting where necessary for rounding errors.

Year	Qtr 1	Qtr 2	Qtr 3	Qtr 4	
2003			0.54	1.38	
2004	1.36	0.69	0.50	1.04	
2005	1.67	0.72	0.54	1.36	
2006	1.40	0.72			
Total	4.43	2.13	1.58	3.78	
Average	1.48	0.71	0.53	1.26	(3.98)
Adjusted Average	1.49	0.71	0.53	1.27	

If this averaging has successfully removed the residual variations, the average ratios should add up to 4.0 units. As they do not, they must be adjusted by multiplying each of them by the ratio 4.00/3.98.

The analysis is completed by calculating the residual variations which, as before, consist of the cyclical and irregular fluctuations:

$$\frac{Y}{T} = S \times R$$

$$\frac{Y}{T \times S} = R$$

The residual variations calculated in this way are shown in the final column of the following table:

Year	**Qtr**	**Sales** (£000)	**Trend**	**Seasonal**	**Residual**
2003	3	150	278	0.53	1.02
	4	400	289	1.27	1.09
2004	1	400	294	1.49	0.91
	2	210	304	0.71	0.97
	3	180	360	0.53	0.94
	4	450	434	1.27	0.82
2005	1	800	478	1.49	1.12
	2	400	554	0.71	1.02
	3	340	629	0.53	1.02
	4	900	661	1.27	1.07
2006	1	950	681	1.49	0.94
	2	500	698	0.71	1.01

The final two columns give the proportional changes for seasonal fluctuations and residual variations and they show that seasonal fluctuations can account for up to 50 per cent of the changes in the trend figures. The residual variations calculated show that the trend and seasonal sales could vary by as much as 18 per cent.

D. FORECASTING

Assumptions

The reason for isolating the trend within a time series is to be able to make a prediction of its future values, and thus estimate the movement of the time series. Before looking at the various methods available to carry out this process, we must state two assumptions that must be made when forecasting:

(a) That Conditions Remain Stable

Those conditions and factors which were apparent during the period over which the trend was calculated must be assumed to be unchanged over the period for which the forecast is made. If they do change, then the trend is likely to change with them, thus making any predictions inaccurate. For example, forecasts of savings trends based on given interest rates will not be correct if there is a sudden change (either up or down) in these rates.

(b) That Extra Factors Will Not Arise

It is sometimes the case that, when trends are predicted beyond the limits of the data from which they are calculated, extra factors will arise which influence the trend. For example, there is a limit to the number of washing machines that can be sold within a country. This capacity is a factor that must be considered when making projections of the future sales of washing machines. Therefore, in forecasting from a time series it must be assumed that such extra factors will not arise.

These assumptions are similar to those mentioned when we looked at the extrapolation of a regression line.

Methods of Forecasting

There are two main methods of forecasting, although both are primarily concerned with short-term forecasts because the assumptions mentioned previously will break down gradually for periods of longer than about a year.

(a) Moving Averages Method

This method involves extending the moving average trend line drawn on the graph of the time series. The trend line is extended by assuming that the gradient remains the same as that calculated from the data. The further forward you extend it, the more *unreliable* becomes the forecast.

When you have read the required trend value from the graph, the appropriate seasonal fluctuation is added to this and allowance is made for the residual variation. For example, consider the Premium Bond sales shown in Figure 9.7. On this figure the moving average trend line stops at the final quarter of 2007. If this line is extrapolated with the same gradient to the first quarter of 2008 then:

2008 1st Qtr: Trend = 750

This is multiplied by the seasonal variation as it is a multiplicative model, i.e. 750 × 149 = 1,118, and the residual variation which varied by as much as ± 18 per cent is added to this. Therefore the final short-term estimate for the sales of Premium Bonds for the first quarter of 2008 is £1,118,000 ± £201,000.

Although fairly easy to calculate, this forecast, like all others, must be treated with caution, because it is based on the value of the trend calculated for the final quarter of 2007, so if this happens to be an especially high or low value then it would influence the trend, and thus the forecast, considerably.

(b) Least Squares Method

If the line of best fit ($y = a + bx$) is used as the trend line and drawn on a time series graph, it can be extended to give an estimate of the trend. Preferably the required value of x can be substituted in the equation to give the trend value. The seasonal fluctuation and residual variations must be added as in the moving averages method.

Using the results of the earlier example involving days lost through sickness at a factory, the trend line was:

$$y = 21.92 + 1.46x$$

where x took all the integer values between 1 and 20.

Now suppose we want to estimate the number of days lost in the first quarter of 2008, i.e. when $x = 21$. The value of the trend would be:

$$Y = 21.92 + 1.46 \times 21$$

$$Y = 52.58$$

$Y = 53$ days, rounded to whole days.

(This result could also be read from the graph in Figure 9.6.)

To this must be added, as it is an additive model, the seasonal fluctuation for a first quarter, which was about 11 days, making a total of 64 days. The residual variation for this series was a maximum of ± 5 days. Therefore the forecast for days lost through sickness for the first quarter of 2008 is between 59 and 69 days.

This forecast again is not entirely reliable, as the trend is depicted by one straight line of a fixed gradient. It is a useful method for short-term forecasting, although like the previous method it becomes more *unreliable* the further the forecast is extended into the future.

There are no hard and fast rules to adopt when it comes to choosing a forecast method. Do not think that the more complicated the method the better the forecast. It is often the case that the simpler, more easily understood methods produce better forecasts, especially when you consider the amount of effort expended in making them. Remember that, whatever the method used for the forecast, it is only an educated guess as to future values.

E. THE Z CHART

We will conclude this chapter with a short description of a particular type of chart which plots a time series, called a Z chart. It is basically a means of showing three sets of data relating to the performance of an organisation over time. The three sets of data are plotted on the same chart and should be kept up-to-date. The graphs are:

(a) The plot of the current data, be it monthly, quarterly or daily.

(b) The cumulative plot of the current data.

(c) The moving total plot of the data.

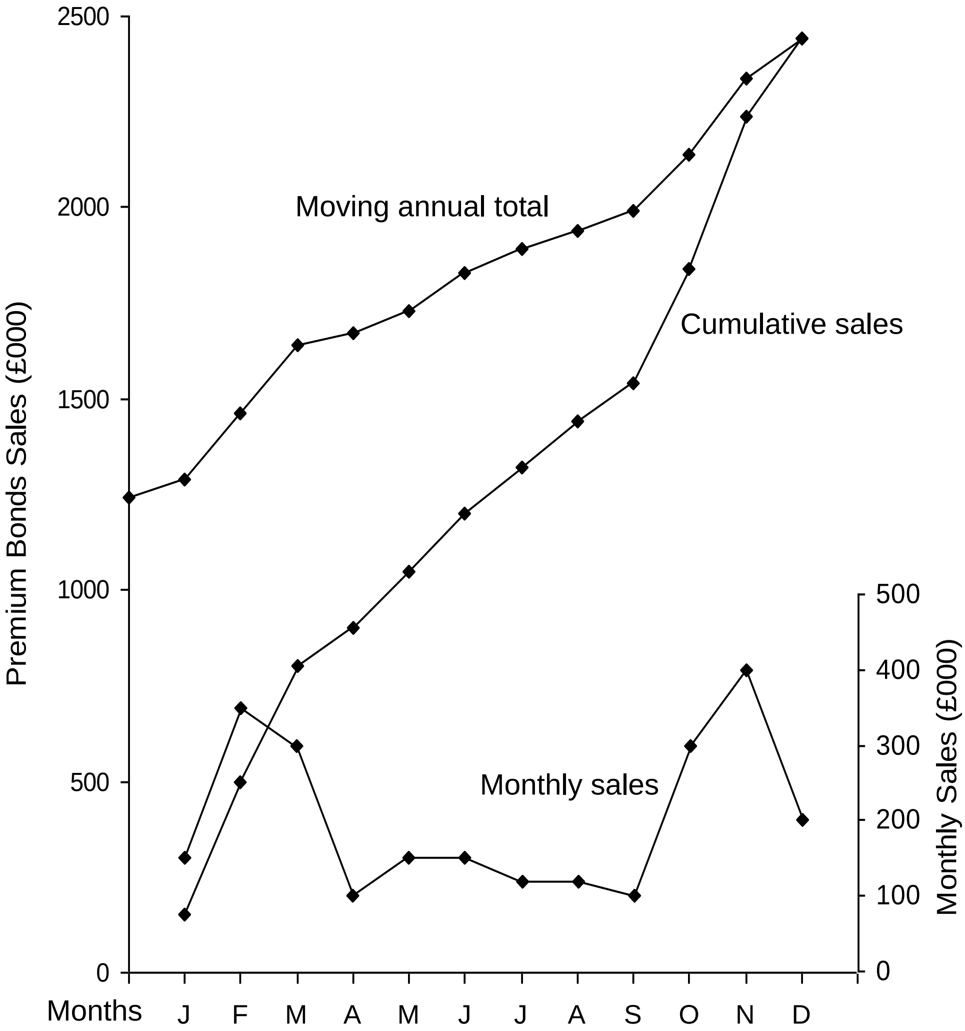
The Z chart is often used to keep senior management informed of business developments. As an example, we will plot a Z chart for the sales of Premium Bonds in 2005 using the data of the table below with the sales broken down into months. The table also shows the cumulative monthly sales and the moving annual totals. Note that the scale used for (a) is shown on the right of the chart and is twice that used for (b) and (c) so that the fluctuations in monthly sales show up more clearly. This is a device often used so that the chart is not too large.

Year	**Month**	**Sales** (£000)	**Cumulative Sales**	**Moving Annual Total**
				1,240
2005	Jan	150	150	1,290
	Feb	350	500	1,460
	Mar	300	800	1,640
	Apr	100	900	1,670
	May	150	1,050	1,730
	June	150	1,200	1,830
	July	120	1,320	1,890
	Aug	120	1,440	1,940
	Sept	100	1,540	1,990
	Oct	300	1,840	2,140
	Nov	400	2,240	2,340
	Dec	200	2,440	2,440

These totals are presented in Figure 9.9. It is called a Z chart because the position of the three graphs on the chart resembles the letter Z.

This is a useful chart because management can see at a glance how production is progressing from one month to the next. It is also possible to compare the current year's performance with a set target or with the same periods in previous years.

Figure 9.9: Z chart for Premium Bond sales



SUMMARY

In this chapter we have discussed the main models used to analyse time series. We began by identifying the various factors into which a time series may be divided in order to use these models, and went on to show how to separate a time series into these constituent factors. This is an important subject and you should particularly note the following points:

- Set out all calculations systematically in tables.
- The layout of the table used for calculation of centred moving averages is very important for all models.
- You must learn thoroughly the method of calculating and adjusting seasonal variations for all models.

Chapter 10

Probability

Contents *Page*

A.	**Introduction**	**145**
	Chance and Uncertainty	145
	Choosing a Scale	145
	Degrees of Probability	145
B.	**Two Laws of Probability**	**146**
	Addition Law for Mutually Exclusive Events	146
	Complementary Events	147
	Multiplication Law for Independent Events	147
	Distinguishing the Laws	149
C.	**Permutations**	**149**
	Listing Possible Results	149
	What is a Permutation?	149
	Permutations of Different Items	150
	Permutations of Some Items Only	150
	Permutations Including Similar Items	151
D.	**Combinations**	**152**
	What is a Combination?	152
	Equivalent Combinations	153
	Complications and Restrictions	153
E.	**Conditional Probability**	**154**
F.	**Sample Space**	**155**
G.	**Venn Diagrams**	**157**
	General Addition Law of Probabilities	157
	Mutually Exclusive Events	162

(Continued over)

General Multiplication Law of Probability 163
Independent and Dependent Events 166

Summary 167

Answers to Questions for Practice 168

A. INTRODUCTION

Chance and Uncertainty

"Probability" is one of those ideas about which we all have some notion, but many of these notions are not very definite. Initially, we will not spend our time trying to get an exact definition, but will confine ourselves to the task of grasping the idea generally and seeing how it can be used. Other words which convey much the same idea are "chance" and "likelihood". Just as there is a scale of, say, temperature, because some things are hotter than others, so there is a scale of probability, because some things are more probable than others. Snow is *more* likely to fall in winter than in summer; a healthy person has *more chance* of surviving an attack of influenza than an unhealthy person.

There is, note, some uncertainty in these matters. Most things in real life are uncertain to some degree or other, and it is for this reason that the *theory of probability* is of great *practical* value. It is the branch of mathematics which deals specifically with matters of uncertainty, and with assigning a value between zero and one to measure how likely it is that an event will occur. For the purpose of learning the theory, it is necessary to start with simple things like coin tossing and dice throwing, which may seem a bit remote from business and industrial life, but which will help you understand the more practical applications.

Choosing a Scale

First of all, let us see if we can introduce some precision into our vague ideas of probability. Some things are very unlikely to happen. We may say that they are very improbable, or that they have a very low probability of happening. Some things are so improbable that they are absolutely impossible. Their probability is so low that it can be considered to be zero. This immediately suggests to us that we can give a numerical value to at least one point on the scale of probabilities. For *impossible things*, such as pigs flying unaided, the *probability is zero*. By similarly considering things which are absolutely certain to occur, we can fix the top end of the scale at 100%. For example, the probability that every living person will eventually die is 100%.

It is often more convenient to talk in terms of proportions than percentages, and so we say that for absolute certainty the probability is 1. In mathematical subjects we use symbols rather than words, so we indicate probability by the letter p and we say that the probability scale runs from $p = 0$ to $p = 1$ (so p can never be greater than 1).

Degrees of Probability

For things which may or may not happen, the probability of them happening obviously lies somewhere between 0 and 1.

First, consider tossing a coin. When it falls, it will show either heads or tails. As the coin is a fairly symmetrical object and as we know no reason why it should fall one way rather than the other, then we feel intuitively that there is an equal chance (or, as we sometimes say, a 50/50 chance) that it will fall either way. For a situation of equal chance, the probability must lie exactly halfway along the scale, and so the probability that a coin will fall heads is 1/2 (or $p = 0.5$). For tails, p is also 0.5.

Next, consider rolling a six-sided die as used in gambling games. Here again this is a fairly symmetrical object, and we know of no special reason why one side should fall uppermost more than any other. In this case there is a 1 in 6 chance that any specified face will fall uppermost, since there are 6 faces on a cube. So the probability for any one face is 1/6 ($p = 0.167$).

As a third and final example, imagine a box containing 100 beads of which 23 are black and 77 are white. If we pick one bead out of the box at random (blindfold and with the box well shaken up) what is the probability that we will draw a black bead? We have 23 chances out of 100, so the probability is 23/100 (or $p = 0.23$).

Probabilities of this kind, where we can assess them from prior knowledge of the situation, are called *a priori* probabilities.

In many cases in real life it is not possible to assess a priori probabilities, and so we must look for some other method. What is the probability that a certain drug will cure a person of a specific disease? What is the probability that a bus will complete its journey without having picked up a specified number of passengers? These are the types of probabilities that cannot be assessed a priori. In such cases we have to resort to experiment. We count the *relative frequency* with which an event occurs, and we call that the probability. In the drug example, we count the number of cured patients as a proportion of the total number of *treated* patients. The probability of cure is then taken to be:

$$p = \frac{\text{number of patients cured}}{\text{number of patients treated}}$$

In a case like this, the value we get is only an *estimate*. If we have more patients, we get a better estimate. This means that there is the problem of how many events to count before the probability can be estimated accurately. The problem is the same as that faced in sample surveys. Probabilities assessed in this way, as observed proportions or relative frequencies, are called *empirical probabilities*.

In cases where it is not possible to assign a priori probabilities and there is no empirical data available to enable empirical probabilities to be computed, probabilities may simply have to be based on people's experiences and personal opinions. Such probabilities are called *subjective probabilities,* and in such cases, it is normally advisable to consult an expert in the field. For example, if a business manager wishes to know the probability that interest rates will rise over the next three months, he or she would be advised to consult a financial economist to gain an expert opinion.

B. TWO LAWS OF PROBABILITY

Addition Law for Mutually Exclusive Events

If a coin is tossed, the probability that it will fall heads is 0.5. The probability that it will fall tails is also 0.5. It is certain to fall on one side or the other, so the probability that it will fall either heads or tails is 1. This is, of course, the *sum* of the two separate probabilities of 0.5. This is an example of the a*ddition law* of probability. We state the addition law as:

> *"The probability that one or other of several mutually exclusive events will occur is the sum of the probabilities of the several separate events."*

Note the expression "mutually exclusive". This law of probability applies only in cases where the occurrence of one event *excludes* the possibility of any of the others. We shall see later how to modify the addition law when events are not mutually exclusive.

Heads automatically excludes the possibility of tails. On the throw of a die, a six excludes all other possibilities. In fact, all the sides of a die are mutually exclusive; the occurrence of any one of them as the top face necessarily excludes all the others.

Example:

What is the probability that when a die is thrown, the uppermost face will be either a two or a three?

The probability that it will be two is 1/6.

The probability that it will be three is 1/6.

Because the two, the three, and all the other faces are mutually exclusive, we can use the addition law to get the answer, which is 2/6, i.e. 1/6 + 1/6, or 1/3.

You may find it helpful to remember that we use the addition law when we are asking for a probability in an either/or situation.

Complementary Events

An event either occurs or does not occur, i.e. we are certain that one or other of these two situations holds. Thus the probability of an event occurring plus the probability of the event not occurring must add up to one, that is:

$$\text{P(A)} + \text{P(not A)} = 1 \qquad (a)$$

where P(A) stands for the probability of event A occurring.

A^1 or $\overline{A}$ is often used as a symbol for "not A". "A" and "not A" are referred to as complementary events. The relationship (a) is very useful, as it is often easier to find the probability of an event not occurring than to find the probability that it does occur. Using (a) we can always find P(A) by subtracting P(not A) from one.

Example:

What is the probability of a score greater than one when one die is thrown once?

Method 1:

Probability of a score greater than $1 = \text{P(score} > 1)$

$$= \text{P(score not} = 1)$$

$$= 1 - \text{P(score} = 1)\text{, using (a)}$$

$$= 1 - \frac{1}{6} = \frac{5}{6}$$

Method 2:

Probability of a score greater than $1 = \text{P(score} > 1)$

$$= \text{P(2 or 3 or 4 or 5 or 6)}$$

$$= \text{P(2)} + \text{P(3)} + \text{P(4)} + \text{P(5)} + \text{P(6)} \text{ using addition law}$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6} \text{ as before.}$$

Multiplication Law for Independent Events

Now suppose that two coins are tossed. The probability that the first coin will show heads is 0.5, and the probability that the second coin will show heads is also 0.5. But what is the probability that *both* coins will show heads? We cannot use the addition law because the two events are not mutually exclusive – a first coin landing heads does not prevent a second coin landing heads. If you did try to apply the addition law, you would get 0.5 + 0.5, which is 1.0, meaning that two heads are bound to result – and that is nonsense! We can get the correct answer by listing all the possible results of tossing two coins:

First Coin	**Second Coin**
Heads	Heads
Heads	Tails
Tails	Heads
Tails	Tails

There are four possible results, one of which is the result we are seeking. We have no reason to suppose that there is anything to favour any particular result, and so the probability that both coins will show heads is 1 in 4 (0.25), i.e. $p = 0.25$. This is an example of the *multiplication law* of probability, which states:

> *"The probability of the **combined occurrence** of two or more independent events is the product of the probabilities of the separate events."*

Note the word "independent". Events are said to be *independent* when the probability of either of them is not affected by the occurrence or non-occurrence of the other. In our coin example, the way one coin falls has absolutely no effect on the way the other one will fall, and so the two events are independent. We may therefore apply the multiplication law, which gives the probability that both heads will occur as 0.5×0.5, i.e. 0.25.

Example:

Two dice are thrown separately. What is the probability that both dice will show a five uppermost?

The events in this case are the showing of a five. The dice have no effect on one another, and so the events are independent.

The probability that the first die shows five is 1/6. The probability that the second die shows five is 1/6.

The probability that both dice show five is $1/6 \times 1/6$, i.e. 1/36.

Try the following examples before looking at the answers (which are given at the end of the chapter).

Questions for Practice 1

1. In a box of 200 items taken from a factory production line, there are 25 faulty items. An inspector picks out an item at random. What is the probability that the selected item is not faulty?

2. In a second box, there are 1,000 items of which 100 are faulty. The inspector picks out an item at random.

   (i) What is the probability that this item is faulty?

   (ii) What is the probability that both items (i.e. one from each box) are not faulty?

3. From a pack of 52 ordinary playing cards, a card is drawn at random. What is the probability that it is *either* a two *or* a seven?

4. If 3 coins are thrown, what is the probability that all 3 will show tails?

Now check your answers with those given at the end of the chapter.

Distinguishing the Laws

Although the above laws of probability are not complicated, you must think carefully and clearly when using them. Remember that events must be *mutually exclusive* before you can use the *addition law*, and they must be *independent* before you can use the *multiplication law*. Another matter about which you must be careful is the listing of equally likely outcomes. Be sure that you list all of them. Earlier we listed the possible results of tossing two coins:

First Coin	**Second Coin**
Heads	Heads
Heads	Tails
Tails	Heads
Tails	Tails

Here there are four equally likely outcomes. Do not make the mistake of saying, for example, that there are only two outcomes (both heads or not both heads), you must list all the possible outcomes. (In this case "not both heads" can result in three different ways, so the probability of this result will be higher than "both heads".)

In this example the probability that there will be one heads and one tails (heads – tails or tails – heads) is 0.5. This is a case of the addition law at work, the probability of heads – tails (1/4) *plus* the probability of tails – heads (1/4). Putting it another way, the probability of different faces is equal to the probability of the same faces – in both cases 1/2.

C. PERMUTATIONS

Listing Possible Results

When we deal with simple things like throwing a few coins or dice, it is easy to make a list of all the possible results. In more complicated cases, it would be impracticable to write out the whole list. Fortunately there are some simple mathematical methods for calculating the number of possible results. These methods are referred to as *permutations* and *combinations*.

What is a Permutation?

The word permutation means a particular sequence or order of arrangement of things. For example, BAC and CBA are both permutations of the first three letters of the alphabet. Permutation problems are concerned with the number of possible sequences into which things can be arranged. There is a basic principle governing such problems:

> *"If one operation can be done in m ways, and if a second operation can be done in n ways, then the two operations can be done in succession in m times n different ways."*

For the purposes of this course we do not need to prove it but only to know and understand it thoroughly. The principle can be extended to any number of operations greater than 2.

Example:

There are three different coloured buses (red, yellow and green) which run between two places. If I want to use a different coloured bus for each direction, in how many different ways can I make the double journey?

Applying the basic principle, we see that the first part of the trip can be done in three ways (red, yellow and green), while the second part of the trip can be done in only two ways (excluding the colour already used). Thus the total number of different possible ways is $3 \times 2 = 6$.

It would be a good idea at this stage for you to try to write out the list of the six possible alternatives.

The principle can be applied to the coin example. The first coin can fall in two possible ways (heads or tails) and the second coin can fall in two possible ways (heads or tails), and so the total number of different possible ways is $2 \times 2 = 4$.

Permutations of Different Items

If we have a group of *different* items, we can now calculate the total number of permutations quite easily. Suppose there are 4 different things which we arrange in a row. Any one thing can be put first, so the first place can be filled in 4 different ways. After the first place has been filled, only 3 things remain, and so the second place can be filled in 3 possible ways. Then the third place can be filled in 2 possible ways and the fourth place in only 1 way. The basic principle tells us that the total number of different arrangements is therefore $4 \times 3 \times 2 \times 1 = 24$. If there had been 5 items, the total number of permutations would have been $5 \times 4 \times 3 \times 2 \times 1 = 120$. You can see the pattern – you simply multiply all the numbers down to 1. There is a special name for this continued product from a given number down to 1; it is called a *factorial*. Thus:

2×1 is factorial 2 (also called 2 factorial), and it is equal to 2

$3 \times 2 \times 1$ is factorial 3, and it is equal to 6

$4 \times 3 \times 2 \times 1$ is factorial 4, and it is equal to 24

and so on.

A special sign is used for a factorial, an exclamation mark, so:

$2! = 2 = 1$

$3! = 3 \times 2 \times 1 = 6$

$7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5{,}040$

and so on.

We can now say that the rule for calculating the total number of permutations of n things is n!

Permutations of Some Items Only

Sometimes we need permutations of only *some* of the items at our disposal. The calculation is very similar, but we stop after the appropriate number of factors instead of taking the multiplication right down to 1. Thus, if we wish to know the number of possible arrangements of *any 3* things taken from a group of 9 things, we calculate $9 \times 8 \times 7$ (i.e. 504). That is just like 9! except that we stop after 3 factors. We use another symbol for this. The number of permutations of 9 things taken only 3 at a time is written as ${}_9P_3$ or 9_P3, and read as "nine P three".

The rule is quite general, and we speak of the number of permutations of n things taken r at a time as ${}_nP_r$. To calculate the value of ${}_nP_r$ we start to work out n! but stop when we have done r factors. Thus ${}_9P_4 = 9 \times 8 \times 7 \times 6$ which is 3,024, and ${}_{100}P_2 = 100 \times 99$ which is 9,900. An alternative method can be seen by taking the ${}_9P_3$ example again and writing:

$${}_9P_3 = 9 \times 8 \times 7 = \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{9!}{6!}$$

You see that the 6 is 9 minus 3, so, for all values of n and r, we can put:

$${}_nP_r = \frac{n!}{(n-r)!}$$

Note that 0! is defined as equal to 1, *not* zero:

$$0! = 1$$

The formula for $_nP_n$ is then

$$= \frac{n!}{(n - n)!} = \frac{n!}{0!} = n!$$

just as we would expect.

Permutations Including Similar Items

So far we have assumed that the n items from which the permutations are taken are all different. But what if some items are *identical*? Take, for example, the word "ACCOUNT". There are 7 letters and therefore 7! possible permutations of those letters. But some of these permutations will look the same because the two letters C are interchanged. So there will be *fewer than* 7! distinguishable permutations. The formula for calculating the number of distinguishable permutations is quite simple, and for this course we do not need to prove it but merely know it. In the example just given, there are two identical letters C, and the total number of distinguishable permutations is:

$$\frac{7!}{2!}$$

A more complicated example would be the word "STATISTIC", where there are 9 letters altogether, but S occurs twice, T occurs 3 times and I occurs twice. Here the total number of distinguishable permutations is:

$$\frac{9!}{2! \times 3! \times 2!}$$

From these two examples you might be able to see the pattern of the general rule. The rule says that if we have n items, of which p are alike of one kind, q are alike of another kind and r are alike of yet another kind, then the total number of distinguishable permutations of the n items is:

$$\frac{n!}{p! \times q! \times r!}$$

and so on if there are more than three groups.

Before we go on to *combinations*, you should make yourself quite familiar with *permutations* so that you do not become confused. Make sure you know what they are – not merely how to calculate them. Attempt the following examples in Questions for Practice 2.

Questions for Practice 2

1. Find the value of

   (i) $\frac{8!}{5!}$

   (ii) $\frac{73!}{72!}$

2. In how many different orders can 6 objects be arranged for checking by an inspector?

3. In how many different ways can 3 ledgers be submitted to 5 auditors if:

   (i) No auditor deals with more than 1 ledger?

(ii) Any auditor may deal with any number of ledgers?

4. Express in factorials only:

(i) 9 × 8 × 7 × 6

(ii) 5 × 6 × 7

5. Write down the formulae for:

(i) $_4P_n$

(ii) $_nP_4$

(iii) $_{2n}P_n$

6. In how many ways can 3 dots and 6 commas be arranged in a line?

Now check your answers with those given at the end of the chapter.

D. COMBINATIONS

What is a Combination?

When dealing with permutations, we are concerned principally with the order or *sequence* in which things occur. Other problems occur in which we need to calculate the number of groups of a certain size *irrespective of their sequence*.

For example, consider the question of how many possible ways there are of choosing a football team of 11 men from a group of 15 club members. Here there is no question of putting the team in a particular sequence, but merely the problem of finding the number of possible different teams. The problem is one of finding the number of *combinations* of 15 things taken 11 at a time. The symbol for this is $_{15}C_{11}$ or $^{15}C_{11}$.

Sometimes you may come across a different symbol,

$$\begin{bmatrix} 15 \\ 11 \end{bmatrix}$$

However, this means the same thing and they are all read as "fifteen C eleven".

The number of combinations of n things taken r at a time is obviously less than the number of permutations, *because each combination can have r! different arrangements*. That gives us the clue to finding a formula for $_nC_r$. There must be r! times as many permutations as there are combinations and so:

$$_nC_r = \frac{n!}{r!(n-r)!}$$

Example:

A factory has 5 identical production lines. During a certain period, there are only enough orders to keep 3 lines working. How many different ways of choosing the 3 lines to be worked are available to the factory manager?

The manager's problem, expressed mathematically, is to find the number of combinations of 5 things taken 3 at a time, that is:

$$_5C_3 = \frac{5!}{3!(5-3)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1}$$

Cancelling out as many factors as possible gives:

$$_5C_3 = 5 \times 2 = 10$$

As in previous examples, it would be a good idea for you to verify this result by writing out all the 10 possible combinations.

Equivalent Combinations

If we select 4 items from 7 items, we leave 3 items behind. For every set of 4 that we choose there must be a set of 3 that we do *not* choose. It follows that:

the number of combinations we can choose, $_7C_4$, must equal

the number of combinations that we do not choose, $_7C_3$.

By applying the reasoning to the general case, we get the important fact that:

$$_nC_r = {}_nC_{n-r}$$

This is very useful sometimes when one of the calculations may be much easier than the other. All that this means is that you should cancel out the *larger* of the two factors in the denominator, as the following examples will show.

Example 1:

Find the value of $_8C_5$

$$_8C_5 = \frac{8!}{5!\,3!}$$

Cancel out the 5! and we get

$$\frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$$

Example 2:

Find the value of $_8C_3$

$$_8C_3 = \frac{8!}{3!\,5!} = 56$$

which is mathematically the same as Example 1 above.

Note: $_nC_r$ is also termed a *binomial coefficient*.

Complications and Restrictions

In some practical problems it is necessary to calculate the number of combinations, subject to certain restrictions. There are no general rules to cover such cases, and so each one must be thought out carefully. The arithmetic is no more difficult than that which we have already done. Here are some further examples, which you should study carefully:

Example 1:

There are 7 clerks in an office. A team of 4 clerks is needed for a special checking job. In how many ways can the team be made up if the longest-serving clerk *must* be included?

If the longest-serving clerk is put in the team, that leaves us with the problem of finding another 3 clerks out of 6. So the answer will be:

$$_6C_3$$

which is 20.

Example 2:

In how many ways can the team be made up if the only restriction is that one particular clerk is not available for this job?

In this case, we have to find the team of 4 from only 6 clerks. The answer is therefore

$$_6C_4$$

which is 15.

Notice that a careful consideration of the restrictions usually enables you to formulate a slightly different question which can then be answered by the usual kind of calculation.

Now try Questions for Practice 3.

Questions for Practice 3

1. In how many ways can a committee of 5 be chosen from 9 candidates so as to include both the youngest and the oldest candidate?
2. In how many ways can the committee be formed if the youngest candidate is excluded?
3. In a bin of 50 items from a cutting machine there are 17 defective items. Calculate the number of possible different samples of 5 which contain no defectives.
4. Calculate, using the data of the above example, the number of possible samples of 5 items from the cutting machine which contain exactly 1 defective item.

Now check your answers with those given at the end of the chapter.

E. CONDITIONAL PROBABILITY

When we dealt with the multiplication law of probability, we insisted that the events must be independent. Now we can apply the same law to non-independent events, provided that we make due allowance for the dependence. The method is best explained by first working an example.

A box contains 13 red beads and 7 white beads. If we make a random selection of 1 bead from the box, then the probability that it will be a white bead is 7/20. Now suppose that this first bead turns out to be a white one. We put it on one side and make another random selection of 1 bead. What is the probability that this second bead will be white? It is not 7/20 this time, because there are only 6 white beads left in the box. There are still 13 red ones, making 19 altogether, so the probability that the second one is white is 6/19.

We can now apply the multiplication law to find out the probability of both the first and second beads being white. It is:

$$\left[\frac{7}{20}\right] \times \left[\frac{6}{19}\right] = \left[\frac{42}{380}\right] = \left[\frac{21}{190}\right]$$

The probability of 6/19 for the second selection is called the *conditional probability* for the second event, on the assumption that the first event has happened. The more general form of the multiplication law is therefore:

> *"The probability of the combined occurrence of two events A and B is the product of the probability of A and the **conditional probability** of B, on the assumption that A has happened."*

The law can be applied to a series of more than two events if care is taken to assess the successive conditional probabilities accurately.

Now attempt Questions for Practice 4.

Questions for Practice 4

1. With the same box of beads as used in the previous paragraphs, what is the probability that the first 2 beads drawn from the box will be red?
2. With the same box again, what is the probability that the first 2 beads will be a red followed by a white?
3. A bin contains 100 snoggle pins made on a new machine. Of these 100 items, 20 are defective. An inspector draws 5 items at random from the bin. What is the probability that all 5 are not defective? (Don't work out all the arithmetic – just show the numbers and the calculations to be done.)
4. Show how to do Question 3 using combinations only, and verify that this gives the same result as before.

Now check your answers with those given at the end of the chapter.

F. SAMPLE SPACE

You need a clear head to perform probability calculations successfully. It helps to have some diagrammatic form of representation, and this is our concern in the final sections of this chapter. First, however, we must introduce more terminology. When we, say, toss a coin three times and note the outcome, we are performing a *statistical experiment*. If we make a list of all possible outcomes of our experiment, we call this a *sample space*. (This is a similar idea to a sampling frame, i.e. a list of a population, mentioned earlier in our discussion of practical sampling methods.) The sample space in the above coin-tossing experiment is:

HHH HTT

THH THT

HTH TTH

HHT TTT

where, for example, THH means that on the first toss we obtained a tail, on the second a head, and on the third a head.

Consider another example. Suppose we have 5 people A, B, C, D, and E. We wish to select for interview a random sample of 2, i.e. each of A, B, C, D, and E must have the same chance of being chosen. What is the sample space, i.e. the list of all possible different samples? The sample space is:

AB BC CD DE

AC BD CE

AD BE

AE

In this example the order of the sample, i.e. whether we choose A followed by B or B followed by A, is not relevant as we would still interview both A and B.

Having identified our sample space, we might be interested in a particular section of it. For instance, in our first example we might want to see in how many of the outcomes we obtained only one head. We call this collection of outcomes an *event*, i.e. we are interested in the event: *obtaining exactly one head*. We often find it convenient to label an event by a capital letter such as A, B, etc. and so we could say event A is obtaining exactly one head.

Looking back at the sample space, we see that there are three outcomes making up this event. If we have a fair coin, the probability of obtaining a head or a tail at any toss is 1/2 and all the outcomes in the sample space are equally likely. There are eight outcomes in all, so we can now deduce that probability of obtaining exactly one head in 3 tosses of a fair coin is:

$$= \frac{\text{no. of outcomes in sample space with one head}}{\text{total no. of outcomes in sample space}} = \frac{3}{8}$$

or, alternatively, we could write:

$$\text{P (event A)} = \frac{\text{no. of outcomes in A}}{\text{total no. of outcomes}} = \frac{3}{8}$$

Note: P(event A) is usually written as p(A), P(A), Pr(A) or Prob. (A).

Questions for Practice 5

Try writing out the details of the following example.

Experiment: rolling one fair die. Sample space is:

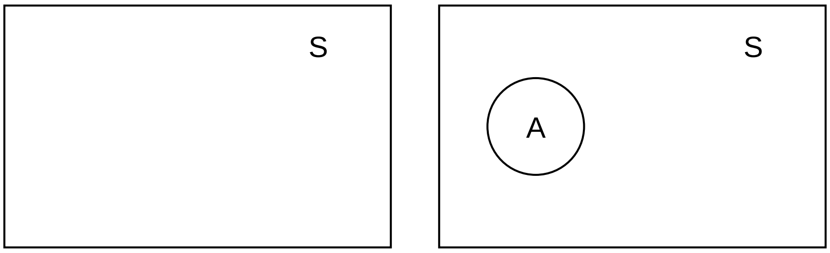
Event A	**P(A)**
1. Obtaining a score greater than 3.	
2. Obtaining an odd number.	
3. Obtaining both a score greater than 3 and an odd number.	

Now check your answers with those given at the end of the chapter.

G. VENN DIAGRAMS

In a Venn diagram, the sample space S is represented by a rectangle, and events in the sample space are denoted by areas within the rectangle (Figure 10.1):

Figure 10.1: Venn diagram – an event A in S



Sample space S A is an event in S

If all the outcomes listed in S are equally likely, then the probability of event A occurring is given by:

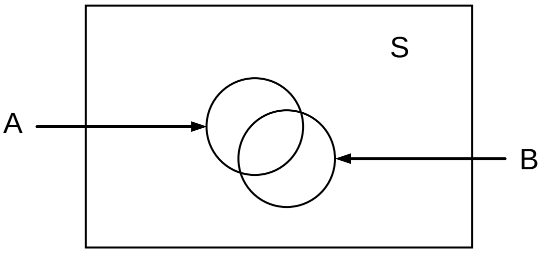
$$P(A) = \frac{\text{number of outcomes in A}}{\text{number of outcomes in S}}$$

$$= \frac{n(A)}{n(S)}$$ where n(A) is shorthand for the number of outcomes in event A.

General Addition Law of Probabilities

If we are interested in two events, A and B, then we have two areas representing A and B inside rectangle S and these *may* overlap (Figure 10.2):

Figure 10.2: Overlapping sets



Consider the three tosses of a coin example which we introduced before.

- Let event A be "obtaining exactly one head",
therefore, event A contains the outcomes (TTH, HTT, THT).
- Let event B be "obtaining a tail on the first toss",
therefore, event B contains the outcomes (TTT, TTH, THT, THH).

In this case A and B overlap because the outcomes THT and TTH are common to both.

We call this overlap "A intersection B" denoted by $A \cap B$, and this is where *both A and B* occur. Thus to evaluate the probability of both A and B occurring together, we need:

$$P(A \text{ and } B) = \frac{\text{no. of outcomes in } A \cap B}{\text{no. of outcomes in S}}$$

For our example:

$$P(\text{A and B}) = \frac{n(A \cap B)}{n(S)} = \frac{2}{8} = \frac{1}{4}$$

i.e. when we toss a coin three times, the probability of obtaining exactly one head and obtaining a tail on the first toss is 1/4.

If we now look at the combined area covered by A and B, we see that within this region we have either event A occurring **or** event B occurring or both events occurring. We call this area "A union B" denoted by $A \cup B$. Thus to evaluate the probability of A or B or both occurring we need:

$$P(\text{A or B or both}) = \frac{\text{no. of outcomes in } A \cup B}{\text{no. of outcomes in } S}$$

$$= \frac{n(A \cup B)}{n(S)} = \frac{5}{8}$$

The 5 events that are in $A \cup B$ are TTH, HTT, THT, TTT, THH. We have to be careful not to count THT and TTH twice. TTH and THT are the events that belong to $A \cap B$. We thus have the result which holds in general that:

$$A \cup B = A + B - A \cap B$$

Thus P(A or B or both)

$$= \frac{\text{no. of outcomes in } (A + B \text{ - } A \cup B)}{\text{no. of outcomes in } S}$$

$$= \frac{\text{no. of outcomes in } A + \text{no. in } B - \text{no. in } A \cap B}{\text{no. in } S}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{n(S)}$$

$$= \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)}$$

$= P(A) + P(B) - P(A \cap B)$ by definition of probability.

We thus have the general law of probabilities for *any* two events A and B:

$P(\text{A or B or both}) = P(A) + P(B) - P(A \cap B)$, i.e.

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Example 1:

If one card is drawn from a pack of 52 playing cards, what is the probability: (a) that it is either a spade or an ace; (b) that it is either a spade or the ace of diamonds?

(a) Let event B be "the card is a spade".

Let event A be "the card is an ace".

We require P(spade or ace [or both])

$= P(\text{A or B})$

$= P(A) + P(B) - P(A \cap B)$

$$P(A) = \frac{\text{no.of aces}}{\text{no. in pack}} = \frac{4}{52}$$

$$P(B) = \frac{\text{no. of spades}}{\text{no. in pack}} = \frac{13}{52}$$

$$P(A \cap B) = \frac{\text{no. of aces of spades}}{\text{no. in pack}} = \frac{1}{52}$$

Therefore, $P(\text{spade or ace}) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$

(b) Let event B be "the card is a spade".

Let event A be "the card is the ace of diamonds".

We require P(spade or ace of diamonds)

$= P(A \text{ or } B)$

$= P(A) + P(B) - P(A \cap B)$

$$P(A) = \frac{\text{no. of aces of diamonds}}{\text{no. in pack}} = \frac{1}{52}$$

$$P(B) = \frac{\text{no. of spades}}{\text{no. in pack}} = \frac{13}{52}$$

$$P(A \cap B) = \frac{\text{no. of spades which are also aces of diamonds}}{\text{no. in pack}} = 0$$

Therefore, $P(\text{spade or ace of diamonds}) = \frac{1}{52} + \frac{13}{52} = \frac{14}{52} = \frac{7}{26}$

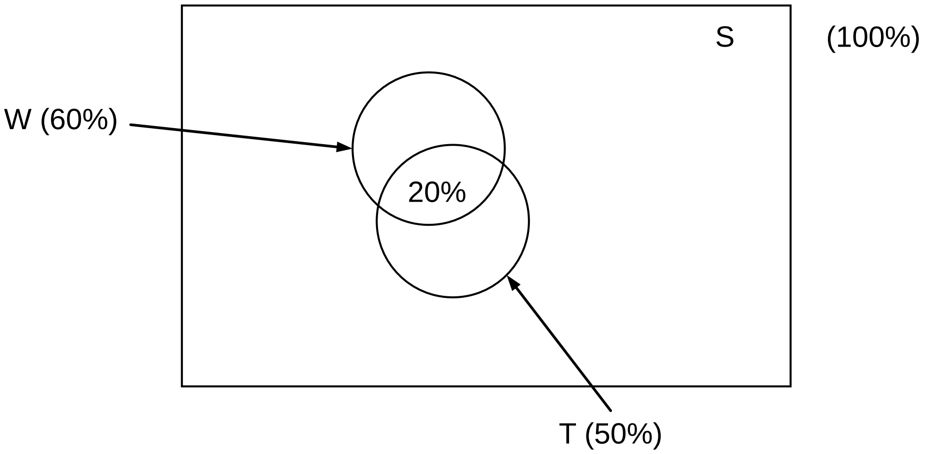
Example 2:

At a local shop 50% of customers buy unwrapped bread and 60% buy wrapped bread. What proportion of customers buy at least one kind of bread if 20% buy both wrapped and unwrapped bread?

- Let S represent all the customers.
- Let T represent those customers buying unwrapped bread.
- Let W represent those customers buying wrapped bread.

The Venn diagram is as shown in Figure 10.3 following.

Figure 10.3: Venn diagram for Example 2



P(buy at least one kind of bread) = P(buy wrapped or unwrapped or both)

= P(T or W)

= P(T) + P(W) − P(T ∩ W)

= 0.5 + 0.6 − 0.2 = 0.9

i.e. nine-tenths of the customers buy at least one sort of bread.

The addition law can be extended *to cover more events*, e.g. for three events A, B and C:

P(at least one of A or B or C occurring) = P(A ∪ B ∪ C) and we find:

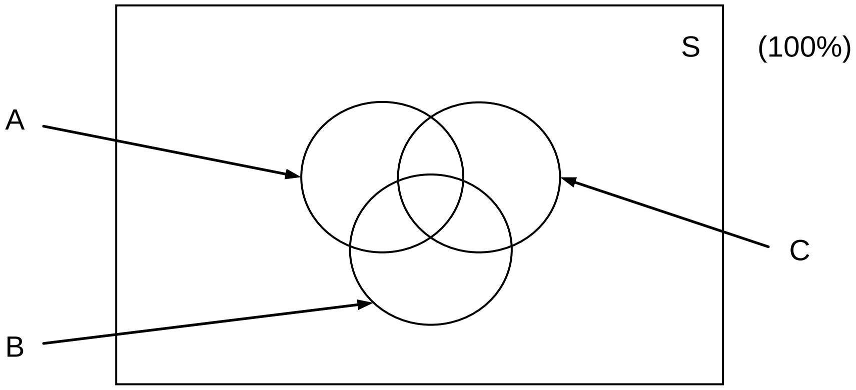
P(A ∪ B ∪ C) = P(A) + P(B) + P(C) − P(A ∩ B) − P(A ∩ C) − P(B ∩ C) + P(A ∩ B ∩ C)

Unless you like learning formulae, do not bother to remember this as you can always solve problems involving three events more easily using the Venn diagram.

Example 3:

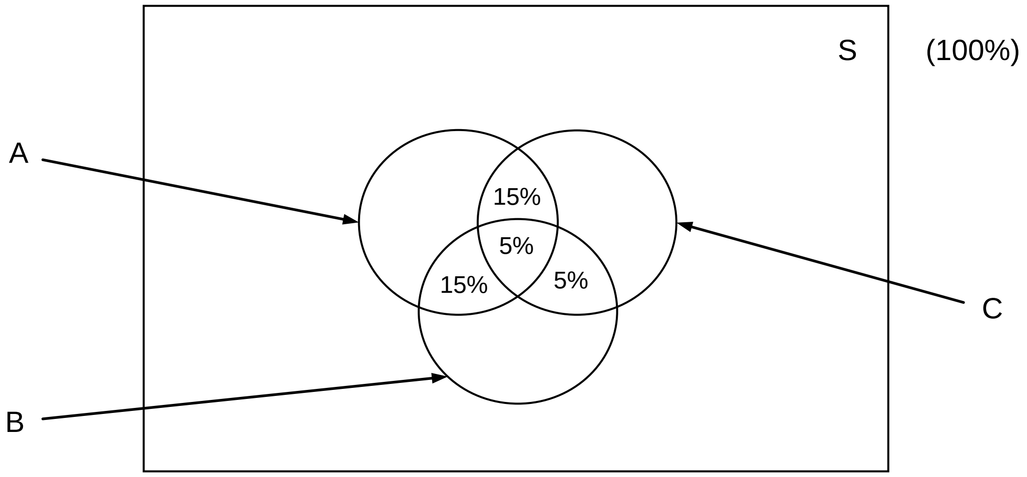
Three magazines A, B and C are published in Townsville. Of the adult population of Townsville 65% read A, 40% read B, 25% read C, 20% read both A and B, 20% read both A and C, 10% read both B and C and 5% read all three. What is the probability that an adult selected at random reads at least one of the magazines? Figure 10.4 shows this example.

Figure 10.4: Venn diagram for Example 3 – 1



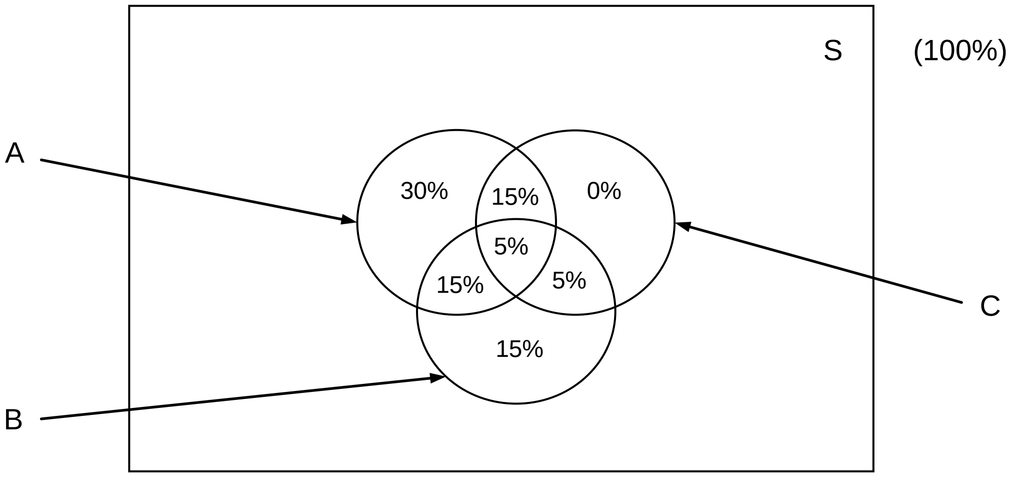
We want P(A or B or C), i.e. we require the proportion of S that is contained within the overlapping circles A, B and C. We take the information given in the question and insert the percentage lying within each section of the diagram. We start at the centre. We know 5% read all three magazines, so we insert 5% where all the three circles overlap. We are told that 20% read both A and B but 5% read C as well so that leaves 15% reading A and B but not C, so we insert 15% where just A and B overlap. Similarly for (A and C) and (B and C). Our diagram is now as shown in Figure 10.5.

Figure 10.5: Venn diagram for Example 3 – 2



We are told 65% read A, but looking at Figure 10.5 you will see that 5% + 15% + 15%, i.e. 35% read A together with at least one other magazine, leaving 30% who read A only. Similarly, 15% read B only and no one reads C only. The completed diagram is shown in Figure 10.6.

Figure 10.6: Venn diagram for Example 3 – 3



The total percentage within A, B and C is 30 + 15 + 15 + 5 + 5 + 15, i.e. 85%. Thus the probability that an adult selected at random reads at least one of the magazines is 0.85.

Now try the following exercises yourself.

Questions for Practice 6

1 In a group of 25 trainees, 16 are males, 12 are university graduates and 10 are male graduates. What is the number of female non-graduates?

2 The probability that a manager reads the *Daily Telegraph* is 0.7. The probability that she reads the *Daily Telegraph* but not the *Financial Times* is 0.6. The probability that she reads neither is 0.2. Find the probability that she reads the *Financial Times* only.

3 Employees have the choice of one of three schemes, A, B or C. They must vote for one but, if they have no preference, can vote for all three or, if against one scheme, they can vote for the two they prefer.

A sample poll of 200 voters revealed the following information:

15 would vote for A and C but not B

65 would vote for B only

51 would vote for C only

15 would vote for both A and B

117 would vote for either A or B, or both A and B, but not C

128 would vote for either B or C, or both B and C, but not A.

How many would vote for:

(i) All three schemes?

(ii) Only one scheme?

(iii) A irrespective of B or C?

(iv) A only?

(v) A and B but not C?

Now check your answers with those given at the end of the chapter.

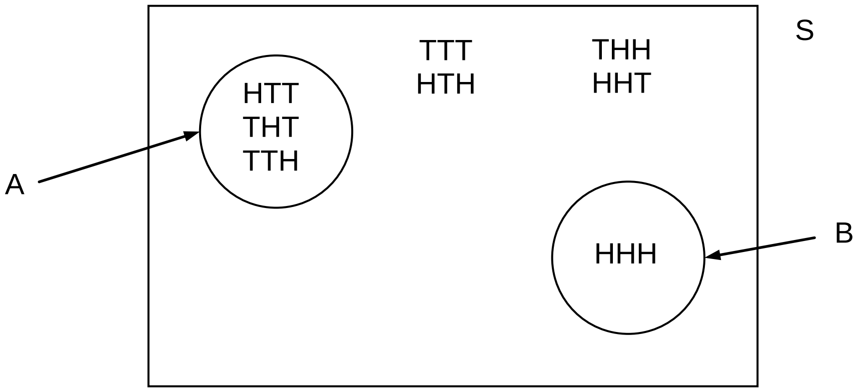
Mutually Exclusive Events

In an earlier section we defined the term "mutually exclusive events", and said that if two events were mutually exclusive then if one occurred the other could not. In a Venn diagram, if two events A and B are mutually exclusive then the areas corresponding to A and B will not overlap. If we return again to the example where we tossed a coin three times, let us consider:

- Event A = (obtaining exactly one head).
- Event B = (obtaining three heads).

These events are mutually exclusive so the Venn diagram is as shown in Figure 10.7:

Figure 10.7: Two mutually exclusive events



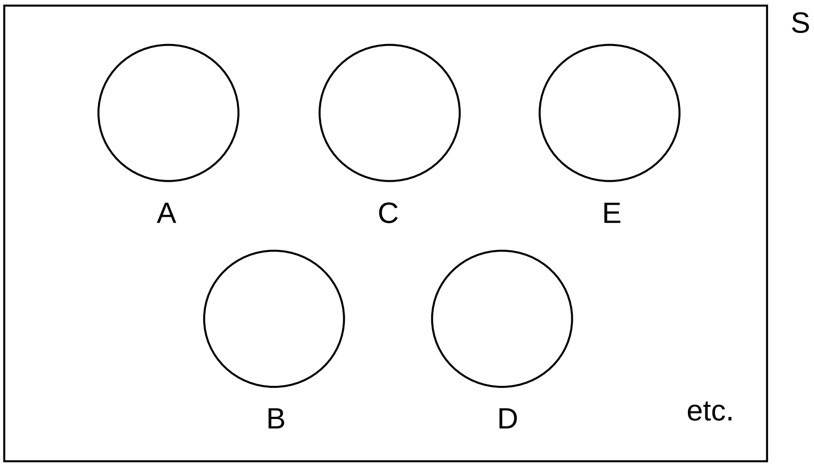
There is no overlap between Events A and B so P(A $\cap$ B) = 0 as we cannot obtain exactly one head and three heads at the same time.

Thus the general addition law simplifies to become:

P(A or B or both) = P(A) + P(B)

This is the simple addition law which we stated previously for mutually exclusive events and it can be generalised for any number of mutually exclusive events (Figure 10.8):

Figure 10.8: Many mutually exclusive events



P(A or B or C or D or E or ...) = P(A) + P(B) + P(C) + P(D) + P(E) + ...

General Multiplication Law of Probability

We have seen before that sometimes we need to work out the probability of an event A occurring, given that event B has already occurred. We called this the *conditional probability* of A given B has occurred. Let us now consider how we can work out such probabilities from the Venn diagram. Considering once more the example where we toss a coin three times, we might want to know the probability of obtaining 3 heads given that the first toss is known to be a head.

Let A = P(obtaining 3 heads) = (HHH)

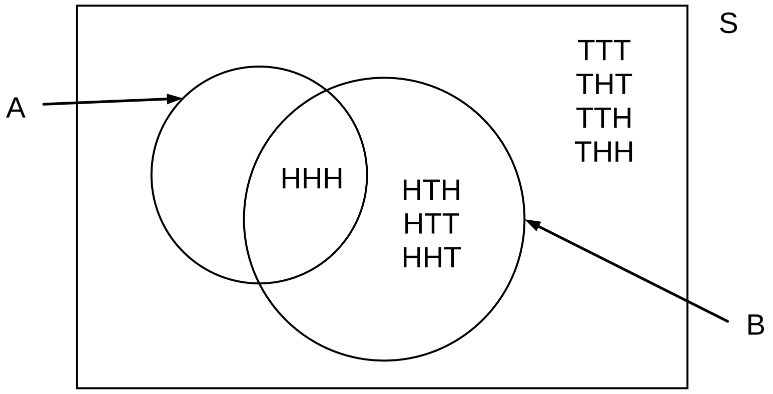
$$P(A) = \frac{1}{8}$$

B = P(first toss a head) = (HTT, HTH, HHT, HHH)

$$P(B) = \frac{4}{8}$$

A and B can be represented on the Venn diagram, Figure 10.9:

Figure 10.9



The conditional probability P(obtaining 3 heads given first toss a head) is written as P(A| B) and read as probability of A given B. Thus required probability

$$= \frac{\text{P(obtainin g 3 heads and first toss head)}}{\text{P(first toss is a head)}} = \frac{\left(\frac{1}{8}\right)}{\left(\frac{4}{8}\right)} = \frac{1}{4}$$

Thus what we are working out is the number of outcomes in the overlap as a fraction of the number of outcomes in B, as we know that B has to have occurred already, i.e.

$$P(A| B) = \frac{n(A \cap B)}{n(B)} = \frac{1}{4}$$

Thus $P(A| B) = \dfrac{P(A \cap B)}{P(B)}$ *(b)*

Example 1:

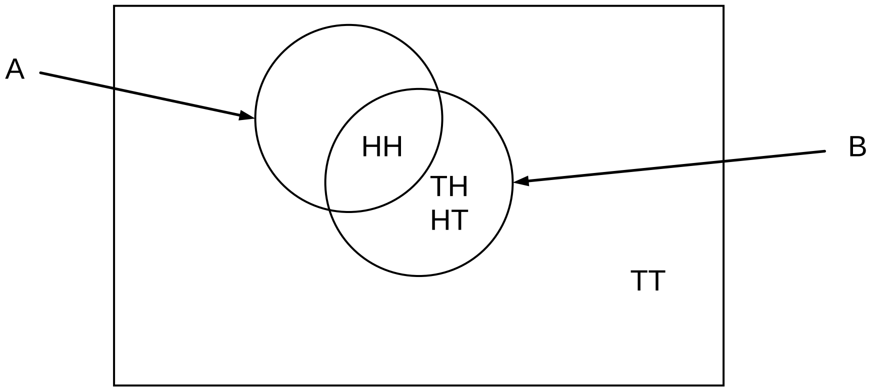
If two coins are tossed, what is the probability that both are heads, given that at least one is a head?

- Let A = P(both are heads) = (HH), and
- B = P(at least one is a head) = (HH, TH, HT).

S = (HH, TH, HT, TT)

The Venn diagram is shown in Figure 10.10:

Figure 10.10: Venn diagram for Example 1



$$P(A) = \frac{1}{4}$$

$$P(B) = \frac{3}{4}$$

$$P(A \cap B) = \frac{1}{4}$$

$$\text{Required probability} = P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\left(\frac{1}{4}\right)}{\left(\frac{3}{4}\right)} = \frac{1}{3}$$

If we rewrite equation (b) we obtain:

$$P(A \cap B) = P(B)P(A|B)$$

i.e. $P(A \text{ } \textit{and} \text{ } B) = P(B)P(A|B)$

Thus the probability of both events A and B occurring is the probability of B occurring *times* the probability of A occurring, given that B has already occurred. The above law is the general multiplication law for conditional probabilities stated in Section E.

Example 2:

A bag contains 5 red and 3 white billiard balls. If two are selected at random without replacement, what is the probability that one of each colour is drawn?

- Let R = (drawing a red ball) and
- W = (drawing a white ball).

As the sampling is without replacement, after selecting one ball we do not return it to the bag before choosing the second ball.

P(R and W) = P(R then W) or (W then R)

= P(R then W) + P(W then R) using addition law for mutually exclusive events

$= P(R)P(W|R) + P(W)P(R|W)$

$$= \left[\frac{5}{8} \times \frac{3}{7}\right] + \left[\frac{3}{8} \times \frac{5}{7}\right] = \frac{15}{56} + \frac{15}{56} = \frac{15}{28}$$

Therefore, probability that one of each colour is drawn is 15/28.

Independent and Dependent Events

For independent events the probability of A occurring does not depend on whether B has already occurred. Thus

$$P(A|B) = P(A)$$

and

$$P(A \cap B) = P(A)P(B)$$

which is the multiplication law for probabilities of independent events which we used earlier in the chapter.

Example:

Two cards are drawn at random from a pack of 52 cards. Find the probability of drawing two aces:

(a) if the sampling is with replacement; and

(b) if the sampling is without replacement.

(a) If the sampling is with replacement, this implies that after we have drawn the first card and looked at it, we put it back in the pack before drawing the second card. Thus the probability of an ace at either draw is

$$\frac{4}{52} = \frac{1}{13}$$

By the multiplication law for independent events, the probability of drawing two aces is

$$\frac{1}{13} \times \frac{1}{13} = \frac{1}{169} = 0.0059$$

(b) If the sampling is without replacement, after we have drawn the first card we do not put it back in the pack before taking the second card.

Thus the probability of an ace at the first draw is $\frac{4}{52}$

However, the probability of an ace at the second draw depends on whether or not we had an ace first time. We thus need to use the multiplication law for conditional probabilities:

Probability of drawing two aces	$=$	P(drawing an ace first time)	$\times$	P(drawing a second ace, given that the first was an ace)

$$= \frac{4}{52} \times \frac{3}{51} = 0.0045$$

i.e. the probability this time is smaller, as was to be expected.

SUMMARY

You must learn the definitions and notation in this chapter so that you are quite sure of the meaning of any question on probability. In particular, you should learn the following formulae:

1. $\text{Relative frequency} = \dfrac{\text{frequency in a particular class}}{\text{total frequency}}$

   $= \dfrac{n(A)}{n(S)} = P(A)$

2. $P(A) + P(\overline{A}) = 1 \Rightarrow P(\overline{A}) = 1 - P(A)$ (the symbol $\Rightarrow$ means "implies").

3. The multiplication law:

   (i) $P(A \cap B) = P(B)P(A|B)$

   (ii) $P(A \cap B) = P(A)P(B)$ if A and B are independent.

4. The addition law

   (i) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   (ii) $P(A \cup B) = P(A) + P(B)$ if A and B are mutually exclusive.

Practice in drawing Venn diagrams will help in solving problems. If you do not see at once how to solve a problem, tackle it in stages:

(a) Define the sample space.

(b) Define the events of interest and make sure that they belong to the sample space.

(c) Decide what type of events you are dealing with.

(d) Try to draw a Venn diagram to illustrate the problem. (Remember this may not always be possible.)

(e) Decide which probability law is required to solve the problem.

Make sure that you explain the method you use.

ANSWERS TO QUESTIONS FOR PRACTICE

Questions for Practice 1

1. Among the 200 items, 25 are faulty and therefore 175 items must be not faulty. In a random selection of one item, the probability that the item is not faulty is 175/200, which is 7/8.

2. There are two parts to this question:

   (i) The probability that the item taken from the second box is faulty is 100/1,000 $=$ 0.1.

   (ii) The probability that the second item is not faulty is:

   $$\frac{1{,}000 - 100}{1{,}000} = \frac{900}{1{,}000} = \frac{9}{10}$$

   The events are independent, so, by multiplication law, the probability is:

   $$\frac{7}{8} \times \frac{9}{10} = \frac{63}{80}$$

3. In the pack of 52 cards, there are 4 twos and 4 sevens, so altogether there are 8 cards favourable to the event we are considering. The probability is therefore 8/52 (or 2/13). Alternatively you could say that picking a two excludes the possibility of picking a seven, so, by the addition law, the choice is:

   $$\frac{4+4}{52} = \frac{8}{52} = \frac{2}{13}$$

4. The probability of tails is 1/2 for each coin. The throws are independent, and so the probability that all 3 will show tails is, by the multiplication law:

   $$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \text{ which is } \frac{1}{8}$$

Questions for Practice 2

1. There are two parts to this question:

   (i) $$\frac{8!}{5!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1} = 8 \times 7 \times 6 = 336$$

   (ii) $$\frac{73!}{72!} = \frac{73 \times 72 \times \text{etc. dow nto} \times 1}{72 \times 71 \times \text{etc. dow nto} \times 1} = 73$$

2. 6! which is 720.

3. There are two parts to this question:

   (i) Any auditor may be chosen to deal with the first ledger, then any one of the remaining 4 may be chosen for the second ledger and so on. The total number of ways is therefore 5 × 4 × 3, which is 60.

   (ii) Any 1 of the 5 auditors may deal with the first ledger, any 1 of the 5 with the second ledger and any 1 of the 5 with the third ledger. The total number of possible allocations of the jobs is therefore $5 \times 5 \times 5 = 125$.

4. There are two parts to this question:

   (i) $9 \times 8 \times 7 \times 6 = \dfrac{9!}{5!}$

   (ii) $5 \times 6 \times 7 = \dfrac{7!}{4!}$

   Notice that any product which is a run of consecutive whole numbers can always be written as the quotient of two factorials.

5. There are three parts to this question:

   (i) ${}_4P_n = \dfrac{4!}{(4-n)!}$

   (ii) ${}_nP_4 = \dfrac{n!}{(n-4)!}$

   (iii) ${}_{2n}P_n = \dfrac{(2n)!}{(2n-n)!} = \dfrac{(2n)!}{n!}$

6. Because dots all look alike and commas all look alike, some of the 9! possible arrangements of the 9 symbols will not be distinguishable. The number of distinguishable arrangements is:

$$\frac{9!}{3!\,6!} = \frac{9 \times 8 \times 7}{3 \times 2 \times 1} = 84$$

Questions for Practice 3

1. If the youngest and the oldest must both be included, then we are left with the problem of choosing 3 out of 7 which is:

$${}_7C_3 = \frac{7!}{3!\,4!} = \frac{7 \times 6 \times 5}{3 \times 2} = 35$$

2. If the youngest is excluded, we have the problem of choosing 5 from 8:

$${}_8C_5 = \frac{8!}{5!\,3!} = \frac{8 \times 7 \times 6}{3 \times 2} = 56$$

3. If there are 17 defective items, then there must be 33 good items. The number of possible samples of 5 items (all good) is therefore:

$${}_{33}C_5 = \frac{33!}{5!\,28!} = \frac{33 \times 32 \times 31 \times 30 \times 29}{5 \times 4 \times 3 \times 2 \times 1} = 237{,}336$$

4. There are 17 possible ways in which the 1 defective item may be chosen. For the remainder of the sample there are ${}_{33}C_4$ ways of choosing the 4 good items. The total number of possible samples is therefore:

$$17 \times {}_{33}C_4$$

   which is 695,640.

Questions for Practice 4

1. The probability that the first bead is red is 13/20. If the first bead is red then the probability that the second bead is red is 12/19. The probability that *both* beads will be red is therefore:

$$\frac{13}{20} \times \frac{12}{19} = \frac{156}{380} = \frac{39}{95}$$

2. The probability that the first bead will be red is 13/20. If it is red, then the probability that the second bead will be *white* is 7/19. The total probability is then:

$$\frac{13}{20} \times \frac{7}{19} = \frac{91}{380}$$

3. The probability that the first one is not defective is 80/100. If the first one is good, then the probability that the second one will be good is 79/99. Continue the reasoning and you get the answer:

$$\frac{80}{100} \times \frac{79}{99} \times \frac{78}{98} \times \frac{77}{97} \times \frac{76}{96}$$

4. The total possible number of samples of 5 is ${}_{100}C_5$. Out of these possible samples, many will be all non-defective. The possible number of all non-defective samples is ${}_{80}C_5$. So the probability of an all-good sample is:

$$\frac{{}_{80}C_5}{{}_{100}C_5} = \frac{80!}{5!\,75!} \div \frac{100!}{5!\,95!} = \frac{80 \times 79 \times 78 \times 77 \times 76}{100 \times 99 \times 98 \times 97 \times 96}$$

Questions for Practice 5

Sample space is (1, 2, 3, 4, 5, 6).

Event A	**P(A)**
1	$\frac{3}{6} = \frac{1}{2}$
2	$\frac{3}{6} = \frac{1}{2}$
3	$\frac{1}{6}$

Questions for Practice 6

1.

Male G 10	Male non-G 6
Female G 2	Female non-G

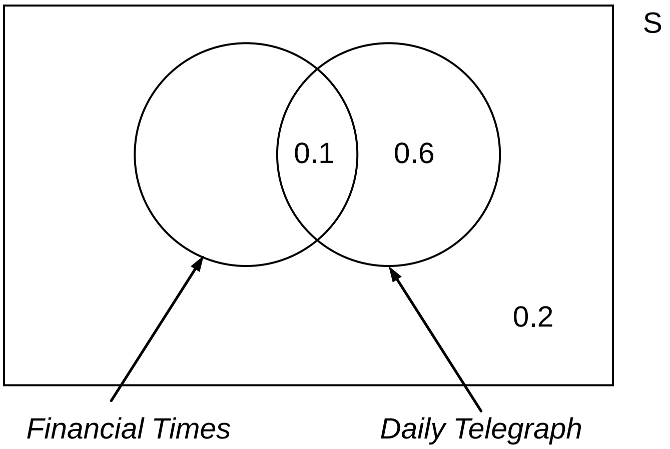
S(25) Fill in the number of male graduates (10) first.

Then fill in the number of male non-graduates (16 − 10).

Next fill in the number of female graduates (12 − 10).

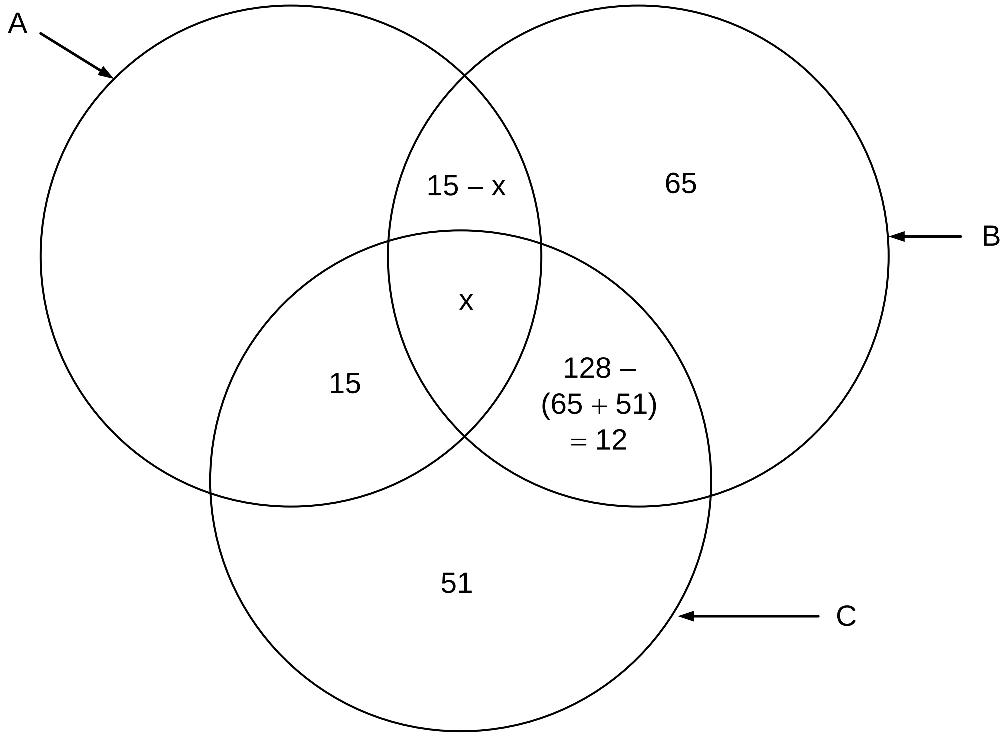
There are 25 trainees altogether, so the number of female non-graduates is 25 − (10 + 6 + 2) = 7.

2. First mark in 0.6 for those reading the *Telegraph* but not the *Financial Times*. Then fill in the overlap for those who read both (0.7 − 0.6). 0.2 of the managers are outside both circles. All the probabilities must add up to 1, so the probability that she reads only the *Financial Times* is 1 − (0.1 + 0.6 + 0.2) = 0.1 (see figure following).



3. Let x be the number voting for all three schemes. We then fill in the numbers in each section in terms of x and use the fact that the sum of the numbers in all the sections must be 200. The figure below contains these values.

Note: we are told that 15 would vote for both A and B and we must assume that this 15 includes those who might vote for C as well, so there are only 15 − x who vote for A and B but not C.



Therefore, 37 + x + 15 – x + 15 + x + 65 + 12 + 51 = 200,

i.e. 195 + x = 200

so x = 5

(i) 5 would vote for all three schemes.

(ii) 42 + 65 + 51 = 158 would vote for only one scheme.

(iii) 42 + 15 – x + x + 15 = 72 would vote for A.

(iv) 42 would vote for A only.

(v) 10 would vote for A and B but not C.

Chapter 11

Binomial and Poisson Distributions

Contents ***Page***

Introduction **174**

A. The Binomial Distribution **175**
Die-Throwing Experiments 175
General Binomial Experiment 179

B. Applications of the Binomial Distribution **182**

C. Mean and Standard Deviation of the Binomial Distribution **184**

D. The Poisson Distribution **184**
Introduction 184
The Exponential Function 185
Formula for Poisson Distribution 185

E. Application of the Poisson Distribution **186**

F. Poisson Approximation to a Binomial Distribution **188**

G. Application of Binomial and Poisson Distributions – Control Charts **191**

Answers to Questions for Practice **195**

Appendix: The Binomial Expansion **199**

INTRODUCTION

First, we should remind ourselves of the distinction between discrete and continuous variables. Discrete variables can only take on a countable number of values (usually whole numbers), while continuous variables can theoretically take on any value on the number line (and include variables such as height, weight and time values).

Next, we need to define a *probability distribution*. A probability distribution is like a frequency distribution, but instead of showing the frequencies, it shows the probabilities associated with all the possible values of a variable:

> *A* ***probability distribution*** *is a graph, table or formula that specifies the probabilities associated with all the possible values of a variable. If the variable in question is a discrete variable, the probability distribution is called a* ***discrete probability distribution****. If the variable in question is a continuous variable, the probability distribution is called a* ***continuous probability distribution****.*

As an example of a discrete probability distribution, consider the simple experiment of throwing a single die and observing which number appears. There are only six possible outcomes of this experiment, each with the same probability (1/6). The probability distribution may be written as in Table 1.

Table 11.1: Throwing a die

Possible outcomes	Probabilities
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

This is a simple rectangular probability distribution. Note that the probabilities add up to 1, which they must do as we have included every possible outcome of the experiment in the table, and it is a certainty that one of these outcomes will occur.

As a second example of a discrete probability distribution, consider the simple experiment of throwing two separate dice and observing the score. Now there are eleven possible outcomes of this experiment. The probability distribution may be written as in Table 11.2.

In looking at this table, you should confirm the probabilities presented in the second column and make sure you understand why 7 is the most likely score. This is a symmetrical probability distribution. Note again that the probabilities sum to 1, since scores of 2 to 12 are the only possible scores when two dice are thrown.

Two of the most important discrete probability distributions, the *binomial distribution* and the *Poisson distribution*, are discussed in this chapter. The most important continuous probability distribution, the *normal distribution*, is considered in Chapter 12.

Table 11.2: Throwing two dice

Possible outcomes	Probabilities
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

A. THE BINOMIAL DISTRIBUTION

The binomial probability distribution is applicable in situations where an experiment, consisting of a certain number of trials satisfies *all of four conditions*. These are:

1. The number of trials must be fixed and finite. This number is usually denoted by n.
2. Every trial must result in one or other of only two mutually exclusive possible outcomes, which, for convenience, we usually label "success" or "failure". Examples are:

   (a) When we roll a die we get a six or we do not get a six.

   (b) A product is either defective or not defective.

   (c) A tossed coin comes down heads or tails.

   (d) A child is either a boy or a girl.

   We must, of course, define which event is the success before the term is used.
3. The probability of a success or failure at each trial must remain constant throughout the experiment. The probability of a success is usually denoted by p, and that of a failure by q.
4. The outcome of every trial must be independent of the outcome of every other trial. For example, if a coin is unbiased, then the probability of obtaining a head on the tenth time it is tossed remains 1/2, even if the previous nine tosses have all resulted in heads.

Provided these four conditions all hold, the binomial distribution enables us to work out the probability of any given number of successes or failures in a specified number of trials.

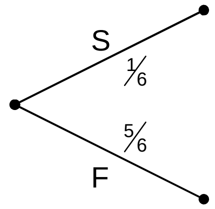
Die-Throwing Experiments

We shall work out from first principles the probabilities for the situation where we throw a fair die and observe whether we obtain a 6 or not. We shall label obtaining a 6 as success and not obtaining a 6 as failure.

We know that for a fair die, $\text{P(success)} = \frac{1}{6}, \text{P(failure)} = \frac{5}{6}$

and we can draw a simple tree diagram (Figure 11.1), where S denotes a success and F denotes a failure:

Figure 11.1: Die-throwing experiments – first throw

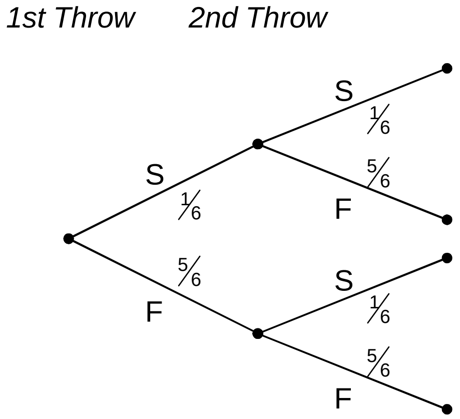


Thus, for one throw of the die we can write:

$$\text{P(0 successes)} = \frac{5}{6}, \ \text{P(1 success)} = \frac{1}{6}$$

Let us now throw the die again. The probability of a success or failure at this second throw remains unchanged at 1/6 or 5/6 respectively. We can extend our tree diagram as follows (Figure 11.2):

Figure 11.2: Die-throwing experiments – second throw



At the end of two throws, the different possible outcomes are as in Table 11.3.

Table 11.3: Outcomes at the end of two throws

1st throw	2nd throw	Number of successes
6	6	2
6	Not a 6	1
Not a 6	6	1
Not a 6	Not a 6	0

We can put our results in Table 11.4.

Table 11.4: Results of die-throwing after two throws

Event	Probability
0 successes	$\left[\frac{5}{6}\right]^2$
1 success	$\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]+\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]$
	$=2\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]$
2 successes	$\left[\frac{1}{6}\right]^2$

We then throw the die once more. The different possible outcomes after 3 throws are best found by looking at the tree, extended one stage more (see Figure 11.4 and Table 11.5):

Figure 11.3: Die-throwing experiments – third throw

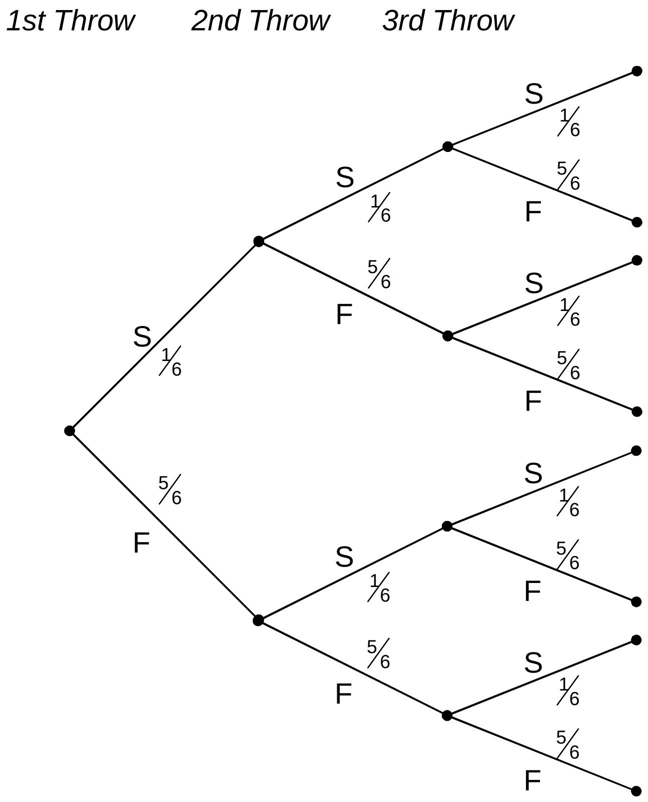


Table 11.5: Results of die-throwing after three throws

Outcomes	Probability
SSS	$\left[\frac{1}{6}\right]^3$
SSF	$\left[\frac{1}{6}\right]^2\left[\frac{5}{6}\right]$
SFS	$\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]$
SFF	$\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]^2$
FSS	$\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]^2$
FSF	$\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]$
FFS	$\left[\frac{5}{6}\right]^2\left[\frac{1}{6}\right]$
FFF	$\left[\frac{5}{6}\right]^3$

and we can condense the results in a second table:

Table 11.6: Condensed results from Table 11.5

Event	Probability
0 successes	$\left[\frac{5}{6}\right]^3$
1 success	$3\left[\frac{5}{6}\right]^2\left[\frac{1}{6}\right]$
2 successes	$3\left[\frac{5}{6}\right]\left[\frac{1}{6}\right]^2$
3 successes	$\left[\frac{1}{6}\right]^3$

Activity 1

Now try to extend the tree diagram one stage further corresponding to 4 throws of the die. List all the possible outcomes in a table with their probabilities and then construct a second table giving the probabilities of 0, 1, 2, 3 and 4 successes.

General Binomial Experiment

In a general experiment where each trial has just two mutually exclusive outcomes, we can denote:

P(success) = p and P(failure) = q

We know that p + q = 1, as one or other of these outcomes is *certain to happen* (see Figure 11.5 and Table 11.7). Hence,

Figure 11.4: General binomial experiment – first trial

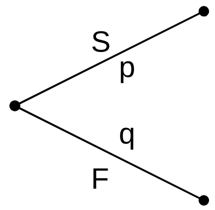


Table 11.7: General binomial experiment – outcome of first trial

Event	Probability
0 successes	q
1 success	p

The trial is performed again (see Figure 11.5, Tables 11.8 and 11.9):

Figure 11.5: General Binomial Experiment – second trial

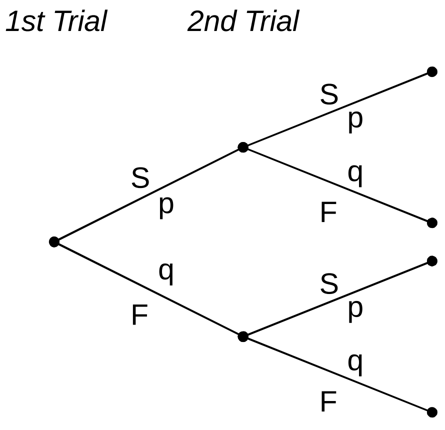


Table 11.8: Outcomes of second trial

Outcomes	Probability
SS	p^2
SF	pq
FS	qp
FF	q^2

Table 11.9: Results of second trial

Event	Probability
0 successes	q^2
1 success	2qp
2 successes	p^2

and for a third time (see Figure 11.6, Tables 11.10 and 11.11):

Figure 11.6: General Binomial Experiment – third trial

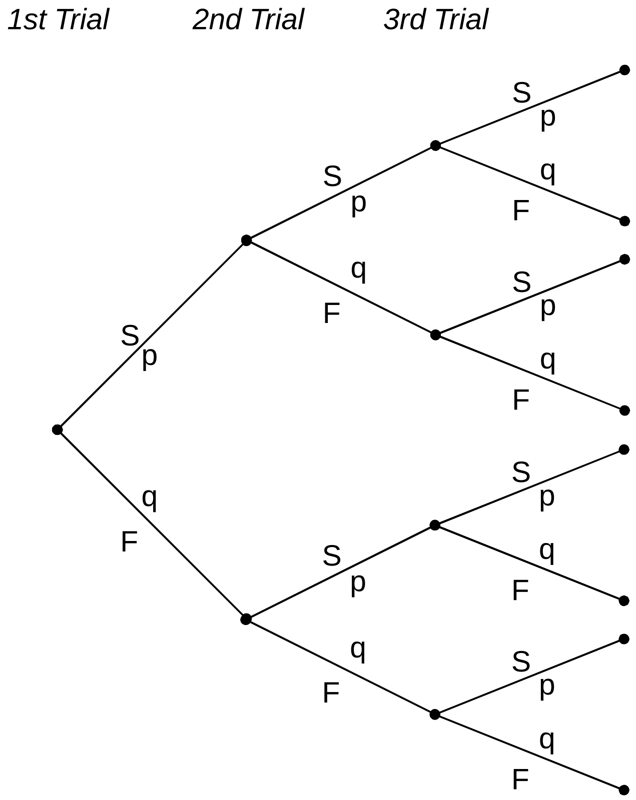


Table 11.10: Outcomes of third trial

Outcomes	Probability
SSS	p^3
SSF	p^2q
SFS	pqp
SFF	pq^2
FSS	qp^2
FSF	qpq
FFS	q^2p
FFF	q^3

Table 11.11: Results of third trial

Event	Probability
0 successes	q^3
1 success	$3q^2p$
2 successes	$3qp^2$
3 successes	p^3

Activity 2

Now look back to the first section of this chapter and write down the binomial expansions of:

$(p + q)^2$ and $(p + q)^3$

You will see that the terms in the expansions are exactly the same as the probabilities we have worked out (shown in Table 11.9 and Table 11.11).

See if you can write down, without using a tree diagram, the probabilities of 0, 1, 2, 3, 4 successes, when we perform this trial 4 times.

Although we can use tree diagrams to work out probabilities when we repeat our trial up to 4 or 5 times, once we have larger numbers of repetitions of the trial, the tree diagrams become unwieldy. It is then much better to use what we have discovered above. That is, if we perform our trial n times, the probabilities of 0, 1, 2, 3 .... up to n successes are given by successive terms in the binomial expansion of $(p + q)^n$, where p is the probability of success and q the probability of failure at any one trial. Note that p and q must remain constant from trial to trial.

From our formula for the binomial expansion, the general term in the expansion of $(p + q)^n$ is ${}_nC_x p^n q^{n-x}$: this gives us the probability of exactly x successes in n trials of the experiment. So we have:

$$P(x) = {}_nC_x p^n q^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n$$

where P(x) represents the probability of x successes in n trials and $q = 1 - p$.

This is the general formula for the binomial probability distribution. We shall see how to apply this formula in the next section, and you will find that it is not quite as fearsome as it may look at first.

B. APPLICATIONS OF THE BINOMIAL DISTRIBUTION

Example 1:

The probability that a match will break on being struck is 0.04. What is the probability that, out of a box of 50:

(a) none will break;

(b) more than 2 will break?

A match will either break or not break when it is struck. Therefore:

$$\text{P(breaking)} = 0.04 = \text{P(success)} = p$$

$$\text{P(not breaking)} = 1 - p = 1 - 0.04 = 0.96 = \text{P(failure)}$$

We have a box of 50 matches, so n = 50.

(a) We require the probability that none will break, i.e. the probability of no successes, P(0).

$\text{P}(0) = {}_nC_0 p^0(1-p)^{n-0}$ using general formula:

$$= \frac{n!}{n!\,0!}(1-p)^n = (1-p)^n$$

$= (0.96)^{50} = 0.1299$ to 4 decimal places.

Therefore, probability none will break = 0.1299

(b) Probability that more than 2 will break = 1 – probability that 2 or less will break.

Probability that 2 or less will break = probability that 0 or 1 or 2 will break

$$= \text{P}(0) + \text{P}(1) + \text{P}(2)$$

We thus need to work out P(1) and P(2):

$$\text{P}(1) = {}_nC_1 p^1(1-p)^{n-1}$$

$$= \frac{n!}{(n-1)!}p(1-p)^{n-1} = \frac{50!}{49!}(0.04)(0.96)^{49}$$

$= 50 \times (0.04)(0.96)^{49} = 0.2706$ to 4 decimal places.

$$\text{P}(2) = {}_nC_2 p^2(1-p)^{n-2}$$

$$= \frac{n!}{(n-2)!2!}p^2(1-p)^{n\text{-}2} = \frac{50!}{48!\,2!}(0.04)^2(0.96)^{48}$$

$$= \frac{50 \times 49}{2}(0.04)^2(0.96)^{48} = 0.2762$$

Therefore, probability that more than 2 will break:

$$= 1 - (0.1299 + 0.2706 + 0.2762)$$

$$= 1 - 0.6767 = 0.3233$$

$= 0.323$ to 3 decimal places.

Example 2:

It has been found that, on average, 5% of the eggs supplied to a supermarket are cracked. If you buy a box of 6 eggs, what is the probability that it contains 2 or more cracked eggs?

An egg is either cracked or not cracked:

$$\text{P(cracked)} = 5\% = 0.05 = \text{P(success)} = p$$

$$\text{P(not cracked)} = 1 - p = 1 - 0.05 = 0.95 = \text{(failure)}$$

We have a box of 6 eggs, so n = 6.

Probability of 2 or more cracked eggs in a box:

= 1 – probability of less than 2 cracked eggs in a box

= 1 – probability of 0 or 1 cracked eggs in a box

= 1 – [P(0) + P(1)].

$$P(0) = {}_nC_0 p^0(1-p)^{n-0} = (1-p)^n = (0.95)^6 = 0.7351$$

$$P(1) = {}_nC_1 p^1(1-p)^{n-1} = \frac{n!}{(n-1)!\,1!}p(1-p)^{n-1}$$

$$= \frac{6!}{5!}(0.05)(0.95)^5 = 6(0.05)(0.95)^5 = 0.2321$$

Therefore, probability of 2 or more cracked eggs in a box:

$$= 1 - (0.7351 + 0.2321)$$

$$= 1 - 0.9672$$

= 0.0328 = 0.033 to 3 decimal places.

Example 3:

A retail sales manager will accept delivery of a large consignment of goods if a random sample of 10 items contains no defectives. If 3% of the producer's total output is defective, what is the probability that delivery of a consignment will be accepted? How would the situation change if the random sample were of only 5 items?

An item is either defective or non-defective. Therefore:

$$\text{P(defective)} = 3\% = 0.03 = \text{P(success)} = p$$

$$\text{P(non-defective)} = 1 - p = 1 - 0.03 = 0.97 = \text{P(failure)}.$$

First, the manager takes a sample of 10, so n = 10.

We require the probability that this sample contains no defectives, i.e. P(0):

$$P(0) = {}_nC_0 p^0(1-p)^{n-0} = (1-p)^n$$

$$= (0.97)^{10}$$

= 0.7374 to 4 decimal places.

Therefore, probability that a delivery will be accepted is 0.7374

Secondly, consider a sample of 5.

$P(0) = (1-p)^n = (0.97)^5 = 0.8587$ to 4 decimal places.

Therefore, probability that delivery will be accepted is 0.8587, which is higher than when a larger sample was taken.

Notes

1. One of the conditions for using the binomial distribution is that the chance of success, p, must be *constant* throughout the series of trials. This means that if we are, say, taking items from a batch and *not* replacing them before the next item is taken, then the binomial distribution does not apply because the batch is fractionally smaller (by 1 item each time). In practice, however, when the batch from which a sample is being taken is very large compared with the sample, the binomial distribution is a satisfactory approximation. As a rough guide, you can consider the batch to be very large if it is more than about 10 times the sample size.

2. Tables are available giving values of $_nC_x$ for various values of n and x. This is particularly useful for large values of n but in examinations usually you are expected to be able to work them out for yourself. Most calculators have keys for calculating the number of combinations and the number of permutations.

C. MEAN AND STANDARD DEVIATION OF THE BINOMIAL DISTRIBUTION

In the case of a theoretical distribution like the binomial distribution there is usually a simple formula to show what the mean and standard deviation ought to be without needing to go through any lengthy calculations.

For any binomial distribution, the mean μ and the standard deviation σ are given by the following formulae:

$$\mu = np$$

$$\sigma = \sqrt{npq} = \sqrt{np(1-p)}$$

Thus, if $n = 3,\ p = \frac{1}{6}, 1 - p = \frac{5}{6}$

therefore:

$$\mu = 3 \times \frac{1}{6} = \frac{1}{2} = 0.5$$

$$\sigma = \sqrt{3 \times \frac{1}{6} \times \frac{5}{6}} = \sqrt{\frac{15}{36}} = 0.645 \text{ to 3 significant figures.}$$

What we are saying is that if we repeat this experiment very many times, the mean number of successes is 0.5. As with a frequency distribution, the mean does not necessarily have to be one of the original values of x.

D. THE POISSON DISTRIBUTION

Introduction

The binomial distribution is useful in cases where we take a fixed sample size, and count the number of successes. Sometimes we do not have a definite sample size, and then the binomial distribution is of no use. For example, if we are studying the occurrence of accidents, we can count how many accidents occur in a month, but it is nonsense to talk about how many accidents did not occur! In such cases we use another theoretical distribution called the *Poisson distribution* (named after the French mathematician).

The Exponential Function

Before we go on to study the Poisson distribution, there is some new mathematics to learn. In mathematics there are a few rather special numbers that are given particular symbols. One of these you probably know very well already; the number π (pi, pronounced "pie") is used to calculate areas of circles. You may also know that π cannot be given an *exact* arithmetical value, but it is very nearly 3.1416.

Another of these special numbers is the exponential number e. Like π, it cannot be given an exact arithmetical value (although mathematicians can calculate it to as many decimal places as they wish). To three decimal places, the value of e is 2.718, which is accurate enough for almost all practical purposes.

Formula for Poisson Distribution

In the example of accidents mentioned above, we could count the number of accidents each month and work out the *mean number* of accidents per month. So, although we do not know the values of n or p, we *do* know the value of the mean np.

Mathematicians can prove that, if you know the value of this mean (let's call it μ) then the theoretical probability that there will be x events (or "successes", if you prefer to keep the same word as before) is:

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

This is the general term of the Poisson distribution.

(Note that sometimes λ (lambda) is used in place of μ.)

Earlier we stated the conditions which must prevail in order that a binomial distribution of events can occur. The Poisson distribution is what is known as a limiting case of the binomial distribution. It is the result of letting n become very large, and p become very small (i.e. a very large number of trials are conducted but the probability of a success in any one particular trial is exceedingly small), but in such a way that the mean number of successes, np, is of a moderate size and constant. What exactly constitutes "moderate size" is subjective, but certainly you need have no cause for concern working with any value less than 10.

In terms of our accident example, we could regard n, the number of trials, as being the number of one-minute intervals in a working month, during any one of which an accident could occur; p would then be the probability of an accident occurring during any particular one-minute interval, and np the mean number of intervals per month during which accidents would occur. This will be identical with the mean number of accidents per month, provided we make the assumption that there are no intervals in which more than one accident occurs (and this is a reasonable assumption).

The Poisson probability formula, as stated here, is obtained from the corresponding binomial formula by letting n tend towards infinity, p tend towards zero, and substituting $\mu = np$.

If the Poisson distribution is applicable in any given situation, this means that all the main conditions affecting the issue (in our example, the accident rate) do not alter. The month-to-month variations would then be due only to random chance causes, and they would be expected to fit a Poisson distribution. Other examples of the Poisson distribution will be given later.

If the Poisson distribution gives a good description of the practical situation, then the events are occurring randomly in time and there is no underlying reason why there should be more accidents in one particular month of the year. Any fluctuations in numbers of accidents are assignable in such cases to random variation.

The mean of a Poisson distribution, μ, is all you need to know when calculating the probabilities – unlike the binomial, where you need two things, n and p. It can be proved that the *variance* of a Poisson distribution is also equal to μ, which means that the standard deviation is the square root of μ. There is no upper limit to x but, in practice, there is a value beyond which the probability is so low that it can be ignored.

E. APPLICATION OF THE POISSON DISTRIBUTION

Applying the Poisson distribution to our accidents example, we might collect data over a year or so and find that the mean number of accidents in a workshop is 2.2 per month. We might now ask ourselves what the probability is of getting a month with no accidents, or only 1 accident, or 2 accidents, and so on. To get the answer, we work out the appropriate terms of the Poisson distribution with mean $\mu = 2.2$.

To find the value of $e^{-2.2}$ there are several options open to you:

(a) Some books of tables give you the values of e^{-x} directly. You will find $e^{-2.2} = 0.1108$.

Other books of tables give e^{-x} for $0 < x < 1$ together with e^{-x} for $x = 1, 2, 3 \ldots$. Thus, to obtain $e^{-2.2}$, we have to use properties of indices, that is:

$$e^{-2.2} = e^{-2} \times e^{-0.2} = 0.13534 \times 0.8187$$

$$= 0.1108 \text{ to 4 significant figures.}$$

(b) Most calculators have the facility to work out e^{-x} directly, and this is the preferred method in the examination.

We can now proceed to work out the Poisson probabilities. We know that:

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

for different values of x.

If you require the probability for just one or two values of x, then this expression may be evaluated fully in each case, e.g.

$$P(6) = \frac{\left(e^{-2.2}\right)(2.2)^{6}}{6!} = \frac{0.1108 \times 113.4}{720} = 0.0174$$

However, if you require a large number of terms of the probability distribution, it is not necessary to evaluate each term separately.

Using the Poisson probability formula, we have:

$$P(x-1) = \frac{e^{-\mu}\mu^{x-1}}{(x-1)!}$$

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!} = \frac{e^{-\mu}\mu^{x-1}}{(x-1)!} \times \frac{\mu}{x} = P(x-1) \times \frac{\mu}{x}$$

Therefore, the probability of x successes can be obtained by multiplying the probability of $(x-1)$ successes by the value $\frac{\mu}{x}$.

This technique is illustrated in Table 11.12, using our accident example in which $\mu = 2.2$.

Table 11.12: Some Poisson distribution values

x	P(x)	
0	$\dfrac{e^{-2.2} \times 2.2^0}{0!} = e^{-2.2}$	$= 0.1108$
1	$P(0) \times \dfrac{2.2}{1} = 0.1108 \times 2.2$	$= 0.2438$
2	$P(1) \times \dfrac{2.2}{2} = 0.2438 \times 1.1$	$= 0.2682$
3	$P(2) \times \dfrac{2.2}{3} = 0.2682 \times 0.7333$	$=.1967$
4	$P(3) \times \dfrac{2.2}{4} = 0.1967 \times 0.55$	$= 0.1081$
5	$P(4) \times \dfrac{2.2}{5} = 0.1081 \times 0.44$	$= 0.0476$
6	$P(5) \times \dfrac{2.2}{6} = 0.0476 \times 0.3667$	$= 0.0174$
7	$P(6) \times \dfrac{2.2}{7} = 0.0174 \times 0.3143$	$= 0.0055$
etc.		

Note that the value obtained for P(6) by this method is identical with that obtained by using the formula.

The values of P(x) are the probabilities that a particular month will have x accidents; or you may consider them as the relative frequencies with which months have 0, 1, 2, 3, etc. accidents. Note that since our mean (2.2) refers to accidents per month, then the probabilities also refer to months. If we had used the mean per week, then the probabilities would have referred to weeks.

F. POISSON APPROXIMATION TO A BINOMIAL DISTRIBUTION

There is another very common use for the Poisson distribution, and that is as an approximation to the binomial distribution. If n is large, the calculations for the binomial distribution are often very tedious, even with a calculator. Fortunately, if we put $\mu = np$, the Poisson terms can, in certain circumstances, be used instead of the binomial terms. The conditions for using the Poisson as an approximation to the binomial are that:

- n should be large;
- p should be small, or $q = 1 - p$ should be small.

Compare these conditions with those given earlier for the Poisson distribution.

There are no fixed rules to say what is meant by large and small, but it has been shown that the Poisson is a good approximation to the binomial if:

$n \geq 10$ and $p \leq 0.01$ *or*

$n \geq 20$ and $p \leq 0.03$ *or*

$n \geq 50$ and $p \leq 0.05$ *or*

$n \geq 100$ and $p \leq 0.08$

Table 11.13 shows some probabilities for comparison.

Table 11.13 Comparison of binomial and Poisson probabilities

x	**Binomial** *n = 100, p = 0.01 np = 1*	**Binomial** *n = 50, p = 0.02 np = 1*	**Poisson** *μ = 1*
0	0.3660	0.3642	0.3679
1	0.3697	0.3716	0.3679
2	0.1849	0.1858	0.1840
3	0.0610	0.0607	0.0613
4	0.0149	0.0145	0.0153
5	0.0029	0.0027	0.0031

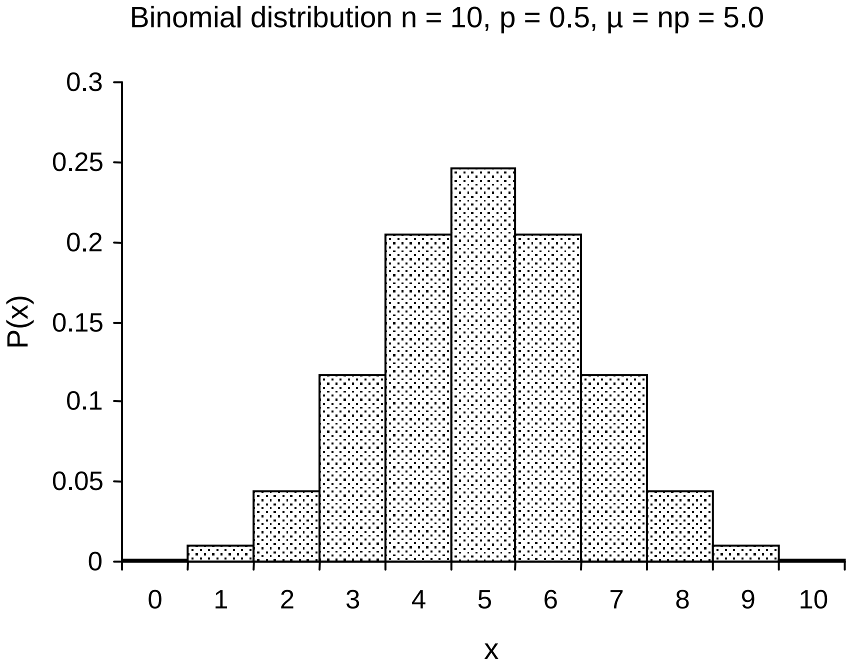
It is also interesting to compare the histograms given by the binomial and Poisson distributions (Figures 11.7 and 11.8).

Whereas in Figure 11.7, where p is not small, the probabilities are noticeably different, in Figure 11.8, where p is smaller although n remains the same, the probabilities are in much closer agreement.

Note that the binomial and Poisson distributions are both discrete, i.e. there are gaps between consecutive values which a discrete variable may take. It is only possible, for example, to have 1, 2, 3, etc. accidents, not 1.1 or 2.7, etc. However, it is essential that in a histogram there should be no gaps between adjacent blocks. This problem is overcome by assuming that each value extends halfway to meet its neighbour in both directions.

Thus in the histograms which follow, each block representing the value 2 has been drawn to cover all values from 1.5 to 2.5, and similarly for other blocks, thus eliminating any gaps which would otherwise be present.

Figure 11.7: Comparative histograms



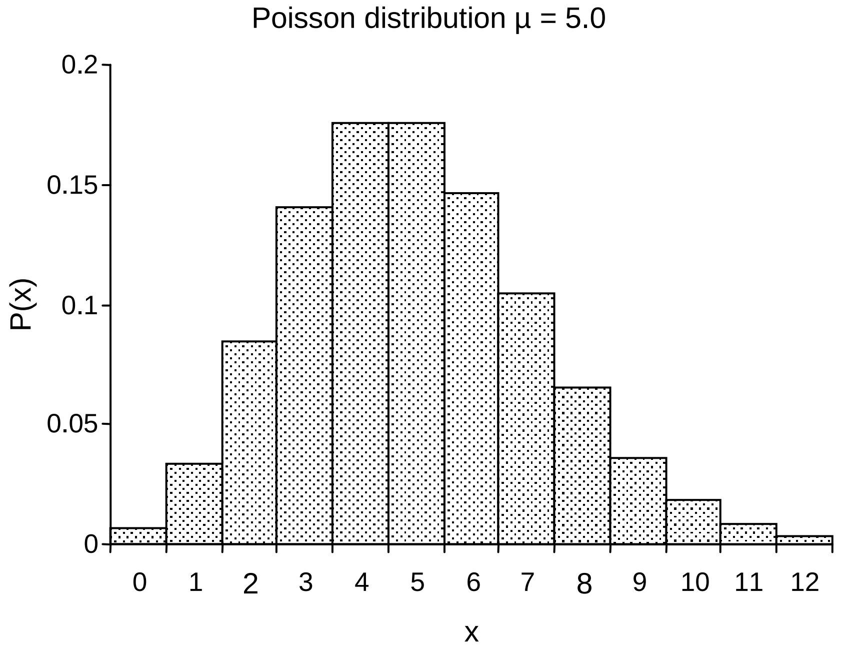
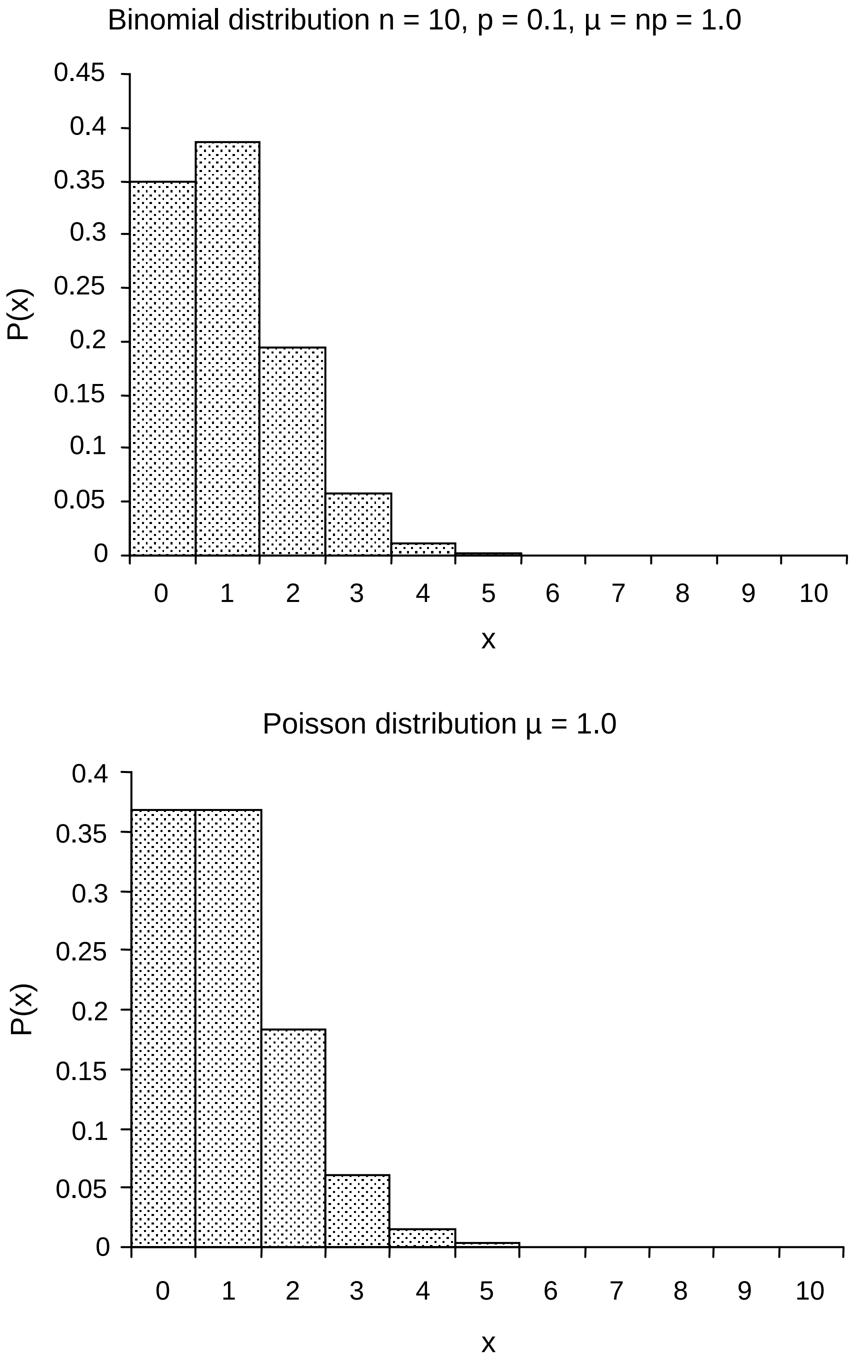


Figure 11.8: Comparative histograms



Binomial distribution n = 10, p = 0.1, μ = np = 1.0

Poisson distribution μ = 1.0

G. APPLICATION OF BINOMIAL AND POISSON DISTRIBUTIONS – CONTROL CHARTS

One of the most important applications of the binomial and Poisson distributions is in *quality control*. You will probably have heard of quality control sampling schemes, where a sample of a product is taken at regular intervals and checked for defectives. Suppose samples of 40 pieces are taken regularly from a production process. Past experience has shown that, in samples of 40, a defective piece can be expected in every other sample. Now let's see how the binomial and Poisson distributions can be used in order to develop a quality control chart.

The average number of defectives in a sample of 40 pieces is 0.5 (one defective unit in every other batch). The probability of finding 0, 1, 2, 3 ... defectives in a given sample could be accurately determined using the binomial expansion but, as the probability of finding a defective is relatively small and the sample of reasonable size, the Poisson distribution can be used as an approximation to the binomial. The individual probabilities for a Poisson distribution with mean 0.5 are as follows:

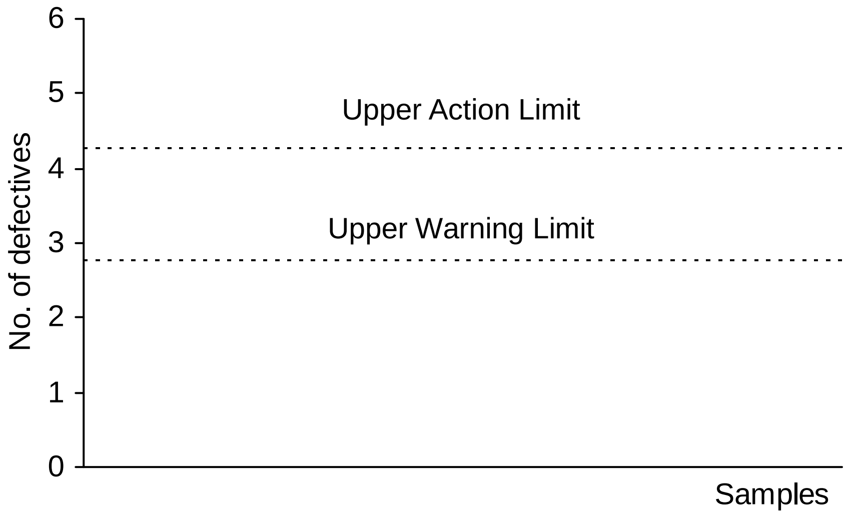
Table: 11.14: Poisson distribution with mean of 0.5

Number of defectives	Probability
0	0.6065
1	0.3033
2	0.0758
3	0.0126
4	0.0016
5	0.0002
	1.0000

(You may wish to practise use of the Poisson distribution by evaluating the probability formula $\frac{e^{-\mu}\mu^x}{x!}$ for $x = 0, 1, 2 \ldots$.)

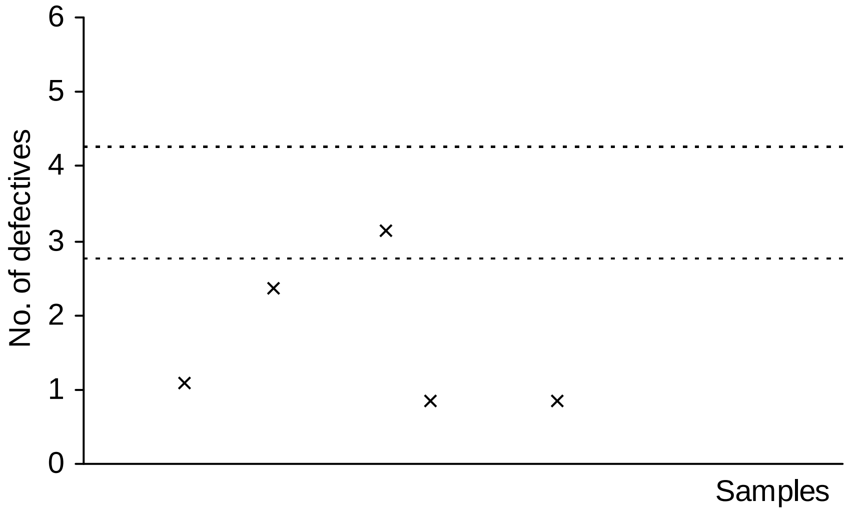
You can see that the chance of a sample containing three or more defectives is about 1.4%, whilst the chance of a sample containing four or more defectives is only 0.2% (about twice in every thousand sample checks). Typically, a control chart would be set up identifying an *upper warning limit*, where the probability of a result arising by chance is 2.5%, and an *upper action limit*, where the probability of a result arising by chance is 0.1%. The control chart for our example would therefore look like Figure 11.9:

Figure 11.9: Control chart upper warning and action limits



When samples are taken, the number of defectives is determined and points are entered on the control chart, as shown in Figure 11.10:

Figure 11.10: Entering sample values on a control chart



One point is above the upper warning limit but, unless more points fall above this line, no action would be taken – one point in every 40 would be expected to fall above the warning limit simply by chance.

Not only can upper action and warning limits be drawn on the chart, but also *lower* limits. At first this may seem strange, in that quality can hardly be too good. However, there are two reasons for identifying results which are better than expected. Firstly, it might be possible for someone to falsify results by making sure that samples chosen all contain good pieces. Secondly, if the process has actually been improved it is obviously important to determine exactly how it has happened, so that the change can be sustained.

Questions for Practice

1. A fair die with 6 sides is thrown 3 times. Show by means of a tree diagram that the probability of obtaining 0, 1, 2 or 3 sixes from the 3 throws is given by the binomial probability function:

$$_3C_x\left[\frac{1}{6}\right]^x\left[\frac{5}{6}\right]^{3-x}$$

where x represents the number of successes.

2. A department produces a standard product. It is known that 60% of defective products can be satisfactorily reworked. What is the probability that in a batch of 5 such defective products, at least 4 can be satisfactorily reworked?

3. What is the probability of passing a batch of 5 units from a machine which averages 20% defectives, when only 2 of the 5 are tested?

4. Calculate the probability that, for 6 telephone lines:

(a) at least 1 of the lines is engaged and

(b) all 6 lines are engaged,

when the probability of 1 line being engaged is 1/4.

5. In a family with 10 children, if the probability of having a male child is the same as that of having a female child, what is the probability that:

(a) 6 of the children will be boys

(b) none will be a girl

(c) at most, 2 will be boys?

6. *This tests your understanding of probability theory and the binomial distribution.*

The World Life Assurance Company Limited uses recent mortality data in the calculation of its premiums. The following table shows, per thousand of population, the number of persons expected to survive to a given age:

Age	0	10	20	30	40	50	60	70	80	90	100
Number surviving to given age	1,000	981	966	944	912	880	748	525	261	45	0

Required:

(a) Use the table to determine the probability that:

(i) A randomly chosen person born now will die before reaching the age of 60.

(ii) A randomly chosen person who is aged 30 now will die before reaching the age of 60.

(iii) A randomly chosen person who is aged 50 now will die before reaching the age of 60.

Comment on the order of magnitude of your three answers.

(b) The company is planning to introduce a life insurance policy for persons aged 50. This policy requires a single payment paid by the insured at the age of 50, so that if the insured dies within the following ten years the dependant will receive a payment of £10,000. However, if this person survives, then the company will not make any payment.

Ignoring interest on premiums and any administrative costs, calculate the single premium that the company should charge to break-even in the long run.

(c) If 12 people each take out a policy as described in (b) and the company charges each person a single premium of £2,000, find the probability that the company will make a loss.

(d) The above table was based on the ages of all people who died in 1986. Comment on the appropriateness to the company when calculating the premiums it should charge.

(e) The above table can be expanded to include survival numbers, per thousand of population, of other ages:

Age	50	52	54	56	58	60
Number surviving to given age	880	866	846	822	788	748

(i) Given that a person aged 50 now dies before the age of 60, use this new information to estimate the expected age of death.

(ii) Calculate a revised value for the single premium as described in part (b), taking into account the following additional information:

- The expected age of death before 60 as estimated at (i).
- A constant interest rate of 8% p.a. on premiums.
- An administration cost of £100 per policy.
- A cost of £200 to cover profit and commissions.

Now check your answers with those given at the end of the chapter.

ANSWERS TO QUESTIONS FOR PRACTICE

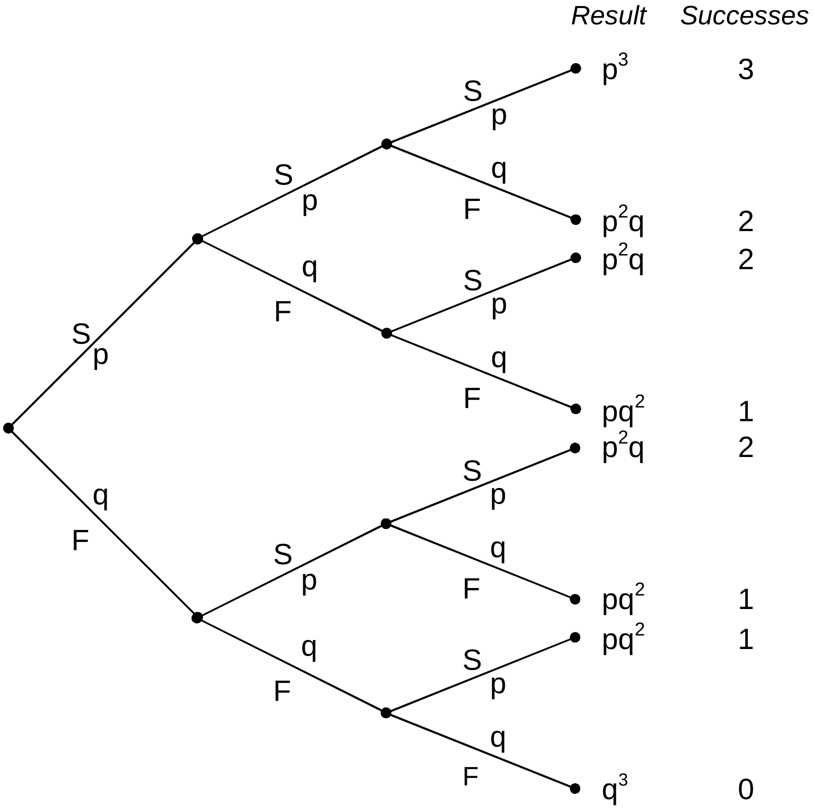
1. Let S represent a success, i.e. throwing a 6, and F represent a failure, i.e. not throwing a 6.

   Let P(S) $= p = \frac{1}{6}$

   P(F) $= q = (1-p)\frac{5}{6}$

   The tree diagram is as shown in Figure 11.11:

   Figure 11.11: Tree diagram



   Summary of results:

   P(3 successes) $= 1 \times p^3 = \left[\frac{1}{6}\right]^3$

   P(2 successes) $= 3 \times p^2q = 3\left[\frac{1}{6}\right]^2\left[\frac{5}{6}\right]$

   P(1 success) $= 3 \times pq^2 = 3\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]^2$

   P(0 successes) $= 1 \times q^3 = \left[\frac{5}{6}\right]^3$

Combining these results gives:

$$p^3 + 3p^2q + 3pq^2 + q^3 = (p + q)^3 \text{ or}$$

$$\left[\frac{1}{6}\right]^3 + 3\left[\frac{1}{6}\right]^2\left[\frac{5}{6}\right] + 3\left[\frac{1}{6}\right]\left[\frac{5}{6}\right]^2 + \left[\frac{5}{6}\right]^3 = \left[\frac{1}{6} + \frac{5}{6}\right]^3$$

which equals the binomial probability function

$$_3C_x \left[\frac{1}{6}\right]^x \left[\frac{5}{6}\right]^{3-x}$$

2. Let p be the probability that an item can be satisfactorily reworked. Then:

$p = 0.6$, $q = (1 - p) = 0.4$, and $n = 5$

P(at least 4 can be reworked)

= (4 or 5 can be reworked)

= P(4) + P(5)

$= {_5C_4}(0.6)^4(0.4) + (0.6)^5$

$= 5(0.6)^4(0.4) + (0.6)^5$

$= 0.2592 + 0.07776 = 0.33696.$

3. Let p be the probability that an item is not defective. Then:

$p = 0.8$, $q = (1 - p) = 0.2$, and $n = 2$

P(Passing batch) = P(2 items tested are both good)

$= P(2) = (0.8)^2 = 0.64$

4. Let p be the probability that a line is engaged. Then:

$p = 0.25$, $q = (1 - p) = 0.75$, and $n = 6$

(a) P(at least 1 line is engaged)

= 1 – P(0 lines are engaged)

$= 1 - P(0) = 1 - (0.75)^6$

= 1 – 0.1780 = 0.8220 to 4 significant figures.

(b) P(all 6 lines are engaged)

$= P(6) = (0.25)^6$

= 0.0002441 to 4 significant figures.

5. Let p be the probability of having a boy. Then:

   p = ½, q = (1 – p) = ½, and n = 10

   (a) P(6 of the children will be boys)

   $$= P(6) = {}_{10}C_6\left[\frac{1}{2}\right]^6\left[\frac{1}{2}\right]^4$$

   $$= \frac{10!}{6!\,4!}\left[\frac{1}{2}\right]^{10} = 210\left[\frac{1}{2}\right]^{10}$$

   = 0.2051 to 4 significant figures.

   (b) P(none will be a girl) = P(10 will be boys) = P(10)

   $$=\left[\frac{1}{2}\right]^{10} = 0.0009766 \text{ to 4 significant figures.}$$

   (c) P(at most 2 will be boys) = P(0 or 1 or 2 will be boys)

   = P(0) + P(1) + P(2)

   $$=\left[\frac{1}{2}\right]^{10} + {}_{10}C_1\left[\frac{1}{2}\right]\left[\frac{1}{2}\right]^9 + {}_{10}C_2\left[\frac{1}{2}\right]^2\left[\frac{1}{2}\right]^8$$

   $$=\left[\frac{1}{2}\right]^{10} + 10\left[\frac{1}{2}\right]^{10} + 45\left[\frac{1}{2}\right]^{10}$$

   $$= 56\left[\frac{1}{2}\right]^{10} = 0.05469 \text{ to 4 significant figures.}$$

6. There are five parts to this question.

   (a) (i) P(die before 60) $= \frac{1{,}000 - 748}{1{,}000} = 0.252$

   (ii) P(die before 60, given already 30) $= \frac{944 - 748}{944} = 0.208$

   (iii) P(die before 60, given already 50) $= \frac{880 - 748}{880} = 0.150$

   (b) Expected value of payout = £10,000 × 0.15 = £1,500

   Premium to break-even = £1,500

(c) A loss will be made if payouts exceed premiums.

Premiums = 12 × £2,000 = £24,000.

A loss is made if three or more die. Using binomial probability distribution:

$$P(x) = {}_nC_x p^x(1-p)^{n-x}$$

$$P(0) = (0.85)^{12} = 0.1422$$

$$P(1) = \frac{12!}{11!\,1!}(0.15)(0.85)^{11} = 0.3012$$

$$P(2) = \frac{12!}{10!\,2!}(0.15)(0.85)^{10} = 0.2924$$

Probability 3 or more die = 1 – P(0) – P(1) – P(2)

= 1 – 0.1422 – 0.3012 – 0.2924 = 0.264

(d) The data relates to people born in the decades prior to 1986. There may be a trend in mortality rates such that people taking out policies now have different life expectancies from those who died in 1986. The statistics therefore need to be used with care.

(e) (i) Assume that those who die by the age of 52, die on their 51st birthday; those who die by age 54 die on their 53rd birthday, and so on.

	Age	×	**Probability of Death**
	51	×	$\frac{14}{132} = 5.409$
	53	×	$\frac{20}{132} = 8.030$
	55	×	$\frac{24}{132} = 10.000$
	57	×	$\frac{34}{132} = 14.682$
	59	×	$\frac{40}{132} = 17.879$
Expected age of death (for someone who dies between ages 50 and 60)			= 56.000

(ii) Average discounted payout: $= \frac{10{,}000}{1.08^6} = £6{,}302.$

Premium for break-even = £100 + £200 + (0.15 × £6,302) = £1,245.30.

APPENDIX: THE BINOMIAL EXPANSION

Some types of situation occur repeatedly in everyday life, and there are various probability distributions that can be used to give us the probabilities associated with *all* the different possible outcomes under particular conditions. Here we shall consider the binomial distribution, but before we can continue with the probability aspect we need to remind ourselves of some algebra.

By multiplication, we can show that:

$$(a + b)^1 = a + b$$

$$(a + b)^2 = a^2 + 2ab + b^2$$

$$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

$$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$

All these products will be seen to fall under the general formula:

$$(a + b)^n = a^n + {}_nC_1a^{n-1}b + {}_nC_2a^{n-2}b^2 + {}_nC_3a^{n-3}b^{3+} \; ... \; + b^n$$

$$= a^n + na^{n-1}b + \frac{n(n-1)}{1\times 2}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{1\times 2\times 3}a^{n-3}b^3 + \; ... + b^n$$

We can check this as follows; if n = 4:

$$(a + b)^4 = a^4 + 4a^{4-1}b + \frac{4(4-1)}{1\times 2}a^{4-2}b^2 + \frac{4(4-1)(4-2)}{1\times 2\times 3}ab^3 + \frac{4(4-1)(4-2)(4-3)}{1\times 2\times 3\times 4}a^{4-4}b^4$$

$$= a^4 + 4a^3b + \frac{4\times 3}{1\times 2}a^2b^2 + \frac{4\times 3\times 2}{1\times 2\times 3}ab^3 + \frac{4\times 3\times 2\times 1}{1\times 2\times 3\times 4}a^0b^4$$

$$= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$

The best way to remember the general formula is in terms of combinations:

$$(a + b)^n = a^n + {}_nC_1a^{n-1}b + {}_nC_2a^{n-2}b^2 + C_3a^{n-3}b^3 + ... + {}_nC_xa^{n-x}b^x + ... + b^n$$

This is what is known as a *binomial expansion*. The binomial coefficients ${}_nC_1$, ${}_nC_2$, etc. are simply combinations:

i.e. $${}_nC_x = \frac{n!}{(n-r)!\,x!} \text{ and}$$

$$n! = n(n-1)(n-2) \; ... \; 1$$

0! is defined to equal 1.

Remember that ${}_nC_x$ is sometimes written $[\genfrac{}{}{0pt}{}{n}{x}]$ or ${}^nC$.

Let us check this second version of the general formula again for n = 4:

From above:

$$ {}_4C_1 = \frac{4!}{(4-1)!\,1!} = \frac{4\times3\times2\times1}{3\times2\times1\times1} = 4 $$

$$ {}_4C_2 = \frac{4!}{(4-2)!\,2!} = \frac{4\times3\times2\times1}{2\times1\times2\times1} = 6 $$

$$ {}_4C_3 = \frac{4!}{(4-3)!\,3!} = \frac{4\times3\times2\times1}{1\times3\times2\times1} = 4 $$

$$ {}_4C_4 = \frac{4!}{(4-4)!\,4!} = \frac{4\times3\times2\times1}{1\times4\times3\times2\times1} = 1 $$

We thus get:

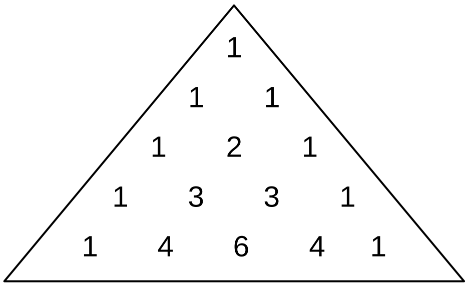
$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$, as before.

If we extract the coefficients from each binomial expansion for successive values of n, we can arrange them in a special triangular form known as *Pascal's Triangle*:

Expansion	**Coefficients**
$(a + b)^1 = a + b$	1 1
$(a + b)^2 = a^2 + 2ab + b^2$	1 2 1
$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$	1 3 3 1
$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$	1 4 6 4 1

The missing expansion is $(a + b)^0$ which equals 1, and this value can be inserted, giving the following triangular figure (Figure 12.1).

Figure 11.12: Pascal's Triangle



1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 is always the outside figure on every line. The total number of figures increases by 1 on each line. The new inside values are found by adding together, in turn, consecutive pairs of figures in the previous row. Thus, from above, the next row would be 1, 1 + 4, 4 + 6, 6 + 4, 4 + 1, 1, i.e. 1, 5, 10, 10, 5, 1. This tells us without any more working that:

$(a + b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$

Check for yourself that the next row is 1, 6, 15, 20, 15, 6, 1 and hence, write down the binomial expansion of $(a + b)^6$.

Chapter 12

The Normal Distribution

Contents *Page*

Introduction **202**

A. The Normal Distribution **202**
Definition 202
Properties 204
Standard Normal Distribution 205

B. Use of the Standard Normal Table **206**

C. General Normal Probabilities **208**

D. Use of Theoretical Distributions **210**
Types of Distribution 210
Use in Statistical Inference 210
Use in this Course 210

Answers to Questions for Practice **211**

Appendix: Areas in the Right-hand Tail of the Normal Distribution **214**

INTRODUCTION

Earlier in the course we discussed the measures (called statistics) used to give a numerical description of the location and variability or dispersion of a set of data. Unless these sets of data are very small, these statistics are calculated from frequency or grouped frequency distributions which are illustrated graphically by histograms or frequency polygons. If the number of classes in a frequency distribution is very large, these polygons can be drawn as smooth frequency curves.

In this chapter we will use these curves, and also some of the concepts of probability from the previous two chapters, to introduce the idea of continuous probability distributions and, in particular, the *normal distribution*, which has wide applications in business and management research.

The calculation of the probability of occurrence of values of continuous variables must be tackled in a slightly different way from that used for discrete variables. Since a continuous variable can take any value on a continuous scale, it never takes an exact value, so we find the probability with which it lies in a given interval.

To find out how to construct a theoretical frequency distribution for a continuous variable, suppose we look at the probability of occurrence of a road accident during the course of a day. Accidents will occur throughout the day with generally a peak during rush hours. It would be impossible to group accidents by each second of the day, so usually they are grouped into hours.

As the intervals chosen are quite small, we have an observed frequency distribution with large frequencies in some intervals and small frequencies in others. From such a frequency distribution, a frequency curve may be plotted.

From the shape of this curve and the value of the statistics calculated from the sample observations taken, we look for a curve the shape and parameters of which appear to fit the population from which the sample has come. This curve is the theoretical frequency distribution of the continuous variable we have observed, i.e. the time at which accidents are likely to occur with a particular probability.

This method for finding a theoretical frequency distribution is also used for discrete variables when the number of values the variable can take is so large that it is best to group them together in some form.

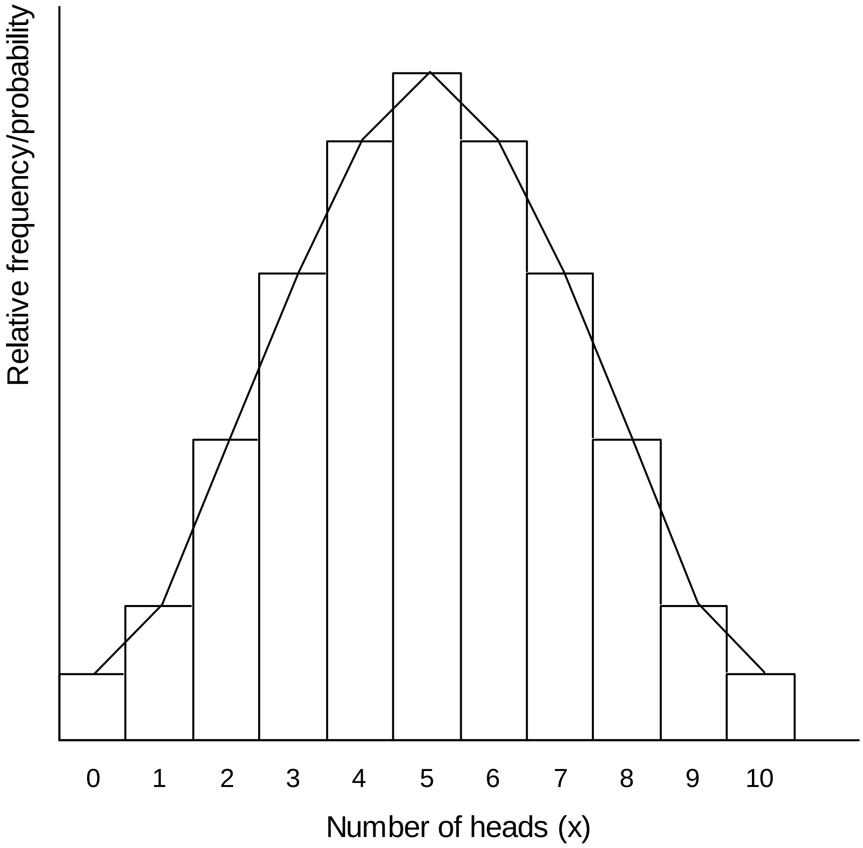
A. THE NORMAL DISTRIBUTION

Definition

The most *important* sort of frequency distribution is one which is *unimodal*, with more frequencies near the centre and fewer frequencies in the tails. This type occurs very often. Consider a coin-tossing trial, but instead of constructing the rectangular frequency distribution from the number of named outcomes as earlier, take the discrete variable, x, as the number of heads occurring at each toss. If the coin is tossed once, x takes the values 0 or 1; if it is tossed twice, x can be 0, 1 or 2; if it is tossed three times, x can be 0, 1, 2 or 3, and so on.

Figure 12.1 shows the relative frequency histogram and polygon for ten tosses of an unbiased coin. Both the histogram and the polygon are symmetrical about $x = 5$, which is thus the value of the mode, median and mean; the relative frequencies decrease symmetrically on each side of $x = 5$ and the value for $x = 0$ and $x = 10$ is small.

Figure 12.1: Histogram and polygon of a symmetrical distribution



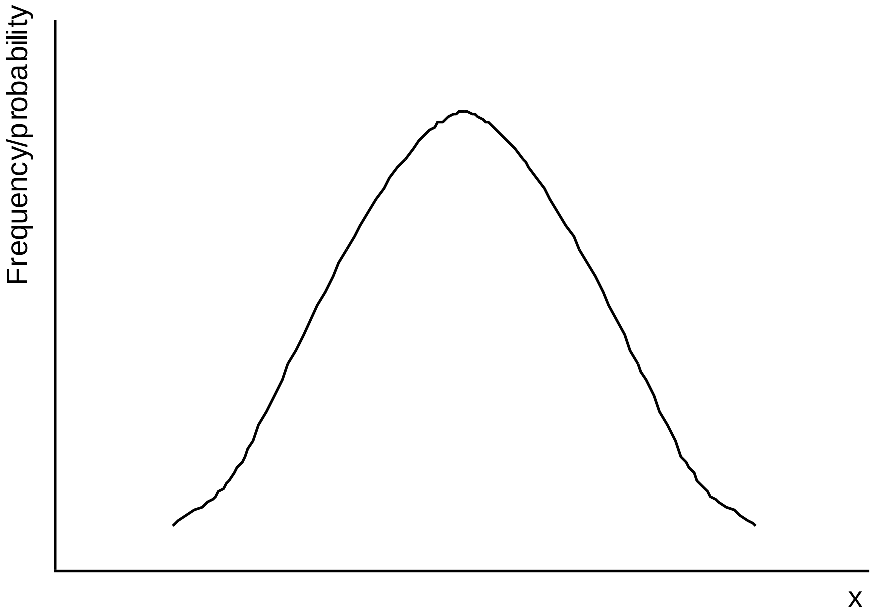
Activity

For your own satisfaction you should construct the relative frequency histograms and polygons for other numbers of tosses, to convince yourself that the same shape relative frequency polygon is obtained each time. You will find that the same shape is obtained if, for example, two dice are thrown and the variable is the total score showing each time.

Examine Figure 12.1 carefully and notice that, if we take the width of each bar of the histogram as one unit, the area of the bar for each value is the relative frequency of that value, and the sum of all the relative frequencies is one unit (1).

This relative frequency distribution is a probability distribution and the probability of occurrence of any value of x is the area of the bar above the value. If you compare the areas of the histogram bars with the area between the boundaries of the bars, the x-axis and the relative frequency polygon, you can see that they are approximately the same size. So you can find the probability of occurrence of any value or values of x either from the histogram or from the polygon. As the number of possible values of x increases, the relative frequency polygon becomes a smooth relative frequency curve, as shown in Figure 12.2.

Figure 12.2: Relative frequency curve of a symmetrical distribution



This relative frequency curve includes the relative frequency of all the values of the variable x, from $-\infty$ to $+\infty$, and the area between the curve and the x-axis is one unit and so it is a theoretical relative frequency distribution or a probability distribution.

This unimodal symmetrical bell-shaped curve is of great importance in statistical work. It is called the *normal distribution* or sometimes the *Gaussian distribution* after the scientist who developed its use for examining random errors of observation in experimental work.

When we consider the relative frequency curves of continuous variables, we discover a similar pattern in the measurements of a great many natural phenomena. For example, the frequency curve obtained from the set of heights of 80 employees, used in an earlier chapter, is unimodal with small frequencies in the tails. Later, we calculated the mean, standard deviation and median of this set of data. Since we were dealing with a comparatively small sample, the values of these measures were empirical, and it is reasonable to assume that the theoretical relative frequency distribution (the probability distribution) deduced from the data would be symmetrical and normal.

The same frequency distribution is found in the populations of dimensions of items from a factory production line and errors of observation in scientific experiments. This wide range of application of the normal distribution accounts for its importance in statistical inference.

Properties

If y is the ordinate of the probability curve, then the normal probability curve can be plotted from the equation $y = \phi(x)$, where x is a continuous variable which can take all values from $-\infty$ to $+\infty$ and $\phi(x)$ is a function of x which is completely defined by two parameters. (The letter ϕ is the lower case Greek phi, pronounced "fi".) (For this course you do not need to know the actual equation that defines this function, which is rather complicated.) We shall denote these two parameters by the Greek letters μ and σ, in other words the mean and standard deviation of the distribution.

By giving these parameters different values, we can obtain a set of curves, each of which has the following properties:

(a) The curve is symmetrical, unimodal and bell shaped.

(b) All the values of y are greater than zero and approach zero as x approaches $\pm\infty$.

(c) It can be proved that the area between the curve and the x-axis is one unit.

(d) It can be proved that:

(i) the mean, mode and median are all equal to the parameter μ.

(ii) the standard deviation is equal to the parameter σ.

(e) $P(x_1 < x < x_2) =$ area under the curve between the ordinates $\phi(x_1)$ and $\phi(x_2)$.

(f) If $z = \frac{x - \mu}{\sigma}$, it can be proved that z has the same normal distribution for *every* pair of values of the parameters μ and σ.

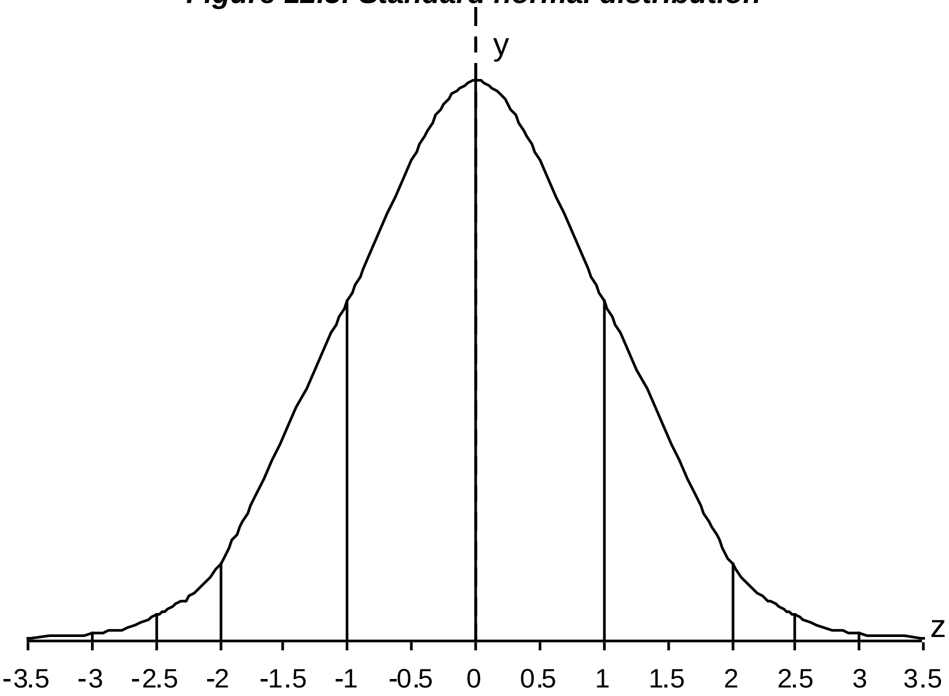
This distribution is called the *standard normal distribution*, and z is called the *z-score* or the *standardised value* of the variable.

Standard Normal Distribution

Since z is normally distributed, the range of z is $-\infty$ to $+\infty$ and the distribution has the properties (a), (b), (c), (d) and (e) listed above. In addition, it can be proved that $\mu = 0$ and $\sigma = 1$.

Figure 12.3 shows the shape and location of the standard normal distribution.

Figure 12.3: Standard normal distribution



Since $\sigma = 1$, the ordinates drawn in Figure 12.3 are 1, 2 and 2.5 standard deviations on each side of the mean. You can see that values of z more than 3 standard deviations from the mean are very unlikely to occur, and that 50% of the values of z lie below zero (the mean), and 50% above zero.

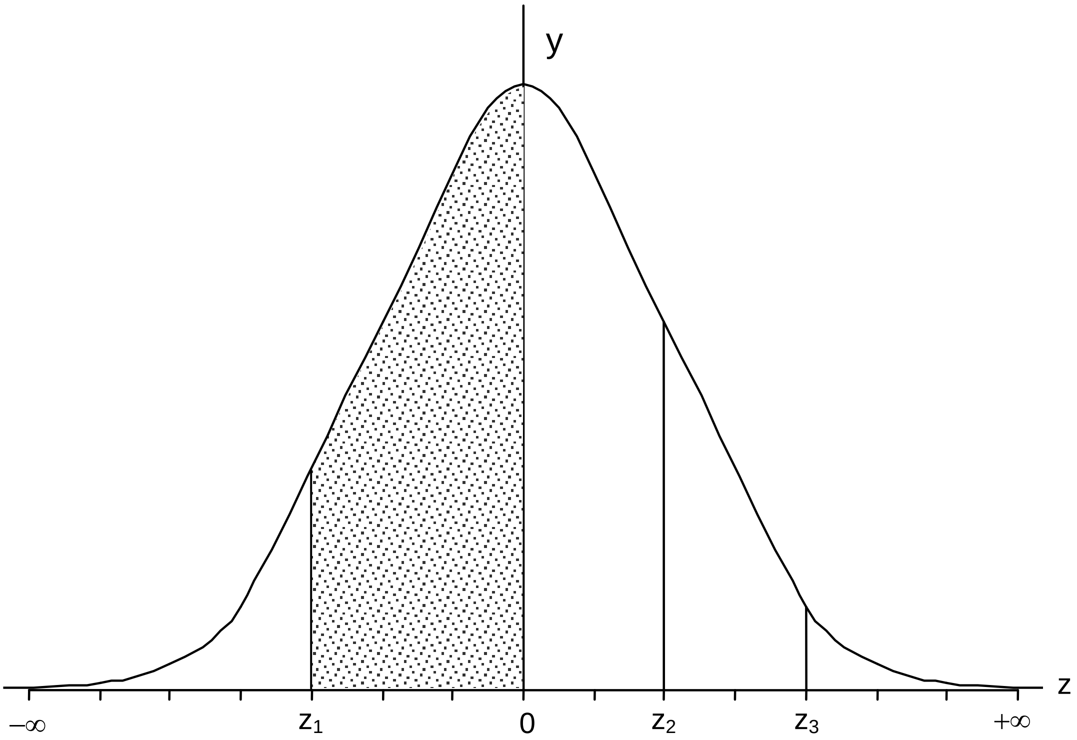
It can be calculated that:

- About 68% of the distribution lies within 1 standard deviation of the mean.
- About 95% of the distribution lies within 2 standard deviations of the mean.
- About 99% of the distribution lies within 2.5 standard deviations of the mean.

Since z is a continuous variable we cannot find the probability that z takes any exact value but only the probability that it lies in a given range of values. In other words, the probability that the value of z lies between any two given values is equal to the area under the curve between the ordinates at these two values.

Figure 12.4 shows the standard normal distribution divided into several areas by the ordinates at z_1, z_2 and z_3.

Figure 12.4: Different areas of the standard normal distribution



The probability that a z-value selected at random will lie between 0 and z_1 is equal to the area under the curve between 0 and z_1 (i.e. the shaded area shown in the graph). Similarly, the probability that z will be less than z_2 is equal to the total area under the curve to the left of the ordinate at z_2. And the probability that z will be greater than z_3 is equal to the area under the curve to the right of the ordinate at z_3.

So how can we find the areas under the standard normal curve between different z-values? The answer is that we can use the table of standard normal probabilities shown in the Appendix to this chapter. This table shows the areas under the standard normal curve *to the right of* given values of z. In the next section, we will examine this table and show, with examples, how it can be used to find the areas under the standard normal curve between any two z-values.

B. USE OF THE STANDARD NORMAL TABLE

The Appendix to this chapter sets out the table showing the standard normal probabilities.

This table gives the areas under the standard normal curve to the right of positive z-values, ranging from 0 to 4. You will notice that as the value of z increases above 3.0, the areas become very small, because the curve is approaching the axis and flattening out, so the probabilities change only slowly and become very close to 0.

To show how to use the table, consider the problem of finding the area under the curve to the right of z = 1.96. First, look down the left-hand column in the table and find 1.9. The other columns in the table give the second decimal place, so now move along the 1.9 row until you reach the .06 column and read the area (0.0250). This tells us that the area under the curve to the right of z = 1.96 is 0.025.

Since the normal curve is symmetrical about the mean, the area to the left of z = -1.96 must also be equal to 0.025. So using the table and the symmetry of the normal curve, we can easily find areas to the right of positive z-values and to the left of negative z-values.

But how can we find the area under the curve between two z-values? To show this, consider the area between -1 and +2.

- The area to the left of -1 is the same as the area to the right of +1. A glance at the table will show that this is equal to 0.1587.
- The area to the right of +2 is 0.02275.
- The total area under the curve is equal to one. So the area between -1 and +2 must be given by:

  Area = 1 – (0.1587 + 0.02275) = 0.81855

In this way, it is possible to use the areas given in the table, together with the symmetry of the curve and the fact that the total area under the normal curve is equal to one, to find any required area under the curve.

Table 12.1 shows the areas to the right of selected z-values taken from the table in the Appendix.

Table 12.1: Selected z-values

z	Area	z	Area	z	Area
0.0	0.5	0.5	0.3085	1.0	0.1587
0.01	0.4960	0.51	0.3050	1.01	0.1562
0.02	0.4920	0.52	0.3015	1.02	0.1539
0.25	0.4013	0.75	0.2266	2.0	0.02275

Remember that these values are the areas under the curve to the right of the selected z-values. Now for some practice.

Questions for Practice 1

1. From Table 12.1 and with the help of Figure 12.4 find:

   (a) P(z < 0.01)

   (b) P(z < 0.52)

   (c) P(z < 1.00)

   (d) P(0.02 < z < 1.02)

   (e) P(z > 0.51).

2. Find:

   (a) $P(z < 0.512)$

   (b) $P(z < 0.006)$

   (c) $P(z < 1.017)$.

3. Find:

   (a) $P(z < -1.02)$

   (b) $P(z > -0.50)$

   (c) $P(-1.01 < z < -0.52)$

   (d) $P(-0.51 < z < 1.00)$.

4. Now use the table in the appendix to find:

   (a) $P(z < 2.86)$

   (b) $P(z > 1.58)$

   (c) $P(z < 2.224)$

   (d) $P(z < -1.83)$

   (e) $P(z > -2.49)$

   (f) $P(-1.6 < z < 1.34)$.

Now check your answers with those given at the end of the chapter.

We now know how to use the standard normal tables to find the probability of occurrence of any required range of values of z. Next we must look at the application of this theory to practical problems where we are dealing with variables which have normal distributions with means not equal to zero and standard deviations not equal to one. So we will now establish the connection between standard normal probabilities and other practical probabilities.

C. GENERAL NORMAL PROBABILITIES

Let the continuous variable x have a normal distribution with mean μ and standard deviation σ. The probability curve for this distribution will have all the properties listed earlier; in particular, it will be symmetrical about the ordinate at μ.

Then $P(x_1 < x < x_2)$ = area under the normal curve between the ordinates at x_1 and x_2. To find this area, we must first transform x_1 and x_2 into z-values by subtracting the mean of x and then dividing by the standard deviation of x. This transformation gives the standardised values of x.

So the standardised values of x_1 and x_2 are:

$$z_1 = \frac{x_1 - \mu}{\sigma} \text{ and } z_2 = \frac{x_2 - \mu}{\sigma}$$

Using the formula, $z = \dfrac{x - \mu}{\sigma}$, it follows that:

- z will be a standard normal variable, and
- $P(z_1 < z < z_2)$ = area under the standard normal curve between the ordinates at z_1 and z_2.

Since z_1 and z_2 are linear functions of x_1 and x_2, the ratios of these two areas to the total areas under the two curves are equal, i.e.

$$\frac{\text{area under curve between } z_1 \text{ and } z_2}{\text{total area under standard normal curve}} = \frac{\text{area under curve between } x_1 \text{ and } x_2}{\text{total area under normal curve}}$$

But the total area under each of these curves is one unit, so the numerators of these two curves must also be equal.

This implies that $P(x_1 < x < x_2) = P(z_1 < z < z_2)$.

Therefore, in dealing with any practical problem, begin by working out the standardised values of the boundaries of the ranges in the problem. Then, using these standardised values and the standard normal table, find the required probability as shown in the earlier questions for practice. To show this, consider the following example:

Suppose that the time taken for an ambulance to reach a medical emergency is normally distributed with a mean of 15 minutes and a standard deviation of 5 minutes. Find the probability that the ambulance will take more than 22.5 minutes to reach a particular medical emergency.

First, we need to transform $x = 22.5$ to a z-value. This gives:

$$z = \frac{x - \mu}{\sigma} = \frac{22.5 - 15}{5} = 1.5$$

Then we use the table to find the area to the right of $z = 1.5$, which is 0.0668. This is the required probability.

Questions for Practice 2

1. A firm of stockbrokers will on average handle 2,500 shares a day with a standard deviation of 250 shares. If the number of shares sold is normally distributed, find the answers to the following questions:

   (a) What is the probability that more than 2,700 shares will be sold in one day?

   (b) What is the probability that less than 1,900 shares are sold on any one day?

   (c) What is the probability that the stockbrokers will sell between 2,300 and 2,550 shares a day?

   (d) What is the probability that they will sell either more than 3,125 or less than 2,000 shares in a day?

2. Computers consist of a number of components one of which is a memory. These memories, produced by an automatic process, have working life which is normally distributed with a mean of 500 hours and a standard deviation of 30 hours. If one thousand of these memories are selected at random from the production line, answer the following questions.

   (a) How many of the memories would you expect to last for longer than 550 hours?

   (b) How many memories would you expect to have a life of between 480 and 510 hours?

   (c) How many memories would you expect to have a life of more than 560 hours or less than 440 hours?

Now check your answers with those given at the end of the chapter.

D. USE OF THEORETICAL DISTRIBUTIONS

Types of Distribution

The theoretical frequency distribution of a variable is in fact the distribution of the whole population of the values that the variable can take. The method of calculating these distributions divides them into *two* types:

- Those populations which consist of a known limited number of values of the variable, so that we can construct their theoretical frequency distributions from known probabilities; the rectangular distribution is an example of this type. Then we can use any sample data to test whether the assumptions we have made are correct.
- Those populations which consist of an unknown or unlimited number of values of the variable, so that we have no prior knowledge of the probabilities. Then we have to use the empirical probabilities calculated from sample data to deduce the population distribution. We have already discussed one very important distribution of this type – the normal distribution.

Use in Statistical Inference

There are *three* main ways in which theoretical distributions and simple data are used in statistical inference. They are:

- To find the shape of a population distribution.
- To estimate the population parameters from the sample statistics.
- To test whether sample data come from a given population or whether two samples come from the same population.

You can read more about this in Chapter 13.

Use in this Course

Although you should appreciate the important part played by theoretical frequency distributions in statistical inference, for this course you will only need to know how to use these distributions to:

- estimate the mean (or proportion) of a population
- test whether a sample comes from a population with a given mean (or proportion)
- test whether two samples come from the same population.

ANSWERS TO QUESTIONS FOR PRACTICE

Questions for Practice 1

1. (a) $P(z < 0.01) = 1 - P(z > 0.01) = 0.5040$

   (b) $P(z < 0.52) = 1 - P(z > 0.52) = 0.6985$

   (c) $P(z < 1.00) = 1 - P(z > 1.00) = 0.8413$

   (d) $P(0.02 < z < 1.02) = P(z > 0.02) – P(z > 1.02) = 0.4920 – 0.1539 = 0.3381$

   (e) $P(z > 0.51) = 0.3050$.

2. (a) $P(z < 0.512) = 1 – P(z > 0.512)$. But since 0.512 has three decimal places, we cannot find $P(z > 0.512)$ directly from the table. Since 0.512 lies between 0.51 and 0.52, $P(z > 0.512)$ will lie between $P(z > 0.51)$ and $P(z > 0.52)$ and we have to interpolate between these two values. The two values of z are close together so we can get a sufficiently accurate value for $P(z > 0.512)$ by assuming that the part of the distribution curve between $z = 0.51$ and $z = 0.52$ is a straight line. Thus, by simple proportion:

   $P(z > 0.512) = P(z > 0.51) – 0.2[P(z > 0.51) – P(z > 0.52)]$

   $= 0.3050 – 0.2(0.3050 – 0.3015)$

   $= 0.3043$

   Therefore, $P(z < 0.512) = 1 – 0.3043 = 0.6957$.

   (b) To find this probability, we need to interpolate between $P(z > 0.00)$ and $P(z > 0.01)$. Since 0.006 is 0.6 of the distance between 0.00 and 0.01, we have:

   $P(z < 0.006) = 1 - P(z > 0.006)$

   $= 1 – [P(z > 0.00) – 0.6[P(z > 0.00) – P(z > 0.01)]]$

   $= 1 – [0.5 – 0.6(0.004)]$

   $= 0.5024$.

   (c) $P(z < 1.017) = 1 – [0.1562 – 0.7(0.0023)] = 1 – 0.1546 = 0.8454$ to 4 decimal places.

3. (a) Look at Figure 12.4 and let $z_1 = –1.02$. Then by symmetry, the area to the left of the ordinate at –1.02 is equal to the area to the right of the ordinate at +1.02

   i.e. $P(z < –1.02) = P(z > 1.02) = 0.1539$.

   (b) This time let z_1 be –0.50. Then, the area to the right of the ordinate at -0.5 is equal to one minus the area to the left of the ordinate at -0.5. This gives:

   $P(z > –0.50) = 1 - P(z < -0.50) = 1- P(z > 0.5) = 1 – 0.3085 = 0.6915$.

   (c) $P(–1.01 < z < –0.52) = P(z < -0.52) – P(z < -1.01) = P(z > 0.52) – P(z > 1.01)$
   $= 0.3015 – 0.1562 = 0.1453$.

   (d) $P(–0.51 < z < 1.00) = 1 – [P(z < -0.51) + P(z > 1.00)]$
   $= 1 – [P(z > 0.51) + P(z > 1.00)] = 1 – (0.3050 + 0.1587) = 0.5363$.

4. (a) $P(z < 2.86) = 1 – P(z > 2.86) = 1 – 0.00212 = 0.99788$

   (b) $P(z > 1.58) = 0.0571$

   (c) $P(z < 2.224) = 1 – P(z > 2.224) = 1 – 0.01308$ *(by interpolation)* $= 0.98692$

(d) $P(z < -1.83) = P(z > 1.83) = 0.0336$

(e) $P(z > -2.49) = 1 - P(z > 2.49) = 1 - 0.00639 = 0.99361$

(f) $P(-1.6 < z < 1.34) = 1 - [P(z > 1.6 + P(z > 1.34)] = 1 - [0.0548 + 0.0901]$
$= 0.8551.$

Questions for Practice 2

1. In all four parts of this question, let x = number of shares sold in a day and since $\mu = 2{,}500$ and $\sigma = 250$, then

$$z = \frac{x - 2{,}500}{250}.$$

(a) $$P(x > 2{,}700) = P\left(z > \frac{2{,}700 - 2{,}500}{250}\right)$$

$= P(z > 0.8) = 0.2119$

(b) $$P(x < 1{,}900) = P\left(z < \frac{1{,}900 - 2{,}500}{250}\right)$$

$= P(z < -2.4) = P(z > 2.4) = 0.0082$

(c) $$P(2{,}300 < x < 2{,}550) = P\left(\frac{2{,}300 - 2{,}500}{250} < z < \frac{2{,}550 - 2{,}500}{250}\right)$$

$= P(-0.8 < z < 0.2)$

$= 1 - [P(z > 0.8) + P(z > 0.2)] = 1 - [0.2119 + 0.4207] = 0.3674$

(d) $$P(x > 3{,}125) = P\left(z > \frac{3{,}125 - 2{,}500}{250}\right)$$

$= P(z > 2.5) = 0.0062$

$$P(x < 2{,}000) = P\left(z < \frac{2{,}000 - 2{,}500}{250}\right)$$

$= P(z < -2) = P(z > 2) = 0.02275$

$P(x < 2{,}000 \text{ or } x > 3{,}125) = P(x < 2{,}000) + P(x > 3{,}125)$ (mutually exclusive events)

$= 0.02275 + 0.0062 = 0.02898.$

2. In all three parts of this question, let x = length of life in hours of a memory and, since μ = 500 and σ = 30,

$$z = \frac{x - 500}{30}.$$

(a) $P(x > 550) = P\left(z > \frac{550 - 500}{30}\right)$

$= P(z > 1.67) = 0.0475$

i.e. the probability that one memory lasts longer than 550 hours is 0.0475.

Therefore, in a random sample of 1,000 memories you would expect 0.0475 × 1,000 to last longer than 550 hours, i.e. 47½.

(b) $P(480 < x < 510) = P\left(\frac{480 - 500}{30} < z < \frac{510 - 500}{30}\right)$

$= P(-0.67 < z < 0.33) = 1 - [P(z > 0.33) + P(z > 0.67) = 1 - [0.2514 + 0.3707]$
$= 0.3779$

i.e. the probability that one memory will last between 480 and 510 hours is 0.3779.

In a random sample of 1,000 memories you would expect 0.3779 × 1,000 of them to last between 480 and 510 hours, i.e. 377.9.

(c) $P(x > 560) = P\left(z > \frac{560 - 500}{30}\right) = P(z > 2)$

$P(x < 440) = P\left(z < \frac{440 - 500}{30}\right) = P(z < -2) = P(z > 2)$

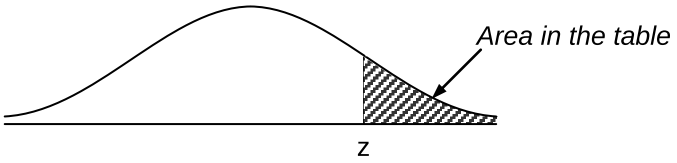
Thus, $P(x < 440 \text{ or } x > 560) = 2 \times P(z > 2) = 2(0.02275) = 0.0455$

i.e. the probability that one memory will last longer than 560 hours or less than 440 hours is 0.0455.

Therefore, in a random sample of 1,000 memories you would expect 0.0455 × 1,000 to last longer than 560 hours or less than 440 hours, i.e. 45.5.

Note that you can very often save time by using the symmetry of the normal curve to reduce the number of calculations.

APPENDIX: AREAS IN THE RIGHT-HAND TAIL OF THE NORMAL DISTRIBUTION



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1496	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1132	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
3.0	.00135									
3.1	.00097									
3.2	.00069									
3.3	.00048									
3.4	.00034									
3.5	.00023									
3.6	.00016									
3.7	.00011									
3.8	.00007									
3.9	.00005									
4.0	.00003									

Chapter 13

Significance Testing

Contents *Page*

A.	**Introduction**	**216**
	Assumptions and Definitions	216
	Notation	216
B.	**Introduction to Sampling Theory**	**217**
	The Sampling Distribution	217
	Central Limit Theorem	218
	Application of Central Limit Theorem	218
C.	**Confidence Intervals**	**219**
	Means	219
	Proportions	220
D.	**Hypothesis Tests**	**221**
	Hypothesis Tests about One Mean	222
	Hypothesis Tests about Two Means	225
	Hypothesis Tests about One Proportion	227
	Hypothesis Tests about Two Proportions	227
E.	**Significance Levels**	**228**
F.	**Small Sample Tests**	**229**
	The t-distribution	229
	Standard Errors for Small Samples	230
	Degrees of Freedom	230
	Let's do the Test	231
	The t Test and Proportions	231
	Difference of Means Test	231
Summary		**232**
Answers to Questions for Practice		**233**

A. INTRODUCTION

We have discussed the practical details of collecting data and selecting samples for statistical surveys, and the theoretical concepts of elementary probability and probability distributions. You need to have a firm understanding of the latter before you can carry out the process of statistical inference, i.e. before you can infer information about a whole population from the data collected in a sample survey. The normal distribution is, as far as you are concerned, the most important distribution.

In this chapter we shall be using information about population means and proportions from the means and proportions calculated from sample data. The inferences about these two measures are all that you require for this course, but similar methods are used in more advanced work for inferring the values of other population measures. We are concerned with two processes in this chapter:

- The estimation of population parameters from sample statistics and the confidence which can be attached to these estimates.
- Tests to decide whether a set of sample data belongs to a given population. This process is known as hypothesis testing.

Before explaining these processes, we need to state some necessary assumptions and definitions and to clarify some notation.

Assumptions and Definitions

The two key assumptions are that:

- the sample analysed is *large*, and
- the sample is a *random sample*.

(a) The meaning of a large sample

In this type of analysis, a sample which contains *more than 30 items* is counted as large. Different techniques must be used in the analysis of small samples, as we consider later in the chapter. Because of the time factor in examinations, you will sometimes be given questions involving smaller samples and, in such cases, either you will be told that the sample is taken from a normal population or you should state, as an extra assumption, that the population is normal.

(b) The meaning of a random sample

A random sample is a sample that has been selected so that any other sample of the same size was *equally likely* to have been selected, or it is a sample that has been selected so that each individual item is equally likely to have been chosen. (These two definitions are equivalent.)

Note that we always assume that we are analysing a random sample, although in practice, because of the nature of the population or a restriction on the cost or time allowed for the survey, some other method of selection has been used. If, in an exam, you are asked to comment on the result of your analysis, you should state that a bias may have been introduced by the method of selection.

Notation

In all work on inference we will keep strictly to the following notation. If you are not careful in the use of notation, you may find that you have calculated two values for the same symbol. The general rule to remember is that Greek letters are used for populations, and Roman letters for samples. These symbols are used universally, so that they need not be defined in each problem.

Table 13.1: Symbols used for population and samples

	Size	**Mean**	**Standard deviation**	**Proportion**
Population	N	μ	σ	π
Sample	n	$\bar{x}$	s	p

You will notice that the rule is broken here with the population size, but the Greek equivalent of N is very easily confused with V.

In addition we need to define a special distribution and state an important theorem. These two items are so important that we will devote the whole of the next section to them.

B. INTRODUCTION TO SAMPLING THEORY

The Sampling Distribution

Suppose you own a factory producing fluorescent light tubes. It would be very useful, both for advertising and for monitoring the efficiency of your plant, to know the average length of life of these tubes. Clearly it is not possible to test every tube, so you take a random sample of size n, measure x hours, the life of each tube, and then calculate $\bar{x}$ and s. These values will be estimates of the population parameters μ and σ.

If you take a number of samples, each one will give you a different value for the statistics and so a different estimate. As x is a variable it will have a probability distribution and, since $\bar{x}$ is a statistic calculated from the sample, it is also a variable and so will have a probability distribution. The distribution of $\bar{x}$ is called the *sampling distribution of the mean* and it has the following properties:

(a) It can be shown that the mean of $\bar{x}$ is equal to the mean of x, i.e. $\mu_{\bar{x}} = \mu$.

(b) It is obvious that the standard deviation of $\bar{x}$ will be smaller than the standard deviation of x because the extreme values of $\bar{x}$ must be smaller than the extreme values of x. The standard deviation of $\bar{x}$ ($\sigma_{\bar{x}}$) depends upon the size of the sample and is defined as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This is called the *standard error of the mean* (SE).

Of course, s will also have a sampling distribution, but that is not included in this course. The standard distribution is less variable than the mean, so its standard deviation will be smaller. In practice, since we rarely know the true value of σ, we can use the following formula for the standard error without significant loss of accuracy:

$$\frac{s}{\sqrt{n}}$$

Central Limit Theorem

The central limit theorem forms the basis of the theory that we use for the statistical inference processes in this chapter. It states that:

(a) If the distribution of x is normal with mean μ and standard deviation σ and samples of size n are taken from this distribution, then the distribution of $\bar{x}$ is also normal with mean μ and standard deviation (SE)

$$\frac{\sigma}{\sqrt{n}}$$

whatever the size of n.

(b) If the distribution of x is not normal, as the size of the sample increases, the distribution of x approaches the normal distribution with mean μ and standard deviation

$$\frac{\sigma}{\sqrt{n}}$$

i.e. for values of n > 30 we can assume that $\bar{x}$ is normal.

Application of Central Limit Theorem

The central limit theorem means that if we have a large sample from a population, the distribution of which we do not know, we can use the standard normal distribution to calculate the probability of occurrence of values of $\bar{x}$ since the standardised value of x, which is

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)},$$

will have the standard normal distribution.

In particular, we can find the length of the interval within which a given percentage of the values of $\bar{x}$ will lie. We are most interested in those intervals for which $\bar{x}$ is the centre, and these intervals are equivalent to the standardised intervals with 0 as the centre.

Example 1:

The statement that about 95% of the values of $\bar{x}$ lie within two standard deviations of the mean implies that exactly 95% of the values will lie within about two standard deviations of the mean. Let z_1 be the exact value; then we can calculate z_1 from the equivalent probability statement:

$P(-z_1 < z < z_1) = 0.95$

This implies that $P(z < z_1 \text{ or } z > z_1) = 0.05$

Therefore, $2P(z > z_1) = 0.05$ and $P(z > z_1) = 0.025$.

Then, from the standard normal table, $z_1 = 1.96$.

Example 2:

Suppose we need the value of z_1 so that exactly 99% of the values will lie in the interval. Then as in Example 1:

$P(-z_1 < z < z_1) = 0.99$

$2P(z > z_1) = 0.01$ and $P(z > z_1) = 0.005$.

From the table, $z_1 = 2.576$ to 3 decimal places (though 2.58 is usually accurate enough).

In the same way you can calculate the z value for any percentage or probability that you choose.

C. CONFIDENCE INTERVALS

Means

Usually, because of the time or cost of sample surveys, we have to base decisions on the analysis of one set of sample data. Having calculated the mean and standard deviation of this sample, we say that the estimated value of the population mean μ is the sample mean $\bar{x}$. This information will be much more valuable if we can also estimate the difference between $\bar{x}$ and the true value of μ. We do this by finding the interval, with centre the calculated value of $\bar{x}$, which will contain a chosen percentage of the other possible values of $\bar{x}$. This is called a confidence interval for μ, and its boundaries are called the *confidence limits*.

Let z be the value of the standard normal corresponding to the chosen percentage; then the confidence interval is defined by:

$$\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

Example:

A sample of 100 fluorescent tubes from the Short Life Tube Company gives a mean length of life of 20.5 hours with a standard deviation of 1.6 hours. Find (a) a 95% confidence interval; (b) a 99% confidence interval for the average length of life of those tubes. Interpret the result in each case.

(a) $\bar{x} = 20.5$, $n = 100$, $s = 1.6$ $(= \sigma)$, $z = 1.96$.

A 95% confidence interval for μ $= \bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$

$$= 20.5 \pm 1.96 \times \frac{1.6}{10} \quad = 20.5 \pm 0.3136$$

$= 20.19$ to 20.81.

(Note that as $\bar{x}$ is given to only 1 decimal place the limits should be given to 2 decimal places.)

This means that for 95% of the possible samples that could be taken, the estimate you would give for μ would lie in the interval 20.19 to 20.81 hours, i.e. you are 95% confident that μ lies between these values.

(b) $\bar{x} = 20.5$, $n = 100$, $s = 1.6$, $z = 2.58$.

A 99% confidence interval for $\mu = \bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$

$$= 20.5 \pm 2.58 \times \frac{1.6}{10} \quad = 20.5 \pm 0.4128$$

$= 20.09$ to 20.91.

This means that you are 99% confident that the true value of μ lies between 20.09 and 20.91 hours.

Proportions

Suppose we need to find the proportion of individuals in a population who possess a certain attribute. For instance, for planning purposes we may require to know:

- The proportion of defective items coming off a production line in a shift.
- The proportion of pensioners in a country.
- The proportion of voters who are in favour of reintroducing the death penalty.
- The proportion of householders in a major city who wish to possess cable television.

Provided the proportion is not expected to be very small, we can use the same technique to find this information as we used for measurements of continuous variables.

The results of a sample survey are used to estimate the population proportion. For the population:

Let $\pi = \dfrac{\text{Number of individuals possessing the attribute}}{\text{N}}$

$$1 - \pi = \frac{\text{Number of individuals not possessing the attribute}}{\text{N}}$$

where N is the population size.

For the sample:

$$p = \frac{\text{Number of individuals possessing the attribute}}{n}$$

$$q = \frac{\text{Number of individuals not possessing the attribute}}{n}$$

and $p + q = 1$

Then p is the estimate of π and, by the Central Limit Theorem, p is normally distributed, so:

- A confidence interval for $\pi = p \pm z$ (SE of p) where z is the value of the standard normal of the chosen percentage and
- The standard error of p can be shown to be $\sqrt{\dfrac{\pi(1-\pi)}{n}}$ which is estimated by $\sqrt{\dfrac{pq}{n}}$.

Thus, the formula for the confidence interval for a proportion may be written as:

$$\pi = p \pm z\sqrt{\frac{pq}{n}}$$

Example

In a sample of 200 voters, 80 were in favour of reintroducing the death penalty. Find a 95% confidence interval for the proportion of all voters who are in favour of this measure.

$$p = \frac{80}{200} = 0.4, \quad q = 1 - p = 0.6, \quad \text{SE of p} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.4 \times 0.6}{200}}$$

Thus, a 95% confidence interval for π

$$
\begin{aligned}
&= p \pm 1.96\sqrt{\frac{pq}{n}} \\
&= 0.4 \pm 1.96 \times 0.0346 \\
&= 0.4 \pm 0.0679 \\
&= 0.332 \text{ to } 0.468
\end{aligned}
$$

Therefore, the proportion of voters in favour lies between 33% and 47%.

D. HYPOTHESIS TESTS

In order to discover whether or not a certain set of data comes from a particular population, we choose one of the descriptive measures of that population and then test to see if it is likely that the sample belongs to that population. To carry out the test we need to formulate two hypotheses:

- The *null hypothesis*, H_o: the sample belongs to the population.
- The *alternative hypothesis*, H_1: the sample does not belong to the population.

The principles behind hypothesis testing are the same whichever measure we choose for the test.

Because the sample measures are variables, we run a risk of making the wrong decision in two ways. We may reject H_o when it is true, making what is called a Type I error, or we may accept H_o when it is false, making what is called a Type II error.

Table 13.2 gives a summary of the possible decisions.

Table 13.2: Types of Error

	Accept H_o	*Reject H_o*
H_o true	Correct decision	Make Type I error
H_o false	Make Type II error	Correct decision

There is no direct algebraic connection between the values of the risks of making the two errors, but as one risk is decreased the other is increased. We use the Type I error in hypothesis tests and the Type II error to find the power of the test, which is outside the scope of this course.

Having formulated the hypotheses, you must decide on the size of the risk of making a Type I error that you are prepared to accept. For example, you may decide that a 5% risk is acceptable (making the wrong decision once in 20 times) or you may require only a 1% risk (wrong once in a 100 times). You then divide all the possible values of the sample measure into two sets, putting (say) 5% of them in the critical or rejection region and 95% in the acceptance region. The percentage chosen is called the level of significance. Next calculate the value of the measure for the sample you are using, see which region it belongs to and state your decision.

Hypothesis Tests about One Mean

The important thing to remember about hypothesis tests is that there should be no personal bias in the decision, i.e. the test must be set up so that the same decision is reached whoever carries out the test. With this in mind, a procedure has been designed that should be followed for every test. The steps are listed below for tests about means, but these steps should be included in tests about any measure.

The necessary steps are as follows:

1. State the null hypothesis, H_o.
2. Decide on the alternative hypothesis, H_1.
3. Choose the level of significance, thus fixing the critical region.
4. Calculate the mean and standard deviation of the sample, if these are not given in the problem, and the standard error of the mean.
5. Calculate the standardised z statistic.
6. Accept or reject the null hypothesis.

Now we will look at these steps in detail:

1. The null hypothesis is $H_0 : \mu =$ the value of the population mean (μ_0 say) given in the problem.
2. The alternative hypothesis depends on the wording of the problem. The wording can suggest one of three possible meanings:

   (a) The sample comes from a population the mean of which is not equal to μ_0, i.e. it may be smaller or larger. Then you take $H_1 : \mu \neq \mu_0$.

   For this alternative you divide the critical region into two equal parts and put one in each tail of the distribution, as shown in Figure 13.1. This is called a *two-tailed test*.

Figure 13.1: Two-tailed test

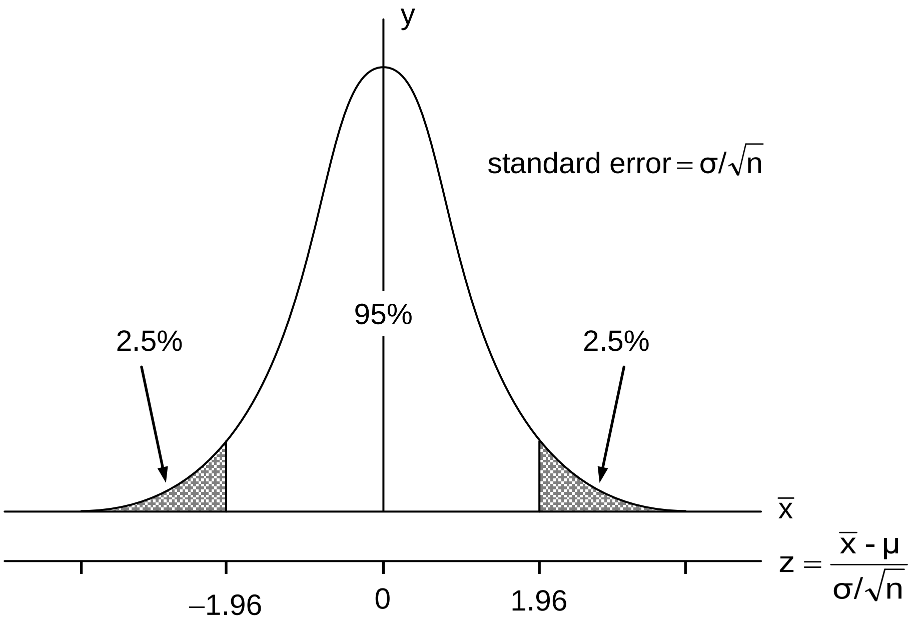


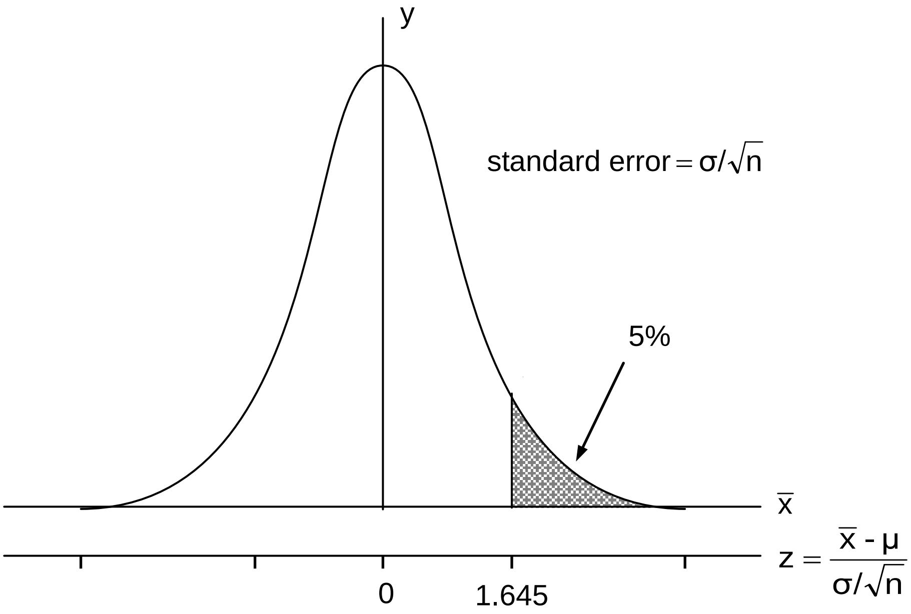
Figure 13.1 shows the critical region (shaded areas) for a two-tailed test when the level of significance is 5%. The z-scale at the bottom shows the critical values of

the z statistic. (Notice that the acceptance region is the region used for a 95% confidence interval for μ.)

(b) The sample comes from a population the mean of which is larger than μ_0. Then you take $H_1 : \mu > \mu_0$ and put the whole of the critical region in the right-hand tail of the distribution, as shown in Figure 13.2. This is called an *upper-tail test*.

(c) The sample comes from a population with μ smaller than μ_0, so take $H_1 : \mu < \mu_0$ and put the whole of the critical region in the left-hand tail of the distribution. This is called a *lower-tail test* and it would be shown in a figure similar to Figure 13.2 with the shaded areas in the left-hand tail and the value –1.645 shown on the z-scale.

Figure 13.2: Upper-tail test



3. Decide what risk of making the wrong decision you are prepared to accept. State this risk as the value of the level of significance and also state the corresponding critical values of z and define the critical region. Table 13.3 shows the critical z values for the four most commonly used levels of significance. These values have been calculated from the standard normal tables.

Table 13.3: Critical z values

Level of significance	U-tail test	L-tail test	Two-tailed test
10%	1.282	–1.282	–1.645 and 1.645
5%	1.645	–1.645	–1.96 and 1.96
1%	2.326	–2.326	–2.576 and 2.576
0.1%	3.09	–3.09	–3.29 and 3.29

You will use these values (particularly the first three) so often that it is worth memorising them.

4. In exam questions you will usually be given $\bar{x}$ and s, but in practice you may have to calculate them from the sample data (see earlier chapters).

   If you are given σ in the problem, calculate $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

   If you are not given σ, use SE $= \frac{s}{\sqrt{n}}$.

5. Calculate $z = \dfrac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

   and state whether the value lies in the critical region or the acceptance region.

6. *Reject H_o if z lies in the critical region, accept H_o if z lies in the acceptance region* and then state the meaning of your decision in terms of the actual problem you are solving. H_o is rejected if z is equal to the critical value used.

   If z lies in the *acceptance region* for the level used, the result of the test is said to be *not significant*.

   If z lies in the *critical region at the 10% or 5% level*, the result of the test is said to be *significant at that level*.

   If z lies in the *critical region at the 1% level*, the result of the test may be said to be *highly significant*.

   If z lies in the *critical region at the 0.1% level*, the result of the test may be said to be *very highly significant*.

In the following example, the steps in the testing procedure are numbered so that you can follow the method easily.

Example:

A sample of 100 fluorescent light tubes from the Short Life Tube Company has a mean life of 20.5 hours and a standard deviation of 1.6 hours. Test:

(a) At the 1% level whether the sample comes from a population with mean 23.2 hours.

(b) At the 5% level whether it comes from a population with mean 20.8 hours.

(c) At the 5% level whether it comes from a population with mean less than 20.8 hours.

Answer:

(a) (1) $H_0 : \mu = 23.2$.

(2) $H_1 : \mu \neq 23.2$.

(3) Level of significance = 1%
critical values of z are –2.576 and 2.576
therefore, critical region is either z < –2.576 or z > 2.576.

(4) $\bar{x} = 20.5$, $s = 1.6$, $n = 100$, so $SE = \frac{1.6}{10} = 0.16$.

(5) $z = \frac{20.5 - 23.2}{0.16} = \frac{-2.7}{0.16} = -16.875 < -2.576$
So z lies in the critical region.

(6) This result is highly significant so reject H_o, i.e. the evidence of this sample suggests that the population mean is not 23.2 hours.

(b) (1) $H_0 : \mu = 20.8$

(2) $H_1 : \mu \neq 20.8$

(3) Level of significance $= 5\%$
critical values of z are -1.96 and 1.96
therefore, critical region is either $z < -1.96$ or $z > 1.96$.

(4) As in (a), SE $= 0.16$

(5) $z = \dfrac{20.5 - 20.8}{0.16} = \dfrac{-0.3}{0.16} = -1.875 > -1.96$
So z does not lie in the critical region.

(6) This result is not significant so accept H_o, i.e. the evidence of this sample suggests that the population mean is 20.8 hours.

(c) (1) $H_0 : \mu = 20.8$

(2) $H_1 : \mu < 20.8$

(3) Level of significance $= 5\%$
critical value of z is -1.645
therefore, critical region is $z < -1.645$.

(4) As in (a), SE $= 0.16$

(5) As in (b), $z = -1.875 < -1.645$

So z lies in the critical region.

(6) This result is significant so reject H_o, i.e. the evidence of this sample suggests that the population mean is less than 20.8.

Questions for Practice 1

A sample of 150 students had an average IQ of 112 with a standard deviation of 9.

(a) At what level of significance would you accept that this sample is taken from a student population with average IQ of 110?

(b) At the 5% level would you accept that the student population had an average IQ greater than 113?

Now check your answers with those given at the end of the chapter.

Hypothesis Tests about Two Means

There are occasions when we are not particularly interested in the population from which a sample is taken, but we need to know whether two samples come from the *same* population. We use the same test procedure to test the difference between the two means.

Using the suffixes 1 and 2 to distinguish between the two samples, the hypotheses become:

$H_0 : \mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0$

H_1 : (i) $\mu_1 \neq \mu_2$, i.e. (two-tailed)

(ii) $\mu_1 > \mu_2$, i.e. $\mu_1 - \mu_2 > 0$ (upper tail)

(iii) $\mu_1 < \mu_2$, i.e. $\mu_1 - \mu_2 < 0$ (lower tail).

SE of $(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$

$= \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$ if σ_1 and σ_2 are not known

$= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ if both populations are known to have the same standard deviations.

Then the test statistic for a difference between two means is given by:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Example:

Two factories are producing visual display units for computers. Use the following sample data to test whether the two production lines are producing units with the same mean life length:

	$\bar{x}$	**s**	**n**
Sample 1	20.5	3.4	125
Sample 2	19.0	2.1	180

Answer:

(1) $H_0 : \mu_1 - \mu_2 = 0$.

(2) $H_1 : \mu_1 - \mu_2 \neq 0$.

(3) Let level of significance = 5%
critical values of z are –1.96 and 1.96
therefore, critical region is z < –1.96 or z > 1.96.

(4) SE $= \sqrt{\frac{(3.4)^2}{125} + \frac{(2.1)^2}{180}} = \sqrt{\frac{11.56}{125} + \frac{4.41}{180}} = 0.34$

(5) $z = \frac{20.5 - 19}{0.34} = \frac{1.5}{0.34} = 4.41 > 1.96$
So z lies in the critical region.

(6) This result is significant so reject H_o, i.e. the test result suggests that the two production lines are producing units with different mean life lengths.

Hypothesis Tests about One Proportion

To test whether a population proportion is equal to some assumed value, π_0, the test statistics is:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$$

Example:

A manufacturer of computers claims that his computers are operational for more than 80% of the time. During the course of a year one computer was operational for 270 days. Test, at the 1% level, whether the manufacturer's claim was justified.

Answer:

(1) H_o: $\pi = 0.8$.

(2) H_1: $\pi > 0.8$.

(3) Level of significance = 1%
critical value of z in a one-tailed test is 2.326
therefore, critical region is z > 2.326.

(4) $p = \frac{270}{365}$, $n = 365$, SE $= \sqrt{\frac{0.8 \times 0.2}{365}} = 0.021 = \sigma_p$ since H_o gives $\pi_0 = 0.8$.

(5) $z = \frac{p - \pi_0}{SE} = \frac{0.74 - 0.8}{0.021} = -2.86 < 2.326$.
So z does not lie in the critical region.

(6) This result is not significant so accept H_o, i.e. the evidence does not support the manufacturer's claim.

Questions for Practice 2

The proportion of drivers who plead guilty to driving offences is usually 60%. Out of 750 prosecutions 400 pleaded guilty. Is this proportion significantly different from usual?

Now check your answers with those given at the end of the chapter.

Hypothesis Tests about Two Proportions

To test for a difference between proportions, the test statistic is:

$$z = \frac{p_1 - p_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where: $\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$\hat{q} = 1 - \hat{p}$

Example:

A market research organisation carried out a sample survey on the ownership of washing machines and concluded that 64% of all households owned a washing machine, out of 200 households sampled. Six months later they repeated the survey, using the same questionnaire, and concluded that 69% owned a washing machine, out of 150 households sampled. Is the difference due to a significant increase in ownership or is it a random sampling error?

The sample data is $p_1 = 0.64$, $n_1 = 200$, $p_2 = 0.69$, $n_2 = 150$.

Answer:

(1) H_o: $\pi_1 = \pi_2$, i.e. $\pi_1 - \pi_2 = 0$

(2) H_1: $\pi_1 < \pi_2$, i.e. $\pi_1 - \pi_2 < 0$

(3) Let the level of significance = 5%
critical value of z is –1.645
therefore, critical region is z < –1.645.

(4) As π_1 and π_2 are not known, we have to estimate the SE of their difference. The best estimate of this SE is found by combining the two samples. Let p = proportion of owners in combined sample.

Number of owners in 1st sample = n_1p_1

Number of owners in 2nd sample n_2p_2

Size of combined sample = $n_1 + n_2$

Therefore, $\hat{p} = \dfrac{n_1p_1 + n_2p_2}{n_1 + n_2}$ and $SE = \sqrt{\dfrac{pq}{n_1} + \dfrac{pq}{n_2}}$

substituting the sample values gives:

$$p = \frac{200 \times 0.64 + 150 \times 0.69}{200 + 150} = \frac{231.5}{350} = 0.66 \text{ and } q = 0.34$$

$$\text{so } SE = \sqrt{\frac{0.66 \times 0.34}{200} + \frac{0.66 \times 0.34}{150}} = 0.051.$$

(5) $$z = \frac{p_1 - p_2}{SE} = \frac{0.64 - 0.69}{0.051} = \frac{-0.05}{0.051} = -0.98 > -1.645$$

so z does not lie in the critical region.

(6) This result is not significant so accept H_o, i.e. the test result suggests that the apparent small rise in ownership is due to sampling error.

E. SIGNIFICANCE LEVELS

It is essential that you realise and can explain the implications of choosing a particular significance level. The most useful way to get a feel for this measure is to think of it as the chance of being wrong when you say there is a change or difference. If we know the cost of being wrong, we can work out the expected value of being wrong:

expectation = P(wrong) × cost of being wrong

If the cost of being wrong is small, we can afford to take a higher risk of being wrong. A good example would be in the early stages of an investigation of a pharmaceutically-active substance which might be developed into a new drug. If we decide on the basis of a test that something may be causing a useful effect and we keep it for further testing, the cost will be low at the start of testing if we later reject it. We would err on the side of keeping the substance in our tests. In this circumstance, a 10% chance of keeping something which later turns out to be useless would be quite acceptable.

Now consider the end of the process when a new drug has been produced. It must be tested extensively for undesirable side effects, because we know that if it is sold and side effects develop, the company selling it may face legal actions which might result in the payment of extremely large damages. The risk of incurring such costs has to be kept very low, and so very small values of the significance level will be used. Just to confuse us, they are sometimes referred to as "very high significance", meaning we have very little doubt that our conclusions are correct.

If we cannot estimate costs in this way, the default value is the 5% significance level. The best way to explain it is to say "In the absence of any information which would make a different significance level a better choice, we use the 5% level."

F. SMALL SAMPLE TESTS

The t-distribution

In the tests we did earlier, the central limit theorem allowed us to state that the means or differences we were testing were normally distributed, so we could calculate the standard deviation of a sample mean or difference from the population standard deviation and the sample size. This is fine if we know the population standard deviation. Often we do not, and all the data we have is represented by a small sample. For example, suppose a supplier claims that his product has a content of a main ingredient of 40%. We measure the content in a sample of 5 bags of the product (chosen at random) and we obtain the following values:

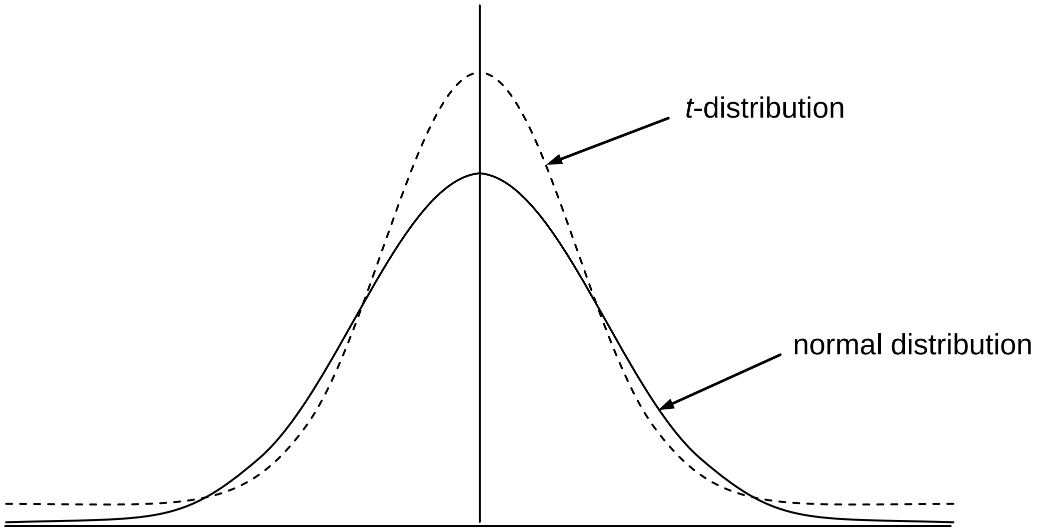
38%, 32%, 44%, 37%, 34% with a mean of 37%.

Does this sample invalidate the supplier's claim?

Before we carry out the test, consider our information. We can use the given figures to estimate the standard deviation of the population of all bags of the product. How reliable is a sample of 5? The answer is not very. So it would seem to make sense to demand a greater difference between the observed mean value of 37% and the claimed value of 40%, than if we could carry out the test with a value of the standard deviation based on a large amount of data. How much allowance for this doubt in the value of the standard deviation must we put into the test? The answer was given by a statistician called W S Gosset. He published his work under the pseudonym of "A Student of Statistics", and used the letter *t* for the test parameter. Ever since, the distribution he derived has been referred to as "Student's *t*-distribution".

Here is how it works: in the tests we have done already we calculated a z value, and tested it against critical values derived from the normal distribution, i.e. + or –1.96 for a two-tailed test at the 5% level. The distribution of *t* which replaces *z* looks like the normal distribution, but has a sharper central peak and higher tails. To get the same tail areas which determine the critical values we have to go further from the mean. In other words we need a larger difference for it to be taken as significant than when we had better information. The difference between the normal distribution and the *t*-distribution with the same mean and standard deviation is shown in Figure 13.3. For a sample of 5 we require values of *t* beyond + or –2.78.

Figure 13.3: The t-distribution



Standard Errors for Small Samples

If we were to take a large number of samples, each of 5 items, from the population of all bags of the products, and for each calculate the standard deviation using the formula we have used so far, and then found the mean of these values, we would get a value for the standard deviation which is lower than the true value. This would be the case, no matter how many samples we averaged the value over. We say that the estimate of the population standard deviation is biased.

We can correct for the bias by multiplying the sample standard deviation *s* by:

$$\sqrt{\frac{n}{(n-1)}}\ .$$

Alternatively, in calculating the standard deviation we use the divisor (n – 1) instead of n in the calculation, which will give us the best estimate of the population standard deviation, σ. This is written $\hat{\sigma}$.

Degrees of Freedom

The divisor (n – 1) is called the "degrees of freedom" (do not worry about why). The degrees of freedom also determine how close the *t*-distribution is to the normal distribution. In fact when the value of the degrees of freedom is greater than about 30, we usually ignore the difference and carry out tests as we did earlier.

To find the critical values for *t* we have to know three things:

- the significance level
- one or two-tailed test?
- degrees of freedom.

Tables are available to give us the values. If you look at a *t* table you will see the familiar values of 1.96 and 1.65 at the bottom of the columns when the degrees of freedom become very large.

Let's do the Test

$H_0 : \mu = 40\%$.

Sample values of x are 38%, 32%, 44%, 37%, 34%.

Sample mean $= \bar{x} = \frac{\sum x}{5} = \frac{185}{5} = 37\%$.

$$\hat{\sigma} = \sqrt{\frac{\Sigma(x - \bar{x})^2}{(n - 1)}} = \sqrt{\frac{(1^2 + 5^2 + 7^2 + 0^2 + 3^2)}{4}}$$

$$= 4.5826.$$

SE of sample mean $= \frac{\sigma}{\sqrt{n}} = \frac{4.5826}{4} = 2.05$

From tables, critical value of *t* at 5% level, two-tailed test with degrees of freedom $= (n - 1) = 4$, is:

$$t = \pm 2.78,$$

$$t = \frac{37 - 40}{2.05} = 1.46.$$

The value of *t* lies well within the range which is consistent with the null hypothesis, so we cannot reject the null hypothesis. In terms of our original question, we do not have enough evidence to say that the average content differs from 40%.

The t Test and Proportions

When testing proportions or differences in proportions you will rarely, if ever, meet a case where the sample sizes are small enough to require the use of the *t*-distribution.

Difference of Means Test

Two groups of students are asked to take a test. The results are:

- Group A scores 45, 87, 64, 92, 38.
- Group B scores 40, 35, 61, 50, 47, 32.

Is there a significant difference in mean score between the two groups?

Group A: mean $= \bar{x} = \frac{326}{5} = 65.2$

$$\hat{\sigma}_A = \sqrt{\frac{2342.8}{4}} = 24.2012.$$

Group B: mean $= \bar{x} = \frac{265}{6} = 44.17$

$$\hat{\sigma}_B = \sqrt{\frac{574.8333}{5}} = 10.7228.$$

H_o : no difference in mean score.

SE of difference in scores $= \sqrt{\frac{\hat{\sigma}_A{}^2}{n_A} + \frac{\hat{\sigma}_B{}^2}{n_B}}$

$$= \sqrt{\frac{24.2012^2}{5} + \frac{10.7226^2}{6}} = 11.67.$$

Degrees of freedom = (n_A – 1) + (n_B – 1) = (5 – 1) + (6 – 1) = 9.

Critical values of *t*, 5% level, two-tailed test = ± 2.26.

$$t = \frac{65.2 - 44.17}{11.67} = 1.80\,.$$

There is no significant difference in scores for the two groups.

You should now realise that to have a significant result when using small samples, large changes or differences are essential. The statistical test will often say "not significant" when there does appear to be a change or difference. In this case it is likely that with more data a significant result will be found.

SUMMARY

This chapter is very important – examination questions are frequently set on confidence intervals and hypothesis tests. Although you may use any levels of significance you like, it is best to keep to the three given here as experience has shown that they are of the most practical use.

You must realise that statistical tests do not prove or disprove hypotheses in the mathematical sense. They only provide evidence in support of one or other of the two hypotheses chosen.

You should be quite sure that you know the following thoroughly:

- How to use the standard normal table.
- The formulae for confidence intervals.
- The procedure for carrying out hypothesis tests.

Whenever we have good information about the variability in a given situation, i.e. σ has been estimated from a large amount of data, we can use a z-test, no matter how small the samples we are testing. It is not the size of sample which decides whether or not we use z or *t*, but whether we used the sample to provide an estimate of σ.

ANSWERS TO QUESTIONS FOR PRACTICE

Questions for Practice 1

The steps in the testing procedure are numbered corresponding to the example given in this section so that you can follow the method easily.

(a) (1) $H_0 : \mu = 110.$

(2) $H_1 : \mu \neq 110.$

(3) (To answer this question you have to test with all four levels of significance, beginning with the largest critical region.)

(i) Level of significance = 10%
critical values of z are –1.645 and 1.645
therefore, critical region is z < –1.645 or z > 1.645.

(ii) Level of significance = 5%
critical values of z are –1.96 and 1.96
therefore, critical region is z < –1.96 or z > 1.96.

(iii) Level of significance = 1%
critical values of z are –2.576 and 2.576
therefore, critical region is z < –2.576 or z > 2.576.

(iv) Level of significance = 0.1%
critical values of z are –3.29 and 3.29
therefore, critical region is z < –3.29 or z > 3.29.

(4) $\bar{x} = 112$, $s = 9$, $n = 150$, $SE = \dfrac{9}{\sqrt{150}} = 0.73$.

(5) $z = \dfrac{112 - 110}{0.73} = \dfrac{2}{0.73} = 2.74$

(i) z > 1.645, so z lies in the critical region

(ii) z > 1.96, so z lies in the critical region.

(iii) z > 2.576, so z lies in the critical region.

(iv) z < 3.29, so z does not lie in the critical region.

(6) H_o would be rejected at the 10%, 5% and 1% levels, but accepted at the 0.1% level, i.e. at the 0.1% level the sample provides evidence that the student population has an average IQ of 110.

(b) (1) $H_0 : \mu = 113.$

(2) $H_1 : \mu > 113.$

(3) Level of significance = 5%
critical value of z is 1.645
therefore, critical region is z > 1.645.

(4) As in (a), SE = 0.73.

(5) $z = \dfrac{112 - 113}{0.73} = \dfrac{-1}{0.73} = -1.37 < 1.645$

so z does not lie in the critical region.

(6) This result is not significant so accept H_o, i.e. the sample evidence suggests that the student population does not have an average IQ greater than 113.

Questions for Practice 2

(1) $H_0 : \pi = 0.6.$

(2) $H_1 : \pi \neq 0.6.$

(3) Let the level of significance = 5%
critical values of z are −1.96 and 1.96
therefore, critical region is z < −1.96 or z > 1.96.

(4) $p = \frac{400}{750} = 0.53$, $n = 750$, $\sigma_p = \sqrt{\frac{0.6 \times 0.4}{750}} = 0.018.$

(5) $z = \frac{0.53 - 0.6}{0.018} = \frac{-0.07}{0.018} = -3.89 < -1.96$

so z lies in the critical region.

(6) This result is significant to reject H_o, i.e. this sample of prosecutions suggests that the proportion of drivers pleading guilty is changing.

Chapter 14

Chi-squared Tests

Contents *Page*

Introduction **236**

A. Chi-squared as a Test of Independence **236**
Formulating the Hypotheses 237
Constructing a Contingency Table 237
Calculating the Chi-squared Statistic of the Sample 238
Determining the Appropriate Number of Degrees of Freedom 239
Ascertaining Whether the Sample Statistic Falls Inside the Acceptance Region 240
Using the Chi-squared Hypothesis Test 240

B. Chi-squared as a Test of Goodness of Fit **241**
Formulating the Null and Alternative Hypotheses 241
Calculating the Expected Frequencies 241
Calculating the Chi-squared Statistic of the Sample 243
Determining the Appropriate Number of Degrees of Freedom 244
Ascertaining Whether the Sample Statistic Falls Inside the Acceptance Region 244

Appendix: Area in the Right Tail of a Chi-squared (X^2) Distribution **245**

INTRODUCTION

In Chapter 13, we learned how to test hypotheses using data from one or two samples. We used one-sample tests to determine if a mean was significantly different from a hypothesised value and two-sample tests to determine if the difference between two means was significant. These tests are known as *parametric tests*, because they involve testing the parameters of a population, such as the mean and proportions. They use the parametric statistics of samples that come from the population being tested. To formulate these tests, we make assumptions about the population, for example, that the population is normally distributed.

There are certain kinds of data that cannot be tested in this way such as: data which was not collected in a random sample and therefore does not have a normal distribution; ordinal data; ranked data; and data from more than two populations. In business, we often encounter data of this type, such as:

- the results of a survey of which brand of washing powder consumers prefer
- an analysis of the arrival of customers at supermarket checkouts
- a survey of employees' attitudes towards performance appraisal in different departments
- a study of whether male staff have been more successful in passing professional examinations than female staff.

For these types of data, it is necessary to use tests which do not make restrictive assumptions about the shape of population distributions. These are known as *non-parametric tests*. Non-parametric tests have certain advantages over parametric tests:

- it is not necessary to assume that the population is distributed in the shape of a normal curve or in any other specific shape
- they can be used with data measured in ordinal or nominal scales, which is often the type of data obtained from business surveys, particularly where self-completion questionnaires are used
- generally, they are quicker to do and easier to understand; sometimes, formal ranking or ordering is not necessary.

But non-parametric methods are often not as precise as parametric tests, because they use less information.

In this chapter, we are going to consider one of the most commonly used non-parametric tests, called the chi-squared test (pronounced "ki-squared"). Critical values in the chi-squared test are denoted by the symbol X^2. The chi-squared test is principally used to determine:

- if two population attributes are independent of each other, or
- if a given set of data is consistent with a particular distribution, known as the "goodness of fit" test.

We will consider each of these versions of the chi-squared test in turn.

A. CHI-SQUARED AS A TEST OF INDEPENDENCE

Managers often need to know whether observable differences between several sample proportions are significant or only due to chance. For example, if the evaluation of data shows that a new method of training staff by using open learning materials results in higher outputs than the old method of on-the-job training with an experienced employee, the personnel manager may decide that the new method of training should be introduced throughout the organisation.

There is a series of steps which need to be undertaken to determine if two attributes are dependent or independent of one another:

(a) formulating the null and alternative hypotheses

(b) constructing a contingency table

(c) calculating the chi-squared statistic of the sample

(d) determining the appropriate number of degrees of freedom

(e) ascertaining whether the sample statistic falls inside the acceptance region.

We will now consider each of these in more detail, taking as an example the results of a survey of staff preferences in different locations towards two car leasing schemes.

Formulating the Hypotheses

In Chapter 13, we examined how to formulate null and alternative hypotheses in order to discover whether a certain set of data came from a particular population. In the same way, we also formulate null and alternative hypotheses to determine whether the two population attributes are independent of each other. In our example, the two hypotheses would be:

- the null hypothesis, H_0, that employees' preferences for the two car leasing schemes are independent of their work location
- the alternative hypothesis, H_1, that employees' preferences for the two car leasing schemes are not independent of their work location.

Having formulated the hypotheses, we then need to decide upon an appropriate level of significance. As we have already learned in Chapter 13, the higher the significance level we use for testing a hypothesis, the greater the probability of a Type I error, that is, of rejecting a null hypothesis when it is true. In our example, let us assume that the organisation wants to test the null hypothesis at a 10% level of significance.

Constructing a Contingency Table

The results of the survey of employees' preferences for the two car leasing schemes in different locations are presented in Table 14.1. This is known as a *contingency table*. It is made up of rows and columns, each showing a basis of classification – the rows classify the information according to preference and the columns classify the information according to location. Because this particular table has 2 rows and 3 columns, it is called a "2 × 3 contingency table" (note that the "Total" row and "Total" column are not counted).

Table 14.1: A contingency table

Preference	**Location A**	**Location B**	**Location C**	**Total**
	No. of staff	**No. of staff**	**No. of staff**	**No. of staff**
Scheme 1	76	66	61	203
Scheme 2	42	35	49	126
Total	118	101	110	329

Calculating the Chi-squared Statistic of the Sample

To evaluate the results of the survey, we start by assuming that there is no connection between the two attributes of preference and location. In other words, we assume that the null hypothesis is correct. If this is so, then we would expect the results of the staff survey to be in proportion at each location. To carry out the test, we therefore need first to calculate what the *expected* frequencies would be, assuming that there is no connection, and then to compare these with the *observed* frequencies.

We calculate the expected frequencies by first combining data from all three locations in order to estimate the overall proportion of employees who prefer each scheme. We start first by combining the data of all those who prefer Scheme 1, as follows:

$$\frac{76+66+61}{118+101+110}=\frac{203}{329}=0.6170$$

If 0.6170 is the estimate of the proportion of employees who prefer Scheme 1, then 0.3830 (1 – 0.6170) will be the estimate of the proportion of employees who prefer Scheme 2. Using these two values, we can estimate the number of employees in each location whom we would expect to prefer each of the two car leasing schemes. These calculations are shown in Table 14.2.

Table 14.2: Preference for car leasing scheme by location

	Location A	**Location B**	**Location C**
Total number sampled	118	101	110
Estimated proportion who prefer 1	× 0.6170	× 0.6170	× 0.6170
Number expected to prefer 1	72.81	62.32	67.87
Total number sampled	118	101	110
Estimated proportion who prefer 2	× 0.3830	× 0.3830	× 0.3830
Number expected to prefer 2	45.19	38.68	42.13

The expected (estimated) and observed (actual) frequencies at each location are brought together in Table 14.3.

Table 14.3: Expected and observed preferences

	Location A	**Location B**	**Location C**
Frequency preferring 1			
Observed	76	66	61
Expected	72.81	62.32	67.87
Frequency preferring 2			
Observed	42	35	49
Expected	45.19	38.68	42.13

We then calculate the chi-squared statistic of the sample, which is expressed as:

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

where: O = observed frequency, and

E = expected frequency.

To obtain X^2, we first subtract E from O for each combination of preference and location shown in Table 14.3. For example, the calculation for those preferring Scheme 1 at Location A is:

$$O - E = 76 - 72.81 = 3.19$$

When all six calculations have been carried out, we then square each of the resulting values. For example, $3.19^2 = 10.18$. Remember that when negative values are squared, the result is always a positive figure, so that, for example, $-6.87^2 = 47.20$.

Next, we divide each squared value by E. For example, 10.18 is divided by 72.81, which gives 0.1398.

Finally, the six resulting values are summed.

The results of these calculations are shown in Table 14.4.

Table 14.4: Calculating X^2

O	E	O – E	$(O - E)^2$	$\frac{(O-E)^2}{E}$
76	72.81	3.19	10.18	0.1398
66	62.32	3.68	13.54	0.2173
61	67.87	–6.87	47.20	0.6954
42	45.19	–3.19	10.18	0.2253
35	38.68	–3.68	13.54	0.3501
49	42.13	6.87	47.20	1.1203
				2.7482

The chi-squared statistic of the sample is therefore 2.7482.

Determining the Appropriate Number of Degrees of Freedom

To use the chi-squared test to ascertain whether the sample statistic falls inside or outside the acceptance region, we first have to calculate the number of *degrees of freedom* (known as *df*) in the contingency table, using the formula:

(number of rows $r - 1$) (number of columns $c - 1$)

As we saw earlier, the contingency table we are using (Table 14.1) has 2 rows and 3 columns, so the appropriate number of degrees of freedom is as follows:

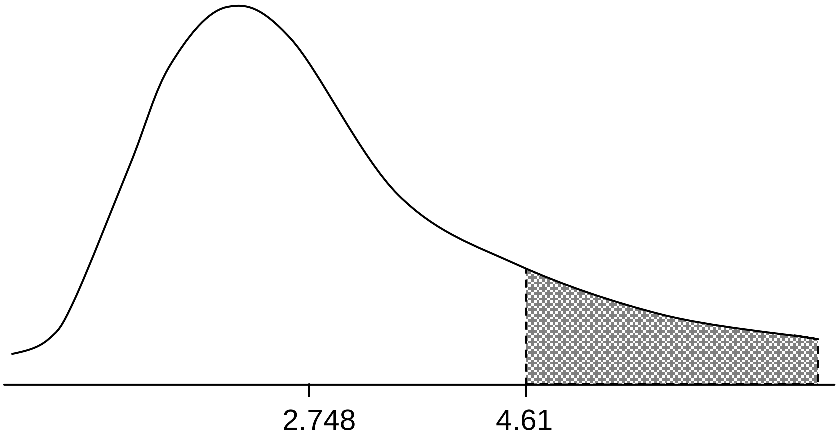
$$df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 1 \times 2 = 2.$$

Ascertaining Whether the Sample Statistic Falls Inside the Acceptance Region

We then use the statistical table for the area in the *right tail of a chi-squared distribution*, set out in the appendix to this chapter, to find the value of the chi-squared statistic. In our example, we set a significance level of 10% and therefore we look in the column headed "0.10". Then we read down to the 2 degrees of freedom row and find that the value of the chi-squared statistic is 4.61.

This means that the region to the right of the chi-squared value of 4.61 contains 10% of the area under the distribution curve, as we can see from Figure 14.1 following. Therefore the acceptance region for the null hypothesis goes from the left tail of the curve along to the chi-squared value of 4.61. The sample chi-squared value which we calculated is 2.7482 and because this value is less than 4.61, it falls within the acceptance region. We can therefore accept the null hypothesis that employees' preference about the two car leasing schemes is independent of their work location.

Figure 14.1: Right tail of the chi-squared distribution



Using the Chi-squared Hypothesis Test

When using the chi-squared hypothesis test, the size of your sample must be large enough to ensure that the theoretically correct distribution (the distribution assumed by the chi-squared table in the appendix) and your sample distribution are similar. If the expected frequencies are too small (less than about 5), the sample value of chi-squared will be overestimated and this will therefore result in too many rejections of the null hypothesis.

When interpreting the chi-squared statistic of a sample, a large value, such as 20, indicates a big difference between observed and expected frequencies, whereas a value of zero indicates that the observed and expected frequencies exactly match. If you find that the sample value of chi-squared is at or near zero, but you believe that a difference between observed and expected values should exist, it is sensible to re-examine the data and the way it was gathered and collated, to make sure that any genuine differences have not been missed.

B. CHI-SQUARED AS A TEST OF GOODNESS OF FIT

The chi-squared test can also be used to determine if there is a similarity – that is, a *good fit* – between the distribution of observed data and a theoretical probability distribution, such as the normal distribution (which we considered in Chapter 12) or the binomial or Poisson distributions (which we considered in Chapter 11). We first perform the chi-squared test to establish whether there is a significant difference between our observed distribution and the theoretical distribution we have chosen; this information then enables us to decide whether the observed data is a sample from our hypothesised theoretical distribution. The chi-squared goodness of fit test is therefore a useful tool for managers, who often need to make decisions on the basis of statistical information. For example, a maintenance manager at a factory may use information about the frequency of breakdowns to decide how many engineers to deploy on each shift.

There is a series of steps which need to be undertaken to determine goodness of fit:

(a) formulating the null and alternative hypotheses

(b) calculating the expected frequencies

(c) calculating the chi-squared statistic of the sample

(d) determining the appropriate number of degrees of freedom

(e) ascertaining whether the sample statistic falls inside the acceptance region.

We will now consider each of these in more detail, taking as an example data collected on the amount of money consumers spend on chocolate each week.

Formulating the Null and Alternative Hypotheses

The sales manager of a confectionery manufacturer has commissioned a survey of the amount of money consumers spend on chocolate each week. The sales manager believes that the variable – the amount of money spent – may be approximated by the normal distribution, with an average of £5.00 per week and a standard deviation of £1.50.

The hypotheses would therefore be:

- the null hypothesis, H_0, that a normal distribution is a good description of the observed data
- the alternative hypothesis, H_1, that a normal distribution is not a good description of the observed data.

Let us assume that the sales manager wants to test the null hypothesis at a 10% level of significance.

Calculating the Expected Frequencies

The results of the consumer survey are shown in Table 14.5.

Table 14.5: Results of consumer survey

Weekly expenditure £	**Number of consumers**
< £2.60	12
£2.60 – £3.79	60
£3.80 – £4.99	82
£5.00 – £6.19	104
£6.20 – £7.39	24
≥ £7.40	18
Total	300

To determine what the expected frequencies would be under a normal distribution, we use the same techniques as we have already used in Chapter 12 to ascertain areas under the normal distribution curve.

First, we start with the formula:

$$z = \frac{x - \mu}{\sigma}$$

where: x = value of the random variable

μ = mean of the distribution of the random variable

σ = standard deviation of the distribution

z = number of standard deviations from x to the mean.

Then we look up the value given for z using the standard normal table set out as the appendix to Chapter 12.

Finally, we multiply this value by the size of the sample to obtain the expected frequency.

For example, to calculate the expected frequency of consumers spending less than £2.60 per week under a normal distribution, we first obtain z as follows:

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{2.59 - 5}{1.5} = \frac{-2.41}{1.5} = -1.61.$$

Then we look up the z value of 1.61 in the appendix to Chapter 12 (we can disregard the minus sign, because a normal distribution is symmetrical), which gives 0.4463. This is the area under a normal curve between the mean and £2.60. Since the left half of a normal curve (between the mean and the left-hand tail) represents an area of 0.5, we can obtain the area below £2.60 by subtracting 0.4463 from 0.5, which gives 0.0537.

Next, we multiply 0.0537 by 300 (the size of the sample population), to obtain the expected frequency. This gives a value of 16.11. Under a normal distribution, the expected frequency of consumers spending less than £2.60 per week on chocolate is therefore 16.11.

The expected frequencies calculated for each class of expenditure are shown in Table 14.6.

Table 14.6: Expected frequencies by class of expenditure

Weekly expenditure *£*	**Observed frequency**	**Normal probability**	**Population**	**Expected frequency**
< £2.60	12	0.0537	× 300	16.11
£2.60 – £3.79	60	0.1571	× 300	47.13
£3.80 – £4.99	82	0.2881	× 300	86.43
£5.00 – £6.19	104	0.2852	× 300	85.56
£6.20 – £7.39	24	0.1560	× 300	46.8
≥ £7.40	18	0.0548	× 300	16.4
Total	300			298.43

Calculating the Chi-squared Statistic of the Sample

We can now calculate the chi-squared statistic of the sample, using the formula:

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

where:

O = observed frequency, and

E = expected frequency.

The calculations are shown in Table 14.7 following.

Table 14.7: Calculating the Chi-squared statistic

O	E	O – E	$(O - E)^2$	$\frac{(O-E)^2}{E}$
12	16.11	–4.11	16.89	1.0484
60	47.13	12.87	165.64	3.5145
82	86.43	–4.43	19.63	0.2271
104	85.56	18.44	340.03	3.9742
24	46.8	–22.8	519.84	11.1077
18	16.4	1.6	2.56	0.1561
				20.03

The chi-squared statistic of the sample is therefore 20.03.

Determining the Appropriate Number of Degrees of Freedom

The number of degrees of freedom is determined by the number of classes (expressed as k), for which we have compared the observed and expected frequencies. However, the actual number of observed frequencies that we can freely specify is $k - 1$, because the last one is always determined by the size of the sample. One additional degree of freedom must also be subtracted from k for each population parameter that has been estimated from the sample data.

In our example, there are six classes, so $k = 6$, but because the total number of observed frequencies must add up to 300, the actual number of frequencies that we can freely specify is only 5. We therefore subtract one degree of freedom from k, so the number that we can freely specify is $k - 1$, which is 5.

Furthermore, we had to use the sample mean to estimate the population mean, so we must subtract another degree of freedom, leaving $k = -2$, which is 4.

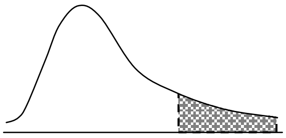
Finally, we also had to use the sample standard deviation to estimate the population standard deviation, so we must subtract yet another degree of freedom, leaving $k - 3$, which is 3.

Ascertaining Whether the Sample Statistic Falls Inside the Acceptance Region

We then use the statistical table for the area in the right tail of a chi-squared distribution, set out in the appendix, to find the value of the chi-squared statistic. In our example, we set a significance level of 10% and therefore we look in the column headed "0.10". Then we read down to the 3 degrees of freedom row and find that the value of the chi-squared statistic is 6.25.

This means that the region to the right of the chi-squared value of 6.25 contains 10% of the area under the distribution curve. Therefore the acceptance region for the null hypothesis goes from the left tail of the curve along to the chi-squared value of 6.25. The sample chi-squared value which we calculated is 20.03 and because this value is greater than 6.25, it falls outside the acceptance region. We therefore cannot accept the null hypothesis that a normal distribution is a good description of the observed frequencies.

APPENDIX: AREA IN THE RIGHT TAIL OF A CHI-SQUARED (χ^2) DISTRIBUTION



Degrees of freedom	0.25	0.10	0.05	0.025	0.01	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	22.7	27.2	30.1	32.9	36.2	38.6	43.8
20	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	28.2	33.2	36.4	39.4	43.0	45.6	51.2
25	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	77.6	85.5	90.5	95.0	100.0	104.0	112.0
80	88.1	96.6	102.0	107.0	112.0	116.0	125.0
90	98.6	108.0	113.0	118.0	124.0	128.0	137.0
100	109.0	118.0	124.0	130.0	136.0	140.0	149.0

Chapter 15

Decision-making

Contents *Page*

Introduction **248**

A. Decision-making Under Certainty **248**
- Outcome Measures 248
- Valuation of Measures 249
- Optimal Outcome 249

B. Definitions **249**

C. Decision-making Under Uncertainty **250**

D. Decision-making Under Risk **252**
- Expected Monetary Value 252
- Decision Trees 253

E. Complex Decisions **255**
- Interpreting Expectations 260

INTRODUCTION

The one activity which distinguishes management from the other functions in an organisation is decision-making. We can study decision-making from many aspects; in many circumstances decisions are taken for reasons which can only be studied by the use of psychology or sociology, but we can also consider what is called "rational decision-making", and this makes use of quantitative measures to a large degree.

The purpose of quantitative methods of presenting and analysing data is to provide the decision-maker with a basis on which to make his or her decisions. This of course tends to make decision-making more difficult. The reason is that, in the absence of good information as to the possible consequences of a decision, decision-making becomes a matter of choosing between those alternative courses of action which are before the decision-maker. A person who makes decisions in a confident way may gain a good reputation as a decision-maker; most people tend to prefer someone who is resolute and confident to someone who hesitates and tries to weigh up the alternatives. Having good leadership qualities in this manner does not, unfortunately, mean that the leader always makes correct decisions. If there is an extended interval between taking a decision and finding out what its ultimate consequences are, then the leader may escape censure for a bad decision. It may take someone studying the history of the events to make a clear judgement, by which time the person who made the decision will have moved on to other things!

When good quantitative information is available and presented in a meaningful way, it becomes a great deal easier to see what the consequences of different courses of action are. They can then be compared and a decision taken on the judgement as to which course of action is likely to have the best outcome. The more that is known about the possible outcomes, the more we have to try to bring the measures of success to a common basis, typically money. Attempts to do so when we are considering what is best for a community rather than just a commercial organisation have led to the "cost-benefit analysis".

It is convenient to classify decision situations under three headings: decisions under certainty; decisions under risk; and decisions under uncertainty. We will look first at decisions under certainty.

A. DECISION-MAKING UNDER CERTAINTY

Here the outcome to a choice of course of action (a strategy) can be evaluated and the same result will always occur for that choice of strategy. As the outcomes are determined, we call this a *deterministic situation*. When all possible strategies can be listed and their outcomes evaluated, the rational decision-maker simply chooses the strategy which gives the best value to the outcome. The fact that there are many real-life situations in which people could take decisions in this manner but do not, either because they do not assess all strategies, or simply by perversity, shows the need to undertake behavioural studies.

There are three problems for the decision-maker and his or her quantitative advisors:

- to find an appropriate measure by which to evaluate outcomes
- to calculate the value of the measure for an outcome
- to identify the best outcome from a large number of outcomes.

Outcome Measures

In most organisational situations there are two popular measures – "How much?" and "How long?". As most commercial decisions resolve to a matter of money or "the bottom line", the first of these is the dominant measure. In non-commercial organisations other measures may apply.

As the concept of organisational decision-making originated in wartime, we can look at the choices facing a military commander when deciding what to do next. One such problem concerned the supply of war materials from the USA to the UK and Europe across the Atlantic Ocean during World War II. The possible measures were to:

- maximise the total tonnage of goods delivered
- minimise the time taken to deliver goods
- minimise the loss of shipping.

They are conflicting objectives to some degree, and the "best" answer to the problem is not easy to see. In the event the answers adopted were related to the shipping available. The very fast ships, such as the liner *Queen Mary*, were used to minimise time, relying on their great speed to evade attack. Most other ships were operated in convoys, to minimise shipping losses, and to satisfy to some degree the total tonnage requirement.

Valuation of Measures

This is where the various mathematical techniques available come into their own. The characteristic of such methods is that there is a "right" answer to the question posed.

Optimal Outcome

Some situations involve a very large number of choices. Even to list them may be a daunting task, and to evaluate them all may be lengthy and costly. Fortunately some excellent methods and algorithms have been developed, which can ensure that we consider all alternatives and identify the "best" outcome. Critical path analysis is of this nature, providing a method of systemising our knowledge so that the effect of choices of order of doing things can be seen. Another very powerful method looks at the assignment of resources between a number of competing uses. This is "linear programming". We do not study linear programming (LP) in this course but you should know that it exists. It is used in areas such as:

- investment decisions
- product mix situations
- product formulation
- production scheduling
- delivery scheduling.

B. DEFINITIONS

Before considering some criteria for making decisions under conditions of uncertainty and risk, we need to establish the following definitions:

(a) Decision Alternatives

The term *"decision alternatives"* refers to the list of choices available to a decision-maker. For example, an investor may have to decide whether to purchase shares or bonds or both, which would represent three decision alternatives.

(b) States of Nature

The future environmental conditions that influence the desirability of decision alternatives are referred to as *states of nature*. For example, whether the stock market is rising, falling or steady would be likely to influence the desirability of investing in shares or bonds, and so would represent three possible states of nature facing an investor. For any given decision problem, the states of nature should be exhaustive and mutually exclusive.

(c) Uncertainty

Under conditions of *uncertainty*, we assume that the possible states of nature are known to the decision-maker, but their probabilities are unknown. Most decisions in business and management are of this type, and probabilities often have to be estimated from past experience, or the states of nature have to be regarded as equally likely.

(d) Risk

Under conditions of *risk*, we assume that the possible states of nature are known and that the probabilities that they will occur are also known.

(e) Payoff Table

A payoff table has columns that represent the possible *states of nature* and rows that represent the *decision alternatives*. Each cell of the table gives the *payoff* (usually a revenue, profit or loss) associated with the given decision alternative and state of nature.

(f) Decision Tree

A decision tree is a diagrammatic representation of a decision problem, which shows the decision alternatives, the possible states of nature, the state of nature probabilities (if known) and the payoffs. It is an alternative way of presenting the information in a payoff table.

C. DECISION-MAKING UNDER UNCERTAINTY

We will investigate the problem of decision-making under uncertainty by examining a simple business problem. Suppose a UK computer manufacturer produces laptop computers for the Italian market in its UK factory, but recognises that future changes in the sterling-euro exchange rate may make the laptops uncompetitive in the Italian market. To reduce this risk, the company is considering the possibility of shifting production to a new factory in France. The *decision alternatives* are:

1. *UK production*: continuing the production of the laptops for the Italian market in the UK.
2. *French production:* shifting production of the laptops for the Italian market to the new factory in France.

The outcomes of the decisions are assumed to depend on what happens to the value of the pound relative to the euro over the next five years. If the pound should weaken against the euro, the profits from shifting production to France are likely to be higher, but if the pound should strengthen against the euro, the profits from shifting production to France are likely to be lower. So in this example, the *states of nature* are:

1. *Weak £*: over the next five years, the pound weakens against the euro.
2. *Stable £*: over the next five years, the pound is unchanged against the euro.
3. *Strong £*: over the next five years, the pound strengthens against the euro.

We assume for now that the probabilities of the pound being weak, stable or strong are unknown. However, the likely payoffs under each state of nature are known and these are shown in the payoff table (Table 15.1).

Table 15.1: Payoff table

		States of Nature		
		Weak £	*Stable £*	*Strong £*
Decision Alternatives	*UK production*	5	7	12
	French production	10	9	8

Payoffs are profits in £ million

We consider four possible decision-making criteria: the *Laplace*, *maximin*, *maximax* and *minimax regret* criteria.

(a) Laplace criterion: choose the decision alternative with the highest mean payoff

According to the Laplace criterion, the various states of nature are assumed to be equally likely, and the decision-maker is required to choose the decision alternative with the highest mean payoff. In our example: if UK production is maintained, the mean payout is £8 million; but if production shifts to France, the mean payoff is £9 million. In this case, the Laplace criterion requires the decision-maker to choose the "French production" decision alternative – i.e. to shift production of the laptops for the Italian market to the new French factory.

(b) Maximin criterion: choose the decision alternative with the highest minimum payoff

The maximin criterion requires the decision-maker to choose the decision alternative with the highest minimum payoff. In other words, the decision-maker is pessimistic and wishes to avoid the lowest possible return. In our example, if production continues in the UK, the minimum possible payoff is £5 million, but if production shifts to France, the minimum payoff is £8 million. Therefore in this case the maximin criterion also requires the decision-maker to choose the "French production" decision alternative – i.e. to shift production of the laptops for the Italian market to the new French factory.

(c) Maximax criterion: choose the decision alternative with the highest maximum payoff

The maximax criterion requires the decision-maker to choose the decision alternative with the highest maximum payoff. In other words, the decision-maker is optimistic and aims for the highest possible return. In our example, if production continues in the UK, the maximum possible payoff is £12 million, but if production shifts to France, the maximum payoff is £10 million. Therefore in this case the maximax criterion requires the decision-maker to choose the "UK production" decision alternative – i.e. to maintain production of the laptops for the Italian market in the UK factory.

(d) Minimax regret criterion: choose the decision alternative with the lowest maximum regret

To apply the minimax regret criterion, we first have to convert the payoff table to a "regret table". Each cell of the regret table contains a "regret" defined as *the difference between the payoff in that cell and the maximum payoff for that particular state of nature*. The minimax regret criterion then requires the decision-maker to choose the decision alternative with the lowest maximum regret. For our example, the regret table is shown in Table 15.2.

From the regret table, we can see that if production continues in the UK, the maximum possible regret is £5 million, but if production shifts to France, the maximum regret is £4 million. Therefore in this case the minimax regret criterion requires the decision-maker to choose the "French production" decision alternative – i.e. to shift production of the laptops for the Italian market to the new French factory.

Table 15.2: Regret table

		States of Nature		
		Weak £	*Stable £*	*Strong £*
Maximum payoff for each state of nature		10	9	12
Decision Alternatives	*UK production*	5	2	0
	French production	0	0	4

Payoffs are profits in £ million

D. DECISION-MAKING UNDER RISK

Under conditions of risk, we assume that decisions are taken with full knowledge of the probabilities associated with the possible states of nature. Let us then suppose that we know that:

- the probability of a relatively weak pound over the next five years is 0.2
- the probability of a relatively stable pound over the next five years is 0.5
- the probability of a relatively strong pound over the next five years is 0.3.

As the three states of nature are exhaustive and mutually exclusive, these three probabilities must sum to one (i.e. it is a certainty that the pound will be relatively weak, stable or strong). These probabilities enable us to calculate the *expected monetary values* of each decision alternative.

Expected Monetary Value

Given a set of n possible monetary values (say, x_1, x_2, ... x_n) and n probabilities associated with each monetary value (say, p_1, p_2, ... p_n), the *expected monetary value (EMV)* is calculated as the sum of the products of the monetary values and their associated probabilities (i.e. $p_1x + p_2x_2 + ... + p_nx_n$). So we can write:

$$\text{EMV} = p_1x_1 + p_2x_2 + \ ... + p_nx_n = \sum p_ix_i \ \text{ for } i = 1, 2, ... n$$

where x_i represents the i^{th} monetary value and p_i represents the probability associated with the i^{th} monetary value.

With regard to decision-making problems, the x's are the payoffs resulting from each decision alternative for a given state of nature and the p's are the state of nature probabilities. Returning to our example, the payoff table with the probabilities included and the calculated EMVs is shown in Table 15.3.

Table 15.3: Payoff table with probabilities and EMVs

		States of Nature			EMVs
		Weak £	*Stable £*	*Strong £*	
Probabilities		0.1	0.5	0.4	
Decision Alternatives	*UK production*	5	7	12	8.8
	French production	10	9	8	8.7

Payoffs are profits in £ million

The EMV of the decision to maintain production in the UK (denoted EMV_{UK}) is calculated as:

$$\text{EMV}_{\text{UK}} = (0.1 \times 5) + (0.5 \times 7) + (0.4 \times 12) = 8.8 \text{ (i.e. £8.8 million)}$$

while the EMV of the decision to shift production to France (denoted EMV_{F}) is calculated as:

$$\text{EMV}_{\text{F}} = (0.1 \times 10) + (0.5 \times 9) + (0.4 \times 8) = 8.7 \text{ (i.e. £8.7 million).}$$

So in this case, where a strong pound has a higher probability (0.4) than a weak pound (0.1), the decision alternative with the higher EMV is "UK production" – i.e. continue producing the laptops for the Italian market in the UK.

However, note that if the probabilities were changed so that a strong pound was less probable than a weak pound, the EMVs would also change. Suppose that:

- the probability of a relatively weak pound over the next five years is 0.4
- the probability of a relatively stable pound over the next five years is 0.5
- the probability of a relatively strong pound over the next five years is 0.1.

Now the EMV of the decision to maintain production in the UK (denoted EMV_{UK}) would be calculated as:

$$\text{EMV}_{\text{UK}} = (0.4 \times 5) + (0.5 \times 7) + (0.1 \times 12) = 6.7 \text{ (i.e. £6.7 million)}$$

and the EMV of the decision to shift production to France (denoted EMV_{F}) would be calculated as:

$$\text{EMV}_{\text{F}} = (0.4 \times 10) + (0.5 \times 9) + (0.1 \times 8) = 9.3 \text{ (i.e. £9.3 million).}$$

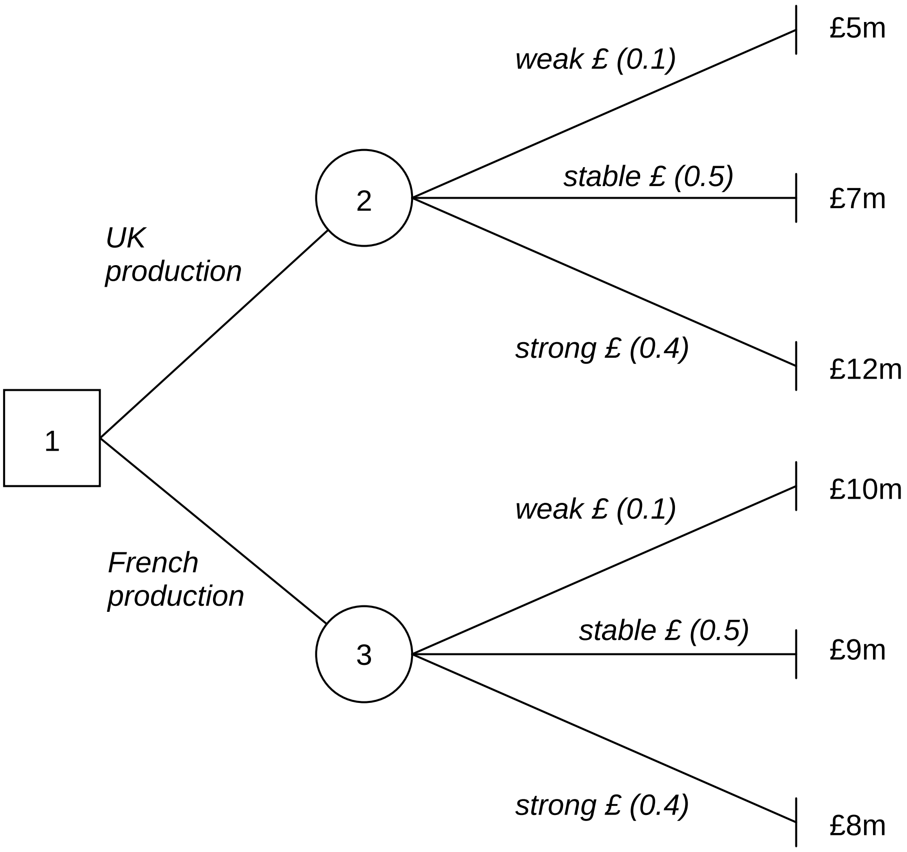
In this case, the decision alternative with the higher EMV would be "French production" – i.e. shift production of the laptops for the Italian market to France.

Decision Trees

A decision problem illustrated in a payoff table (such as that in Table 15.3) may also be illustrated by means of a *decision tree*, which shows diagrammatically the decision alternatives, the states of nature, the payoffs and the payoff probabilities.

The decision tree for the decision problem facing our UK computer manufacturer is shown in Fig. 15.1.

Figure 15.1: Decision tree for the UK computer manufacturer



In a decision tree, the decisions and states of nature are depicted by branches and the points at which decisions are taken or at which states of nature occur are depicted by nodes. There are three types of nodes:

- *decision nodes* that depict decision points (represented by square boxes)
- *chance nodes* that depict points at which states of nature occur (represented by circles)
- *payoff nodes* that depict the final payoffs (represented by vertical lines).

The computer manufacturer has to decide between UK production and French production of the laptops for the Italian market. The point at which this decision is taken is depicted by decision node 1 (i.e. the square box labelled 1 on the left-hand side of the decision tree). If UK production is chosen, the payoff depends on the state of nature that occurs. This is depicted by the chance node 2 (i.e. the circle labelled 2 in the diagram). Similarly, if French production is chosen, the payoff also depends on the state of nature that occurs. This is depicted by the chance node 3 (i.e. the circle labelled 3 in the diagram). Finally, the vertical lines at the right-hand side of the decision tree depict the final payoffs associated with each decision and each state of nature. The state of nature probabilities are also shown on the appropriate branches of the decision tree. The decision tree provides the same information as the payoff table and can be used in the same way to compute the EMVs for each decision alternative.

E. COMPLEX DECISIONS

Some decisions are taken as part of a set, and if there is a mix of decisions and ranges of outcomes, it can become difficult to see what is happening. The problem of tendering for contracts is of this type.

As an example, consider the case of a small builder who is invited to tender for a contract. The builder knows that there is a second tender coming up, more valuable than the first, but that unless he puts in a tender for the first he will not be invited to tender for the second. If he wins the first contract he must complete it and cannot bid for the second. He knows that his costs on the contracts will depend on the weather; also that his chance of winning the contract will depend on his bid.

He makes estimates as follows:

Contract A

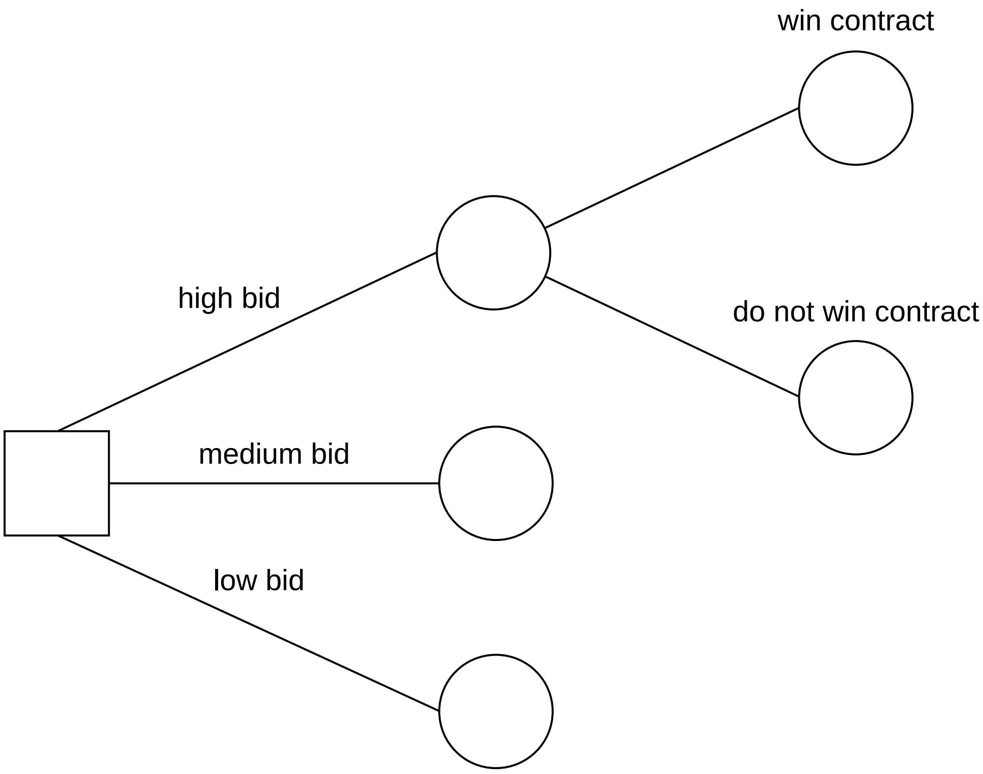
Bid level high:	0.2 chance of winning the contract
Bid level medium:	0.6 chance of winning the contract
Bid level low:	0.9 chance of winning the contract
Bid level high:	Profit: £60,000
Bid level medium:	Profit: £40,000
Bid level low:	Profit: £25,000
Effect of weather:	
Very bad weather	Reduce profits by £15,000
Poor weather	Reduce profits by £5,000
Probability of very bad weather	= 0.2
Poor weather	= 0.4

Contract B

Bid level high:	0.3 chance of winning the contract
Bid level medium:	0.7 chance of winning the contract
Bid level low:	0.9 chance of winning the contract
Bid level high:	Profit: £100,000
Bid level medium:	Profit: £80,000
Bid level low:	Profit: £60,000
Effect of weather:	
Very bad weather	Reduce profits by £20,000
Poor weather	Reduce profits by £10,000
Probability of very bad weather	= 0.2
Poor weather	= 0.4

It costs £5,000 to prepare a bid, which has been allowed for in all the profit figures. What should he do? We can draw a decision tree to review the choices and possible outcomes. A partial decision tree for this problem is shown in Figure 15.2.

Figure 15.2: Section of a tree only

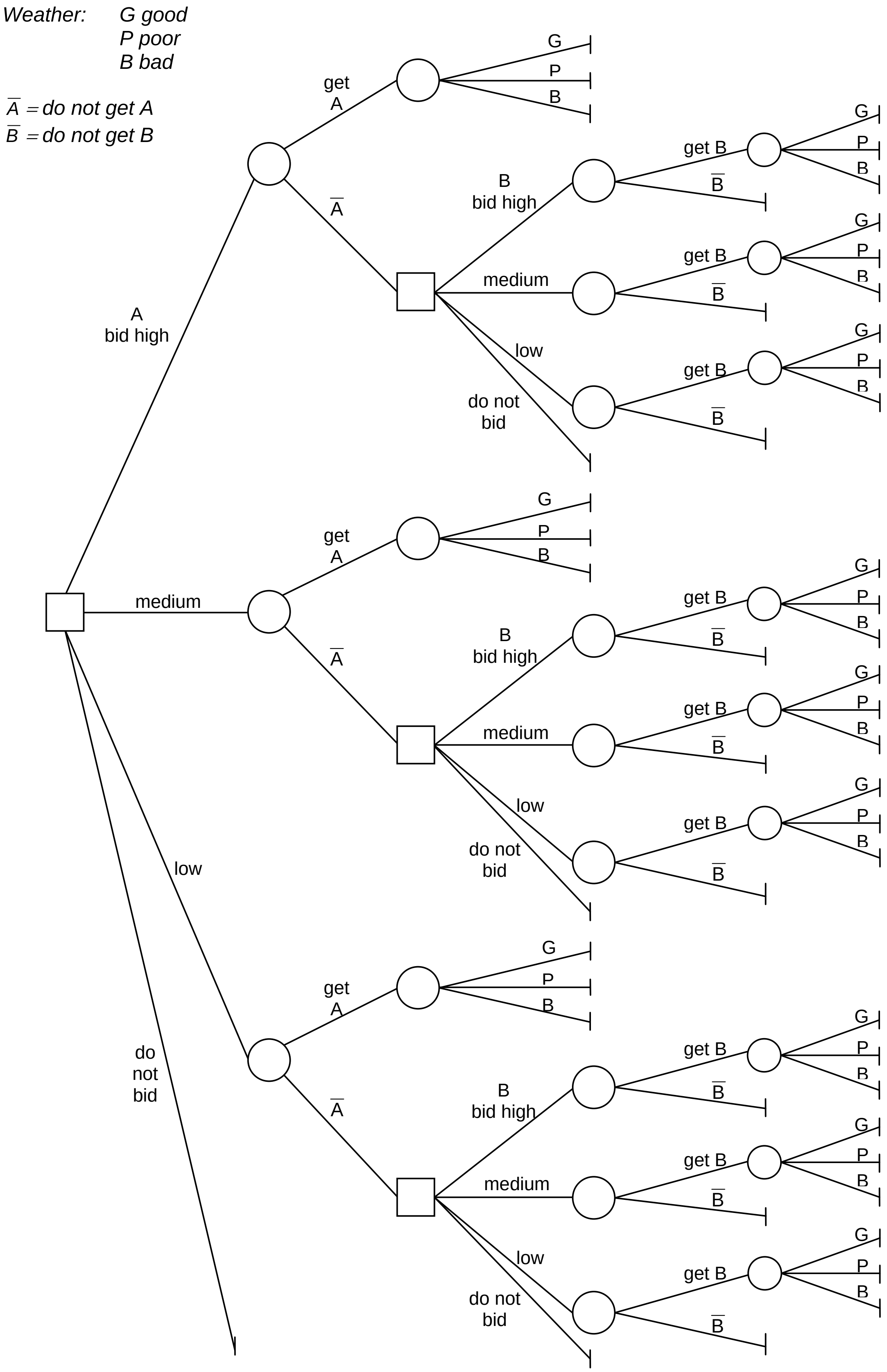


Activity 1

Without looking at Figure 15.3, try to draw the complete tree for yourself on the basis of the information given in the text.

Then check your version against Figure 15.3.

Figure 15.3: Completed tree



For each "twig" on the extreme right of the tree in Figure 15.3, we can calculate the value of the outcome. For example, take the sequence:

1. bid high for first contract; do not get first contract
2. bid high for second contract; get second contract
3. experience poor weather.

The values are:

1. –£5,000 – cost of first bid
2. £100,000 – profit with high bid
3. –£10,000 – reduction for poor weather.

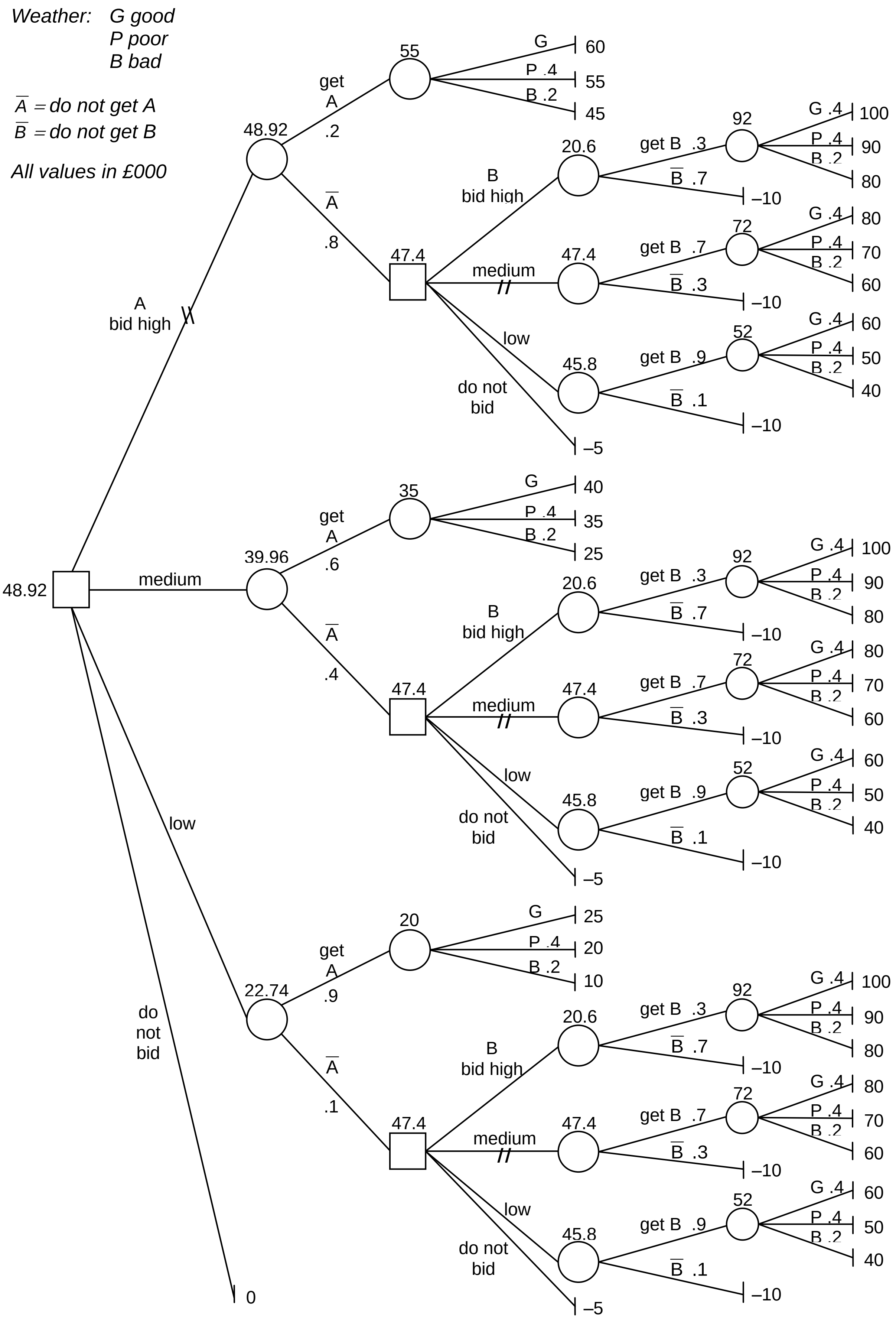
Hence net profit is £85,000.

Activity 2

Now work out the values for all the other twigs in Figure 15.3.

Then check your answer with the tree in Figure 15.4.

Figure 15.4: Completed tree showing values of outcomes



We can now work through the tree from right to left and, at each chance node, find the expectation for that node and, for each decision node, choose the strategy which gives the highest expectation.

Check the tree carefully so that you can see where all the figures come from. Note that decisions are shown by putting a cross stroke on the branches chosen.

We find that the best decision is to make a high bid for the first contract; if this is not successful, make a medium bid for the second contract. The expectation of profit with this choice of strategies is £48,920.

You may notice that the second part of the tree, concerning the second contract, repeated itself three times. We could have evaluated this as a separate decision tree and entered the result, which was medium bid each time, in a simpler tree for the first bid. Sometimes successive decisions interact; the profit on the second contract might have been linked to what happened in the first, in which case the tree must be analysed as a unit.

Interpreting Expectations

It is important to realise what an expectation, such as the answer to the last problem, actually means.

When the builder has completed a contract, the profit will be one of the figures at the ends of the twigs in the tree. The exact figure will depend on the builder's decisions, which contract the builder gets and the weather. The figure of £48,920 we found is the average profit the builder would make if it were possible to repeat the contract many times, following the rule we derived. Of course in real life this is not possible, so why bother? The answer is that if a company is in a situation where similar decisions have to be taken frequently, then it is reasonable to take an average view. The company can afford to lose money on some contracts as long as on average, taken over many contracts, it makes a profit. The analysis is then valid. In other words, for a large construction company which bids for many contracts the analysis is suitable. Compare this to the situation of our small builder. If he is short of cash he may be worried that if he does not get the contract, he will be unable even to pay the cost of preparing the bid.

Chapter 16

Applying Mathematical Relationships to Economic and Business Problems

Contents ***Page***

A. Using Linear Equations to Represent Demand and Supply Functions 262

The Demand Function 262

The Determination of Equilibrium Price and Quantity 263

Shifts in the Demand and Supply Functions 265

B. The Effects of a Sales Tax 267

C. Breakeven Analysis 269

Calculation of Breakeven Point 269

Formulae 269

D. Breakeven Charts 271

Information Required 271

Cost/Volume Chart 271

Profit/Volume Chart 273

Margin of Safety 274

Assumptions and Limitations of Breakeven Charts 274

E. The Algebraic Representation of Breakeven Analysis 275

Using Linear Equations to Represent Cost and Revenue Functions 275

The Breakeven Point 275

Changes in the Cost and Revenue Functions 277

Calculating Profit at Different Output Levels 277

Answers to Question for Practice 279

A. USING LINEAR EQUATIONS TO REPRESENT DEMAND AND SUPPLY FUNCTIONS

The Demand Function

First, if you feel your algebra is a little rusty, then you might find it useful to revisit the appendix to Chapter 4 before continuing with this chapter.

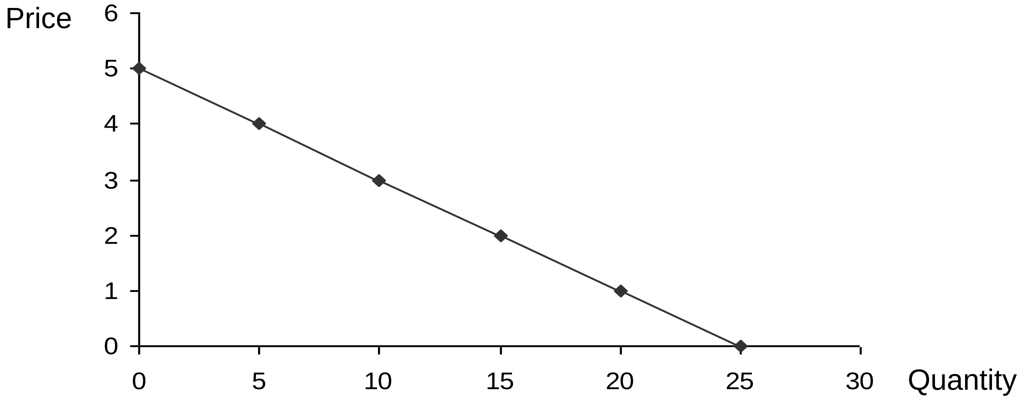
You are already familiar with the concept of the demand curve, which shows how much of a product will be demanded by consumers at a range of possible prices. The demand curve is obtained by taking data on the amount demanded at different prices, known as the demand schedule, and plotting the resulting values on a graph. An example of a demand schedule is shown in Table 16.1.

Table 16.1: A demand schedule

Price	Quantity
1	20
2	15
3	10
4	5
5	0

The demand schedule is shown in graphical form, that is, in the form of a demand curve, in Figure 16.1. (Confusingly, a depiction of a demand or supply schedule is always known as a "curve", even when it is a straight line!)

Figure 16.1: A demand curve



Because the demand curve is a straight line, we can express it in the form of a linear equation. The line has a negative slope and therefore it is expressed as:

$$q = a - bp$$

where: q = quantity demanded

p = price

and a and b are constants.

For example, the demand function shown above can be expressed as:

$$q = 25 - 5p$$

Sometimes, the equation for the demand function may be given, and from this, we can determine the quantity that will be demanded at any particular price. For example, if the demand function is:

$$q = 50 - 2p$$

we can calculate the values of q at a range of different values of p. These are shown in the next table.

Table 16.2: Quantities from a known demand function

Price	Quantity
5	40
10	30
15	20
20	10
25	0

The equation can also be written to express price as a function of quantity, so we can use it to determine the price at which particular quantities will be demanded. In the example above, the equation would be rewritten as:

$$p = 25 - 0.5q$$

The Determination of Equilibrium Price and Quantity

Supply functions can also be expressed as equations, in the same way as demand functions. As you will recall from your studies of Economics, in a competitive market we assume that price is one of the most important influences on the quantity supplied: as the price of a product increases, the quantity supplied will increase. We can therefore state that supply is a function of price and we can express the supply function mathematically as:

$$q = c + dp$$

where: $q =$ quantity supplied

$p =$ price

and c and d are constants.

Note the $+$ sign in the equation, because the supply curve has a positive slope.

If we consider the equations for the demand and supply functions together, we can determine the equilibrium price and quantity. The two equations contain like terms, so we can treat them as simultaneous equations and proceed to solve them for the unknown variables, p and q. For example, let us consider the following demand and supply functions:

Demand function:

$$q^d = 50 - 2p$$

Supply function:

$$q^s = 2 + 4p$$

The equilibrium condition is that quantity demanded is equal to quantity supplied, which can be expressed as:

$$q^d = q^s$$

(Note that here the superscripts d and s are not powers (indices) of q; they are just used to distinguish the quantity demanded from the quantity supplied.) Considering these as simultaneous equations, we can proceed to solve them, as follows:

$$50 - 2p = 2 + 4p$$

$$4p + 2p = 50 - 2$$

$$6p = 48$$

$$p = 8$$

The equilibrium price is therefore £8.

The equilibrium quantity can then be determined by substituting $p = 8$ into one of the equations. Let us take the equation of the demand function as an example:

$$q^d = 50 - 2p$$

$$q^d = 50 - (2 \times 8) = 50 - 16 = 34$$

The equilibrium quantity is therefore 34 units.

We can check this by performing the same operations on the equation of the supply function:

$$q^s = 2 + 4p$$

$$q^s = 2 + (4 \times 8) = 2 + 32 = 34$$

The operations which we have performed can be expressed algebraically as follows:

$q^d = a + bp$ where b is less than 0

$q^s = c + dp$ where d is greater than 0

$$q^d = q^s$$

To obtain the equilibrium price:

$$a + bp = c + dp$$

which can be rearranged to give:

$$bp - dp = c - a$$

$$p = \frac{c - a}{b - d}$$

The equilibrium quantity can now be obtained from either the demand or supply equation. Using the demand equation, we have:

$$q^d = a + b\left(\frac{c - a}{b - d}\right)$$

$$= \frac{a(b - d) + b(c - a)}{b - d}$$

$$= \frac{ab - ad + bc - ba}{b - d}$$

$$= \frac{bc - ad}{b - d}$$

Shifts in the Demand and Supply Functions

Changes in the demand for a product are brought about by a number of different influences. It is assumed that the price of the product is the principal influence, but there are also others, such as the prices of other products (complementary goods and substitutes), income, tastes and expectations, which were explored in your study of economics. We also noted that because a normal graph can only depict the relationship between two variables – in this case, quantity and price – a change in the quantity demanded which is caused by a change in one of the other influences has to be shown graphically by a shift in the whole demand curve.

In the same way that the demand function can be expressed mathematically through an equation or a graph, shifts in the demand function can also be expressed through equations or graphs. Let us consider again the following demand and supply functions:

Demand function

$$q^d = 50 - 2p$$

Supply function

$$q^s = 2 + 4p$$

The demand schedule for a range of possible prices can be calculated from the equation, as shown in Table 16.3.

Table 16.3: Demand schedule

Price	Quantity
1	48
2	46
3	44
4	42
5	40
6	38
7	36
8	34
9	32
10	30

Let us suppose that there is then a change in consumer tastes away from the product. As a result, the quantity demanded at any particular price will be less than before. The new equation for demand is:

$$q^d = 44 - 2p$$

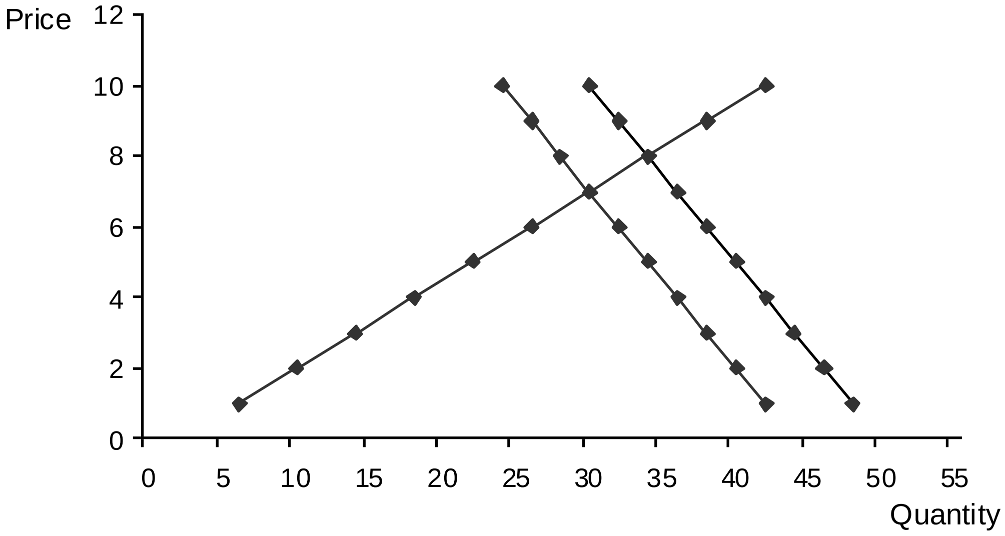
The new demand schedule is shown in Table 16.4.

Table 16.4: New demand schedule

Price	Quantity
1	42
2	40
3	38
4	36
5	34
6	32
7	30
8	28
9	26
10	24

The old and the new demand schedules are shown in graphical form in Figure 16.2 below. As you can see, the demand curve has shifted to the left, showing that at each particular price, demand for the product will be less. We can read the new equilibrium price and quantity off the graph – the new equilibrium price is £7 and the new equilibrium quantity is 30.

Figure 16.2: Old and new demand schedules



We can also use equations to find the new equilibrium price and quantity, in exactly the same way as we did to find the old equilibrium price and quantity.

The new demand function is:

$$q^d = 44 - 2p$$

The supply function is:

$$q^s = 2 + 4p$$

The equilibrium condition is :

$$q^d = q^s$$

Considering these as simultaneous equations, we can proceed to solve them, as follows.

$$44 - 2p = 2 + 4p$$

$$4p + 2p = 44 - 2$$

$$6p = 42$$

$$p = 7$$

The new equilibrium price is therefore £7.

The equilibrium quantity can then be determined by substituting $p = 7$ into one of the equations. Let us take the equation of the demand function as an example:

$$q^d = 44 - 2p$$

$$q^d = 44 - (2 \times 7) = 44 - 14 = 30$$

The new equilibrium quantity is therefore 30 units.

If we want to analyse the effects of shifts in the demand curve to the right, or the effects of shifts in the supply curve, we can use exactly the same method.

B. THE EFFECTS OF A SALES TAX

To show the effects of imposing a sales tax on a product, reconsider the original demand and supply functions:

Demand function

$$q^d = 50 - 2p$$

Supply function

$$q^s = 2 + 4p$$

which may be written in terms of price as:

Demand function

$$p = 25 - 0.5q^d$$

Supply function

$$p = -0.5 + 0.25q^s$$

We know that the equilibrium price is £8 and the equilibrium quantity is 34 units.

Now suppose that a sales tax of £4.50 per unit sold is imposed on this product and collected from the suppliers. That is, for every unit sold, the suppliers have to pay £4.50 in tax to the government. The supply curve shows the amounts per unit that suppliers must receive to induce them to supply different quantities. Thus, before the tax was imposed, suppliers were willing to supply 42 units at a price of £10 per unit, and 50 units at a price of £12 per unit (use the demand equation above to confirm these figures). After the sales tax is imposed however, they will only be willing to supply 42 units at a price of £14.50, and 50 units at a price of £16.50 (because they have to send £4.50 per unit to the government to pay the tax). In other words, the supply curve will have shifted vertically upwards by the full amount of the tax. The new supply curve will be:

Supply function after tax

$$p = (-0.5 + 4.5) + 0.25q^s$$

Notice that this is obtained by simply adding the tax to the constant term in the equation.

In general, when a sales tax of £X per unit is imposed on suppliers of an item of merchandise, the supply curve for the item shifts vertically upwards by the full amount of the tax. So if the supply equation for the item of merchandise is written as $p = A + Bq_s$, then the supply equation after the sales tax will become $p = (A + X) + Bq^s$.

The supply curves, before and after the sales tax, together with the demand curve, are shown in Figure 16.3. It is clear from the graph that the equilibrium price rises to £11 and the equilibrium quantity falls to 28 units. These new equilibrium values can also be found algebraically. At the new equilibrium, where $q^d = q^s = q$, we have:

$$25 - 0.5q = (-0.5 + 4.5) + 0.25q = 4 + 0.25q$$

Rearranging gives:

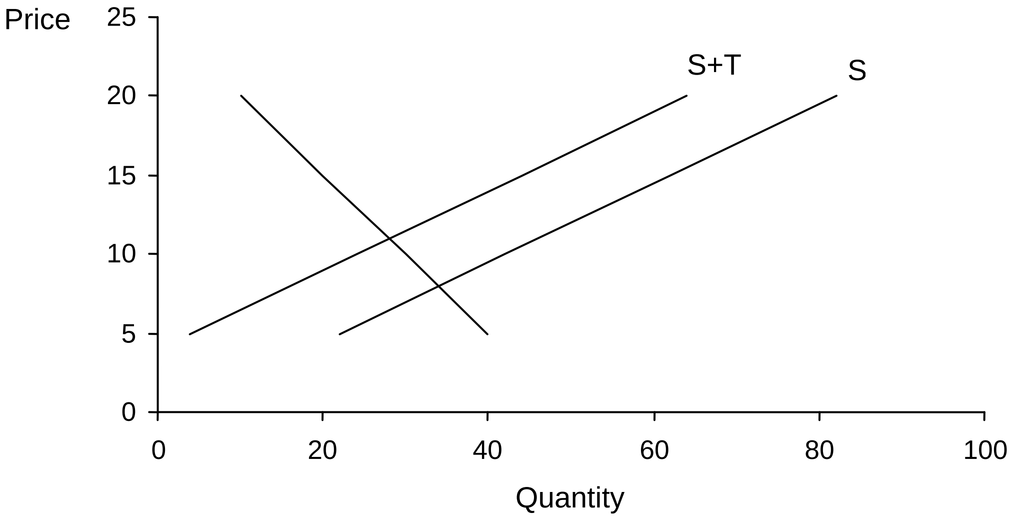
$$0.75q = 21$$

$$q = 28$$

Substituting this into the demand or new supply function gives $p = 11$.

Since a tax of £4.50 per unit was imposed on suppliers, but the equilibrium price has only increased by £3 (i.e. from £8 to £11), we can say that the share of the tax borne by consumers of the product is two-thirds (i.e. £3 divided by £4.50). The other third is borne by the producers. These shares represent the *effective incidence* of the tax and will depend on the shapes of the demand and supply curves.

Figure 16.3: New equilibrium after a sales tax is imposed



C. BREAKEVEN ANALYSIS

For any business, there is a certain level of sales at which there is neither a profit nor a loss. Total income and total costs are equal. This point is known as the *breakeven point*. It is easy to calculate, and can also be found by drawing a graph called a *breakeven chart*.

Calculation of Breakeven Point

Example:

The organising committee of a Christmas party have set the selling price at £21 per ticket. They have agreed with a firm of caterers that a buffet would be supplied at a cost of £13.50 per person. The other main items of expense to be considered are the costs of the premises and discotheque, which will amount to £200 and £250 respectively. The variable cost in this example is the cost of catering, and the fixed costs are the expenditure for the premises and discotheque.

Answer

The first step in the calculation is to establish the amount of contribution per ticket:

	£
price of ticket (sales value)	21.00
less catering cost (marginal cost)	13.50
contribution per ticket	7.50

Now that this has been established, we can evaluate the fixed costs involved. The total fixed costs are:

	£
premises hire	200
discotheque	250
total fixed expenses	450

The organisers know that for each ticket they sell, they will obtain a contribution of £7.50 towards the fixed costs of £450. Clearly it is necessary only to divide £450 by £7.50 to establish the number of contributions that are needed to break even on the function. The breakeven point is therefore 60 – i.e. if 60 tickets are sold there will be neither a profit nor a loss on the function. Any tickets sold in excess of 60 will provide a profit of £7.50 each.

Formulae

The general formula for finding the breakeven point (BEP) is:

$$\text{BEP} = \frac{\text{fixed costs}}{\text{contribution per unit}}$$

If the breakeven point (BEP) is required in terms of sales *revenue*, rather than sales *volume*, the formula simply has to be multiplied by selling price per unit, i.e.:

$$\text{BEP (sales revenue)} = \frac{\text{fixed costs}}{\text{contribution per unit}} \times \text{selling price per unit}$$

In our example about the party, the breakeven point in revenue would be $60 \times £21 = £1{,}260$. The committee would know that they had broken even when they had £1,260 in the kitty.

Suppose the committee were organising the party in order to raise money for charity, and they had decided in advance that the function would be cancelled unless at least £300 profit would be made. They would obviously want to know how many tickets they would have to sell to achieve this target.

Now, the £7.50 contribution from each ticket has to cover not only the fixed costs of £450, but also the desired profit of £300, making a total of £750. Clearly they will have to sell 100 tickets (£750 ÷ £7.50).

To state this in general terms:

$$\text{volume of sales needed to achieve a given profit} = \frac{\text{fixed costs} + \text{desired profit}}{\text{contribution per unit}}.$$

Suppose the committee actually sold 110 tickets. Then they have sold 50 more than the number needed to break even. We say they have a *margin of safety* of 50 units, or of £1,050 ($50 \times$ £21), i.e.:

margin of safety = sales achieved – sales needed to break even.

It may be expressed in terms of sales volume or sales revenue.

Margin of safety is very often expressed in percentage terms:

$$\frac{\text{sales achieved} - \text{sales needed to break even}}{\text{sales achieved}} \times 100\%$$

i.e. the party committee have a percentage margin of safety of:

$$\frac{50}{110} \times 100\% = 45\%.$$

The significance of the margin of safety is that it indicates the amount by which sales could fall before a firm would cease to make a profit. If a firm expects to sell 2,000 units, and calculates that this would give it a margin of safety of 10%, then it will still make a profit if its sales are at least 1,800 units (2,000 less 10% of 2,000), but if its forecasts are more than 10% out, then it will make a loss.

The profit for a given level of output is given by the formula:

(output × contribution per unit) – fixed costs.

It should not be necessary for you to memorise this formula, since when you have understood the basic principles of marginal costing you should be able to work out the profit from first principles.

Question for Practice

Using the data from the first example, what would the profit be if sales were:

(a) 200 tickets?

(b) £2,100 worth of tickets?

Now check your answers with those given at the end of the chapter.

D. BREAKEVEN CHARTS

A number of types of breakeven chart are in use. We will look at the two most common types:

- cost/volume charts
- profit/volume charts.

Information Required

(a) Sales Revenue

When we are drawing a breakeven chart for a single product, it is a simple matter to calculate the total sales revenue which would be received at various outputs.

As an example, take the following figures:

Table 16.5: Output and sales revenue

Output *(Units)*	**Sales Revenue** *(£)*
0	0
2,500	10,000
5,000	20,000
7,500	30,000
10,000	40,000

(b) Fixed Costs

We must establish which elements of cost are fixed in nature. The fixed element of any semi-variable costs must also be taken into account.

We will assume that the fixed expenses total £8,000.

(c) Variable Costs

The variable elements of cost must be assessed at varying levels of output.

Table 16.6: Output and variable costs

Output *(Units)*	**Variable costs** *(£)*
0	0
2,500	5,000
5,000	10,000
7,500	15,000
10,000	20,000

Cost/Volume Chart

The graph is drawn with level of output (or sales value) represented along the horizontal axis and costs/revenues up the vertical axis. The following are the stages in the construction of the graph:

(a) Plot the *sales line* from the above figures.

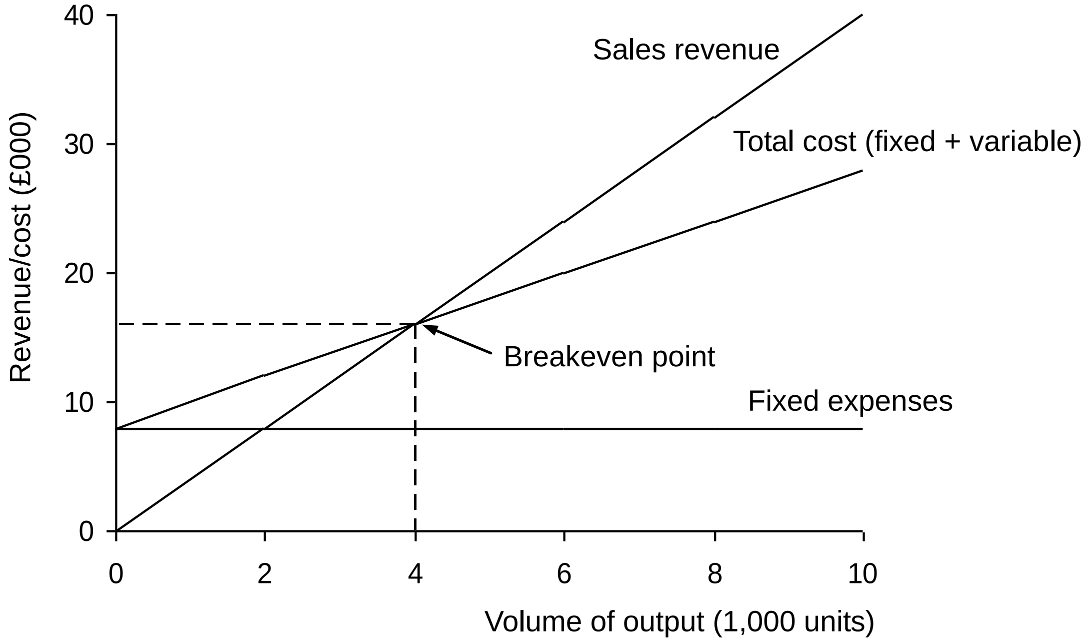
(b) Plot the *fixed expenses line*. This line will be parallel to the horizontal axis.

(c) Plot the *total expenses line*. This is done by adding the fixed expense of £8,000 to each of the variable costs above.

(d) The *breakeven point* is represented by the meeting of the sales revenue line and the total cost line. If a vertical line is drawn from this point to meet the horizontal axis, the breakeven point in terms of units of output will be found.

The graph is illustrated in Figure 16.4, a typical cost/volume breakeven chart.

Figure 16.4: Cost/volume breakeven chart



Note that although we have information available for four levels of output besides zero, one level is sufficient to draw the chart, provided we can assume that sales and costs will lie on straight lines. We can plot the single revenue point and join it to the origin (the point where there is no output and therefore no revenue). We can plot the single cost point and join it to the point where output is zero and total cost equals fixed cost.

In this case, the breakeven point is at 4,000 units, or a revenue of £16,000 (sales are at £4 per unit).

This can be checked by calculation:

sales revenue = £4 per unit

variable costs = £2 per unit

thus, contribution = £2 per unit

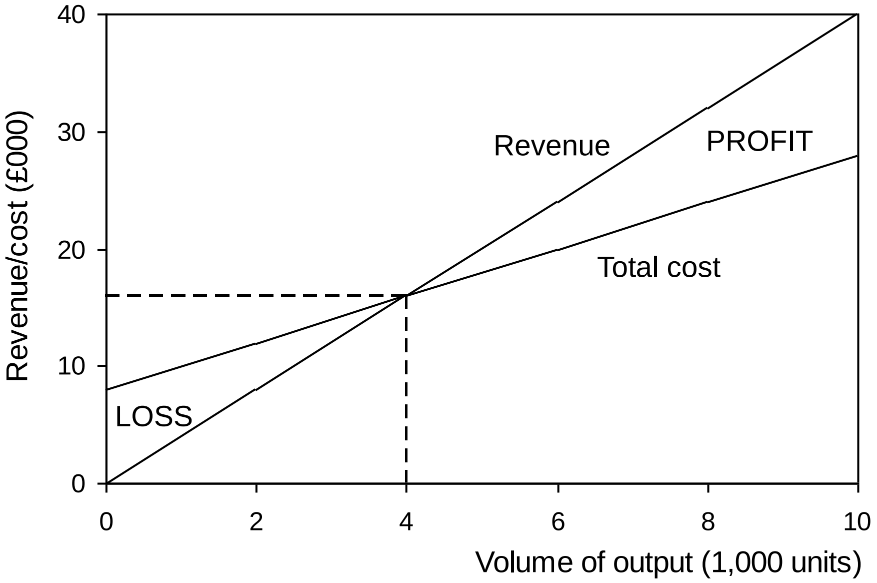
fixed costs = £8,000

breakeven point = fixed costs ÷ contribution per unit

= 4,000 units.

The relationship between output and profit or loss is shown in Figure 16.5, a typical cost/volume chart.

Figure 16.5: Cost/volume breakeven chart



Profit/Volume Chart

With this chart the *profit line* is drawn, instead of the revenue and cost lines. It does not convey quite so much information, but does emphasise the areas of loss or profit compared with volume.

The contribution line is linear, so we need only two plotting points again.

When the volume of output is zero, a loss is made which is equal to fixed costs. This may be one of our plotting points. The other plotting point is calculated at the high end of the output range, that is:

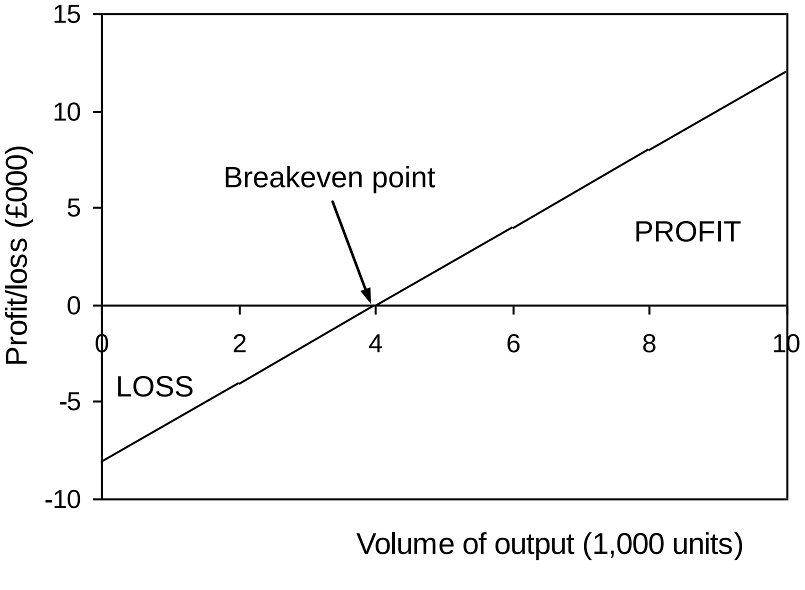
when output = 10,000 units

revenue = £40,000

total costs = £(8,000 + 20,000) = £28,000

profit = £(40,000 – 28,000) = £12,000 (see Figure 16.6).

Figure 16.6: Profit/volume chart

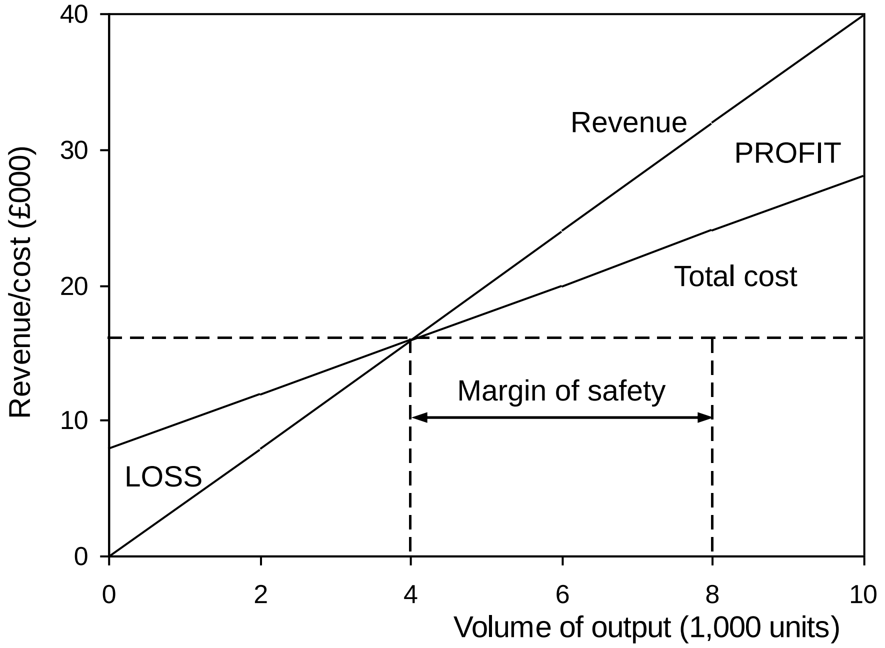


When drawing a breakeven chart to answer an exam question, it is normal to draw a cost/volume chart unless otherwise requested in the question. The cost/volume chart is the more common type, and does give more detail.

Margin of Safety

If management set a level of budgeted sales, they are usually very interested in the difference between the budgeted sales and the breakeven point. At any level between these two points, some level of profit will be made. This range is called the *margin of safety* (see Figure 16.7), where the level of activity is budgeted (planned) at 8,000 units.

Figure 16.7: Margin of safety



Assumptions and Limitations of Breakeven Charts

- It is difficult to draw up and interpret a breakeven chart for more than one product.
- Breakeven charts are accurate only within fairly narrow levels of output. This is because if there was a substantial change in the level of output, the proportion of fixed costs could change.
- Even with only one product, the income line may not be straight. A straight line implies that the manufacturer can sell any volume the manufacturer likes at the same price. This may well be untrue: if the manufacturer wishes to sell more units the price may have to be reduced. Whether this increases or decreases the manufacturer's total income depends on the elasticity of demand for the product. The sales line may therefore curve upwards or downwards, but in practice is unlikely to be straight.
- Similarly, we have assumed that variable costs have a straight line relationship with level of output – i.e. variable costs vary directly with output. This might not be true. For instance, the effect of diminishing returns might cause variable costs to increase beyond a certain level of output.
- Breakeven charts hold good only for a limited time.

Nevertheless, within these limitations a breakeven chart can be a very useful tool. Managers who are not well versed in accountancy will probably find it easier to understand a breakeven chart than a calculation showing the breakeven point.

E. THE ALGEBRAIC REPRESENTATION OF BREAKEVEN ANALYSIS

Using Linear Equations to Represent Cost and Revenue Functions

We have already seen how equations can be used to represent demand and supply functions and hence to determine equilibrium price and quantity. Similarly, equations can be used to represent cost and revenue functions and to calculate profit and output.

Let us consider a simple example. Table 16.5 shows the sales revenue which is yielded at different levels of output – it is a *revenue schedule*. The schedule is depicted graphically in Figure 16.4, where we can see that it takes the form of a straight line. We already know that a relationship which when plotted on a graph produces a straight line is a linear function, and hence can be described by means of a linear equation. It therefore follows that the revenue schedule we are considering is a linear function and can be described by a linear equation.

We know that the general form of a linear function is:

$$y = a + bx$$

where: a = the point where the line crosses the vertical axis

b = the slope of the line.

We also know that for any two points, we can obtain the gradient of a straight line by using the following formula:

$$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\text{differencein y co-ordinates}}{\text{differencein x co-ordinates}}$$

From Figure 16.4, we can see that the line crosses the vertical axis at 0.

To find the gradient, we perform the following calculation:

$$\frac{20{,}000 - 10{,}000}{5{,}000 - 2{,}500} = \frac{10{,}000}{2{,}500} = 4$$

We can therefore state the equation for revenue (R) as follows:

$$R = 4q$$

where: q = output.

This is known as the *revenue function*.

We can also perform a similar calculation to find the equation of the total cost line – the *cost function* – depicted in Figure 16.4. Remember that we need first to sum fixed costs (set at £8,000) and variable costs (shown in Table 16.6) to obtain values for total costs; then we can carry out the calculation as before.

The Breakeven Point

We have already seen that the breakeven point corresponds to the volume of output at which total revenue equals total cost. At this point, profit is zero; beyond this point, any increase in output will yield a profit.

In algebraic terms, profit can be expressed as:

$$\Pi = Pq - (F + Vq)$$

where: Π = profit

P = unit selling price

q = sales volume in units

F = total fixed costs

V = unit variable cost.

The breakeven point at which total revenue equals total cost and profit equals zero can be expressed as:

$$Pq_b - (F + Vq_b) = 0$$

where: q_b = breakeven volume.

We can rearrange the equation to express breakeven volume as:

$$q_b = \frac{F}{P - V}$$

where: $P - V$ is the contribution per unit.

Therefore the breakeven point equals total fixed costs (F) divided by the contribution per unit ($P - V$). To convert q_b into breakeven sales (Y), we multiply both sides of the q_b formula by P, as follows:

$$Y = Pq_b = \frac{PF}{P - V}$$

This can also be expressed as:

$$Y = \frac{F}{1 - V/P}$$

where: $1 - V/P$ = contribution ratio.

This formula gives us breakeven sales.

Let us consider an example of a company that produces a product which sells for 50 pence per unit. Total fixed costs amount to £10,000 and the variable cost per unit is 30 pence.

The unit contribution (or the excess of unit sales price over unit variable cost) is:

$$P - V = 0.50 - 0.30 = 0.20$$

The breakeven point is:

$$q_b = \frac{10{,}000}{0.20} = 50{,}000 \text{ units.}$$

The contribution ratio is:

$$1 - V/P = 1 - \frac{0.30}{0.50} = 40\%$$

Breakeven sales:

$$Y = \frac{10{,}000}{0.40} = £25{,}000$$

This can also be expressed as:

$$Y = Pq_b = 0.50 \times 50{,}000 \text{ units} = £25{,}000.$$

Changes in the Cost and Revenue Functions

We can use the breakeven formulae above to analyse the effect of changes in the cost and revenue functions – that is, in the parameters and variables, such as the unit selling price, variable costs and fixed costs. Let us consider each of these in turn.

A reduction in the unit selling price will decrease the contribution and hence increase the breakeven volume. If we assume that the unit price is reduced from 50 pence to 40 pence, while all the other variables remain unchanged, we can find the new breakeven point as follows:

$$q_b = \frac{10{,}000}{0.40 - 0.30} = 100{,}000 \text{ units}$$

and

$$Y = 100{,}000 \times 0.40 = £40{,}000$$

or

$$Y = \frac{10{,}000}{1 - (0.30/0.40)} = £40{,}000.$$

An increase in the unit variable cost will decrease the unit contribution and increase the breakeven volume. If we assume that the price of raw materials increases by 10 pence per unit, while the other variables remain unchanged, we can find the new breakeven point as follows:

$$q_b = \frac{10{,}000}{0.50 - 0.40} = 100{,}000 \text{ units}$$

and

$$Y = 100{,}000 \times 0.50 = £50{,}000$$

or

$$Y = \frac{10{,}000}{1 - (0.40/0.50)} = £50{,}000.$$

Similarly, a decrease in unit variable cost will decrease the breakeven volume.

An increase in total fixed costs will increase breakeven volume, while a decrease in total fixed costs will decrease breakeven volume. If we assume that fixed costs increase by £2,000, while the other variables remain unchanged, we can find the new breakeven point as follows:

$$q_b = \frac{10{,}000 + 2{,}000}{0.50 - 0.30} = 60{,}000 \text{ units}$$

and

$$Y = 60{,}000 \times 0.50 = £30{,}000$$

or

$$Y = \frac{12{,}000}{1 - (0.30/0.50)} = £30{,}000.$$

Calculating Profit at Different Output Levels

We have already seen that profit at breakeven point equals zero. Therefore, the profit for any volume of output greater than breakeven equals the profit generated by the additional output beyond the breakeven volume. We can express profit for any given sales volume (q_1) as:

$$(q_1 - q_b) \times (P - V).$$

In our example, the breakeven volume is 50,000 units. Let us assume that we now want to find the profit generated by sales of 70,000 units. Using the formula above:

$$(70{,}000 - 50{,}000) \times (0.50 - 0.30) = £4{,}000.$$

The profit generated by sales of 70,000 units is therefore £4,000.

ANSWERS TO QUESTION FOR PRACTICE

(a) We already know that the contribution per ticket is £7.50.

Therefore, if they sell 200 tickets, total contribution is 200 × £7.50 = £1,500.

Out of this, the fixed costs of £450 must be covered; anything remaining is profit.

Therefore profit = £1,050. (Check: 200 tickets are 140 more than the number needed to break even. The first 60 tickets sold cover the fixed costs; the remaining 140 show a profit of £7.50 per unit. Therefore profit = 140 × £7.50 = £1,050, as before.)

(b) £2,100 worth of tickets is 100 tickets since they are £21 each.

	£
total contribution on 100 tickets =	750
less fixed costs	450
profit	300