INTROSTAT

Les Underhill and Dave Bradfield

Department of Statistical Sciences
University of Cape Town

June 2014

Introduction

IntroSTAT apart, there seem to be two kinds of introductory Statistics textbooks. There are those that assume no mathematics at all, and get themselves tied up in all kinds of knots trying to explain the intricacies of Statistics to students who know no calculus. There are those that assume lots of mathematics, and get themselves tied up in the knots of mathematical statistics.

IntroSTAT assumes that students have a basic understanding of differentiation and integration. The book was designed to meet the needs of students, primarily those in business, commerce and management, for a course in applied statistics.

IntroSTAT is designed as a **lecture-book**. One of our aims is to maximize the time spent in **explaining** concepts and **doing** examples. It is for this reason that three types of examples are included in the chapters. Those labeled A are used to motivate concepts, and often contain explanations of methods within them. They are for use in lectures. The B examples are worked examples — they shouldn't be used in lectures — there is nothing more deadly dull than lecturing through worked examples. Students should use the B examples for private study. The C examples contain problem statements. A selection of these can be tackled in lectures without the need to waste time by the lecturer writing up descriptions of examples, and by the students copying them down.

There are probably more exercises at the end of most chapters than necessary. A selection has been marked with asterisks (*) — these should be seen as a minimum set to give experience with all the types of exercises.

Acknowledgements ...

We are grateful to our colleagues who used previous versions of IntroSTAT and made suggestions for changes and improvements. We have also appreciated comments from students. We will continue to welcome their ideas and hope that they will continue to point out the deficiencies. Mrs Tib Cousins undertook the enormous task of turning edition 4 into TEX files, which were the basis upon which this revision was undertaken. Mrs Margaret Spicer helped us proofread the text; Dr Derek Chalton and Mr Tim Low suggested the corrections and improvements that are incorporated into the second edition. We are grateful to have remaining errors pointed out to us.

This volume is essentially the 1996 edition of Introstat with some minor and major corrections of errors, and was reset in LATEX from the the original plain TEX version.

Contents

Chapter 1

EXPLORING DATA

KEYWORDS: Data summary and display, qualitative and quantitative data, pie charts, bar graphs, histograms, symmetric and skew distributions, stem-and-leaf plots, median, quartiles, extremes, five-number summary, box-and-whisker plots, outliers and strays, measures of spread and location, sample mean, sample variance and standard deviation, summary statistics, scatter plots, contingency tables, exploratory data analysis.

Facing up to uncertainty ...

We live in an uncertain world. But we still have to take decisions. Making good decisions depends on how well informed we are. Of course, being well informed means that we have useful information to assist us. So, having useful information is one of the keys to good decision making.

Almost instinctively, most people gather information and process it to help them take decisions. For example, if you have several applicants for a vacant post, you would not draw a number out of a hat to decide which one to employ. Almost without thinking about it, you would attempt to gather as much relevant **information** as you can about them to help you compare the applicants. You might make a short-list of applicants to interview, and prepare appropriate questions to put to each of them. Finally you come to an **informed** decision.

Sometimes the available information is such that we feel it is easy to make a good decision. But at other times, so much confusion and uncertainty cloud the situation that we are inclined to go by "gut-feeling" or even by guessing. But we can do better relying on our instinct and guess work. This books aims to equip you with some of the necessary skills to "outguess" the competition. Or, putting it less brashly, to help you to make consistently sound decisions.

As the world becomes more technologically advanced, people realize more and more that information is valuable. Obtaining the information they need might just require a phone call, or maybe a quick visit to the library. Sometimes, they might need to expend more energy and extract some information out of a database. Or worse, they might have to design an experiment and gather some data of their own. On other occasions, the information might be hidden in historical records.

Usually, **data** contains information that is not self-evident. The message cannot be extracted by simply eye-balling the data. Ironically, the more valuable the information, the more deeply it usually lies buried within the data. In these instances, statistical

tools are needed to extract the information from the data. Herein lies the focus of this book.

For example, consider the record of share prices on the Johannesburg Stock Exchange. Hidden in this data lies a wealth of information — whether or not a share is risky, or if it is over- or underpriced. This data even contains traces of our own emotions — whether our sentiments are mawkish or positive, risk prone or risk averse — and our preferences — for higher dividends, for smaller companies, for blue-chip shares. Little wonder that there is a multitude of financial analysts out there trying to analyse share price data hoping that they might unearth valuable information that will deliver the promise of better profits.

Just as the financial analysts have an insatiable appetite for information on which to base better investment decisions, so in every field of human endeavour, people are analysing information with the objective of improving the decisions they take.

One of the essential set of ideas and skills needed to extract information from data, to interpret this information, and to take decisions based on it, is the subject called Statistics. Not everyone is willing, or has the foresight, to master a course in the science of Statistics. We are fortunate that this is true — otherwise statisticians would not hold the monopoly on superior decision-making!

You have already made at least one good decision — the decision to do a course in Statistics.

What Statistics is (and is not!) ...

Most people seem to think that what statisticians do all day is to count, to add and to average. The two kinds of "statisticians" that most frequently impinge on the general public are really parodies of statisticians: the "sports" statistician and the "official" statistician. Statistics is not what you see at the bottom of the television screen during the French Open Tennis Championships: **statistique**, followed by a count of the number of double faults and aces the players have produced in the match so far! Nor is statistics about adding up dreary columns of figures, and coming to the conclusion, for example, that there were 30 777 000 sheep in South Africa in 1975. That sort of count is enough to put anyone to sleep!

If statisticians do not count in the 1–2–3 sense, in what sense do they count? What is statistics? We define statistics as the **science of decision making in the face of uncertainty**. The emphasis is not on the collection of data (although the statistician has an important role in advising on the data collection process), but on taking matters one step further — interpreting the data. Statistics may be thought of as data-based decision making. Perhaps it is a pity that our discipline is called Statistics. A far better name would have been Decision Science. Statistics really comes into its own when the decisions to be made are not clear-cut and obvious, and there is uncertainty (even after the data has been gathered) about which of several alternative decisions is the best one to choose.

For example, the decision about which card to play in a game of bridge to maximize your chance of winning, or the decision about where to locate a factory so as to maximize the likelihood that your company's share of the market will reach a target value, are not simple decisions. In both situations, you can gather as much data as you can (the cards in your hand, and those already played in the first case; proximity to raw materials and to markets in the second), and take a best possible decision on the basis of this data, but there is still no guarantee of success. In both cases, your opposition may react in unexpected ways, and you risk defeat.

In the above sentences, the words "uncertainty", "chance", "likelihood" and "risk" have appeared. All these terms are qualitative and open to many interpretations. Before the statistician can get down to his or her real job (of taking decisions in the face of uncertainty), this nebulous concept of uncertainty has to be put onto a firm footing by being quantified in an objective way. Probability Theory is the branch of mathematics that achieves this quantification of uncertainty.

Therefore, before you can become a statistician, you have to learn a hefty chunk of Probability Theory. This material is included in chapters 2 to 7. Chapters 8 to 12 deal with the science of data-based decision making.

However, in the remainder of this chapter, we aim to give you insight into what is to come in the later chapters, to give you a feeling for data, and to explore "data-based decision making" using intuitive concepts.

Display, summarize and interpret ...

Before getting deeply involved with tackling any situation or problem in daily life, it is wise to take a step back and take a glimpse at "the big picture" — and so it is with Statistics. As a starting point, statisticians make a "quick and dirty" summary of the data they are about to analyse in order to get a "feel" for what they are dealing with.

The initial overview usually involves: constructing a visual display of the data in a graph or perhaps a table; summarizing the data with a few pertinent "key" numbers; and gaining insight into the "potential" of the data.

What do we mean by data? ...

Data is information. There are data drips and data floods, and statisticians have to learn to deal with both. Usually, there is either too little or too much data! When data comes in floods, the problem is to extract the salient features. When data comes in drips, the problem is to know what are valid interpretations.

Besides the various amounts of data, there are different types of data. For the moment, we need to distinguish between **qualitative** and **quantitative** data. Qualitative data is usually non-numerical (nominal), and arises when we classify objects using labels or names as categories: for example, make of car, colour of eyes, gender, nationality, profession, cause of death, etc. Sometimes the categories are semi-numerical (ordinal): for example, size of companies categorized as small, medium or large.

Quantitative data, on the other hand, is always numerical, and these data values can be ranked or ordered. Quantitative data usually arises from counting or measuring: for example, flying time between airports, number of rooms in a house, salary of an accountant, cost of building a school, volume of water in a dam, number of new car sales in a month, the size of the AIDS epidemic, etc..

Visual displays of qualitative data ...

Two efficient ways of displaying qualitative data are the **pie chart** and the **bar chart**.

Table 1.1: MBA Student Data

First degree	GMAT score	First degree	GMAT score	First degree	GMAT score
1. Engineering	610	28. Engineering	710	55. Arts	500
2. Engineering	510	29. Science	600	56. Arts	620
3. Engineering	610	30. Science	550	57. Arts	550
4. Engineering	580	31. Science	540	58. Arts	600
5. Engineering	720	32. Science	620	59. Arts	520
6. Engineering	620	33. Science	650	60. Arts	520
7. Engineering	540	34. Science	500	61. Commerce	550
8. Engineering	500	35. Science	590	62. Commerce	520
9. Engineering	750	36. Science	630	63. Commerce	560
10. Engineering	640	37. Science	660	64. Commerce	560
11. Engineering	550	38. Science	570	65. Commerce	600
12. Engineering	650	39. Science	600	66. Commerce	540
13. Engineering	600	40. Science	630	67. Commerce	550
14. Engineering	600	41. Science	500	68. Commerce	650
15. Engineering	510	42. Science	580	69. Commerce	510
16. Engineering	570	43. Science	560	70. Commerce	560
17. Engineering	620	44. Science	550	71. Medicine	590
18. Engineering	590	45. Arts	560	72. Medicine	700
19. Engineering	660	46. Arts	550	73. Medicine	640
20. Engineering	550	47. Arts	500	74. Medicine	680
21. Engineering	560	48. Arts	510	75. Medicine	580
22. Engineering	630	49. Arts	570	76. Other	550
23. Engineering	540	50. Arts	510	77. Other	680
24. Engineering	560	51. Arts	660	78. Other	540
25. Engineering	650	52. Arts	500	79. Other	640
26. Engineering	540	53. Arts	710	80. Other	620
27. Engineering	680	54. Arts	510	81. Other	450

Example 1A: Table 1.1 lists data on a class of 81 Master of Business Administration (MBA) students. The table shows each student's faculty for their first degree, in either Arts, Commerce, Engineering, Medicine, Science or "other". Also given are their test scores for an entrance examination known as the GMAT, a test commonly used by business schools worldwide as part of the information to assist in the selection process. Our brief is to construct a visual summary of the distribution of students within the various first-degree categories in the table.

Firstly, we note that first-degree category, the data that we are being asked to display, is qualitative data. Appropriate display techniques include the frequency table, the pie chart and the bar graph.

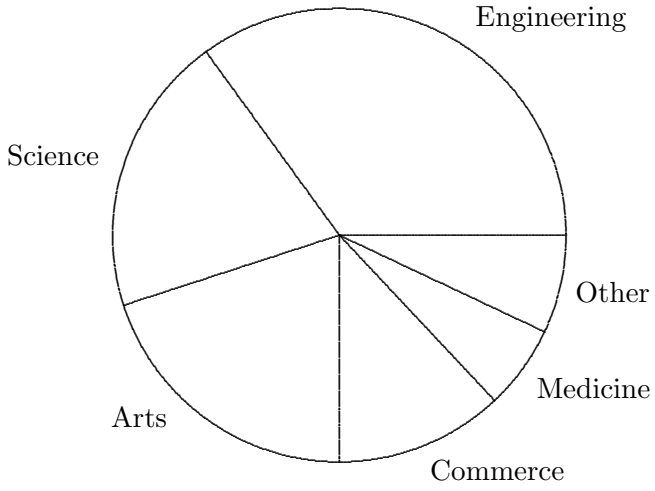
Secondly, we find the **frequency distribution** of the qualitative data by counting the number of students within each category. At the same time, we calculate **relative frequencies** by dividing the frequency in each category by the total number of observations. We repeat the relative frequencies to a convenient small number of decimal places.

First degree	Frequency	Relative frequency
Engineering	28	0.35
Science	16	0.20
Arts	16	0.20
Commerce	10	0.12
Medicine	5	0.06
Other	6	0.07
Total	81	1.00

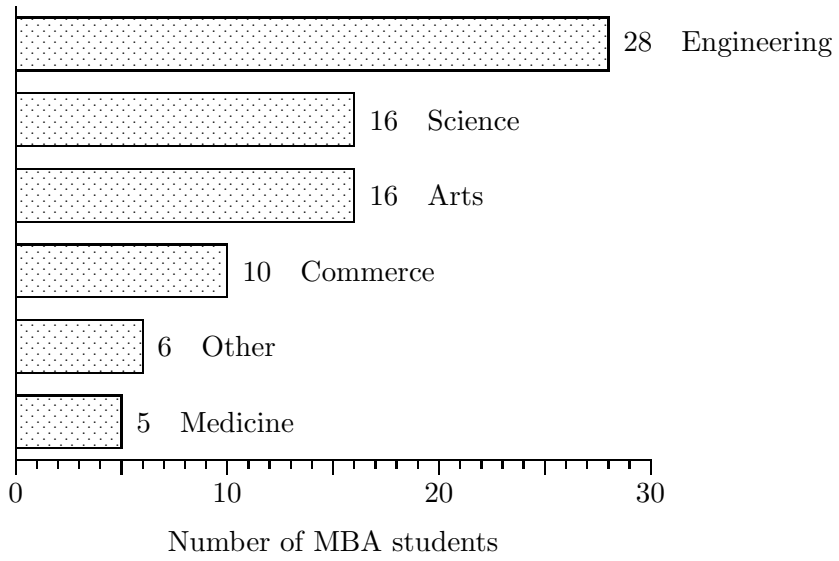
Thirdly, we plot the pie chart and the bar graph.

Pie chart: the actual construction of a pie chart is straightforward! We have arranged the segments in anti-clockwise order, starting at "three o'clock", but there is no hard-and-fast rule about where to start and which direction to choose. The pie chart communicates most effectively if the relative frequencies are arranged in decreasing order of size. Rotate the textbook through 90^o and 180^o and note that the pie chart may appear different to its original form, even though it is actually the same graphical figure.

Pie chart showing proportions of M.B.A. students



Bar graph: We may use horizontal or vertical bars to convey the relative frequencies of categories. Our first example was horizontal bars. Notice that there is no quantitative scale along the vertical axis of the bar graph, that the "bars" are not connected, and that the widths of the bars have no particular relevance. Because there is no quantitative ordering of the **nominal** categories, we are free to arrange them as we please. As for the pie chart, it is generally most effective to arrange the bars for nominal categories in decreasing order of relative frequency; this choice of ordering makes comparison easier, and also tends to highlight the important features of the data. Relative frequencies could also have been used in the construction of the bar graph. We would obtain similar bars but the horizontal axis would reflect proportions.



Visually the most striking impact of both the pie chart and the bar graph is that engineers form the largest proportion of this class of MBA students. Next, we might ask for a reason for this feature. A plausible explanation is that engineers are not exposed to much management and administrative training during their undergraduate years, and that they make up for this fact by doing an MBA. A second explanation is that the data was extracted during recessionary times, when engineers were not in demand — perhaps they were "investing in themselves" by doing an MBA while projects were scarce. How would you set about investigating whether this latter explanation is correct?

Our diagrams have provided some insight into this data. Sometimes, all we achieve is to demonstrate the obvious. At other times, our charts and diagrams will reveal completely unexpected phenomena. Careful interpretation is then needed, frequently with the help of the "experts" from the discipline from which the data comes.

Example 2C: Table 1.2 gives the composition of the "All-Share Index" of the Johannesburg Stock Exchange as at 2 January 1990. The breakdown of the All-Share Index reflects that it is composed of seven "major sector" indices, namely Coal, Diamonds, All-Gold, Metals & Minerals, Mining Financial, Financial, and Industrial Indices.

(a) Construct a bar chart showing the number of shares in each of the seven major sector indices that contribute to the All-Share Index.
(b) Now construct a bar chart showing the relative weightings of each of the seven major sectors as a percentage of the All-Share Index, and comment on any differences you find from (a) above.

Table 1.2: Composition of the All Share Index on the JSE

Subsidiary Sector Index	No. of shares in subsidiary index	Percentage weighting in All-Share Index	Major sector index	
Coal	2	0.82	Coal	
Diamonds	1	8.27	Diamonds	
Gold — Rand and others	5	1.39		
Gold — Evander	2	0.89		
Gold — Klerksdorp	3	5.45	All-Gold	
Gold — OFS	4	3.99		
Gold — West Wits	3	7.27		
Copper	1	0.55		
Manganese	1	0.91	Metal	
Platinum	2	5.00	&	
Tin	1	0.01	minerals	
Other metals & minerals	3	0.26		
Mining houses	3	16.79	Mining	
Mining holding	3	7.13	financial	
Banks & financial services	5	2.80		All-
Insurance	4	2.15		Share
Investment trusts	3	1.02	Financial	
Property	12	0.47		
Property trusts	11	0.97		
Industrial holdings	6	9.94		
Beverages & hotels	2	3.43		
Building & construction	6	0.92		
Chemicals	2	3.46		
Clothing, footwear & textiles	10	0.62		
Electronics, electr. & battery	7	1.39		
Engineering	7	0.95		
Fishing	1	0.08		
Food	4	2.14		
Furniture & household goods	5	0.30	Industrial	
Motors	6	0.43		
Paper & packaging	3	2.26		
Pharmaceutical & medical	3	0.48		
Printing & publishing	3	0.18		
Steel & allied	2	1.99		
Stores	10	2.16		
Sugar	1	0.43		
Tobacco & match	1	2.48		
Transportation	3	0.22		

VISUAL DISPLAYS OF QUANTITATIVE DATA I : HISTOGRAMS ...

Histograms are a time-honoured and familiar way of displaying quantitative data. A histogram differs from a bar chart in two ways. The histogram has bars that are not separated from the other thats is, there is no gap between adjacent bars. The bars within a histogram do not correspond to named categories, as in the bar chart. In the histogram the bars correspond to intervals on the number line. These intervals are constructed so that they are all of equal length. The length of the interval is selected so that it is easy to construct the intervals on the number line, but also to ensure we have a suitable number of intervals. We demonstrate the construction procedure by means of an example.

Example 3A: Referring back to the data on GMAT scores in Example 1A, draw a histogram for the distribution of the GMAT scores for the students.

We recommend a four-step histogram procedure for quantitative data:

1. Determine the **size of the sample**[1], i.e. the number of observations. We have $n = 81$ students — and throughout this book we reserve use of the symbol n for the concept of sample size, the number of observations we are dealing with! Find the smallest and largest numbers in the sample. Call these $x_{\min}$ and $x_{\max}$, respectively. The smallest GMAT score was from student 81, who scored 450, and the largest was the 750 achieved by student 9:

$$x_{\min} = 450$$

$$x_{\max} = 750$$

2. Choose **class intervals** that cover the range from $x_{\min}$ to $x_{\max}$. Here are two guidelines that help determine an approximate **length** L of the class intervals: the first is due to Mr Sturge, the second is used by the computer package GENSTAT. If the class intervals are made too narrow, the histogram looks "spikey", and if intervals are too wide, the histogram is "blurred". Sturge says: use class intervals of approximate length L where

$$L = \frac{x_{\max} - x_{\min}}{1 + \log_2 n} = \frac{x_{\max} - x_{\min}}{1 + 1.44 \log_e n}$$

GENSTAT says: use class intervals of approximate length L where

$$L = \frac{x_{\max} - x_{\min}}{\sqrt{n}}.$$

For our data, Sturge says

$$\frac{x_{\max} - x_{\min}}{1 + 1.44 \log_e n} = \frac{750 - 450}{1 + 1.44 \log_e 81} = 40.94,$$

while GENSTAT says

$$L = \frac{x_{\max} - x_{\min}}{\sqrt{n}} = \frac{750 - 450}{\sqrt{81}} = 33.33.$$

[1] We consider a **sample** to be a small number of elements taken from the **population of interest**. Each element in the sample provides an observed value for the numerical variable of interest, these values become our dataset. We hope that the sample is **representative** of the population as a whole, so that the sample data represent the population data. Then conclusions drawn from the sample data will be valid also for the complete population data. We consider methods of obtaining a representative sample in Chapter 11.

As a general rule, avoid choosing class intervals which are of awkward lengths. Multiples of 2, 5 and 10 are most frequently used. Feel free to choose intervals between half and double those suggested by the Sturge or GENSTAT guidelines.

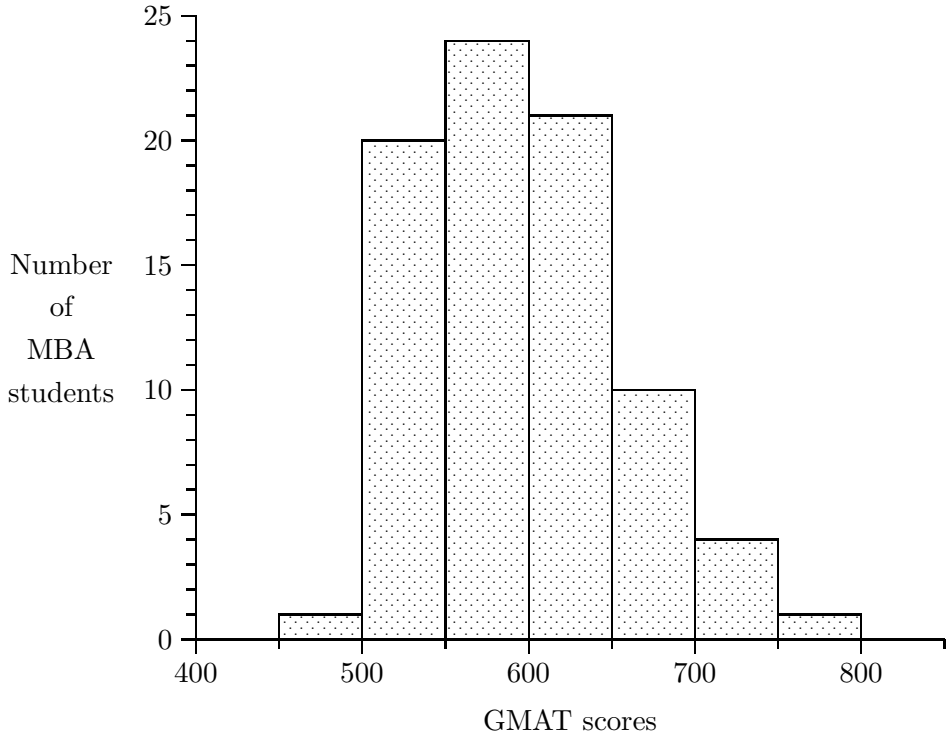
All the class intervals should be the same width. Resist the temptation to make the class intervals wider over that part of the range where the data is sparser — unequal widths have the effect of destroying the visual message of the histogram. For this example, $L = 50$ is a sensible choice for the width of the class interval. It is convenient to start our class intervals at 450, and carry on in steps of 50 as far as is necessary to include the maximum observed value. Thus the boundaries of the class intervals are at 450, 500, 550, 600, 650, 700, 750, and 800. We also need to agree that scores that fall on any boundary will be allocated to the higher of the two class intervals, so strictly the class intervals are 450–499, etc., as shown in the frequency distribution table below.

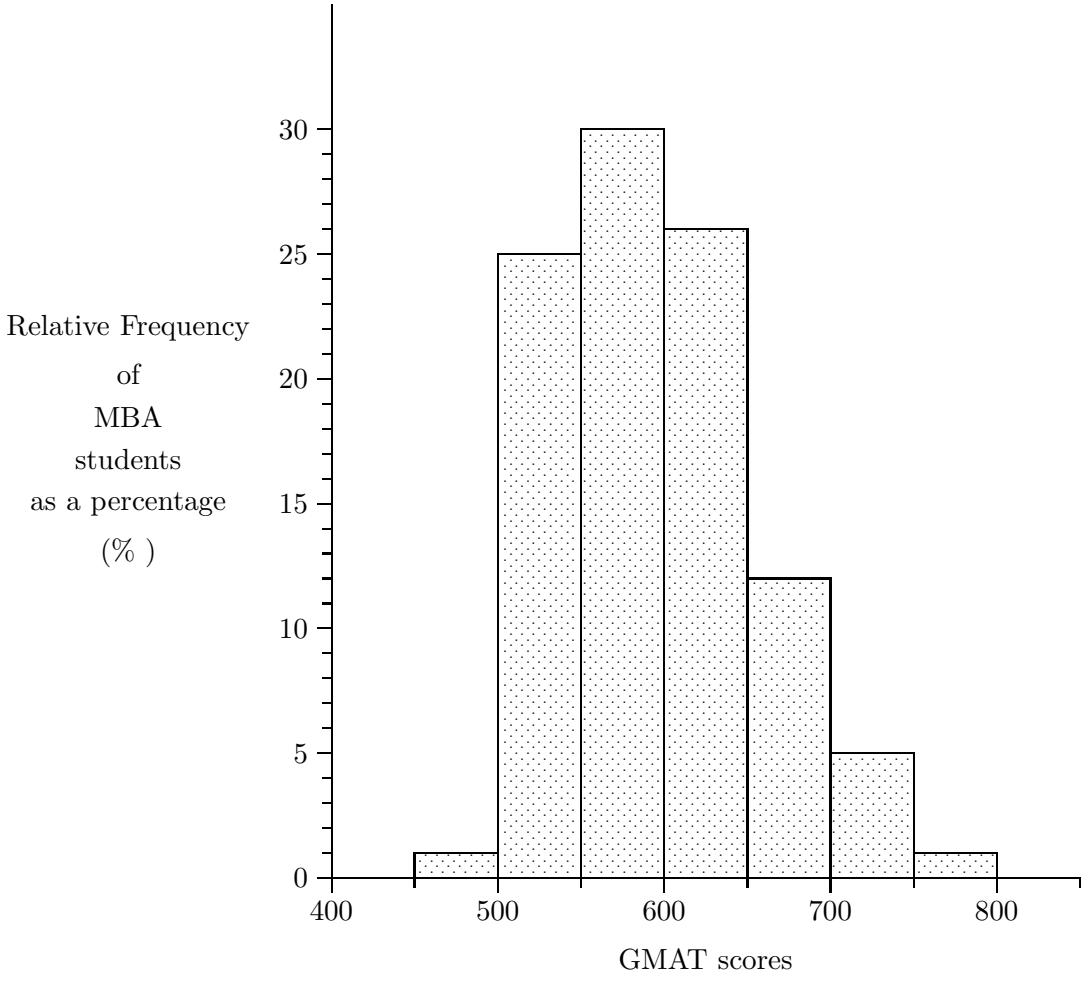
3. Count the number of GMAT scores within each class interval. The most convenient way to do this counting is to set up a **tick sheet**, and to make one pass through the data allocating each score to its class interval. This process sets up a **frequency distribution**:

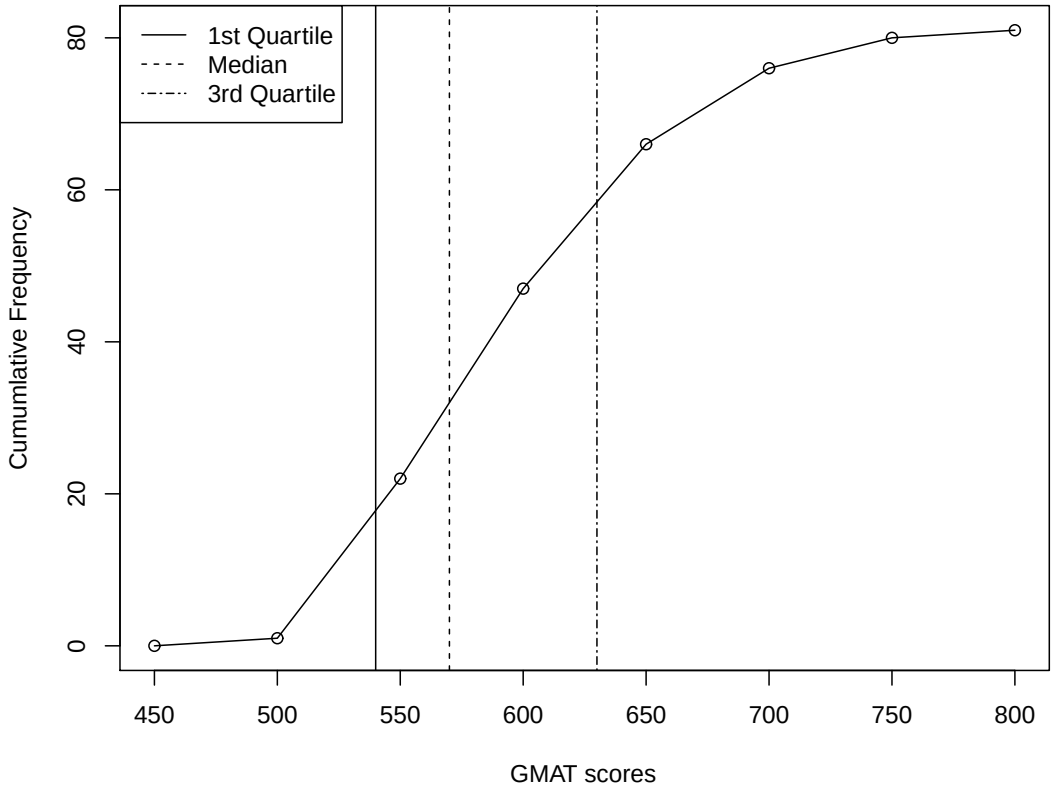
Class interval	frequency
450–499	1
500–549	21
550–599	25
600–649	19
650–699	10
700–749	4
750–799	1
Total	81

In some textbooks and computer programmes the class intervals are called **bins** and interval length L is called the **bin width**. Thus the class frequency is called the **bin frequency**.

4. Plot the histogram, choosing suitable scales for each axis:


Number of MBA students
25
20
15
10
5
0
400
500
600
700
800
GMAT scores


Relative Frequency of MBA students as a percentage (%)
30
25
20
15
10
5
0
400
500
600
700
800
GMAT scores



The striking feature of the histogram for the GMAT data on the previous page is that it is not **symmetric** but is **skewed to the right**, which means that it has a long **tail** stretching off to right. The terms in bold are technical, jargon terms, but their meanings are obvious.

A seasoned statistician would expect a distribution of test scores (or examination results) to have a **tail** at both ends of the frequency distribution. In the above display, there has been a **truncation** of the distribution at 500 (apart from a single score of 450). We would infer that the acceptance criterion on the MBA programme is a GMAT score of 500 or more. In reality there **is** a tail on the left, but it is suppressed by the fact that applicants who achieved these lower scores were not accepted into the MBA programme. In the light of this information, a statistician would also query the score of 450. Is it an error in the data? Maybe it should be 540, and there has been a transcription error. But a more plausible explanation is that the student was outstanding in some other aspect of the selection process — maybe the personal interview was very impressive!

Example 4B: The risks taken by investors when they invest in the stock exchange are of considerable interest to financial analysts. Investors associate the risks of investments with how volatile (or varying) the price changes are. Analysts measure volatility of price changes using the "standard deviation" — a statistical measure of variability that we will learn about later in this chapter. The table below reports the standard deviations (or riskiness) of a sample of 75 shares listed on the Johannesburg Stock Exchange. The units of the data values are per cent per month. Construct a suitable histogram of the data.

23	22	17	18	21	25	23	25	12	23	27	14	28	9	23
19	23	11	16	11	15	15	12	12	12	21	13	11	13	13
27	20	17	8	13	28	14	9	13	11	23	23	10	12	12
26	25	11	12	20	22	21	9	13	19	19	13	14	15	17
17	10	25	26	11	12	25	22	12	11	22	20	14	10	23

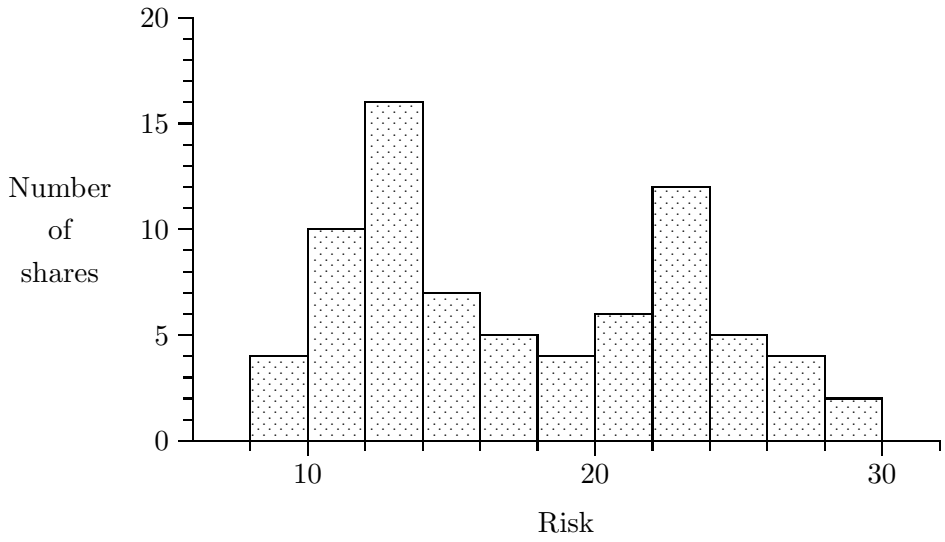
1. The sample size is $n = 75$, the extreme values are $x_{\min} = 8$, $x_{\max} = 28$.

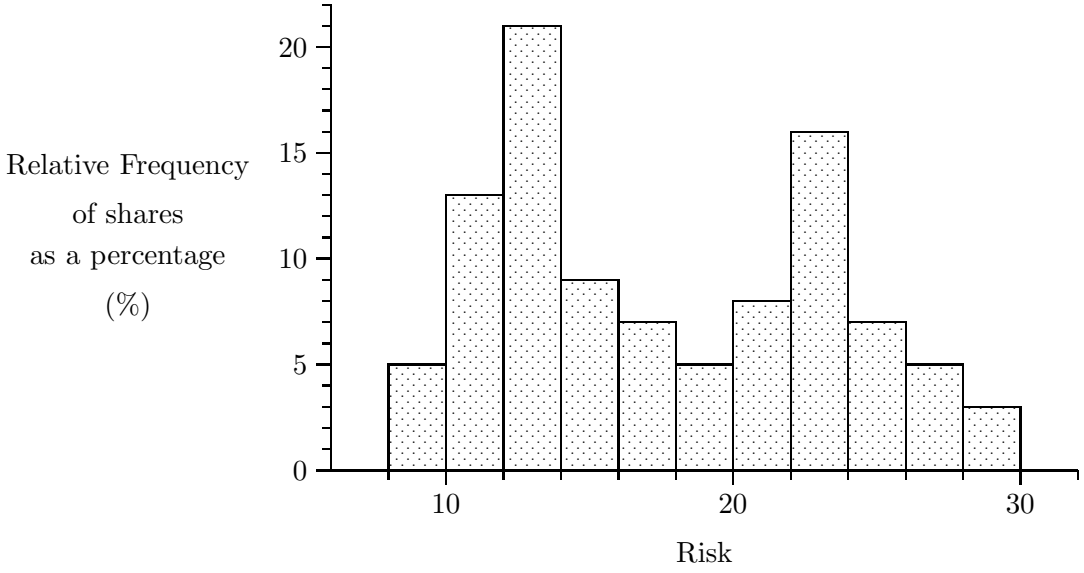
2. Sturge says:
$$L = (28 - 8)/(1 + 1.44 \log_e 75) = 2.8,$$
while GENSTAT says:
$$L = (28 - 8)/\sqrt{(75)} = 2.3.$$
So a sensible length for the class interval is 2, and we use class interval boundaries at 8, 10, 12, ..., 28. Effectively, the class intervals are 8–9, 10–11, ..., 28–29.

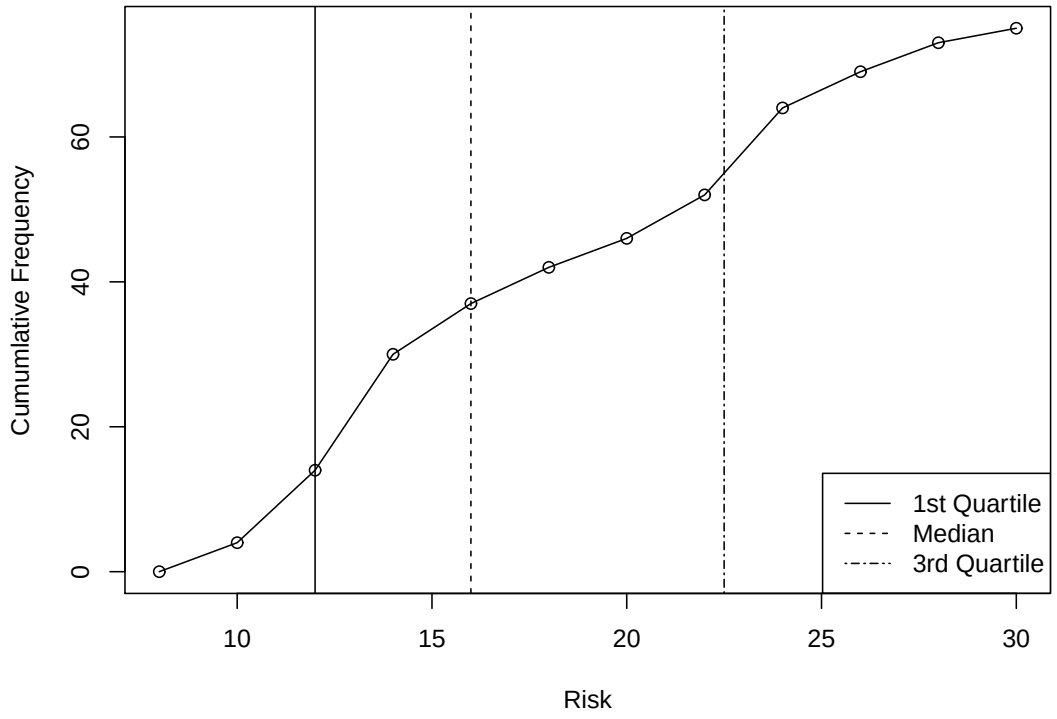
3. Count the number of share standard deviations falling into each class:

Class interval	Frequency
8–9	4
10–11	10
12–13	16
14–15	7
16–17	5
18–19	4
20–21	6
22–23	12
24–25	5
26–27	4
28–29	2
Total	75

Finally, we plot the histogram:


20
15
10
5
0
Number
of
shares
10
20
30
Risk


20
15
10
5
0
Relative Frequency
of shares
as a percentage
(%)
10
20
30
Risk



The striking feature of this histogram is that is has two clear peaks. In statistical jargon, it is said to be **bimodal**. The visual display of this information has thus revealed information which was not at all obvious from even a careful search through the 75 values in the table of data. The financial analyst now needs an explanation for the bimodality. Further investigation revealed that "gold shares" were predominantly responsible for the peak on the right, while "industrial shares" were found to be responsible for the peak on the left. The histogram reveals that gold shares generally have a substantially higher risk than industrial shares. In layman's terms, we conclude that gold shares are generally more "volatile" than industrial shares.

Example 5C: Plot a histogram to display the examination marks (as percentages) of 25 students and comment on the shape of the histogram:

68	72	39	50	69	52	51	50	41	52	65	37	45
78	48	55	53	61	71	42	57	34	57	66	87	

Example 6C: A company that produces timber is interested in the distribution of the heights of their pine trees. Construct a histogram to display the heights, in metres, of the following sample of 30 trees:

18.3	19.1	17.3	19.4	17.6	20.1	19.9	20.0	19.5	19.3
17.7	19.1	17.4	19.3	18.7	18.2	20.0	17.7	20.0	17.5
18.5	17.8	20.1	19.4	20.5	16.8	18.8	19.7	18.4	20.4

Visual displays of quantitative data II : stem-and-leaf plots ...

Stem-and-leaf plots are a relatively new display technique. The visual effect is very similar to that of the histogram; however, they have the advantage that additional information is represented — the original data values can be extracted from the display. Thus stem-and-leaf plots can be used as a means of **data storage**.

We learn how to construct a stem-and-leaf plot by means of an example, using an old dataset. The procedure is simple.

Example 7A: At the end of 1983/84 English football season, the points scored by each club were as follows:

Arsenal	63	Nottingham Forest	71
Aston Villa	60	Notts County	41
Birmingham City	48	Queens Park Rangers	73
Coventry City	50	Southampton	71
Everton	59	Stoke City	50
Ipswich Town	53	Sunderland	52
Leicester City	51	Tottenham Hotspur	61
Liverpool	79	Watford	57
Luton Town	51	West Bromwich	51
Manchester United	74	West Ham United	60
Norwich City	50	Wolverhampton	29

To produce the stem-and-leaf plot for the points scored by all the teams, we split each number into a "stem" and a "leaf". In this example, the natural split is to use the "tens" as stems, and the "units" as leaves. Because the numbers range from the 20s to the 70s, our stems run from 2 to 7. We write the in a column:

stems	leaves
2	
3	
4	
5	
6	
7	

We now make one pass through the data. We split each number into its "stem" and its "leaf", and write the "leaf" on the appropriate "stem". The first number, the 63 points scored by Arsenal, has stem "6" and leaf "3". We write a "3" as a leaf on stem "6":

stems	leaves
2	
3	
4	
5	
6	3
7	

Aston Villa's 60 points become leaf "0" on stem "6". Birmingham City's 48 points are entered as leaf "8" on stem "4". After the first six scores have been entered, we have:

stems	leaves
2	
3	
4	8
5	093
6	30
7	

Continue until leaves have been entered for all 22 numbers:

stems	leaves	count
2	9	1
3		0
4	81	2
5	0931100271	10
6	3010	4
7	94131	5
		22

We append a third column in which, we enter the count of the number of leaves on each stem, add up the counts, and check that we have entered the right number of leaves!

The final step is to sort the leaves on each stem from smallest to largest, and to add a cumulative count column:

stems	sorted leaves	count	cum. count
2	9	1	1
3		0	1
4	18	2	3
5	0001112379	10	13
6	0013	4	17
7	11349	5	22

What have we created? Essentially, we have a histogram on its side, with class intervals of length 10. But in addition, we have retained all the original information. In a histogram, we would only have known that five teams scored between 70 and 79 points; now we know that there were scores of 71 (two teams), 73, 74, and Liverpool's league-winning 79!

Example 8A: For the data of example 1A, produce and compare stem-and-leaf plots of the GMAT scores of students with Engineering and Arts backgrounds.

All the GMAT scores ended in a zero, so this common feature gives us no useful information about variability of the scores; therefore we may use the hundreds as "stems" and the tens as "leaves". For both categories of students, we would then have only three stems, "5", "6" and "7". Looking back at the histogram display of GMAT scores (example 3A), we note that we used class intervals of width 50 units. We can create this width class interval in the stem-and-leaf plot as demonstrated below.
Engineering:

stems	leaves	count
5·	140144	6
5⋆	8579566	7
6·	11240023	8
6⋆	5658	4
7·	21	2
7⋆	5	1
		28

Arts:

stems	leaves	count
5·	01101022	8
5⋆	6575	4
6·	20	2
6⋆	6	1
7·	1	1
7⋆		0
		16

In this approach, we split each 100 into two stems; the first is labelled "·" and encompasses the leaves from 0 to 4, the second is labelled"⋆" and includes the leaves from 5 to 9.

The final step is to sort the leaves for each stem.

ENGINEERING			ARTS		
stems	sorted leaves	cum. count	stems	sorted leaves	cum. count
5·	011444	6	5·	00011122	8
5⋆	5566789	13	5⋆	5567	12
6·	00112234	21	6·	02	14
6⋆	5568	25	6⋆	6	15
7·	12	27	7·	1	16
7⋆	5	28	7⋆		16

Again, a skewness to the right is evident in both displays.

Striking, too, is the observation that students with an engineering background tend to have GMAT scores in the upper 500s and lower 600s, whereas the majority of arts background students have scores in the 500s. Although the sample sizes are small, this contrast in pattern seems marked enough to suggest that amongst MBA students, the engineers perform better on the GMAT test , on average, than the arts students.

If splitting "stems" into two parts seems inadequate for the data set on hand, here is a system for splitting them into five!

Example 9B: Produce a stem-and-leaf plot for the risk data of Example 4B.

As for the histogram, it would be sensible to use stems of width 2. Each stem is therefore split into five: 0 and 1 are denoted "·", 2 and 3 are denoted "t", 4 and 5 are denoted "f", 6 and 7 are denoted "s", and 8 and 9 are denoted "⋆". Notice the convenient mnemonics — English is a marvellous language!

Arts:

stems	sorted leaves	count	cum. count
⋆	8999	4	4
1·	0001111111	10	14
1t	2222222223333333	16	30
1f	4444555	7	37
1s	67777	5	42
1⋆	8999	4	46
2·	000111	6	52
2t	222233333333	12	64
2f	55555	5	69
2s	6677	4	73
2⋆	88	2	75

Note that we have presented the stem-and-leaf plot with the leaves already sorted.

Example 10C: Produce a stem-and-leaf plot for the examination marks of another group of 25 students.

50	79	53	85	50	53	65	58	43	45	48	51	54
72	71	61	51	72	53	39	67	27	43	69	53	

Example 11C: The maximum temperatures (°C) at 20 towns in southern Africa one summer's day are given in the following table. Produce a stem-and-leaf plot.

Pietersburg	30	Windhoek	32
Pretoria	20	Cape Town	22
Johannesburg	28	George	21
Nelspruit	30	Port Elizabeth	18
Mmabatho	33	East London	17
Bethlehem	30	Beaufort West	23
Bloemfontein	31	Queenstown	22
Kimberley	31	Durban	26
Upington	28	Pietermaritzburg	27
Keetmanshoop	28	Ladysmith	30

Example 12C: In order to assess the prices of the television repair industry, a faulty television set was taken to 37 TV repair shops for a quote. The data below represents the quoted prices in rands. Construct a stem-and-leaf plot and comment on its features.

60	55	158	38	48	120	85	245	90	60	49	38	98
185	200	150	140	75	125	125	125	145	200	145	94	165
105	75	75	120	36	150	120	176	60	78	28		

Five-number data summaries — median, lower and upper quartiles, extremes ...

At the beginning of this chapter, we said that statisticians obtained a feel for the data they were about to analyse in two ways. We have now dealt with the first way, that of constructing a visual display. Now we move on to the second way, by computing a few "key" numbers which summarize the data.

Our aim now is to reduce a large batch of data to just a few numbers which we can grasp simultaneously, and thus help us to understand the important features of the data set as a whole.

It is useful now to introduce the concepts of **rank**. In a numerical dataset of size n sorted from smallest to largest, the smallest number is said to have **rank 1**, the second smallest **rank 2**, ..., the largest **rank n**. We call the smallest number $x_{(1)}$ (so $x_{(1)} = x_{\min}$), the second smallest $x_{(2)}$, ..., and the largest is $x_{(n)}$ (so $x_{(n)} = x_{\max}$). We use $x_{(r)}$ for the **number with rank** r. The cumulative count column of a stem-and-leaf plot makes it easy to find the observation with any given rank.

We use $x_{(r+\frac{1}{2})}$ to denote the number half-way between the numbers with rank r and rank $r+1$:

$$x_{(r+\frac{1}{2})} = \frac{x_{(r)} + x_{(r+1)}}{2}.$$

We say that $x_{(r+\frac{1}{2})}$ is the number with rank $r+\frac{1}{2}$. Such numbers are called **half-ranks**.

We define the **median** of a batch of n numbers as the number which has rank $(n+1)/2$. We use $x_{(m)}$ to denote the median. If n is an odd number, then the rank of the median will be a whole number, and the median will be the "middle number" in the data set. But if n is even, then the rank will be a half-rank, and will be the average of the "two middle numbers" in the data set.

The **lower quartile** is defined to be the number with rank $l = ([m]+1)/2$ where m is the rank of the median. The notation $[m]$ means that if m is something-and-a-half,

we drop the half! The alternative to doing this is having to define "quarter ranks"! The **upper quartile** has rank $u = n - l + 1$. The lower and upper quartiles are denoted $x_{(l)}$ and $x_{(u)}$, respectively. The extremes, the smallest and largest values in a data set, have ranks 1 and n, and we agreed earlier to call them $x_{(1)}$ and $x_{(n)}$, respectively.

These five-numbers provide a useful summary of the batch of data, called, with complete lack of imagination, the **five-number summary**. We write them from smallest to largest:

$$(x_{(1)}, x_{(l)}, x_{(m)}, x_{(u)}, x_{(n)}).$$

Example 13A: Find the five-number summary for the end-of-season football points of Example 7A.

An easy way to find the five-number summary is to use the stem-and-leaf plot.

stems	sorted leaves	count	cum. count
2	9	1	1
3		0	1
4	18	2	3
5	0001112379	10	13
6	0013	4	17
7	11349	5	22

Because $n = 22$, the median has rank $m = (n + 1)/2 = (22 + 1)/2 = 11\frac{1}{2}$. We need to average the numbers with ranks 11 and 12. From the cumulative count, we see that the last leaf on stem 4 has rank 3, and the last leaf on stem 5 has rank 13. Counting along stem 5, we find that 53 is the number with rank 11 and 57 has rank 12. Thus the median is the average of these two numbers $(53 + 57)/2) = 55$; we write $x_{(m)} = 55$. Half the teams scored below 55 points, half scored above 55 points.

The lower quartile has rank $l = ([m] + 1)/2 = ([11\frac{1}{2}] + 1)/2 = (11 + 1)/2 = 6$. The observation with rank 6 is 50, thus $x_{(l)} = 50$. The upper quartile has rank $u = n - l + 1 = 22 - 6 + 1 = 17$. The observation with rank 17 is 63, thus $x_{(u)} = 63$. The five-number summary is:

$$(29, 50, 55, 63, 79).$$

Why is this a big deal? Because it tells us that ...

1. Half the teams scored below 55 points, half scored above 55 points, because 55 is the median.

2. Half the teams scored between 50 and 63 points, because these two numbers are the lower and upper quartiles.

3. A quarter of the scores lay between 29 and 50, a quarter between 50 and 55, a quarter between 55 and 63, and a quarter between 63 and 79.

4. All the scores lay between 29 and 79.

Example 14B: Find the five-number summaries for GMAT scores of both engineering and arts students. Use the stem-and-leaf plot of example 8A.

ENGINEERING				ARTS			
stems	sorted leaves	count	cum. count	stems	sorted leaves	count	cum. count
5·	011444	6	6	5·	00011122	8	8
5⋆	5566789	7	13	5⋆	5567	4	12
6·	00112234	8	21	6·	02	2	14
6⋆	5568	4	25	6⋆	6	1	15
7·	12	2	27	7·	1	1	16
7⋆	5	1	28	7⋆		0	16

For the engineers, the median has rank $m = (28 + 1)/2 = 14\frac{1}{2}$. Thus $x_{(m)} = (600 + 600)/2 = 600$. The lower quartile has rank $l = ([m] + 1)/2 = ([14\frac{1}{2}] + 1)/2 = 7\frac{1}{2}$, and the upper quartile rank $n - l + 1 = 22 - 7\frac{1}{2} + 1 = 21\frac{1}{2}$. So $x_{(l)} = (550+550)/2 = 550$, and $x_{(u)} = (640 + 650)/2 = 645$. The five-number summary is

$$(500, 550, 600, 645, 750).$$

For the arts students, the median has rank $m = (16 + 1)/2 = 8\frac{1}{2}$. Thus $x_{(m)} = (520 + 550)/3 = 535$. The lower quartile has rank $l = ([m] + 1)/2 = ([8\frac{1}{2}] + 1)/2 = 4\frac{1}{2}$, and the upper quartile rank $n - l + 1 = 16 - 4\frac{1}{2} + 1 = 12\frac{1}{2}$. So $x_{(l)} = 510$, and $x_{(u)} = 585$. The five-number summary is

$$(500, 510, 535, 585, 710).$$

For the engineers, the median GMAT score was 600; by contrast, for arts students, it was only 535. The central 50% of engineers obtained scores in the interval from 550 to 645, while the central 50% of arts students were in a downwards-shifted interval, 510 to 585. This comparison of the central 50% ranges, reinforces our earlier interpretation that engineers tend to have higher GMAT scores than arts students.

Example 15C: Find the five-number summaries for the data of (a) Example 10C, (b) Example 11C, and (c) Example 12C.

Visual displays of quantitative data III : box-and-whisker plots ...

Five-number summaries can be displayed graphically by means of **box-and-whisker** plots. (This ridiculous name was invented by the American statistician who invented the method, **John Tukey** , who also invented both the name "stem-and-leaf plot" and the plot itself! John Tukey was not only an inventor of crazy names; he also made an enormous impact on the theory and practice of the discipline Statistics.) Once again, we will use an example to describe how to produce a box-and-whisker plot.

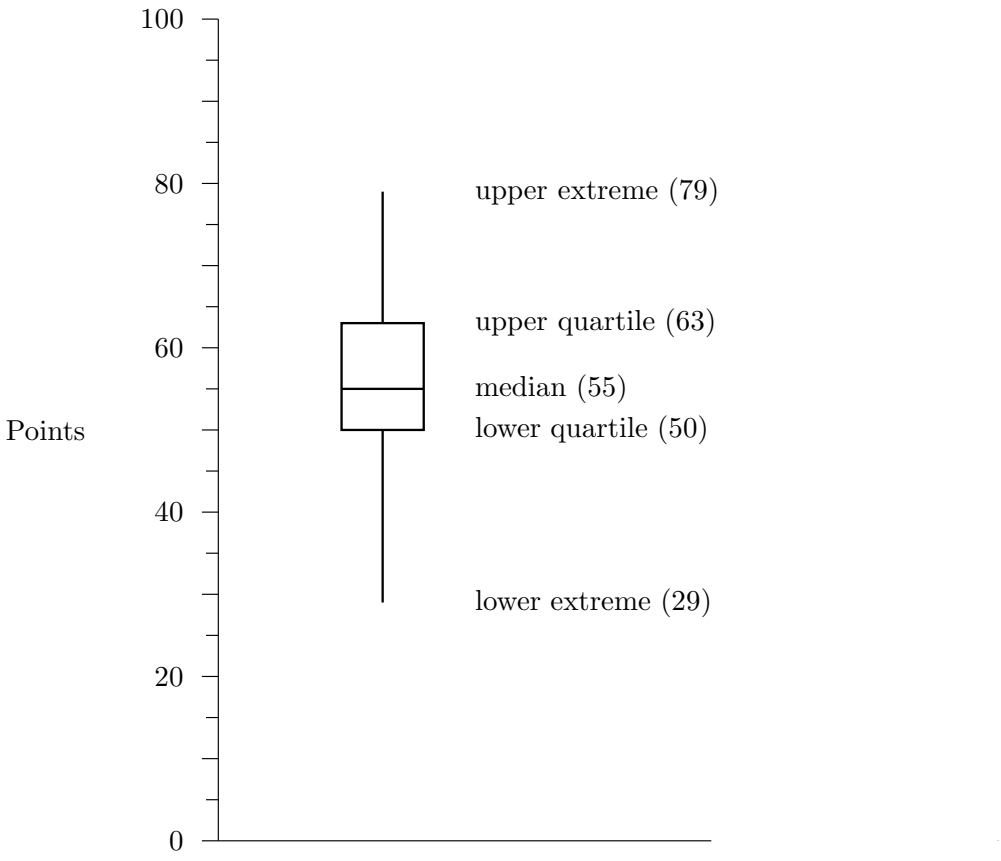


Figure 1.1: Football team points ($n = 22$)

Example 16A: Produce a box-and-whisker plot for the football team points of Example 7A, using the five-number summary $(29, 50, 55, 63, 79)$ computed in Example 13A. The procedure is simple:

1. Draw a vertical axis which covers at least the complete range of all the data values.
2. Draw a "box" from the lower to the upper quartile.
3. Draw a line across the box at the median.
4. Draw "whiskers" from the box out to the extremes.

Applied to the five-number summary $(29, 50, 55, 63, 79)$, this procedure yields the box-and-whisker plot in Figure 1.1.

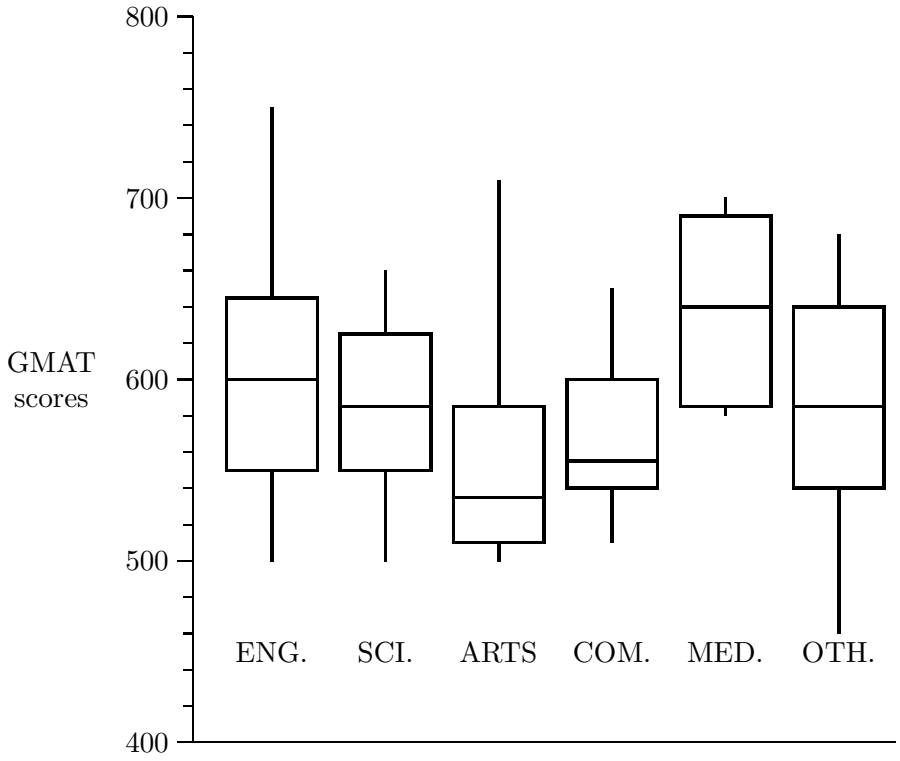
Box-and-whisker plots are especially useful when we wish to compare two or more sets of data. To achieve this comparison, we construct the plots side-by-side. It is essential to use the same vertical scale for all the plots that are to be compared.

Example 17B: Draw a series of box-and-whisker plots to compare the GMAT scores of each category of MBA students.

We computed the five-number summaries of the GMAT scores for engineering and arts students in Example 14B. The five-number summaries for all the categories of the data in example 1A are:

Engineering	(500, 550, 600, 645, 750)	n=28
Science	(500, 550, 585, 625, 660)	n=16
Arts	(500, 510, 535, 585, 710)	n=16
Commerce	(510, 540, 555, 600, 700)	n=10
Medicine	(580, 585, 640, 690, 700)	n=6
Other	(460, 540, 585, 640, 680)	n=5

The box-and-whisker plots, shown side-by-side, reveal the differences between the various categories of students.



We see from a comparison of the box-and-whisker plots that the students in this class with a medical background had the highest median GMAT score, followed by engineers, with arts students having the lowest median. The skewness to the right (now shown as a long whisker pointing upwards!) which we commented on earlier for the class as a whole, is also evident for engineering, science, arts and commerce students, the categories for which the sample sizes were large.

Outliers and strays...

In many data sets, there are one or more values that appear to be very different to the bulk of the observations. Intuitively, we recognize these values because they are a long way from the median of the data set as a whole. We can make our stem-and-leaf plots more informative by plotting and labelling some of the outlying values in such a way that they are highlighted and our attention immediately drawn to them. These outlying values may be valid but they could well represent errors that have crept into the data, either when the observation was made, or when the numbers were transcribed from one sheet of paper to another, or when they were entered into a computer, or even when they were being transferred from one computer to another. On the other hand, if outlying values might represent genuine observations, they may be of special interest and

importance. In any event, these observations need to be checked, and either confirmed or rejected. It is useful to have rules that will aid us to identify such observations.

Outliers are those observations which are greater than

$$x_{(m)} + 6(x_{(u)} - x_{(m)})$$

or less than

$$x_{(m)} - 6(x_{(m)} - x_{(l)})$$

and we label them boldly on the box-and-whisker plot.

Less outlying values called **strays** are those observations which are not outliers but are greater than

$$x_{(m)} + 3(x_{(u)} - x_{(m)})$$

or less than

$$x_{(m)} - 3(x_{(m)} - x_{(l)})$$

and we label them less boldly on the plot.

The largest and smallest observations which are not strays are called the **fences** (more Tukeyisms!). When outliers and strays are being portrayed in a box-and-whisker plot, the convention is to take the whiskers out as far as the fences, not the extremes. This strategy helps to isolate and highlight the outlying values.

In any event, it is sometimes helpful to identify a few values of special interest or importance in a box-and-whisker plot.

Example 18A: The university computing service provides data on the amount of computer usage (hours) by each of 30 students in a course:

Student no.	Usage	Student no.	Usage	Student no.	Usage
AD483	53	AM044	2	AS677	36
CI144	7	CS572	25	EK817	20
FV246	38	GM337	36	GR803	33
HN050	48	JK314	84	JR894	154
JV670	31	KM232	35	LJ419	44
LW032	48	MA276	69	MJ076	95
PH544	4	PS279	60	RR676	18
SA831	51	SC186	47	SS154	37
TB864	11	VO822	41	WG794	34
WB909	73	YG007	38	ZP559	125

Is the lecturer justified in claiming that particular students appear to be making excessive use of the computer (playing games?) while the usage of others is so low that she is suspicious that they are not doing the computing work themselves?

The stem-and-leaf plot is

stems	sorted leaves	cum. count
0	247	3
1	18	5
2	05	7
3	134566788	16
4	14788	21
5	13	23
6	09	25
7	3	26
8	4	27
9	5	28
10		28
11		28
12	5	29
13		29
14		29
15	4	30

The five-number summary is (2, 31, 38, 53, 154).

The outliers were those observations greater than

$$x_{(m)} + 6(x_{(u)} - x_{(m)}) = 38 + 6(53 - 38) = 128$$

or less than

$$x_{(m)} - 6(x_{(m)} - x_{(l)}) = 38 - 6(38 - 31) = -4.$$

There was only one outlier, the usage of 154 hours by student JR894.

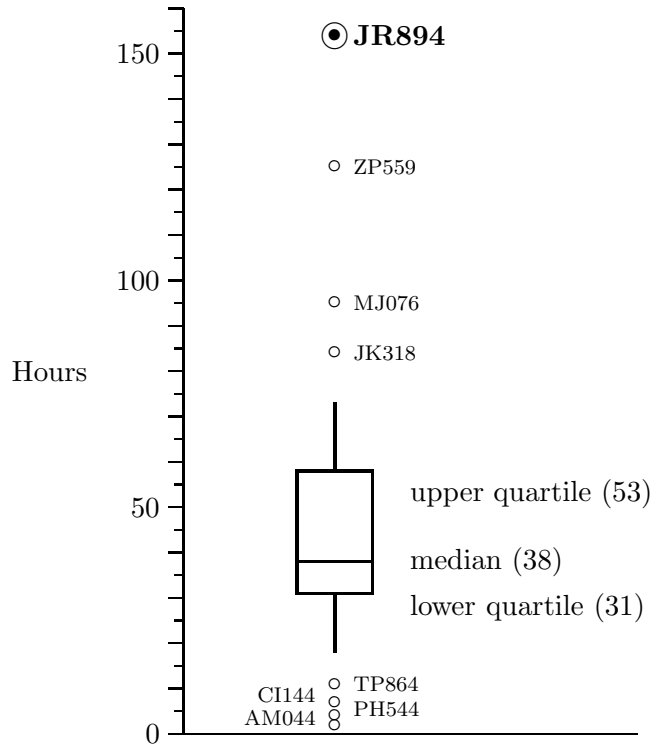
The strays were those observations which were not identified as outliers but were greater than

$$x_{(m)} + 3(x_{(u)} - x_{(m)}) = 38 + 3(53 - 38) = 83$$

or less than

$$x_{(m)} - 3(x_{(m)} - x_{(l)}) = 38 - 3(38 - 31) = 17.$$

There are seven strays: four students (AM044 (2 hours), PH544 (4 hours), CI144 (7 hours), TB864 (11 hours)) are at the low usage end, and three (JK314 (84 hours), MJ076 (95 hours) and ZP559 (125 hours) are at the high usage end. The fences are the outermost observations that were not strays, and are the 18 hours and 73 hours. The box-and-whisker plot, with the outlier labelled boldly, and strays merely labeled



The lecturer now has a list of students whose computer utilization appears to be suspicious.

Example 19C: A company that produces breakfast cereals is interested in the protein content of wheat, the basic raw material of its products. The protein content(percentages of mass) of 29 samples of wheat (percentages) was recorded as follows:

9.2	8.0	10.9	11.6	10.4	9.5	8.5	7.7	8.0	11.3	10.0	12.8	8.2	10.5	10.2
11.9	8.1	12.6	8.4	9.6	11.3	9.7	10.8	83	10.8	11.5	21.5	9.4	9.7	

Confirm the statistician's conclusion that the values 83 and 21.5 are outliers.

The statistician asked that these values should be investigated. Checking back to the original data, it was discovered that 83 should have been 8.3, and 21.5 should have been 12.5. Transposed digits and misplaced decimal points are two of the most frequent types of error that occur when data is entered into a computer.

Example 20C: A winery is concerned about the possible impact of "global warming" on the grape crop. It was able to obtain some interesting historical rainfall data going back to 1884 from a wine-producing region. The rainfall (mm) in successive Januaries at Paarl for the 22-year period 1884–1905 were recorded as follows:

Year	Rain	Year	Rain	Year	Rain
1884	2.6	1892	37.8	1900	3.0
1885	4.9	1893	.0	1901	145.1
1886	16.3	1894	.0	1902	39.7
1887	21.6	1895	52.3	1903	105.9
1888	6.1	1896	4.1	1904	17.8
1889	.0	1897	6.4	1905	10.6
1890	.0	1898	15.8		
1891	1.1	1899	27.7		

(a) Produce a stem-and-leaf plot.
(b) Find the five-number summary.
(c) Draw the box-and-whisker plot, showing outliers and strays, if any.

"Statistics" in Statistics ...

Within the discipline Statistics, we give a precise technical definition to the concept, a **statistic**. A statistic is any quantity determined from the data values of a sample. Thus the median is "a statistic", and so are the other four numbers that make up a five-number summary. These quantities are examples of **summary statistics,** because they endeavour to summarize specific aspects of the information concealed within the sample data. We now learn about a further bunch of "statistics".

Measures of location and spread ...

We use the term **measure of location** to describe any statistic that purports to locate the "middle", in some sense, of the data set. For example, confronted by a collection of data on house prices, we would use a measure of location to answer the question: *What is the typical price of a house?*

The next questions might be: *How much variability is there in house prices? What is the difference between the price of a cheap house and that of an expensive house?* **Measures of spread** are designed to provide answers to these two questions. In the next few sections we consider a few of the most important measures of location, and then some measures of spread.

The sample median

The **median**, which we denoted $x_{(m)}$, locates the "middle" of the data in the sense that half the observations from the sample are smaller than the median and half are larger than the median. To find the median it is necessary to **sort** or rank the data values from the smallest value to the largest. Remember that if the sample size n is an odd number, the median is the "middle" observation, but if n is even, the sample median is the average of the "two middle" observations.

The sample mean...

The **sample mean** is, with good justification, the most important measure of location. It is found by adding together all the values of a variable in the sample data, and dividing this total by n, the sample size. We introduce a subscript notation to describe a sample of size n. We denote the first observation we make on the variable X as x_1, the second x_2, ..., the nth x_n. Then the sample mean of the X values, almost universally denoted $\bar{x}$ (pronounced, "x bar"), is defined to be

$$\begin{aligned}\bar{x} &= (x_1 + x_2 + \cdots + x_n)/n \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

The sample mean locates the "middle" of the batch of data values for the variable X in a special way. It is equivalent to hanging a 1 kg mass at points x_1, x_2, ... x_n along a ruler (of zero mass), and then $\bar{x}$ is the point at which the ruler balances. (The masses in particular need not be be 1 kg, but they must all be equal!)

The mean is much easier to calculate than the median. The mean requires a single pass through the data, adding up the values. In contrast, the data needs to be sorted before the median can be computed, an operation which often requires several passes through the data.

Example 21A: Find the sample mean of the dividend yields of 15 shares in the paper and packaging sector of the Johannesburg Stock Exchange. Also find the median. Compare the mean and the median. The yields are expressed as percentages.

Copi	3.3	E. Haddon	7.6	Pr. Paper	6.7
Caricar	8.4	Kohler	7.1	Prs. Sup	2.9
Coates	10.7	Metal Box	6.6	Sappi	7.5
Consol	6.0	Metaclo	8.6	Trio Rand	8.2
DRG	9.6	Nampak	5.8	Xactics	3.0

We sum the 15 dividend yields and divide by 15:

$$\begin{aligned}\bar{x} &= (3.3 + 8.4 + 10.7 + \cdots + 8.2 + 3.0)/15 \\ &= 6.80(\%)\end{aligned}$$

The stem-and-leaf plot for these 15 data values is shown below:

stems	sorted leaves	count	cum. count
2	9	1	1
3	03	2	3
4		0	3
5	8	1	4
6	067	3	7
7	156	3	10
8	246	3	13
9	6	1	14
10	7	1	15

The median has rank $m = (15 + 1)/2 = 8$, and thus $x_{(m)} = 7.1\%$. In this example, there is little difference between the two measures of location. But this is not always the case ...

Example 22A: Find the mean and the median of the weekly volume of the same 15 shares as in Example 21A. The weekly volume is the number of shares traded in a week.

Copi	2 300	E. Haddon	0	Pr. Paper	700
Caricar	2 100	Kohler	100	Prs. Sup	0
Coates	3 100	Metal Box	111 400	Sappi	40 600
Consol	1 200	Metaclo	700	Trio Rand	84 100
DRG	31 800	Nampak	100	Xactics	45 900

The sample mean is $\bar{x} = (2\,300 + 2\,100 + \cdots + 84\,100 + 45\,900)/15 = 21\,607$ (shares traded per week).

Sort the data, locate the middle (8th) value, and find that the median is $x_{(m)} = 2\,100$ (shares traded per week).

The mean is just over 10 times larger than the median. What has gone wrong? Nothing, it is simply just that the mean and median locate the "middle" of the data according to a different set of rules! In this example, the mean has been dragged upwards by a few large values, so that only five of the fifteen numbers are larger than the mean. But even if a million Metal Box shares had been traded during the week, the median would have remained the same! The median is an example of a measure of location which is said, in statistical jargon, to be **robust**. The mean is not robust, being **sensitive** to outlying values in the data set. Because the mean is not robust, it is important to be aware of possible outliers in any sample of data for which the mean is being computed.

The mean and the median tend to be close to each other when the distribution of the values is symmetric and there are no outlying observations. The mean and median differ increasingly as the distribution of the data becomes more and more skew. The observations in the long tail of a skew distribution drag the mean in the direction of the tail. The sample mean of a very skew distribution might give a totally misleading impression of the "middle" of the data set.

There are no hard-and-fast rules which state when to use the sample mean and when to use the median as a measure of location. In general terms, the median is good for most sets of data. The sample mean is most useful when the data has a symmetric distribution. Data with a long tail to the right can be made more symmetric by taking logarithms or taking square roots of all the data values. Such manipulations to the original data values are called **transformations**.

The sample mean has mathematical advantages over the median. The mean is a FAR easier statistic for the mathematical statisticians to use in algebraic manipulations with than the median. A vast amount of statistical theory has been developed for the sample mean, and for this reason it is the predominant measure of location used in sophisticated statistical methods.

Measures of spread ...

Measures of spread give insight into the variability of a set of data. Two measures of spread can be defined in an obvious way from the five-number summary. They are: the **range** R, defined as

$$R = x_{(n)} - x_{(1)},$$

and the **interquartile range** I, defined as

$$I = x_{(u)} - x_{(l)}.$$

The range is unreliable as a measure of spread because it depends only on the smallest and largest values in the sample, and is thus as sensitive as it can possibly be to outlying values in the sample. It is the ultimate example of a non-robust statistic! On the other hand, the interquartile range is the length of the interval covering the central half of the data values in the sample, and it is not sensitive to outliers in the data. The interquartile range is a robust measure of spread.

The **sample variance** and its square root, the **sample standard deviation**, have the same advantage, easier algebraic manipulation, over the range and interquartile range that the mean had over the median. Therefore the sample variance is frequently the only measure of spread calculated for a set of data.

The sample variance, denoted by s^2, is defined by the formula

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

In words, it is the sum of the squared differences between each data value and the sample mean, with this sum being divided by one less than the number of terms in the sum. The sample standard deviation, denoted by s, is the square root of the sample variance.

It is a nuisance to have these two measures of spread, s and s^2, one of which is simply the square root of the other. Why have both? The standard deviation is the easier of the two measures of spread to use intuitively , largely because it is measured in the same units as the original data. The variance is measured in "squared units", an awkward quantity to visualize or keep in mind. For example, if data consists of prices measured in rands, the sample variance has units "squared rands" (whatever that means!), but the standard deviation is in "rands". Even worse, if the data consists of percentages, the sample variance has units "$\%^2$", whereas the standard deviation has the intelligible units "%". But mathematical statisticians prefer to work with the variance — not having to deal with a square root in the algebra makes their lives simpler and neater. So the two equivalent measures of spread co-exist side by side, and we just have to come to terms with both of them.

Example 23A: Compute the sample variance s^2 and the standard deviation s for the dividend yields of the 15 shares of Example 21A.

We have computed $\bar{x} = 6.8\%$. So

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{(15-1)}\big[(3.3-6.8)^2 + (8.4-6.8)^2 + (10.7-6.8)^2 + \cdots \\ &\qquad \cdots + (8.2-6.8)^2 + (3.0-6.8)^2\big] \\ &= \frac{1}{14}\big[(-3.5)^2 + (1.6)^2 + (3.9)^2 + \cdots + (1.4)^2 + (-3.8)^2\big] \\ &= \frac{1}{14}\big[12.25 + 2.56 + 15.21 + \cdots + 1.96 + 14.44\big] \\ &= \frac{1}{14}\big[75.62\big] \\ &= 5.40\% \end{aligned}$$

The standard deviation is $s = \sqrt{5.40} = 2.32\%$.

The variance and the standard deviation are always positive. This fact is guaranteed, because all the terms in the sum are squared, which makes them positive, even though some of the individual differences are negative.

The variance can be calculated more efficiently by a **short-cut formula**. The "short cut" involves reducing the number of subtractions needed to calculate the variance, from n to just 1 subtraction. Examine the following steps carefully:

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

The third term involves adding $\bar{x}^2$ to itself n times. So it is equal to $n\,\bar{x}^2$. But $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$, so

$$n\,\bar{x}^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2$$

The second term in the sum above can also be rewritten:

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

Substituting these expressions for the second and third terms yields

$$(n-1)s^2 = \sum_{i=1}^{n} x_i^2 - \frac{2}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2 + \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2$$
$$= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2.$$

Thus the short-cut formula for s^2 is

$$s^2 = \frac{1}{n-1}\Big[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2\Big].$$

Look carefully at this formula. There is now only **one** subtraction, whereas the original formula involved n subtractions.

Example 24A: Calculate the sample variance of the 15 dividend yields again, this time using the short-cut formula.

We need $\sum_{i=1}^{n} x_i$, the sum of the data values, given by

$$\sum_{i=1}^{n} x_i = 3.3 + 8.4 + \cdots + 3.0 = 102.0;$$

and $\sum_{i=1}^{n} x_i^2$, the **sum of squares** of the data values, i.e. square them first, then add them, like this:

$$\sum_{i=1}^{n} x_i^2 = 3.3^2 + 8.4^2 + \cdots + 3.0^2 = 769.22$$

Then

$$s^2 = \frac{1}{n-1}\Big[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2\Big]$$
$$= \frac{1}{14}\Big[(769.22 - \frac{1}{15}(102.0)^2\Big] = 5.40$$

as before.

If the data has a symmetric unimodal distribution with no outliers, then the **standard deviation** has the following approximate interpretation. The interval from one standard deviation below the sample mean to one standard deviation above it, $(\bar{x}-s, \bar{x}+s)$, should contain about two-thirds of the observations. Thus the sample mean and the sample standard deviation together provide a "two-number summary" of the data set. Many data sets are summarized by these two statistics — the sample mean provides a measure of location and the sample standard deviation a measure of spread.

However, the sample variance and the sample standard deviation have the disadvantage that, like the mean, they are sensitive to outliers. They are sensitive in two ways. First of all, the outlier distorts the mean, so all the differences $(x_i - \bar{x})$ are misleading. Secondly, if x_j, the jth data value, is an outlier, then the term $(x_j - \bar{x})$ will be large relative to the other differences, and, once it is squared, it can make a disproportionately large contribution to the sum of squared differences.

Note that the intervals $(x_{(1)}, x_{(n)})$, $(\bar{x}-s; \bar{x}+s)$, and $(x_{(l)}, x_{(u)})$ cover exactly 100%, approximately 68%$\approx$ two-thirds, and exactly 50% of the observations, respectively. But it is not possible to make direct comparisons between the range, the standard deviations and the interquartile range.

Example 25B: Calculate the sample means, sample standard deviations, medians, interquartile ranges and ranges of the GMAT scores for each faculty category of Example 1A. Comment on the results.

For the sample mean and variance, we need the quantities $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$. For the category "Engineering", we have

$$\sum_{i=1}^{n} x_i = 619 + 510 + \cdots + 710 = 16\,850$$

$$\sum_{i=1}^{n} x_i^2 = 610^2 + 510^2 + \cdots + 710^2 = 10\,254\,100$$

Then

$$\bar{x} = 16\ 850/28 = 601.8$$

and

$$s^2 = \frac{1}{27}(10\,254\,100 - (16\,850)^2/28) = 4\,222.6.$$

The standard deviation is $s = \sqrt{4\,222.6} = 65.0$.

The remaining measures of spread and location can readily be obtained from the five-number summaries of Example 17B. For example, for "Engineering", the lower and upper quartiles were $x_{(l)} = 550$ and $x_{(u)} = 645$, and the interquartile range is $I = 645 - 550 = 95$. The extremes were $x_{(1)} = 500$ and $x_{(n)} = 750$, so the range $R = 750 - 500 = 250$.

Calculation of these summary statistics for the remaining five categories yields the table:

	Location		Spread		
First degree	$\bar{x}$	$x_{(m)}$	s	I	R
Engineering	601.8	600	65.0	95	250
Science	583.1	585	48.5	75	160
Arts	555.6	535	62.8	75	210
Commerce	567.0	555	45.7	60	140
Medicine	638.0	640	53.1	105	120
Other	581.7	585	80.1	100	220

In commenting on this table, we inspect first the measures of location. The sample means show that students with first degrees in medicine had the highest mean GMAT score (638.0), followed by engineering students (601.8), and then commerce (567.0). The lowest mean was recorded for arts students (555.6). The medians follow the same pattern, and apart from arts, the sample means and medians are relatively close. In the box-and-whisker plots in Example 17B, we saw that the distribution of GMAT scores for arts appeared to be strongly skewed to the right. Hence the difference between the sample mean (555.6) and the median (535) for this category of students is consistent with the earlier evidence of skewness.

For the measures of spread, it is evident that the category "Other" has the largest standard deviation (80.1), followed by engineering (65.0). The smallest standard deviation was for commerce (45.7). The interquartile ranges (I) and ranges (R) follow a broadly similar pattern. A plausible explanation as to why the category "Other"

should have the largest standard deviation (and the second largest interquartile range and range) is that it encompasses a wide diversity of students, not belonging to any of the single faculty categories.

The conclusions reached here provide a partial description of this MBA class of 81 students. If the set of MBA students were "representative" of all MBA students at all universities, we might be able to generalize the statements. Another concern that we would have to address before we could generalize the results, relates to issues of sample size. Could the differences in the measures of location and spread we observed here occur just because we had an unusually bright group of, for example, medical students in this MBA class? We will defer further consideration of these statistical issues until chapter 8! In order to prepare ourselves for taking that kind of decision we have to learn some probability theory.

Example 26C: Calculate the sample mean and standard deviation, the median, the range and interquartile range of Paarl rainfall data (Example 20C).

Example 27C: (a) Suppose that the sample mean and standard devation of the n numbers $x_1, x_2, \ldots, x_n$ are $\bar{x}$ and s. An additional observation x_{n+1} becomes available. Show that the updated mean $\bar{x}^\star$ is

$$\bar{x}^\star = \frac{n\bar{x} + x_{n+1}}{n+1}$$

and the updated standard deviation $s^\star$ is

$$s^\star = \sqrt{\frac{1}{n}\big[(n-1)s^2 + n(\bar{x} - \bar{x}^\star)^2 + (x_{n+1} - \bar{x}^\star)^2\big]}.$$

(b) The sample mean of nine numbers is 4.8 with standard deviation 3.0. A 10th observation is made. It is 6.8. Update the mean and the standard deviation.

Exploratory data analysis ...

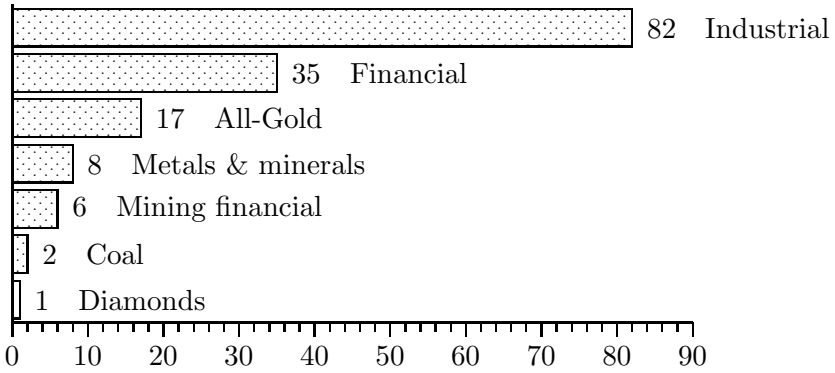
The techniques we have learnt in this chapter have largely been aiming at getting a feel for a sample of data, a process somewhat grandly called **exploratory data analysis**. In the age of instant arithmetic and the personal computer, the temptation is to use the statistical methods of chapters 8 to 12 and beyond, and to accept the answers uncritically.

We have seen in this chapter that the presence of one or more outliers in a sample can have a pretty devastating effect on the sample mean and the sample standard deviation, the most frequently used summary statistics of all. Likewise, we have seen how skewness affects these statistics. Most statistical methods make a variety of assumptions — many of these can be checked out, visually at least, by the exploratory data analysis techniques described in this chapter. Many of these techniques have become part of the data analysis software of statistical packages. You are strongly encouraged to use them **before** you do more complex statistical analyses.

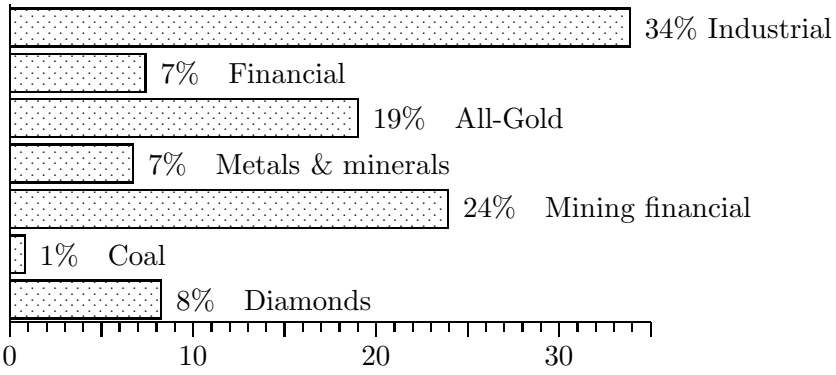
Solutions to examples ...

2C The frequencies and relative frequencies, from which the bar graphs are constructed, are given in the table.

Major sector	(a) Frequency	(b) Percentage of All-Share Index
Coal	2	0.82%
Diamonds	1	8.27%
All-Gold	17	18.99%
Metals & minerals	8	6.73%
Mining financial	6	23.92%
Financial	35	7.41%
Industrial	82	33.86%
Total	151	100%



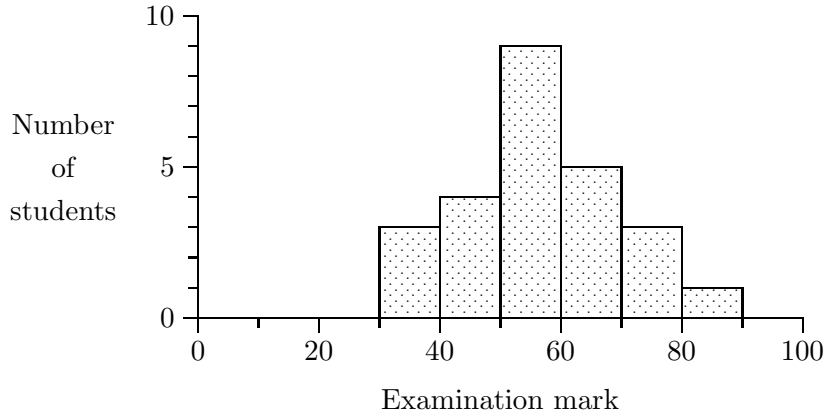
(a) Number of shares



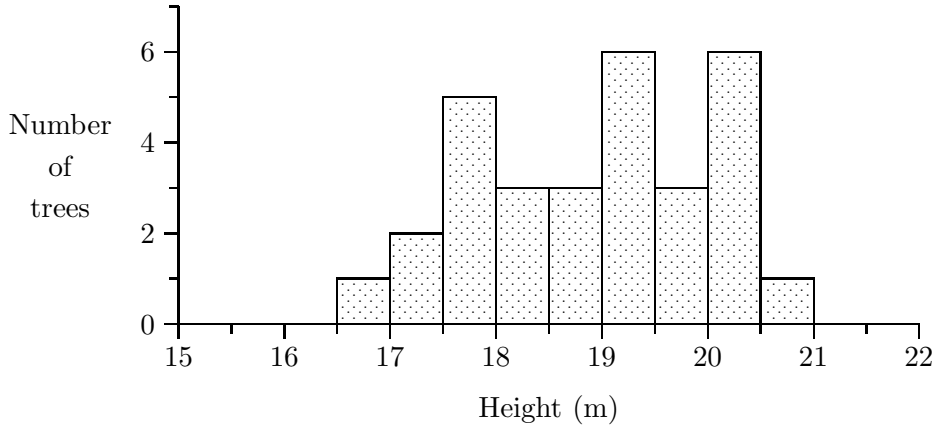
(b) Percentage of All-Share Index

Keeping the ordering of the shares the same in both charts highlights the fact that the All-Share Index does not give equal weighting to each share. Especially striking is the large weighting that the Mining financial sector has in the All-Share Index in relation to the small number of shares. In fact, the single share Anglo-American had a weighting of 8.95% in the All-Share Index!

5C We chose a class interval of 10.



6C We used a class interval of 0.5.



10C

stems	sorted leaves		cum. count
2	7	1	1
3	9	1	2
4	3358	4	6
5	0011333348	10	16
6	1579	4	20
7	1229	4	24
8	5	1	25

11C

stems	sorted leaves	count	cum. count
1∗	78	2	2
2·	01223	5	7
2∗	67888	5	12
3·	00001123	8	20

12C

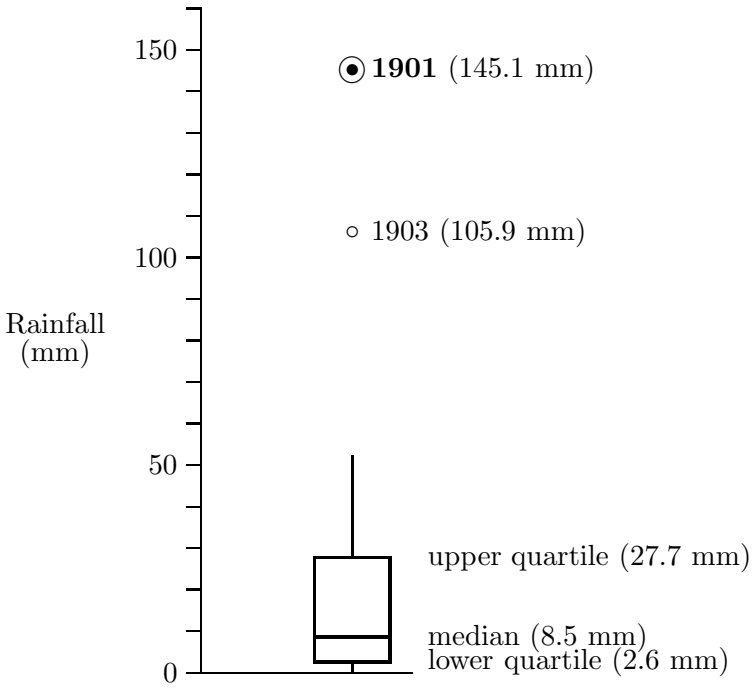
stems	sorted leaves	count	cum. count
0 ·	28,36,38,38,48,49	6	6
0∗	55,60,60,60,75,75,78,85,90,94,98	12	18
1 ·	05,20,20,20,25,25,25,40,45,45	10	28
1∗	50,50,58,65,76,85	6	34
2 ·	00,00,45	3	37

15C (a) (27, 50, 53, 67, 85)
(b) (17, 22.5, 28, 30, 33)
(c) (28, 60, 105, 145, 245)

20C

stems	sorted leaves	cum. count
0	0.0,0.0,0.0,0.0,1.1,2.6,3.0,4.1,4.9,6.1,6.4	11
1	0.6,5.8,6.3,7.8	15
2	1.6,7.7	17
3	7.8,9.7	19
4		19
5	2.3	20
6		20
7		20
8		20
9		20
10	5.9	21
11		21
12		21
13		21
14	5.1	22

(b) (0.0, 2.6, 8.5, 27.7, 145.1)



26C $\bar{x} = 235.8$ $s = 365.7$ $x_m = 85$ $R = 1451$ $I = 251.$

Exercises ...

1.1 As a cartoon strip matures, it is likely to change in subtle ways. In this exercise, we want to look at the pattern of word usage in Shultz's *Peanuts*, comparing the period 1959/60 with 1975/76. The tables below show the number of words per cartoon strip in the two periods. Produce stem-and-leaf plots and find five-number summaries for both periods. Draw side-by-side box-and-whisker plots, and discuss how the number of words per cartoon strip has changed. Also draw cumulative frequency graphs with each case. Number of words in 66 *Peanuts* cartoon strips from 1959/60, reprinted in *You're a winner, Charlie Brown.*

51	35	35	30	41	52	55	38	41	32	49
44	55	42	40	5	15	27	0	16	28	14
23	26	45	63	58	28	37	59	24	35	43
46	43	53	35	34	51	47	43	22	45	23
30	28	36	29	76	40	32	59	32	49	29
39	26	34	35	60	41	47	40	46	45	4

Number of words in 66 *Peanuts* cartoon strips from 1975/76, reprinted in *Let's hear it for dinner, Snoopy.*

37	40	44	44	49	35	34	29	35	37	33
39	49	39	35	35	52	47	45	44	52	28
22	34	36	52	50	39	32	21	20	17	22
32	46	63	39	60	40	30	29	17	30	42
30	45	38	43	30	47	28	39	45	45	32
35	45	41	18	20	33	47	17	40	37	19

1.2 If you are a salesperson, it is very easy to get pessimistic because on many days no sales are made. The days when everthing goes well keep you going. The daily sales of a second-hand car salesperson are tabled below. Display them as a histogram. Calculate also the sample mean, standard deviation and median. Is the median a helpful measure of central tendency for this data?

0	0	1	0	0	0	1	0	0	2	0	1	3	0	1	2	0	0
0	4	0	0	1	0	0	2	0	0	3	0	0	1	0	1	1	1
0	0	0	1	1	0	1	0	1	2	0	0	0	4	0	1	0	2
0	0	1	1	0	3	0	5	0	1	0	0	1	2	0	1	0	0
0	0	1	1	0	0	0	0	1	5	2	0	2	0	0	2	0	1
0	3	1	1	6	0	0	0	0	2	1	3	3	0	0	0	1	0

1.3 Service is an important factor in making a business profitable. For a sample of 25 tables of customers, the manager of a restaurant kept track of how long they waited from arrival to receipt of their main course. The following waiting times (minutes) were recorded:

34	24	43	56	74	45	23	43	56	67	30	19	36
32	65	36	24	54	39	43	67	54	32	18	97	

(a) Display the data as a stem-and-leaf plot, compute the five-number summary and draw the box-and-whisker plot.
(b) Find the sample mean and standard deviation. What proportion of the data values is within one standard deviation of the mean?
(c) Comment on the shape of the distribution of the data, and attempt to interpret it.

1.4 Water is a crucial resource in a generally arid country like South Africa. The mean annual runoff (millions of m^3) of 63 rivers in the old Cape Province of South Africa is given in the table below. Find the five-number summary, and attempt to draw the box-and-whisker plot. Take logarithms of each data value, and repeat the exercise. Discuss the effect of the logarithmic transformation.

River	Mean annual runoff	River	Mean annual runoff	River	Mean annual runoff
Kei	1001	Kasuka	5	Groot Brak	29
Quko	41	Kariega	15	Klein Brak	45
Kwenxura	25	Bushmans	38	Hartenbos	5
Kwelera	32	Boknes	13	Gouritz	744
Gqunube	35	Sundays	29	Kafferkuils	141
Buffalo	82	Coega	13	Duiwenhoks	131
Goba	6	Swartkops	84	Breë	1893
Gxulu	6	Yellowwoods	45	Heuningnes	78
Ncera	6	Gamtoos	485	Uilskraal	65
Tyolumnqa	25	Kabeljou	27	Kleinriviersvlei	96
Kiwane	2	Seekoei	27	Botriviersvlei	116
Keiskamma	133	Krom	105	Palmiet	310
Cqutywa	2	Klipdrif	35	Wildevoëlvlei	38
Bira	6	Storms	69	Sout	38
Mgwalana	5	Elandsbos	67	Diep	43
Mtati	4	Keurbooms	160	Berg	235
Mpekweni	2	Knysna	110	Verlorevlei	102
Fish	479	Goukamma	44	Wadrifsoutpan	19
Kleinemond	6	Swartvlei	73	Jakkals	10
Riet	4	Touw	30	Olifants	1217
Kowie	23	Kaaimans	59	Orange	9344

(Data from Noble and Hemens, *S.A. National Scientific Programmes, Report No 34*, 1976.)

1.5 This is an exercise in robustness! Calculate the median, interquartile range, range, sample mean and sample standard deviation for the following 12 data values:

10.8	9.7	14.1	12.3	10.9	8.9	11.7	12.6	11.2	10.5	8.3	131

Which value looks suspiciously like an outlier? Put its decimal point back in the right place, and recompute the summary statistics. Which of these statistics change, and which remain the same? For which statistic is the percentage change the largest? (The percentage change is the difference between the "correct" and the "biased" values, divided by the correct value, multiplied by 100.)

1.6 Heathrow Airport in London is one of the world's busiest airports. The time of touchdown for planes arriving between 17h30 and 19h30 on 17 October 1991 is recorded (to the second) in the table below. Compute the inter-arrival times in seconds and present appropriate summary statistics. What do you think the target inter-arrival time is? Was there any apparent difference between the first and the

second hour of observation? How frequently did "glitches" (= irregularities) occur?

17h30:07	17h59:36	18h29:51	19h01:41
32:46	18h01:24	31:51	03:33
34:14	03:10	34:04	06:10
37:13	04:26	36:40	07:29
38:56	05:47	37:52	09:25
40:27	08:49	40:41	10:11
41:37	10:27	42:23	12:29
43:21	11:24	43:59	13:41
44:24	12:51	45:20	15:33
45:50	15:34	46:42	17:26
47:10	16:52	48:38	19:14
48:58	18:10	50:44	21:03
50:03	19:24*	52:02	22:24
51:49	21:48	53:44	24:15
52:50	24:20	55:05	25:38
54:32	25:41	56:44	26:59
55:42	26:51	58:11	19h29:26
17h57:21	18h28:22	19h00:06	

* This was the arrival of the supersonic jet *Concord*.

1.7 You have a sample of data $x_1, x_2, \ldots, x_n$, with sample mean $\bar{x}$ and sample variance s^2.

(a) Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.
(b) Add some constant b (i.e. add any number) to each of the x_i. How does this change the sample mean and variance?
(c) Now multiply each of the x_i by a constant c. How does this change the sample mean and variance?
(d) Transform the x_i to $y_i = c \times x_i + b$. What are the sample mean and variance of $y_1, y_2, \ldots, y_n$?

1.8 During a year, a company had the following advertising expenditures: Magazine, R11 000; Newspaper, R45 000; Pamphlets, R8 000; Radio, R34 000; Television, R110 000; Miscellaneous, R5 000. Construct a bar graph that provides an effective visual display of the breakdown of advertising expenditure.

1.9 Take any sets of data which interest you, and apply the exploratory data analysis techniques of this chapter to them.

Solutions to exercises

1.1

1959/60

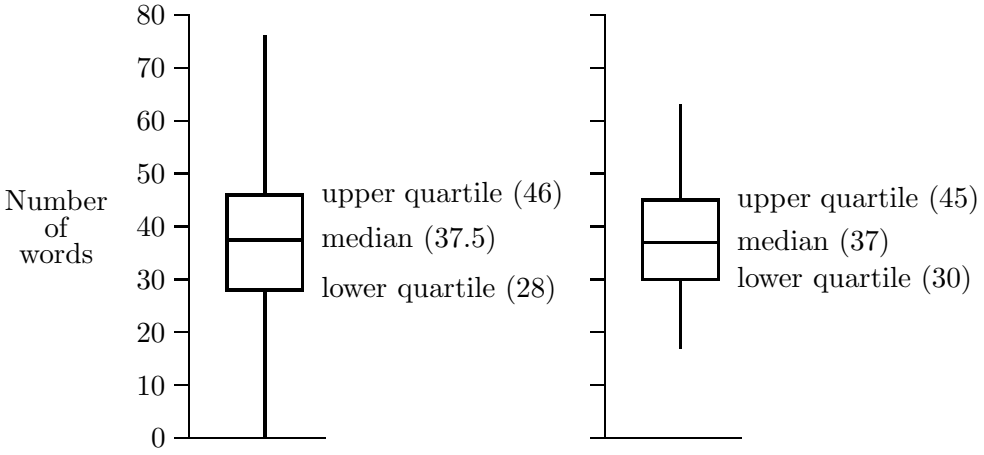
stems	sorted leaves	count	cum. count
0	045	3	3
1	1456	4	7
2	233466788899	12	19
3	0022244555556789	16	35
4	00011123334555667799	20	55
5	12355899	8	63
6	03	2	65
7	6	1	66

1975/76

stems	sorted leaves	count	cum. count
0		0	0
1	77789	5	5
2	01228899	8	13
3	000002223344555556777899999	27	40
4	00012344455555677799	20	60
5	0222	4	64
6	03	2	66
7		0	66

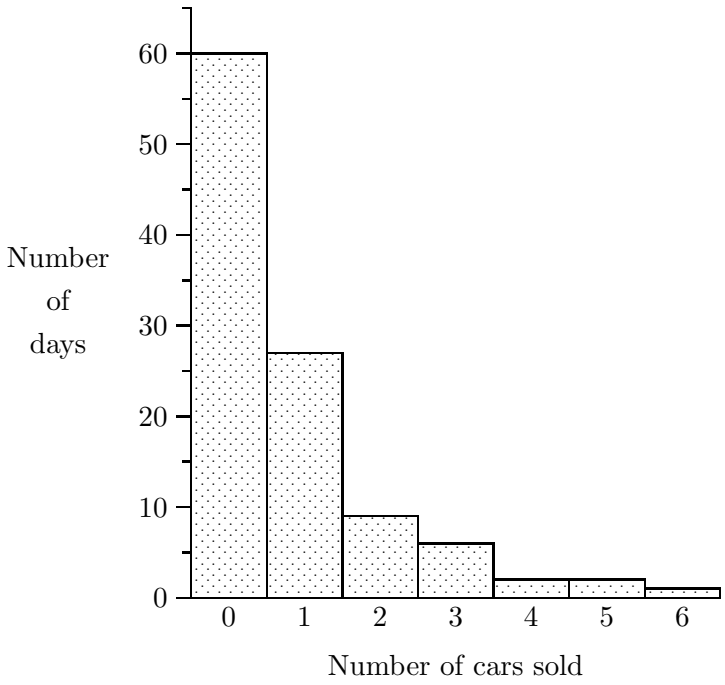
The five-number summaries are (0, 28, 37.5, 46, 76) for the early period and (17, 30, 37, 45, 63) for the late period.

The box-and-whisker plots, plotted side-by-side are:



The medians are similar during both periods, but there appears to be less variability during 1975/76 than during 1959/60 — both the range and the interquartile range are shorter.

1.2

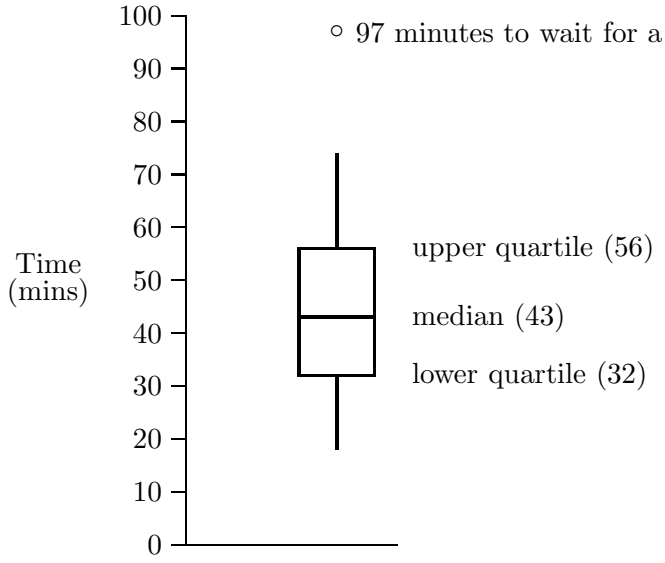


$\bar{x} = 0.82$, $s = 1.24$, $x_m = 0$
No, the median is not much use here — more than half the data values are zero! Even though the data is very skew, the mean is more interesting than the median.

1.3 (a) The stem-and-leaf plot is

stems	sorted leaves	count	cum. count
1	89	2	2
2	344	3	5
3	0224669	7	12
4	3335	4	16
5	4466	4	20
6	577	3	23
7	4	1	24
8		0	24
9	7	1	25

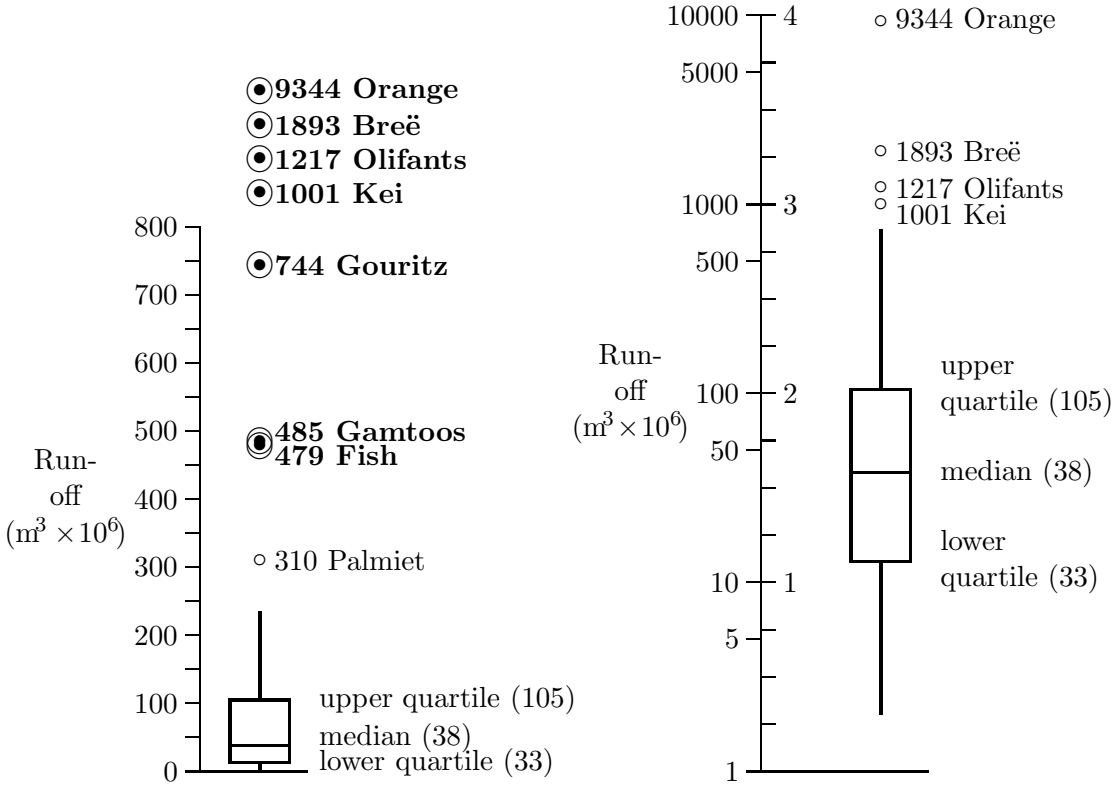
Five-number summary (18, 32, 43, 56, 97). The value of 97 is a stray.



(b) $\bar{x} = 44.4$, $s = 19.3$. 15 out of 25 observations (60%) lie within one standard deviation of the mean, i.e. in the interval $(25.1, 63.7)$.

(c) Apart from the stray (which should be investigated), the data show relatively little skewness to the right. In the context, the interquartile range (24 minutes) is probably wider than desirable, and efforts should be made to make services times more consistent.

1.4 Five-number summary (2, 13, 38, 105, 9344), strays greater than 239, outliers greater than 440. On taking logarithms (base 10), the five-number summary is (0.30, 1.11, 1.58, 2.02, 3.97), strays exceed 2.9, and there are no outliers.

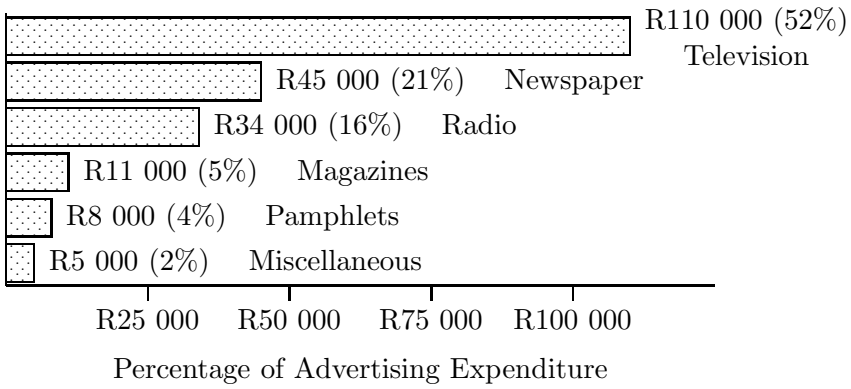


In the plot on the left, the runoffs from four rivers cannot be plotted to scale. The inner scale on the plot on the right shows logarithms to base 10. Notice how

dramatic the effect is — always look at scales on plots to see if data has been transformed.

1.5 With the outlier (131), $x_m = 11.05$, $I = 2.35$, $R = 122.7$, $\bar{x} = 21.00$, and $s = 34.68$. With the outlier corrected (13.1), $x_m = 10.85$, $I = 2.35$, $R = 5.8$, $\bar{x} = 11.18$, and $s = 1.71$. The outlier affects the range, standard deviation and sample mean (in that order).

1.8 The bar graph is most effective if the expenditures are arranged in order:



Chapter 2

SET THEORY

KEYWORDS: Set, subset, intersection, union, complement, empty and universal sets, mutually exclusive sets; pairwise mutually exclusive and exhaustive sets.

Why do we have to do set theory? ...

Simply because one of Murphy's Laws states that before you can do anything, you have to do something else. Before we can do "statistics" we have to do "probability theory", and for that we need some "set theory". So here we go.

Definition of sets ...

We define a set A to be a collection of distinguishable objects or entities. The set A is determined when we can either (a) list the objects that belong to A or (b) give a rule by which we can decide whether or not a given object belongs to A.

Example 1A: (a) If we say, "The letters e, f, g belong to the set A", then we write

$$A = \{e,\ f,\ g\}$$

(b) If we say, "The set B consists of real numbers between 1 and 10 inclusive", then we write

$$B = \{x \mid 1 \leq x \leq 10\}.$$

We read this by saying: "The set B consists of all real numbers x **such that** x is larger than or equal to 1 but is less than or equal to 10."

Because the object e belongs to the set A we write

$$e \in A$$

and we say: "e is an element of A". Because e does not belong to B, we write

$$e \notin B$$

and we say: "e is not an element of B".

Note, firstly, that if $C = \{1,\ 3,\ 5,\ a\}$ and $D = \{a,\ 1,\ 5,\ 3\}$ then $C = D$. The order in which we list the elements of a set is irrelevant. Secondly, if $E = \{a,\ b,\ c,\ a\}$ and $F = \{a,\ b,\ c\}$ then $E = F$. The set E contains only the distinguishable elements a, b and c.

Example 2B:

(a) Express in set theory notation: the set U of numbers which have square roots between 1 and 4.
(b) Write out in full all the elements of the set $Z = \{(x, y) \mid x \in \{1,\ 2,\ 3,\ 4\}, y = x^2\}$.

(a) Because the square roots of numbers between 1 and 16 belong to U, we write $U = \{x \mid 1 \leq x \leq 16\}$.
(b) $Z = \{(1, 1),\ (2, 4),\ (3, 9),\ (4, 16)\}$.

Example 3C: Which of the following statements are correct and which are wrong?

(a) $\{3,\ 3,\ 3,\ 3\} = \{3\}$
(b) $6 \in \{5,\ 6,\ 7\}$
(c) $C = \{-1,\ 0,\ 1\}$
(d) $F = \{x \mid 4 < f < 5\}$
(e) $\{1,\ 2,\ 7\} = \{7,\ 2,\ 1,\ 7\}$
(f) If $H = \{2,\ 4,\ 6,\ 8\}$, $J = \{1,\ 2,\ 3,\ 4\}$ and $K = \{2x \mid x \in H\}$, then $K = J$
(g) $\{1\} \in \{1,\ 2,\ 3\}$.

Subsets ...

Suppose we have two sets, G and H, and that every element of G also belongs to H. Then we say that "G is a **subset** of H" and we write $G \subset H$. We can also write $H \supset G$ and say "H contains G". If every element in G does not also belong to H, we write $G \not\subset H$ and say "G is not a subset of H."

Example 4A: Let $G = \{1,\ 3,\ 5\}$, $H = \{1,\ 3,\ 5,\ 9\}$ and $J = \{1,\ 2,\ 3,\ 4,\ 5\}$. Then clearly $G \subset H$, $H \not\subset J$, $J \supset G$.

Note that the notation $\subset$, $\supset$ for sets is analogous to the notation $\leq$, $\geq$ for ordinary numbers (rather than the notation $<$, $>$). The "round end" of the subset notation tells you which of the sets is "smaller" (in the same way as the "pointed end" shows which of two numbers is smaller).

Our definition of subset has a curious (at first sight) but logical consequence. Because every element in G belongs to G, we can write $G \subset G$. For numbers, we can write $2 \leq 2$.

If $H \subset G$ and $G \subset H$, then, obviously, $H = G$. For numbers, $x \leq 2$ and $x \geq 2$ together imply that $x = 2$.

Example 5C: Let $V = \{v \mid 0 < v < 5\}$, $W = \{0,\ 5\}$, $X = \{1,\ 2,\ 3,\ 4\}$, $Y = \{2,\ 4\}$, $Z = \{x \mid 1 \leq x \leq 4\}$. Which of the following statements are true, and which are false:

(a) $V = W$
(b) $Y \subset X$
(c) $W \supset V$
(d) $Z \supset X$
(e) $X = Z$
(f) $Z \not\subset V$
(g) $Y \subset W$
(h) $Y \in Z$

Intersections ...

Suppose that $L = \{a,\ b,\ c\}$ and $M = \{b,\ c,\ d\}$. Then $L \not\subset M$ and $M \not\subset L$. But if we consider the set $N = \{b,\ c\}$, then we see that $N \subset L$ and $N \subset M$, and that no other set of which N is a subset has this property. This leads us to the idea of intersection.

The **intersection** of any two sets is the set that contains precisely those elements which belong to both sets. For the sets, L, M and N above we write $N = L \cap M$ and read this "N equals L intersection M". The intersection of two sets M and N can be thought of as the set containing those elements which belong to **both** M **and** to N.

Example 6A: If $P = \{x \mid 0 \leq x \leq 10\}$ and $Q = \{x \mid 5 < x < 20\}$, find $P \cap Q$. Is $5 \in P \cap Q$? Is $10 \in P \cap Q$?

Paying careful attention to the endpoints,

$$P \cap Q = \{x \mid 5 < x \leq 10\}.$$

No, $5 \notin P \cap Q$, but, yes, $10 \in P \cap Q$.

The empty set, mutually exclusive sets ...

What happens if $L = \{a,\ b,\ c\}$ and $R = \{d,\ e,\ f\}$? If we want $L \cap R$ to be a set, then we must introduce a new concept, the **empty set**, the set that has no members. This is a sensible concept: consider the set of English-speaking fish, or consider the set of real numbers whose square is negative. We reserve the symbol $\emptyset$ to denote the empty set. We use this symbol for no other purpose. We write and read this as $L \cap R = \emptyset$, "the intersection of sets L and P is the empty set".

Pairs of sets whose intersection is the empty set are said to be **mutually exclusive sets** (or **disjoint sets)**. Thus L and R are mutually exclusive.

The universal set, the sample space ...

Another reserved symbol is the letter S. It is used for the set containing all objects under consideration. Thus if, in a particular problem, the only objects of interest are the colours of a traffic light, then $S = \{\text{red, amber, green}\}$. The set S is known to mathematicians as the **universal set**. In statistical jargon the set S is called the **sample space**.

Unions

The concept **union** contrasts with the concept intersection. The **union** of two sets A and B is the set that contains the elements that belong to A **or** to B. Here we use the word "or" in an inclusive sense — we do not exclude from the union those elements that belong to **both** A **and** B.

If $A = \{1,\ 2,\ 3\}$ and $B = \{2,\ 3,\ 4,\ 5\}$ then the union of A and B is the set $C = \{1,\ 2,\ 3,\ 4,\ 5\}$. We write

$$C = A \cup B$$

and say "C equals A union B".

Example 7A: If $P = \{x \mid 0 \leq x \leq 10\}$ and $Q = \{x \mid 5 < x < 20\}$, find $P \cup Q$.

The union includes all the elements of both set P and set Q:

$$P \cup Q = \{x \mid 0 \leq x < 20\}.$$

Complements ...

Our final concept from set theory is that of the **complement** of a set. Given the sample space S, we define the complement of a set A to be the set of elements of S which are not in A. The complement of A is written $\overline{A}$, and is **always relative to the sample space** S.

If $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$, $A = \{1,\ 3,\ 5\}$ and $B = \{2,\ 4,\ 6\}$ then $\overline{A} = \{2,\ 4,\ 6\}$. We write

$$\overline{A} = B$$

and say "the complement of A equals B" or, more briefly, "A complement equals B".

Example 8A: If $S = \{x \mid 0 \leq x \leq 1\}$ and $D = \{x \mid 0 < x < 1\}$, find $\overline{D}$.

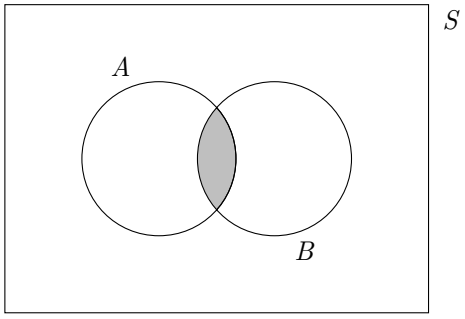
Because the set D excludes the endpoints of the interval from zero to one, $\overline{D} = \{0,\ 1\}$.

Example 9C: If the sample space S contains the letters of the alphabet, i.e. $S = \{a,\ b,\ c, \ldots,\ x,\ y,\ z\}$, the set A contains the vowels, the set B contains the consonants, the set C contains the first 10 letters of the alphabet $C = \{a,\ b,\ c, \ldots, h,\ i,\ j\}$ pick out the true and false statements in the following list:

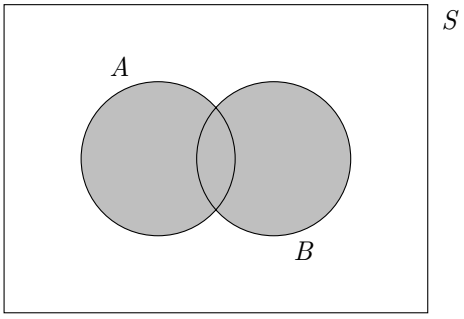
(a) $A \cup B = S$
(b) $A \cap B = \emptyset$
(c) $S \subset S$
(d) $A \cap C = \{a,\ e,\ i\}$
(e) $\overline{A} \subset B$
(f) $\overline{A} = B$
(g) $S \cap B = B$
(h) $A \cup \overline{A} = S$
(i) $\overline{C} \cap A = \{o,\ u\}$
(j) $(\overline{A \cup C}) = \overline{A} \cap \overline{C}$
(k) $A \cap C \subset C$
(l) $\overline{S} = \emptyset$

Venn diagrams ...

A pictorial representation of sets that helps us solve many probems in set theory is known as the **Venn diagram**. In the diagrams below think of all the "points" in the rectangle as being the sample space S, and all the points inside the circles for A and B as the sets A and B respectively. The shaded area in the diagram on the left then represents $A \cap B$, the set of points belonging to A **and** B. Similarly the diagram on the right is a visual representation of $A \cup B$, the set of points belonging to A **or** B. Recall once again the special, inclusive meaning we give to "or". When drawing Venn diagrams it is helpful to associate "intersection" with "and" and "union" with "or".



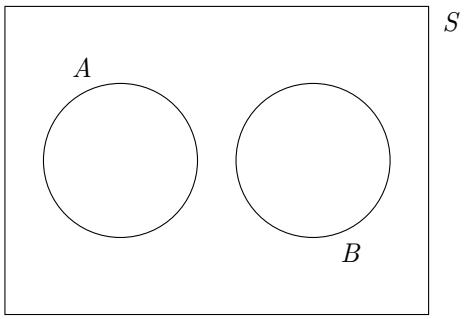
$A \cap B$



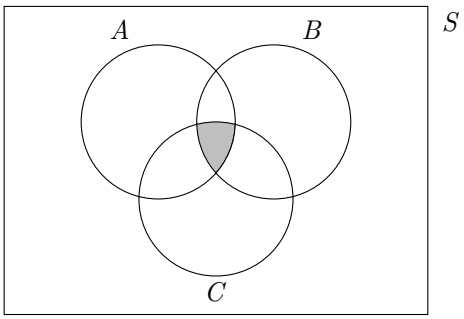
$A \cup B$

The diagram on the left below shows how to depict two mutually exclusive sets in a Venn diagram.

Venn diagrams are usually only useful for up to three sets: the area shaded in the diagram on the right is $A \cap B \cap C$.



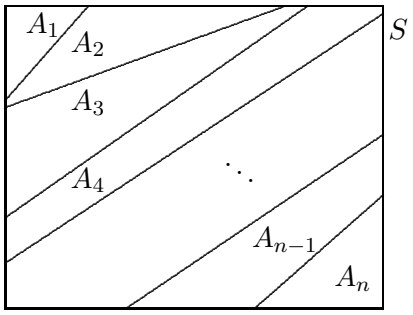
$A \cap B = \emptyset$

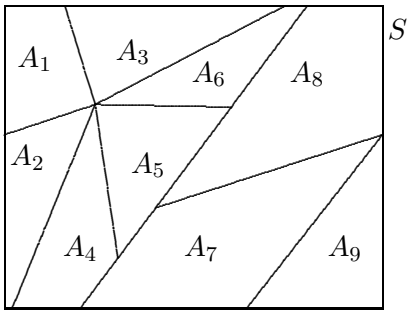


$A \cap B \cap C$

Pairwise mutually exclusive, exhaustive sets ...

If a family of sets A_1, A_2, ..., A_n are such that any pair of them is **mutually exclusive**, i.e. $A_i \cap A_j = \emptyset$ if $i \neq j$, and if $A_1 \cup A_2 \cup \ldots \cup A_n = S$, i.e. the union of the sets "**exhausts**" the sample space, then the family of sets A_1, A_2, ..., A_n are said to be **pairwise mutually exclusive and exhaustive**. If we represent such a family of sets on a Venn diagram, the sets must cover the sample space, and they must be disjoint. Here are two examples:

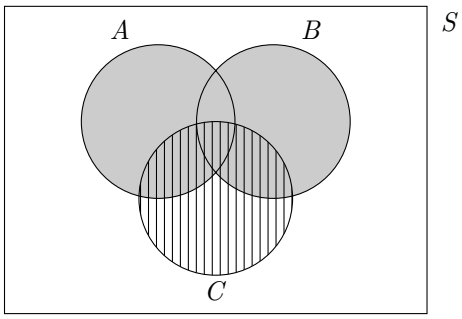




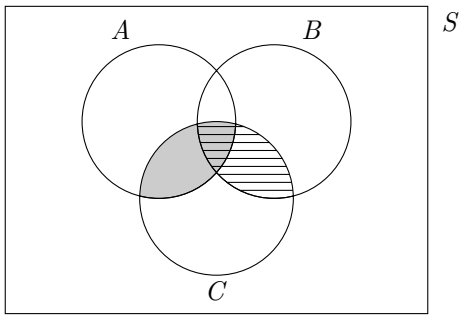
Using Venn diagrams ...

Example 10B: Draw Venn diagrams to show that $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

In the left-hand Venn diagram, the grey-shaded shaded area is $A \cup B$, and the vertically shaded area is C. Their intersection $(A \cup B) \cap C$ is shaded both grey and vertically. In the right-hand Venn diagram, the two shaded areas are $A \cap C$ and $B \cap C$. Their union is the same as the area shaded both grey and vertically in the left-hand diagram.



$(A \cup B) \cap C$



$(A \cap C) \cup (B \cap C)$

Example 11C: Draw Venn diagrams to show that the following are true:

(a) $A \cup B = A \cup (B \cap \overline{A})$
(b) $(\overline{A} \cap C) \cup (\overline{B} \cap A) = (A \cup C) \cap (\overline{A \cap B})$
(c) The sets $A \cap B$, $\overline{A} \cap C$, $\overline{B} \cap A$ and $(\overline{A \cup C})$ form a family of pairwise mutually exclusive and exhaustive sets.

Example 12C: Draw Venn diagrams to determine which of the following statements are true.

(a) $(\overline{A \cap B}) = \overline{A} \cap \overline{B}$
(b) $(A \cap \overline{B}) \cup (\overline{A} \cap B) \subset A \cup B$
(c) $(A \cup B) \cap \overline{C} = (A \cap \overline{C}) \cup (B \cap \overline{C})$
(d) $(\overline{C} \cap A) \cup (\overline{C} \cap B) = (\overline{C \cup (A \cap B)})$
(e) $[(A \cup B) \cap C] \cup [(A \cup C) \cap B] = [(\overline{A} \cup \overline{B} \cup \overline{C}) \cap (A \cup B) \cap C] \cup (A \cap B)$
(f) If the sets A_1, A_2, A_3, and A_4 are pairwise mutually exclusive and exhaustive, and B is an arbitrary set, then

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \cup (A_4 \cap B).$$

Solutions to examples ...

3C (a), (b), (c) and (e) are correct; (d) should read either $F = \{x \mid 4 < x < 5\}$ or $F = \{f \mid 4 < f < 5\}$. For (f), check that the following statement is correct: if H and J are as given, and if $K = \{2x \mid x \in J\}$ then $K = H$. For (g), note that we never use the $\in$-notation with a set on the left hand side.

5C Only (b) and (d) are true.

9C All are true.

11C All are true.

12C (b) (c) (e) and (f) are true. For (a), check that $(\overline{A \cap B}) = \overline{A} \cup \overline{B}$ is true.

Easy exercises ...

2.1 Let S be $\{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$, the set of all possible outcomes when a die is thrown and the number of dots on the uppermost face recorded. Describe in words the following sets:

(a) $\{6\}$
(b) $\{1,\ 2,\ 3,\ 4\}$
(c) $\{1,\ 3,\ 5\}$
(d) $\{2,\ 4,\ 6\}$
(e) $\{5,\ 6\}$
(f) $\overline{\{6\}}$

*2.2 If $S = \{0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9\}$, and $A = \{0,\ 1,\ 2\}$, $B = \{3,\ 4,\ 5,\ 6,\ 7\}$, $C = \{7,\ 8\}$, and $D = \{2,\ 4,\ 6,\ 8\}$, which of the following statements are true?

(a) A and B are mutually exclusive
(b) $\overline{B} = \{0,\ 1,\ 2,\ 8,\ 9\}$
(c) $A \cup B \cup C \cup D = \{0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8\}$
(d) $D \subset (B \cup C)$
(e) $\overline{A} \cap \overline{B} \cap \overline{C} \cap \overline{D} = \{9\}$
(f) $A \cup (B \cap D) = (A \cap B) \cup (A \cap D)$.

2.3 Let S denote the set of all companies listed on the Johannesburg Stock Exchange.

Let $A = \{x \mid x$ is in the gold mining sector$\}$,
let $B = \{x \mid x$ has annual turnover exceeding R10 million$\}$,
let $C = \{x \mid x$ has financial year ending in June$\}$,
let $D = \{x \mid$ the share price of x is higher now than six months ago$\}$.
Describe in words the following sets:

(a) $A \cup B$,
(b) $A \cap D$,
(c) $A \cap C \cap D$,
(d) $B \cap (C \cup A)$,
(e) $\overline{A}$,
(f) $\overline{C} \cup D$,
(g) $\overline{B \cap C}$
(h) $(\overline{B \cap A}) \cup (C \cap D)$.

*2.4 If A, B and C are subsets of a universal set S, draw Venn diagrams to determine which of the following statements are true.

(a) $A \cup \overline{A} = S$
(b) $A \cap \overline{A} = \emptyset$
(c) $\overline{A \cup B} = \overline{A} \cap \overline{B}$
(d) $\overline{A \cap B} = \overline{A} \cup \overline{B}$
(e) $A \cap (B \cup \overline{C}) = (A \cap B) \cup (A \cap \overline{C})$
(f) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
(g) $A \cup B \cup C = S$
(h) $A \cap B \cap C = \emptyset$
(i) $A \cup B \supset A \cap B$
(j) $A \cap (B \cup C) \subset A \cup (B \cap C)$

2.5 If $S = \{1,\ 2,\ 3\}$, list all the subsets of S.

*2.6 Draw a series of Venn diagrams representing three sets, and shade in the following areas.

(a) $\overline{A} \cap B \cap C$
(b) $(B \cap A) \cup (A \cap C)$
(c) $\overline{A \cup B \cup C}$
(d) $(A \cup B \cup C) \cap (\overline{B} \cap C)$.

More difficult exercises ...

*2.7 Let $B_1, B_2, \ldots, B_n$ be n disjoint subsets of S such that

$$\cup_{i=1}^{n} B_i = S \quad \text{and} \quad B_i \cap B_j = \emptyset \quad \text{for} \quad i \neq j.$$

Let A be any other subset of S.

Use a Venn diagram to show that

$$A = \cup_{i=1}^{n} (A \cap B_i).$$

[Notation: $\cup_{i=1}^{n} B_i$ means $B_1 \cup B_2 \cup \ldots \cup B_n$.]

2.8 Show that if the set S has n elements, then S has 2^n subsets. [Hint: Use the binomial theorem.]

2.9 Let A and B be two events defined on a sample space S. Depict the following events in Venn diagrams:

(a) $C = (A \cap \overline{B}) \cup (\overline{A} \cap B)$
(b) $D = (\overline{A \cup B}) \cup (A \cap B)$
(c) What can you say about events C and D?

Solutions to exercises ...

2.1 (a) The number six is obtained.
(b) A number less than or equal to four is obtained.
(c) An odd number is obtained.
(d) An even number is obtained.
(e) A number greater than or equal to 5 is obtained.
(f) A number other than 6 is obtained.

2.2 All are true except (d) and (f).

2.3 (a) Set of companies either in the gold mining sector or with turnovers exceeding R10 million.
(b) Set of gold mining companies whose share price is higher now than six months ago.
(c) Set of gold mining companies with a financial year ending in June whose share price is higher now than six months ago.
(d) Set of all companies which have an annual turnover exceeding R10 million and which are either gold mining companies or companies with financial years ending in June (or both).
(e) Set of companies not in the gold mining sector.
(f) Set of companies which either do not have a financial year ending in June or have a share price which is higher now than six months ago.
(g) Set of companies which either do not have an annual turnover exceeding R10 million or do not have financial year ending in June.

(h) Set of companies which either do not have an annual turnover exceeding R10 million or are not in the gold mining sector or both have a financial year ending in June and have a share price which is higher now than six months ago.

(Notice how difficult it is to express unambiguously in words the meaning of a few mathematical symbols.)

2.4 All are true, except (g) and (h).

2.5 $\emptyset$, $\{1\}$, $\{2\}$, $\{3\}$, $\{1,2\}$, $\{1,3\}$, $\{2,3\}$, $\{1,2,3\}$.

2.9 (c) C and D are mutually exclusive.

Chapter 3

PROBABILITY THEORY

KEYWORDS: Random experiments, sample space, events, elementary events, certain and impossible events, mutually exclusive events, probability, relative frequency, Kolmogorov's axioms, permutations, combinations, conditional probability, Bayes' theorem, independent events.

New wine in old wineskins ...

In the mathematical sciences, in contrast to most other disciplines, we prefer not to coin new words for new concepts. We rather prefer to give new meanings to old words. In this chapter we ask you to put aside your intuitive ideas of what constitutes an "experiment" or an "event" and replace them with the new meanings statisticians have given them.

Random experiments, sample spaces, trials ...

To statisticians, a **random experiment** is a procedure whose **outcome** in a particular performance or trial cannot be predetermined. Although we cannot foretell what the outcome of any single repetition of the experiment will be, we must be able to list the set of all possible outcomes of the experiment. In general, random experiments must be capable, in theory at least, of indefinite repetition. It must also be possible to observe the outcome of each repetition of the experiment. The set of all possible outcomes of a random experiment is called the **sample space** of the random experiment. We usually use the letter S to denote the sample space. Each repetition of the procedure for the random experiment is called a **trial**, and gives rise to **one and only one** of the possible outcomes.

Example 1A: The following are examples of random experiments and their sample spaces.

(a) We toss a coin. We can list the set of possible outcomes: $S = \{\text{heads}, \text{tails}\}$. We can repeat the experiment endlessly, and we can observe the result of every trial.

(b) A phone number is chosen at random. The number is dialled, and the person who answers is asked whether he/she is currently watching television. If the telephone is unanswered after 45 seconds, the outcome, "no reply", is recorded. The set of possible outcomes, the sample space, is $S = \{\text{yes}, \text{no}, \text{won't say}, \text{number engaged}, \text{no reply}\}$.

(c) A light bulb is allowed to burn until it burns out. The lifetime of the bulb is recorded. The possible outcomes are the set of non-negative real numbers (i.e. the set of positive numbers plus zero — the bulb might not burn at all). The sample space is thus $S = \{t \mid t \geq 0\}$.

(d) A die is placed in a shaker, which is agitated violently, and thrown out onto the table. The dots on the upturned face are counted. The sample space is $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$.

(e) In a survey of traffic passing a particular point on Boulevard East, a time period of one minute is chosen at random, and the number of vehicles that pass the point in the minute is counted. The possible outcomes are the integers, including zero: $\{0,\ 1,\ 2,\ 3, \ldots\}$.

(f) A geologist takes rock samples in a mine in order to determine the quality of the iron-ore to be mined. The analytical laboratory reports the proportion of iron in the ore. The sample space is $S = \{p \mid 0 \leq p \leq 1\}$.

Events ...

An **event** is defined to be any subset of the sample space of S. Thus if $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ then the sets $A = \{1,\ 3,\ 5\}$ and $C = \{3,\ 4,\ 5,\ 6\}$ are events. A is clearly the event of getting an odd number. C is the event of getting a number greater than or equal to 3.

The empty set is the set containing no elements. It is often denoted by $\{\}$ or by $\emptyset$.

By the definition of subsets, $\emptyset$ and S are both subsets of S and are thus events. They have special names. The event $\emptyset$ is called the **impossible event**, and S is called the **certain event**.

An **elementary event** is an event with exactly one member, the events $D = \{3\}$ and $E = \{5\}$ are elementary events. A and C are not elementary events, nor are $\emptyset$ and S.

Given two events A and B we say that A and B are **mutually exclusive** if $A \cap B = \emptyset$, that is no elementary events are contained in the intersection of A and B, e.g. if $A = \{1,\ 3,\ 5\}$ and $B = \{2,\ 4,\ 6\}$ then $A \cap B = \emptyset$. This makes sense, because A is the event of getting an odd number, B is the event of getting an even number — and obviously you cannot get an odd number **and** an even number on the same throw of a die.

When does an event "occur"? ...

We give a special meaning to the concept, "the occurrence of an event". We say that the event $A = \{1,\ 3,\ 5\}$ has occurred if the outcome of a trial is **any** member of A. So if we toss a die and get a 3, then we can say that the event A has occurred — we have obtained an odd number. Simultaneously, the events C and D, as defined above, have also occurred.

Example 2B: What is the sample space for each of the following random experiments?

(a) A game of squash is played and the score at the end of the first set is noted.

(b) We record the way in which a batsman in cricket ends his innings.

(c) An investor owns two shares which she monitors for a month. At the end of the month she records whether they went up, down or remained unchanged.

(a) Squash is played to 9 points, with "deuce" at 8-all, in which case the player who reached 8 first decides whether to play to 9 points or to 10 points. Thus the sample space is $S = \{$9-0, 9-1, 9-2, 9-3, 9-4, 9-5, 9-6, 9-7, 9-8, 10-8, 10-9$\}$.

(b) The set of ways in which a batsman's innings can end is given by $S = \{$bowled, caught, leg before wicket, run out, stumped, hit wicket, not out, retired, retired hurt, obstruction, timed out$\}$.

(c) It is convenient to let U = up D = down and N = no change. Then $S = \{$ UU, UD, UN, DU, DD, DN, NU, ND, NN$\}$, where, for example, DU means "first share down, second share up".

Notice how we construct the most detailed possible sample space — the set of outcomes {both up, one up & one down, one up & one unchanged, one down & one unchanged, both unchanged, both down} is not acceptable because each of these could represent several distinguishable outcomes. For example, "one up & one down" could represent either "UD" or "DU".

Example 3C: A random experiment consists of tossing 3 coins of values R1, R2 and R5 and observing heads and tails. Which of the following is the correct sample space?

(a) $S = \{$3 heads,2 heads,1 head,0 heads$\}$.

(b) $S = \{$3 heads,2 heads 1 tail,1 head 2 tails,3 tails$\}$.

(c) $S = \{$HHH,HHT,HTH,THH,HTT,THT,TTH,TTT$\}$ where, for example, HTH means "heads on R1, tails on R2 and heads on R5".

Example 4B: Refer to the random experiments (a) to (c) of Example 2B, and give the subsets of S that correspond to the following events.

(a) (i) The squash game is won by 5 or more points.

(ii) The game goes to deuce.

(b) When the batsman ended his innings the bails were dislodged from the wickets.

(c) None of the shares decline.

In each case we simply list the outcomes that favour the occurrence of the event in which we are interested. The answers are:

(a) (i) {9-0, 9-1, 9-2, 9-3, 9-4} (ii) {9-8, 10-8, 10-9}

(b) {bowled, run out, stumped, hit wicket}

(c) {UU, UN, NU, NN}.

Example 5C: A salesperson, after calling on a client, records the outcome: sale made (S), or no sale made (N). List the sample space of outcomes in one afternoon if

(a) two clients are visited

(b) three clients are visited.

(c) Suppose now that three outcomes are recorded: sale made (S), sales potential good (P), no sale ever likely to be made (N). List the sample space if two clients are visited.

Example 6C. A party of five hikers, three males and two females, walk along a mountain trail in single file.

(a) What is the sample space S?

(b) Find the subset of S that correspond to the events:

- U : a female is in the lead
- V : a male is bringing up the rear

- W : females are in the second and fourth positions.

(c) Find the subsets of S that correspond to $\overline{U}$, $U \cap W$, $V \cap W$, and $U \cap \overline{V}$.

Kolmogorov, father of probability ...

Andrey Nikolaevich Kolmogorov was a Russian mathematician who, in 1933, published the axioms of probability, and established the theoretical foundation for the rigorous mathematical study of probability theory.

KOLMOGOROV'S AXIOMS OF PROBABILITY

Suppose that S is the sample space for a random experiment. For all events $A \subset S$, we define the probability of A, denoted $\Pr(A)$, to be a real number with the following properties:

1. $0 \leq \Pr(A) \leq 1$ for all $A \subset S$
2. $\Pr(S) = 1$
3. If $A \cap B = \emptyset$ (i.e. if A and B are mutually exclusive events) then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

A consequence of the Kolmogorov axioms is that $\Pr(\emptyset) = 0$. The function $\Pr(A)$ provides a means of attaching probabilities to events in S. The first two axioms tell us that probabilities lie between zero and one, and that these extreme probabilities occur for the impossible and certain events, respectively. The probabilities of all other events are graded between these two extremes — unlikely events have probabilities close to zero, and events which are nearly certain have probabilities close to one. If an event is as likely to occur as it is not to occur, then it has probability 0.5. Thus for an unbiased coin, for which "heads" and "tails" are equally likely, the probability of the event "heads" is equal to the probability of the event "tails" is equal to 0.5!

This function $\Pr(A)$ is almost certainly a new kind of function to you. The functions you have seen before, e.g. $y = 3(x^2 + 5)$, take one real number, x, and map them onto another real number, y. If it helps you, you can think of the function $y = f(x)$ as a kind of mincing machine — you put a **number** (x) in, you get another **number** (y) out. Now you must think of the function $\Pr(A)$ as a new kind of mincing machine — you put a **set** (A) in, and out pops a **number** between zero and one (inclusive of these end limits)!

Relative frequencies ...

To try to get some insight into the concept of probability, consider a random experiment on some sample space S repeated **infinitely many times**. Let's start by doing n trials of the random experiment and counting the number of times r that some event $A \subset S$ occurs during the n trials. Then we define r/n to be the **relative frequency** of the event A. Obviously, $0 \leq r/n \leq 1$. Thus relative frequencies and probabilities both lie between zero and one.

We can think of the probability of the event A as the relative frequency of A as n, the number of trials of the random experiment gets very large. In symbols

$$\Pr(A) = \lim_{n \to \infty} \frac{r}{n}.$$

If you toss a fair coin, then the probability of "heads" is equal to the probability of "tails", i.e. $\Pr(H) = \Pr(T) = 0.5$. If you tossed the coin 10 times you might observe 6 heads, a relative frequency of $6/10 = 0.6$. But if you tossed it 100 times you might observe 53 heads, relative frequency $53/100 = 0.53$. If you kept going for a few hours more, and tossed it 1000 times you might observe 512 heads, giving a relative frequency of $512/1000 = 0.512$. **As the number of trials increases, the relative frequency tends to get closer and closer to the "true" probability.**

A class experiment — birthdays in April, May and June ...

Almost exactly a quarter of the days of the year fall into April, May or June ($91/365.25 = 0.249$, allowing for leap years every fourth year). Thus we expect the probability that an individual's birthday falls into one of these three months to be pretty close to 0.249. Let's do an experiment within the class, and fill in this table.

	Number of students (n)	Number with birthdays in April, May, June (r)	Relative frequency (r/n)
Front row			
Front three rows			
Whole class			

Do the relative frequencies get closer to the "true" probability as n gets larger?

Some useful theorems ...

We consider several theorems that follow immediately from the axioms of probability.
Theorem 1. Let $A \subset S$. Then $\Pr(\overline{A}) = 1 - \Pr(A)$.
Proof: We write S as the union of two mutually exclusive events:

$$A \cup \overline{A} = S.$$

Because A and $\overline{A}$ are mutually exclusive, i.e. $A \cap \overline{A} = \emptyset$, we can use axiom 3 to state

$$\Pr(A \cup \overline{A}) = \Pr(A) + \Pr(\overline{A}).$$

But $A \cup \overline{A} = S$, and $\Pr(S) = 1$, by Kolmogorov's 2nd axiom, so $\Pr(A \cup \overline{A}) = 1$. Therefore

$$\Pr(A) + \Pr(\overline{A}) = 1$$

and

$$\Pr(\overline{A}) = 1 - \Pr(A),$$

as required

Theorem 2. If $A \subset S$ and $B \subset S$ then $\Pr(A) = \Pr(A \cap B) + \Pr(A \cap \overline{B})$.
Proof: Write A as the union of the two mutually exclusive sets:

$$A = (A \cap B) \cup (A \cap \overline{B}).$$

Clearly,

$$(A \cap B) \cap (A \cap \overline{B}) = \emptyset.$$

Therefore, using axiom 3,

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap \overline{B})$$

Notice that theorem 2 may also be expressed as

$$\Pr(A \cap \overline{B}) = \Pr(A) - \Pr(A \cap B).$$

Theorem 3. The Addition Rule. For any arbitrary events A and B,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

Proof: Write $A \cup B$ as the union of two mutually exclusive sets:

$$A \cup B = B \cup (A \cap \overline{B})$$

Because B and $A \cap \overline{B}$ are mutually exclusive, we can again apply axiom 3 and say

$$\Pr(A \cup B) = \Pr(B) + \Pr(A \cap \overline{B}).$$

But, by theorem 2, $\Pr(A \cap \overline{B}) = \Pr(A) - \Pr(A \cap B)$. The result follows.

Theorem 4. If $B \subset A$, then $\Pr(B) \leq \Pr(A)$.
Proof: If $B \subset A$ then we can write A as the union of two mutually exclusive sets,

$$A = B \cup (A \cap \overline{B})$$

and

$$\begin{aligned} \Pr(A) &= \Pr(B) + \Pr(A \cap \overline{B}) \\ &\geq \Pr(B) \end{aligned}$$

because $\Pr(A \cap \overline{B}) \geq 0$ as all probabilities are non-negative.

Theorem 5 If $A_1, A_2, \ldots, A_n$ are pairwise mutually exclusive, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\Pr(A_1 \cup A_2 \cup \ldots \cup A_n) = \Pr(A_1) + \Pr(A_2) + \ldots + \Pr(A_n),$$

or, in a more concise notation,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(A_i).$$

Proof: The proof is by repeated use of axiom 3. The events $(A_1 \cup A_2 \cup \ldots \cup A_{n-1})$ and A_n are mutually exclusive. Thus

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \Pr\left(\bigcup_{i=1}^{n-1} A_i\right) + \Pr(A_n)$$

Next, the events $(A_1 \cup A_2 \cup \ldots \cup A_{n-2})$ and A_{n-1} are mutually exclusive. Thus

$$\Pr\left(\bigcup_{i=1}^{n-1} A_i\right) = \Pr\left(\bigcup_{i=1}^{n-2} A_i\right) + \Pr(A_{n-1}),$$

so that

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \Pr\left(\bigcup_{i=1}^{n-2} A_i\right) + \Pr(A_{n-1}) + \Pr(A_n).$$

Continue the process, and the result follows.

Example 7A: If $\Pr(A) = 0.5$, $\Pr(B) = 0.6$ and $\Pr(A \cap B) = 0.3$, find $\Pr(\overline{B})$, $\Pr(A \cap \overline{B})$ and $\Pr(A \cup B)$.

By theorem 1, $\Pr(\overline{B}) = 1 - \Pr(B) = 1 - 0.6 = 0.4$.

By theorem 2, $\Pr(A \cap \overline{B}) = \Pr(A) - \Pr(A \cap B) = 0.5 - 0.3 = 0.2$.

By theorem 3, $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.5 + 0.6 - 0.3 = 0.8$

Example 8C: Is it possible for events in the same sample space to have probabilities $\Pr(A) = 0.8$, $\Pr(\overline{B}) = 0.6$ and $\Pr(A \cap \overline{B}) = 0.7$?

Example 9C: In the sample space S, let $\Pr(A) = 0.7$, $\Pr(B) = 0.5$, $\Pr(C) = 0.1$, $\Pr(A \cup B) = 0.9$, and $\Pr((A \cup B) \cap C) = 0$. Depict the events A, B and C on a Venn diagram and find the probability of the events $A \cap B$, $A \cap C$, $A \cup B \cup C$, $\overline{B} \cap C$, and $\overline{(A \cap B)}$.

Example 10C: If we know that $\Pr(A \cup B) = 0.6$ and $\Pr(A \cap B) = 0.2$, can we find $\Pr(A)$ and $\Pr(B)$?

Equally probable elementary events ...

All probability problems, in theory at least, can be solved by making use of theorem 5. The elementary events whose union make up the sample space S are always mutually exclusive because if one elementary event occurs, no other elementary event occurs. Therefore, if we knew the probabilities of all the elementary events, we would also be able to compute the probabilities of any event in S. By theorem 5, the probability of any event is simply the sum of the probabilities of the elementary events that make up the event.

There is a wide class of problems for which we **do** know the probabilities of **all** the elementary events in a sample space. These are the problems for which it is reasonable to assume that **all the elementary events are equally likely.** If there are N elementary events contained in S, and each one has the same probability of occurring, then the probability of each and every elementary event must be $1/N$.

Equally probable elementary events occur in many games of chance — coins and dice are assumed to be unbiased; when a card is drawn from a pack of 52 playing cards, the probability of any particular card is assumed to be $1/52$. Let A be some event in this scenario. Then A must consist of the union of elementary events, each of probability $1/N$. If we could determine the **number of elementary events contained in** A, then we could write

$$\begin{aligned}\Pr(A) &= \frac{\text{number of elementary events contained in } A}{\text{number of elementary events contained in } S} \\ &= \frac{n(A)}{n(S)} = \frac{n(A)}{N},\end{aligned}$$

where we define the function $n(A)$ to mean the count of the number of elementary events contained in A. Clearly, $n(S) = N$.

Example 11A: Consider tossing a fair die. Then $S = \{1, 2, 3, 4, 5, 6\}$ and $N = 6$. Let $A = \{1, 3, 5\}$ the event of getting an odd number. Find $\Pr(A)$.

The number of elementary events contained in A is $n(A) = 3$. So

$$\Pr(A) = \frac{n(A)}{N} = \frac{3}{6} = \frac{1}{2},$$

which your intuition should tell you is correct!

COMPUTING PROBABILITIES WHEN THE ELEMENTARY EVENTS ARE ALL EQUIPROBABLE

If all the N elementary events contained in a sample space have probability $1/N$, the following three steps enable the probability of any event A in the sample space to be found:

1. Determine the sample space S made up out of elementary events. Determine the number, $N = n(S)$ of elementary events contained in S. You might have to list and count them, or you might be able to use one of the "counting rules" given below. If the N elementary events are equally probable, then each one occurs with probability $1/N$.
2. Determine A, the subset of S, the event whose probability is to be found. Count the number of elementary events contained in A — suppose $n(A)$ elementary events make up the event A.
3. Then $\Pr(A) = n(A)/N$.

Example 12A: 100 people bought tickets in a charity raffle. 60 of them bought the tickets because they supported the charity. 75 bought tickets because they liked the prize. No one who neither supported the charity nor liked the prize bought a ticket.

(a) What is the probability that the prize-winning ticket was bought by someone who liked the prize?
(b) What is the probability that the prize was won by someone who did not support the charity?
(c) What is the probability that the prize was won by someone who both supported the charity and liked the prize?

(a) Let A and B be the events "liked the prize" and "support the charity" respectively. To find $\Pr(A)$, we apply the three steps as follows:

1. $N = 100$
2. $n(A) = 75$
3. Therefore $\Pr(A) = n(A)/N = 75/100 = 0.75$.

(b) To find $\Pr(\overline{B})$, we only have to go through steps 2 and 3.

1. $n(\overline{B}) = 40$

2. Thus $\Pr(\overline{B}) = n(\overline{B})/N = 40/100 = 0.4$

(c) We now need $\Pr(A \cap B)$:

1. $n(A \cap B) = 60 + 75 - 100 = 35$
2. $\Pr(A \cap B) = n(A \cap B)/N = 35/100 = 0.35$.

Example 13C: A pack of playing cards contains 52 cards, 13 belonging to each of the four suites "Spades", "Hearts", "Diamonds" and "Clubs". Within each suite the 13 cards are labelled: Ace, 2, ..., 10, Jack, Queen, King. Let D be the event that a randomly selected card is a diamond, and K be the event that the card is a king, and B be the event that the card has one of the numbers from 2 to 10. Find $\Pr(D)$, $\Pr(K)$, $\Pr(B)$, $\Pr(D \cap K)$, $\Pr(\overline{B} \cup D)$, $\Pr(\overline{B} \cap K)$ and $\Pr(D \cup K \cup B)$.

Example 14C: The seats of a jet airliner are arranged in 55 rows (numbered 1 to 55) of 10 seats (lettered A to K, leaving out I). In each row, seats C, D, G and H are on aisles, and A and K are window seats. Smoking is permitted in rows 45 to 55 inclusive. If a passenger is assigned a seat at random, what is the probability of being allocated

(a) an aisle seat?
(b) a seat in the smoking section?
(c) a window seat in the non-smoking section?
(d) a window seat in row 1?

Permutations and combinations ...

Even in problems in which all the elementary events are equally probable, it is usually impractical to list and to count all the elementary events contained in the sample space or in the event of interest. The theory of combinations and permutations frequently comes to the rescue, and enables the number of elementary events contained in sample spaces and events to be determined quite easily. This theory is summarized in a series of eight "counting rules" given later.

Permutations of n objects ...

Recall that a set is just a group of objects, and that the order in which the objects are listed is irrelevant. We now consider the number of different ways all the objects in a set may be arranged in order. A set containing n distinguishable objects has

$$n(n-1) \times \cdots \times 3 \times 2 \times 1 = n! \quad (\text{"}n \text{ factorial"})$$

different **orderings** of the objects belonging to the set. We can see this by thinking in terms of having n slots to fill with the n objects in the set. Each slot can hold one object. We can choose an object for the first slot in n ways; there are then $n-1$ objects available for the second slot, so we can select an object for the second slot in $n-1$ ways, leaving $n-2$ objects available for the third slot, ... , until the last remaining object has to placed in final slot. We say that there are $n!$ distinct **arrangements** (technically, we call each arrangement or ordering a **permutation**) of the n objects in the set.

Example 15A: If the set $A = \{1, 2, 3\}$, list all the possible permutations.

There are $3! = 3 \times 2 \times 1 = 6$ distinct arrangements of the objects in A. They are:

$$1\,2\,3 \quad 1\,3\,2 \quad 2\,1\,3 \quad 2\,3\,1 \quad 3\,1\,2 \quad 3\,2\,1.$$

Permutations of n objects taken r at a time ...

Suppose now that we have a set containing n objects, and that we have r $(0 < r \leq n)$ slots to fill. In how many ways can we do this, assuming that each object is "used up" once it is allocated to a slot? We number the slots from 1 to r and fill each in turn. We can choose any of the n objects to fill the first slot. Having filled the first slot there are $n - 1$ objects available, any of which may be chosen for the second slot. Therefore, the first two slots can be filled in $n(n-1)$ ways. The first three slots can be filled in $n(n-1)(n-2)$ ways.

By the time we have filled the $(r-1)$st slot and are ready for the rth slot, we have used $r - 1$ members of our set and therefore have $n - (r-1) = n - r + 1$ members left to choose from. Hence, the r slots can be filled in

$$\begin{aligned} &n(n-1) \times \cdots \times (n-r+1) \\ &\quad = \frac{n(n-1) \times \cdots \times (n-r+1) \times (n-r)(n-r-1) \times \cdots \times 3 \times 2 \times 1}{(n-r)(n-r-1) \times \cdots \times 3 \times 2 \times 1} \\ &\quad = \frac{n!}{(n-r)!} \text{ ways} \end{aligned}$$

Thus there are $n!/(n-r)!$ ways of ordering r elements taken from a set containing n elements using each element at most once. Note that we are (a) choosing r objects and (b) arranging them. We are here involved in two processes, choosing **and** arranging. The number of ways of choosing and arranging r objects out of n distinguishable objects is called the **number of permutations of n objects taken r at a time** and is denoted by $(n)_r$ ("n permutation r").

$$(n)_r = \frac{n!}{(n-r)!}$$

This formula is also valid for $r = n$ if we adopt the convention that $0! = 1$.

Example 16A: The focusing mechanism on Ron's camera is bust, so that he can only take pictures of people at a distance of 2 metres, so he only takes pictures of 3 people at a time. How many different pictures (a rearrangement of the same people is considered a different picture) are possible if 10 people are present?

This is the same as asking for the number of permutations of 10 objects taken 3 at a time, given by

$$(10)_3 = \frac{10!}{(10-3)!} = 10 \times 9 \times 8 = 720.$$

Example 17A: Suppose (as happened in South Africa in 1994) that 19 political parties contested an election. One party wanted the ballot papers to have the parties listed in random order. Another said it was impractical. How many different orderings would have been possible?

This is equivalent to asking: "How many permutations of 19 objects taken 19 at a time are there?" The answer is:

$$\begin{aligned} (19)_{19} &= \frac{19!}{(19-19)!} = \frac{19!}{0!} = \frac{19!}{1} \\ &= 121\,645\,100\,000\,000\,000 = 1.216451 \times 10^{17}, \end{aligned}$$

roughly 25 million different ballot papers per man, woman and child on planet earth!

Combinations of n objects taken r at a time ...

Now suppose we want merely to count the number of ways of choosing r elements out of the n elements in our set **without** regard to the arrangement of the chosen elements. In other words, we want to determine the number of r-element subsets that we can form. We call this the **number of combinations of n objects taken r at a time**, and denote it by the symbol $\binom{n}{r}$ ("n combination r").

To find the value of $\binom{n}{r}$, we reason as follows. When we found the number of permutations of n objects taken r at a time, we divided the process into two operations — **choosing** r objects and then **arranging** them. **We are now only interested in the first operation.** We recall that a subset having r objects can be arranged in $r!$ permutations. A little reflection will convince you that there are therefore $r!$ times more permutations than combinations; that is

$$\binom{n}{r} \times r! = (n)_r = n!/(n-r)!$$

Therefore the formula for $\binom{n}{r}$ is given by

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

We have discovered that the number of r-element subsets that can be formed from a set containing n elements is $\binom{n}{r}$.

Example 18A: In how many ways can a 9 man work team be formed from 15 men?

The problem asks **only** for the number of ways of **choosing** 9 men out of 15:

$$\binom{15}{9} = \frac{15!}{9!\,6!} = 5005.$$

Example 19A: How many different bridge hands can be dealt from a pack of 52 playing cards?

A bridge hand contains 13 cards — what matters is only the group of cards (even though you might arrange them in a convenient order). Therefore, bridge hands consist of combinations of 52 objects taken 13 at a time:

$$\binom{52}{13} = 635\,013\,559\,600.$$

At 15 minutes per bridge game, there are enough different bridge hands to keep you going for about 20 million years continuously.

Example 20B: From 8 accountants and 5 computer programmers, in how many ways can one select a committee of

(a) 3 accountants and 2 computer programmers?
(b) 5 people, subject to the condition that the committee contain at least 2 computer programmers and at least two accountants.

(a) We can choose 3 accountants from 8 in $\binom{8}{3}$ ways. We can choose 2 computer programmers from 5 in $\binom{5}{2}$ ways. We multiply the results, because for every group of 3 accountants that we choose, we can choose one of $\binom{5}{2}$ different groups of

computer programmers. Thus we can choose a committee of 3 accountants and 2 computer programmers in

$$\binom{8}{3}\binom{5}{2} = 56 \times 10 = 560 \quad \text{ways.}$$

(b) The total possible number of ways of forming the committee is:
The total number of ways of forming a committee composed of 3 accountants and 2 computer programmers **plus** the number of ways of composing a committee of 2 accountants and 3 computer programmers:

$$\binom{8}{3}\binom{5}{2} + \binom{8}{2}\binom{5}{3} = 560 + 280 = 840 \quad \text{ways.}$$

Permutations, with repetitions ...

We now suppose that we have n types of objects and r slots, and that we have at least r objects of each type available. We can thus fill the first slot with any of the n types of objects, there are still n types of objects available for the second slot, ... Because there are at least r objects of each type, there are still objects of each of the n types available for the final, rth slot.

Thus the **number of permutations of n types of objects taken r at a time, allowing repetitions** is

$$n \times n \times \ldots \times n = n^r.$$

Example 21A: How many four digit numbers can be made from the 10 digits from 0 to 9, if repetitions are permitted?

We have four slots to fill. But because all of the 10 digits remain available to fill every slot, this can be done in $10^4 = 10000$ ways. This makes sense, because there are 10 000 numbers from 0 (actually 0 000) to 9 999.

Example 22B:

(a) How many four letter words can be made with a 26-letter alphabet — including all nonsense words?
(b) It is proposed to adopt a system of motor car number plates which uses three letters of the alphabet (excluding I and O) followed by three digits. How many number plates are possible?

(a) Because words like BEER, with repeated letters, are permissible, the potential number of 4 letter words is $26^4 = 456\ 976$.
(b) Clearly, because number plates like BBB444 with repetitions are permissible, the number of possible number plates is $24^3 \times 10^3 = 13\ 824\ 000$, or nearly 14 million.

Combinations, with repetitions ...

Once again, we have n types of objects, with at least r available of each type. **The number of selections of r objects, allowing repetitions**, is given by

$$\binom{n+r-1}{r}.$$

The proof of this result is included as an exercise at the end of the chapter.

Example 23A: A company needs to purchase four new vehicles. How many selections of makes are possible if they select any of seven different makes (Volkswagen, Toyota, Nissan, Honda, Mazda, Ford, Uno)?

Clearly, a repetition of any of the makes is possible. However, the order of the makes is unimportant. The number of combinations of 7 objects taken 4 at a time, allowing repetitions, is given by

$$\binom{n+r-1}{r} = \binom{7+4-1}{4} = \binom{10}{4} = 210.$$

Example 24B: A supermarket sells 10 types of jam. You buy three tins. How many combinations are possible?

Assuming the supermarket has at least 3 tins of every type of jam, the number of combinations of 10 jams taken 3 at a time, allowing repetitions (i.e. you could buy more than one tin of one type of jam), is

$$\binom{n+r-1}{r} = \binom{10+3-1}{3} = \binom{12}{3} = 220.$$

Counting rules ...

The discussion above can be summarized into several "counting rules":

Counting Rule 1: The number of distinguishable arrangements of n distinct objects, not allowing repetitions is $n!$.

Counting Rule 2: The number of ways of ordering r objects chosen from n distinct objects, not allowing repetitions is

$$(n)_r = \frac{n!}{(n-r)!}\ .$$

Counting Rule 3: The number of ways of choosing a set of r objects from n distinct objects, not allowing repetitions, is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}\ .$$

Counting Rule 4: The number of permutations of r objects, chosen from n distinct objects, **allowing repetitions** is

$$n^r.$$

Counting Rule 5: The number of combinations of r objects, chosen from n distinct objects, **allowing repetitions** is

$$\binom{n+r-1}{r}.$$

Counting rules 1 to 5 relate to scenarios in which there are a total of n distinguishable objects. They can be compressed into a two-way table:

	Without repetition	With repetition
	rules 1 and 2	rule 4
Permutations	$(n)_r = \frac{n!}{(n-r)!}$	n^r
	rule 3	rule 5
Combinations	$\binom{n}{r} = \frac{n!}{(n-r)!\,r!}$	$\binom{n+r-1}{r}$

Counting rule 1 is the special case of counting rule 2, with $r = n$.

We add three further useful counting rules which we will state, and leave the proofs as exercises.

Counting Rule 6: The number of distinguishable arrangements of n items, of which n_1 are of one kind and $n_2 = n - n_1$ are of another kind is

$$\frac{n!}{n_1!\,n_2!} = \binom{n}{n_1} = \binom{n}{n_2}.$$

Here it is assumed that the n_1 items of the first kind are indistinguishable from each other, and the n_2 items of the second kind are indistinguishable from each other.

Before we move onto the final two counting rules, we define a generalization of the binomial coefficient, known as the **multinomial coefficient**. We let

$$\binom{n}{n_1\ n_2\ \ldots\ n_k} = \frac{n!}{n_1!\,n_2!\,\ldots\,n_k!},$$

where $\sum_{i=1}^{k} n_i = n$, and call this the multinomial coefficient. The sum of the numbers in the "bottom row" of a multinomial coefficient must be equal to the number at the "top"! Notice that in multinomial coefficient notation, the binomial coefficient would therefore have to be written with two numbers in the "bottom row":

$$\binom{n}{r} = \binom{n}{r\ (n-r)}.$$

Counting Rule 7: The number of ways of choosing k *combinations* of sizes $n_1, n_2, \ldots, n_k$ from a set of n items ($\sum n_k = n$) is given by the multinomial coefficient

$$\binom{n}{n_1\ n_2\ \ldots\ n_k}.$$

Counting Rule 8: The number of distinguishable arrangements of n items of k types, of which n_1 are of the first type, n_2 are of the second type, $\ldots, n_k$ are of the kth type is given by

$$\binom{n}{n_1\ n_2\ \ldots\ n_k}.$$

Using the counting rules ...

Example 25A: A furniture store is displaying a large couch in the window. The window dresser has six cushions of which four are yellow and two are red. In how many ways can the cushions be arranged on the couch (in a row, of course)?

We require arrangements of 6 items, where 4 are of one kind and 2 are of a second kind. Counting rule 6 tells us that there are

$$\binom{6}{4} = \binom{6}{2} = 15 \text{ possible arrangements.}$$

Example 26A: A chain of nine hardware stores wishes to test-market a new product in four of the nine stores. How many selections of four stores are possible?

The stores are distinguishable, we cannot have a repetition of the same store, and the order of the stores is irrelevant! This is a classic application of counting rule 3. There are

$$\binom{9}{4} = \frac{9!}{4!\,5!} = 126$$

different selections.

Example 27A: A committee of 12 is to be split into three subcommittees, having three, four and five members, respectively. In how many ways can the subcommittees be formed?

By counting rule 7, the number of combinations of sizes 3, 4 and 5 chosen from a set of 12 items is given by

$$\binom{12}{3\ 4\ 5} = \frac{12!}{3!\,4!\,5!} = 27\,720.$$

Example 28B: A clothing store has designed a series of seven different advertisements, labelled A–G. A local newspaper offers a special rate if advertisements are placed on the first, second and third pages of the next weekend edition.

(a) How many different arrangements of the advertisements are possible, assuming that the same advertisement is not repeated.

(b) If the marketing manager decides to allocate the advertisements randomly, and decides not to use the same advertisement more than once, what is the probability that advertisements A and B appear on the first and second pages respectively?

The advertisements are distinguishable, repetitions are not allowed, arrangements are important, so the solution requires application of counting rule 2.

(a) The number of arrangements is $(7)_3 = 7!/4! = 210$.
(b) For the third page, one of C, D, E, F or G must be selected. Hence Pr(A and on first and B on second page) $= 5/210 = 0.024$.

Example 29B: A wealthy investor decides to give four of her 12 investments to her daughter. Five of her investments are in gold-mining companies, the remaining seven in various industrial companies. Her daughter is given the opportunity to select the four companies at random.
(a) How many different sets of companies could the daughter be given?

(b) What is the probability that the daughter gets a poorly diversified portfolio of investments, with either four gold-mining companies, or four industrial companies?

(a) The 12 companies are distinguishable, repetitions are not possible, and arrangements are irrelevant, so counting rule 3 is appropriate. The number of combinations of 12 companies taken four at a time is $\binom{12}{4} = 495$.

(b) The number of ways of selecting four gold-mining companies is $\binom{5}{4} = 5$, and therefore $\Pr(4-\text{gold-mining companies}) = 5/495$. The number of ways of selecting four industrial companies is $\binom{7}{4} = 35$, with probability 35/495. The probability of an undiversified portfolio is the sum of these two probabilities: $(5 + 35)/495 = 0.081$, about "one chance in 12".

Example 30B: At a party, there are substantial stocks of five brands of beer — Castle, Lion, Ohlssons, Black Label and Amstel. One of the party-goers grabs two cans without looking. How many different combinations of 2 beers are possible?

The brands are distinguishable, repetitions are permitted, but the ordering is unimportant, so counting rule 5 is used. The number of ways of selecting two cans from five brands allowing repetitions is

$$\binom{5+2-1}{2} = \binom{6}{2} = 15.$$

Note, that not all of these 15 outcomes are equally probable!

Example 31B: A group of 20 people is to travel in three light aircraft seating 4, 6 and 10 people respectively. What is the probability that three friends travel on the same plane?

The total number of ways of choosing combinations of sizes 4, 6 and 10 from a group of size 20 is, by counting rule 7, given by

$$\binom{20}{4\ 6\ 10} = \frac{20!}{4! \times 6! \times 10!} = 38\ 798\ 760.$$

If the three friends travel in the four-seater aircraft, the remaining 17 must be split into groups of sizes 1, 6 and 10. This can be done in $\binom{17}{1\ 6\ 10}$ ways. Similarly, if they travel in the six-seater, the other 17 must be split into groups of 4, 3 and 10,and if they travel in the 10-seater, the others must be in groups of sizes 4, 6 and 7. Thus the total number of ways the three friends can be together is

$$\binom{17}{1\ 6\ 10} + \binom{17}{4\ 3\ 10} + \binom{17}{4\ 6\ 7} = 4\ 900\ 896 \quad \text{ways.}$$

Thus, $\Pr(3\ \text{friends together}) = (4\,900\,896)/(38\,798\,760) = 0.1263$.

Example 32B: A bridge hand consists of 13 cards dealt from a pack of 52 playing cards. What is the probability of being dealt a hand containing exactly 5 spades?

The cards are distinguishable, repetitions are not possible, and the arrangement of the cards is irrelevant (the order in which you are actually dealt the cards does not make any difference to the hand). So counting rule 3 is the one to use to determine that the total number of possible hands is $\binom{52}{13}$.

Applying counting rule 3 twice more, the number of ways of drawing 5 spades from the 13 in the pack and the remaining 8 cards for the hand from the 39 clubs, hearts and diamonds is $\binom{13}{5}\binom{39}{8}$. Hence

$$\Pr(5 \text{ spades}) = \frac{\binom{13}{5}\binom{39}{8}}{\binom{52}{13}} = 0.1247.$$

Example 33B: If there are r people together, what is the probability that they all have different birthdays (assuming that leap years don't exist)?

To determine the total number of ways they can have birthdays we use counting rule 4, the dates are distinguishable, repetitions are possible, and we think in terms of going through the r people in some order. The total number of ways is 365^r.

The number of ways they can have **different** birthdays is given by counting rule 2, which doesn't allow repetitions: $(365)_r$. So

$$\Pr\left(\text{all different birthdays}\right) = \frac{(365)_r}{365^r}.$$

If $n = 23$ this probability is 0.493, which is just less than one-half. The probability of the complementary event, that there is at least one pair of shared birthdays, is therefore 0.507, marginally over 0.5. This means that, on average, in every second group of 23 people there will be shared birthdays.

Example 34B: Participants in a market research survey are given a set of 16 cards, each having a picture of a well known car model. The participants are asked to sort the cards into three piles:

Pile 1 — the 3 models rated best of all.

Pile 2 — the 5 models rated above average.

Pile 3 — the 8 models rated below average.

In how many ways can the three piles be formed?

The 16 cards are distinguishable, but once they are in their pile their ordering is irrelevant, and repetitions are impossible. So use counting rule 7: the sorting task can be performed in

$$\binom{n}{r_1\ r_2\ r_3} = \binom{16}{3\ 5\ 8} = 720\,720 \text{ ways.}$$

Example 35B: To open a certain bicycle combination lock you have to get all five digits (between 0 and 9) correct. What is the probability that a thief is successful on his first "combination"?

A combination lock is not a combination lock at all — it should be called a permutation lock! You don't only have to hit the right digits, you have to get them **in exactly the right order!** Clearly, the probability is $1/10^5 = 0.00001$.

Example 36C: There are 33 candidates for an election to a committee of three. What is the probability that Jones, Smith and Brown are elected?

Example 37C: A group of eight students fill the front row at Statistics lectures daily. They decide to keep attending lectures until they have exhausted every possible arrangement in the front row. For how many days will they attend lectures?

Example 38C: A young investor is considering the purchase of a portfolio of three shares from the "Building and construction" sector of the stock exchange. He chooses three shares at random from the 25 shares currently listed in this sector.

(a) How many ways can shares be selected for the portfolio?
(b) What is the probability that Everite, Grinaker and L.T.A. (three shares in this sector) are selected?
(c) What is the probability that Grinaker is one of the selected shares?

Example 39C: A firm of speculative builders has bought three adjoining plots. The company builds houses in seven styles. It is concerned about the visual appearance of the houses from the street. So they ask their drafting section to sketch all possible selections of street views.

(a) How many sketches are required if (i) no repetitions of styles are allowed, and if (ii) they allow repetitions of styles?
(b) If they choose one sketch at random from those in part (a)(ii), what is the probability that all the houses will be of different styles?
(c) In order to determine the materials required, the quantity surveying department is concerned only with the three styles which might be built (and not on which plot they are built on). How many combinations of styles must they be prepared for if (i) no repetitions of styles are allowed, and if (ii) repetitions are allowed?

Example 40C: Two new computer codes are being developed to prevent unauthorized access to classified information. The first consists of six digits (each chosen from 0 to 9); the second consists of three digits (from 0 to 9) followed by two letters (A to Z, excluding I and O).

(a) Which code is better at preventing unauthorized access (defined as breaking the code in one attempt)?
(b) If both codes are implemented, the first followed by the second, what is probability of gaining access in a single attempt?

Example 41C: A housewife is asked to rank five brands of washing powder (A, B, C, D, E) in order of preference. Suppose that she actually has no preference, and her ordering is arbitrary. What is the probability that

(a) brand A is ranked first?
(b) brand C is ranked first and brand D is ranked second?

Conditional probability ...

Conditional probabilities provide a method for updating or revising probabilities in the light of new information. On Monday, the weather forecaster might say the probability of rain on Thursday is 50% (weather forecasters have not heard of Kolmogorov's axioms, and insist on giving probabilities as percentages!), on Tuesday he might revise this probability in the light of additional information to 70%, and on Wednesday, with even more reliable information, he might say 60%. In statistical jargon, we would say that each forecast was **conditional** on the information available up to that point in time.

Example 42A: We draw one card from a pack of 52 cards. The probability that it is the King of Clubs is 1/52. Suppose now that someone draws the card for you, and tells you that the card is a club. **Now** what is the probability that it is the King of Clubs? Obviously 1/13. We have reduced our sample space from the set of 52 cards to the set of 13 clubs.

CONDITIONAL PROBABILITY

Let A and B be two events in a sample space S. Then the **conditional probability** of the event B given that the event A has occurred, denoted by $\Pr(B \mid A)$, is

$$\Pr(B \mid A) = \Pr(A \cap B) / \Pr(A)$$

provided that $\Pr(A) \neq 0$. $\Pr(B \mid A)$ is read "the probability of B given A".

The conditional probability $\Pr(B \mid A)$ may be thought of as a reassessment of the probability of B given the information that some other event A has occurred.

Example 42A continued: We reconsider our example using this definition.

$$\begin{aligned}\Pr(\text{King of clubs} \mid \text{clubs}) &= \frac{\Pr(\text{King of clubs } \cap \text{clubs})}{\Pr(\text{clubs})} \\ &= \frac{\Pr(\text{King of clubs})}{\Pr(\text{clubs})}\end{aligned}$$

The event "King of clubs" is a **subset** of the events "clubs" — so the intersection of these two events is the event "King of clubs". $\Pr(\text{clubs})$ is the probability of drawing a club — there are 13 ways of doing this. Hence $\Pr(\text{clubs}) = 13/52$. Therefore

$$\begin{aligned}\Pr(\text{King of clubs} \mid \text{clubs}) &= \frac{1/52}{13/52} \\ &= 1/13,\end{aligned}$$

as before.

Example 43C: You, a woman with a medical background, are one of 198 applicants for an M.B.A. programme of whom 81 will be selected. You hear, along the grapevine, on good authority that there were 70 woman applicants, of whom 38 were selected. Assess your probabilities of being accepted before and after you receive the grapevine information. Use the definition of conditional probability.

Example 44B: Suppose A and B are two events in a sample space, and that $\Pr(A) = 0.6$, $\Pr(B) = 0.2$ and $\Pr(A \mid B) = 0.5$.

Find

(a) $\Pr(B \mid A)$ (c) $\Pr(\overline{A} \cap B)$
(b) $\Pr(A \cup B)$ (d) $\Pr(B \mid \overline{A})$.

In this type of problem a useful first step always is to simplify as many conditional probabilities into absolute probabilities as possible.

From the given information we note that

$$\begin{aligned}\Pr(A \mid B) &= \Pr(A \cap B) / \Pr(B) \\ 0.5 &= \Pr(A \cap B)/0.2.\end{aligned}$$

Thus

$$\Pr(A \cap B) = 0.1.$$

We are now in a position to tackle the questions asked.

(a) $\Pr(B \mid A) = \Pr(B \cap A)/\Pr(A) = 0.1/0.6 = 0.17$

(b) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.6 + 0.2 - 0.1 = 0.7$

(c) $\Pr(B) = \Pr(A \cap B) + \Pr(\overline{A} \cap B)$ by theorem 2. Therefore

$$\Pr(\overline{A} \cap B) = \Pr(B) - \Pr(A \cap B) = 0.2 - 0.1 = 0.1.$$

(d) And finally,

$$\begin{aligned}\Pr(B \mid \overline{A}) &= \Pr(B \cap \overline{A})/\Pr(\overline{A}) \\ &= \Pr(B \cap \overline{A})/(1 - \Pr(A)) \\ &= 0.1/(1 - 0.6) = 0.25.\end{aligned}$$

Example 45B: Show that $\Pr(B \mid A) + \Pr(\overline{B} \mid A) = 1$.

$$\begin{aligned}\Pr(B \mid A) + \Pr(\overline{B} \mid A) &= \Pr(B \cap A)/\Pr(A) + \Pr(\overline{B} \cap A)/\Pr(A) \\ &= \frac{\Pr(B \cap A) + \Pr(\overline{B} \cap A)}{\Pr(A)} \\ &= \Pr(A)/\Pr(A) \qquad \text{by Theorem 2} \\ &= 1\end{aligned}$$

Example 46C: Is it possible for events A and B in a sample space to have the following probabilities?

$$\Pr(A) = 0.5 \quad \Pr(B) = 0.8 \quad \Pr(A \mid B) = 0.2.$$

Example 47C: The probability that a first year student passes Economics if he passes Statistics is 0.5, the probability that he passes Statistics if he passes Economics is 0.8, and the (unconditional) probability that he passes Statistics is 0.7. The Statistics results come out first, and the student finds he has failed. What is now the conditional probability of passing Economics?

Example 48C: Show that for any three events A, B and C in a sample space S

$$\Pr(A \cap B \cap C) = \Pr(A \mid B \cap C)\Pr(B \mid C)\Pr(C).$$

Bayes' Theorem ...

For any two events A and B there are two conditional probabilities that can be considered:

$$\Pr(B \mid A) = \Pr(A \cap B)/\Pr(A)$$
$$\Pr(A \mid B) = \Pr(A \cap B)/\Pr(B).$$

A very useful tool for finding conditional probabilities is Bayes' theorem, which connects $\Pr(B \mid A)$ with $\Pr(A \mid B)$, named in honour of Rev. Thomas Bayes, who did pioneering work in probability theory in the 1700's.

Bayes' Theorem. If A and B are two events, then

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\ \Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid \overline{A})\Pr(\overline{A})}$$

Proof: Recall the definition of conditional probability

$$\Pr(A \mid B) = \Pr(A \cap B)/\Pr(B)$$

and theorem 2

$$\Pr(B) = \Pr(A \cap B) + \Pr(\overline{A} \cap B).$$

Substituting, we have

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(A \cap B) + \Pr(\overline{A} \cap B)}.$$

We note that

$$\Pr(A \cap B) = \Pr(B \mid A)\Pr(A)$$

and

$$\Pr(\overline{A} \cap B) = \Pr(B \mid \overline{A})\Pr(\overline{A}).$$

Therefore

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid \overline{A})\Pr(\overline{A})}.$$

Example 49B: A television manufacturer cannot produce the full quota of tubes it requires, so it purchases 20% of its needs from an outside supplier. The quality manager has determined that 6% of the tubes produced in house are defective, and that 8% of the purchased tubes are defective. He finds the tube of a randomly selected television to be defective. What is the probability that the tube was produced by the company.

- Let D be the event "tube defective",

- C be the event "produced by the company".

We are given $\Pr(D \mid C) = 0.06$, $\Pr(D \mid \overline{C}) = 0.08$ and $\Pr(C) = 0.8$. We want to find $\Pr(C \mid D)$. By Bayes' theorem

$$\Pr(C \mid D) = \frac{\Pr(D \mid C)\Pr(C)}{\Pr(D \mid C)\Pr(C) + \Pr(D \mid \overline{C})\Pr(\overline{C})}$$
$$= \frac{0.06 \times 0.8}{0.06 \times 0.8 + 0.08 \times 0.2} = 0.75.$$

Example 50B: You feel ill at night and stumble into the bathroom, grab one of three bottles in the dark and take a pill. An hour later you feel really ghastly, and you remember that one of the bottles contains poison and the other two aspirin.

Your handy medical text says that 80% of people who take the poison will show the same symptoms as you are showing, and that 5% of people taking aspirin will have them.

$$\begin{aligned} \text{Let } B \text{ be the event } &\text{“having the symptoms”} \\ A \text{ be the event } &\text{“taking the poison”} \\ \text{Then } \overline{A} \text{ is the event } &\text{“taking aspirin”}. \end{aligned}$$

What is the probability that you took the poison given that you have got the symptoms, i.e. what is $\Pr(A \mid B)$?

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid \overline{A})\Pr(\overline{A})}.$$

From the information supplied by the handy medical text

$$\Pr(B \mid A) = 0.8 \quad \text{and} \quad \Pr(B \mid \overline{A}) = 0.05.$$

From our groping round in the dark, we conclude that

$$\Pr(A) = 1/3 \quad \text{and} \quad \Pr(\overline{A}) = 2/3.$$

Thus

$$\Pr(A \mid B) = \frac{0.8 \times 1/3}{0.8 \times 1/3 + 0.05 \times 2/3} = 0.89.$$

We recommend that you call the doctor!

Example 51C: A trick coin with two tails is put in a jar with three normal coins. One coin is selected at random and tossed. If the result is a tail, what is the probability that the trick coin was selected?

Example 52C: Let A and B be two events with positive probability. Which of the following statements are always true, and which are not, in general, true?

(a) $\Pr(A \mid B) + \Pr(A \mid \overline{B}) = 1.$
(b) $\Pr(A \mid B) + \Pr(\overline{A} \mid B) = 1.$
(c) $\Pr(A \mid B) + \Pr(\overline{A} \mid \overline{B}) = 1.$

Example 53C: Let A and B be two mutually exclusive events and let C be any other event. Show that

(a) $\Pr(A \cup B \mid C) = \Pr(A \mid C) + \Pr(B \mid C).$
(b) $\Pr(C \mid A \cup B) = \dfrac{\Pr(A \cap C) + \Pr(B \cap C)}{\Pr(A) + \Pr(B)}.$

Example 54C: The miners are out on strike, with a list of demands. Negotiators reckon that if management meets one of the demands, the probability that the strike will end is 0.85. But if this demand is not met, the probability that the strike will continue is 0.92. You assess the probability that management will agree to meet the demand as 0.3. Later you hear that the strike has ended. What is the probability that demand was met?

Example 55C: A well is drilled as part of an oil exploration programme. The probability of the well passing through shale is 0.4. If the well passes through shale, the probability of striking oil is 0.3. If it does not pass through shale, the probability drops to 0.1.

(a) Given that oil was found, what is the probability that it did not pass through shale?
(b) Given that oil was not found, what is the probability that it passed through shale?

Example 56C: We have been using the kiddie version of Bayes' theorem. Prove the adult version. Let $A_1, A_2, \ldots, A_n$ be a set of mutually exclusive and exhaustive events in S. Let B be any other event. Then

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i)\Pr(A_i)}{\Pr(B \mid A_1)\Pr(A_1) + \Pr(B \mid A_2)\Pr(A_2) + \cdots + \Pr(B \mid A_n)\Pr(A_n)} .$$

Example 57C: Jim has applied for a bursary for next year. His estimates of the probabilities of getting each grade of result, and his probabilities of getting the bursary given each grade are given in the table.

Grade	1st	Upper 2nd	Lower 2nd	3rd	Fail
Pr(getting grade)	0.20	0.15	0.50	0.10	0.05
Pr(getting bursary|grade)	0.90	0.75	0.40	0.15	0.00

You subsequently hear that he was awarded the bursary. What is the probability

(a) that he got a first class pass?
(b) that he failed?
(c) that he got either an upper second or a lower second?

Example 58C: A family has two dogs (Rex and Rover) and a cat called Garfield. None of them is fond of the postman. If they are outside, the probabilities that Rex, Rover and Garfield will attack the postman are 30%, 40% and 15%, respectively. Only one is outside at a time, with probabilities 10%, 20% and 70%, respectively. If the postman is attacked, what is the probability that Garfield was the culprit?

Independent events ...

The intuitive feeling is that independent events have no effect upon each other. But how do we decide whether two events A and B are independent? If the occurrence of event A has nothing to do with the occurrence of event B, then we expect the conditional probability of B given A to be the same as the unconditional probability of B:

$$\Pr(B \mid A) = \Pr(B).$$

The information that event A has occurred does not change the probability of B occurring. If $\Pr(B \mid A) = \Pr(B)$, then, using the definition of conditional probability,

$$\frac{\Pr(B \cap A)}{\Pr(A)} = \Pr(B),$$

or,

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B).$$

This leads us to definition of independent events.

Events A and B are **independent** if

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B).$$

In words, the probability of the intersection of independent events is the product of their individual probabilities.

The definition can be extended to the independence of a series of events: if events $A_1, A_2, \ldots, A_n$ are independent, then

$$\Pr(A_1 \cap A_2 \cap \ldots \cap A_n) = \Pr(A_1) \times \Pr(A_2) \times \cdots \times \Pr(A_n).$$

Many students initially have a conceptual difficulty separating the concepts of events that are **mutually exclusive** and events that are **independent.** It helps (a little) to realize that "mutually exclusive" is a concept from set theory (chapter 2) and can be represented on a Venn diagram. But "independence" is a concept in probability theory (chapter 3), and cannot be represented in a Venn diagram.

Note that independent events are **never** mutually exclusive. For example, the events "the gold price goes up today" and "it rains in Cape Town this afternoon" are conceptually independent — they have nothing to do with each other. However, the gold price might go up today and it might rain in Cape Town this afternoon — the intersection of these events is not empty, and they are therefore not mutually exclusive. On the other hand, if you toss a coin, the events "heads" and "tails" are mutually exclusive. Someone tells you he tossed a coin and got "heads". In the light of this information, your assessment of the probability of getting "tails" is instantly adjusted downwards from 0.5 to zero! Mutually exclusive events are certainly not independent events!

Try making up your own examples to clear up the difference between the two concepts.

Example 59A: Let A be the event that a microchip is manufactured perfectly. Let B be the event that the chip is installed correctly. If $\Pr(A) = 0.98$ and $\Pr(B) = 0.93$ what is the probability that the installed chip functions perfectly?

We require $\Pr(A \cap B)$. Because manufacture and installation may be considered independent, we have:

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B) = 0.98 \times 0.93 = 0.9114.$$

Example 60B: A four-engined plane can land safely even if three engines fail. Each engine fails, independently of the others, with probability 0.08 during a flight. What is the probability of making a safe landing?

Let A_i be the event that engine i fails. Then the event "safe landing" can be written as $(\overline{A_1 \cap A_2 \cap A_3 \cap A_4})$, the complement of the event "all engines fail".

$$\begin{aligned}\Pr(\overline{A_1 \cap A_2 \cap A_3 \cap A_4}) &= 1 - \Pr(A_1 \cap A_2 \cap A_3 \cap A_4) \\ &= 1 - \Pr(A_1) \times \Pr(A_2) \times \Pr(A_3) \times \Pr(A_4) \\ &= 1 - 0.08^4 = 0.999\ 959\ 040.\end{aligned}$$

Quite safe! On average, about one flight in 24 414 will crash!

Example 61C: An orbiting satellite has three panels of solar cells, which function independently of each other. Each fails during the mission with probability 0.02. What is the probability that there will be adequate power output during the entire mission if

(a) all three panels must be active?
(b) at least two panels must be active?

Example 62C: The probability that a certain type of air-to-air missile will hit its target is 0.4. How many missiles must be fired simultaneously if it is desired that the probability of at least one hit will exceed 0.95?

Example 63C: A test pilot will have to use his ejector seat with probability 0.08. The probability that the ejector seat works is 0.97. The probability that his parachute opens is 0.99. Assuming these events to be independent, calculate the probability that the test pilot survives the flight.

Example 64C: Suppose that a fashion shirt comes in three sizes and five colours. The three sizes (and the percentage of the population who purchase each size) are: small (30%), medium (50%), and large (20%). Market research indicates the following colour preferences: white (6%), blue (26%), green (36%), orange (18%), and red (14%). The management of a store expects to sell 1000 of these shirts. How many shirts of each size and colour should they order? Assume independence.

Example 65C: Some financial academics argue that the day-to-day movements of share prices are statistically independent. Assume, hypothetically, that the share *De Beers* has a probability of 0.55 of rising on any given trading day. What is the probability that it rises on three successive trading days?

Example 66C: The probability that the rand will weaken against the dollar tomorrow is 0.53. The probability that you will wake up late tomorrow is 0.42.

(a) What is the probability that, tomorrow, the rand weakens against the dollar **and** you wake up late?
(b) What is the probability that, tomorrow, the rand weakens against the dollar **or** you wake up late?

Example 67C: The probability that you will be able to answer the question in the examination on Chapter 3 of *IntroSTAT* is 0.65. The probability that you enter the numbers into your calculator correctly is 0.94. The probability that your calculator operates correctly is 0.99. The probability that you copy the answer correctly from your calculator to your answer book is 0.97. What is the probability that you get the question correct?

Solutions to examples ...

3C Alternative (c) is correct, it lists all the elementary events.

5C (a) $S = \{SS, SN, NS, NN\}$.
(b) $S = \{SSS, SSN, SNN, NNN, NNS, NSN, NSS, SNS\}$
(c) $S = \{SS, SN, SP, NS, NP, NN, PP, PS, PN\}$

6C (a) $S = \{MMMFF, MMFMF, MFMMF, FMMMF, FMMFM, FMFMM, FFMMM, MFMFM, MFFMM, MMFFM\}$

(b) $U = \{FMMMF, FMMFM, FMFMM, FFMMM\}$
$V = \{FMMFM, FMFMM, FFMMM, MFMFM, MFFMM, MMFFM\}$
$W = \{MFMFM\}$

(c) $\overline{U} = \{MMMFF, MMFMF, MFMMF, MFMFM, MFFMM, MMFFM\}$
$U \cap W = \emptyset$
$V \cap W = \{MFMFM\}$
$U \cap \overline{V} = \{FMMMF\}$

8C $\Pr(A \cup B) = 1.1$, therefore impossible.

9C C is disjoint from $A \cup B$, but $A \cup B$ and C exhaust S.
$\Pr(A \cap B) = 0.3$, $\Pr(A \cap C) = 0$, $\Pr(A \cup B \cup C) = 1.0$, $\Pr(\overline{B} \cap C) = 0.1$ $\Pr(\overline{A \cap B}) = 0.7$.

10C No, all we can say is $\Pr(A) + \Pr(B) = 0.8$.

13C $\Pr(D) = 0.25$, $\Pr(K) = 0.077$, $\Pr(B) = 0.692$, $\Pr(D \cap K) = 0.019$, $\Pr(\overline{B} \cup D) = 0.481$, $\Pr(\overline{B} \cap K) = 0.077$ and $\Pr(D \cup K \cup B) = 0.827$.

14C (a) 0.40 (b) 0.20 (c) 0.16 (d) 0.004

36C $1/\binom{33}{3} = 0.000183$

37C $8! = 40\,320$ days $= 110.5$ years!

38C (a) $\binom{25}{3} = 2\,300$ (b) $\frac{1}{2300} = 0.0004$ (c) $\binom{24}{2}/2300 = 0.12$

39C (a) (i) $(7)_3 = 210$ (ii) $7^3 = 343$ (b) $(7)_3/(7^3) = 0.612$
(c) (i) $\binom{7}{3} = 35$ (ii) $\binom{7+3-1}{3} = 84$

40C (a) $1/(10^6) = 0.000\,001\,000$ and $1/(10^3 \times 24^2) = 1/(576\,000) = 0.000\,001\,736$. Six digits are better than three digits followed by two letters.
(b) $1/(10^6 \times 10^3 \times 24^2) = 0.1736 \times 10^{-11}$

41C (a) $4!/5! = \frac{1}{5}$ (b) $3!/5! = \frac{1}{20}$

43C Before, 0.409; after 0.543.

46C $\Pr(A \cap B) = 0.16$ and thus $\Pr(A \cup B) = 1.14$, which is impossible.

47C $\Pr(\text{passing economics} \mid \text{failed statistics}) = 0.0875/0.3 = 0.2917$

51C By Bayes' theorem, $\Pr(\text{trick coin} \mid \text{tails}) = \frac{1}{4} \Big/ \left(\frac{1}{4} + \frac{1}{2} \times \frac{3}{4}\right) = 0.4$

52C Only (b) is correct.

54C $\Pr(\text{demands met} \mid \text{strike ended}) = \frac{0.85 \times 0.3}{0.85 \times 0.3 + 0.08 \times 0.7} = 0.820$

55C (a) $\frac{0.1 \times 0.6}{0.1 \times 0.6 + 0.3 \times 0.4} = 0.333$
(b) $\frac{0.7 \times 0.4}{0.7 \times 0.4 + 0.9 \times 0.6} = 0.341$

57C (a) 0.3547 (b) 0.0 (c) 0.6158

58C $\frac{0.15\times0.7}{0.3\times0.1+0.4\times0.2+.15\times.7} = 0.488$

61C (a) $0.98^3 = 0.9412$ (b) $0.9412 + 3 \times 0.02 \times 0.98^2 = 0.9988$

62C $0.6^n \leq 0.05$ for $n \geq 6$

63C Pr(survives) $= 0.92 + 0.08 \times 0.97 \times 0.99 = 0.9968$
On average, the pilot will be killed in one test flight in 315!

	white	blue	green	orange	red
Small	18	78	108	54	42
Medium	30	130	180	90	70
Large	12	52	72	36	28

64C

65C $0.55^3 = 0.166$

66C (a) $0.53 \times 0.42 = 0.223$ (b) $0.53 + 0.42 - 0.223 = 0.727$

67C $0.65 \times 0.94 \times 0.99 \times 0.97 = 0.587$.

Exercises ...

*3.1 A breakfast cereal manufacturer packs one of five pictures (a, b, c, d, e) in each box of cereal. If you buy two boxes, what is the sample space for the random experiment whose outcome is the two pictures in the boxes?

*3.2 A small town has three grocery stores (1, 2 and 3). Four ladies living in this town each randomly and independently pick a store in which to shop. Give the sample space of the experiment which consists of the selection of the stores by the ladies. Then define the events:

A: all the ladies choose Store 1

B: half the ladies choose Store 1 and half choose Store 2

C: all the stores are chosen (by at least one lady).

*3.3 Let A, B and C be three arbitrary events. Find expressions for the events

(a) only A occurs
(b) both A and B but not C occur
(c) all three events occur
(d) at least one occurs
(e) at least two occur
(f) exactly one occurs
(g) exactly two occur
(h) no more than two occur
(i) none occur.

*3.4 Let A and B be two events defined on a sample space S. Write down an expression for each of the following events, express their probabilities in terms of $\Pr(A)$, $\Pr(B)$ and $\Pr(A \cap B)$, and evaluate their probabilities if $\Pr(A) = 0.3$, $\Pr(B) = 0.4$ and $\Pr(A \cap B) = 0.2$:

(a) either A or B occurs
(b) both A and B occur
(c) A does not occur
(d) A occurs but B does not occur
(e) neither A nor B occurs
(f) exactly one of A or B occurs.

3.5 Show that $\Pr(A \cup \overline{B}) = 1 - \Pr(B) - \Pr(A \cap B)$.

3.6 Prove that $\Pr(A \cap B) \leq \Pr(A) \leq \Pr(A \cup B)$ for any events A and B.

3.7 If A, B and C are events in a sample S, which of the following assignments of probabilities are impossible?

(a) $\Pr(A) = 0.7 \quad \Pr(B) = 0.9 \quad \Pr(C) = \ 0.3 \quad \Pr(A \cap B) = 0.4.$
(b) $\Pr(A) = 0.2 \quad \Pr(B) = 0.5 \quad \Pr(C) = \ 0.3 \quad \Pr(A \cap B) = 0.25.$
(c) $\Pr(A) = 0.3 \quad \Pr(B) = 0.8 \quad \Pr(C) = -0.1 \quad \Pr(A \cap B) = 0.2.$
(d) $\Pr(A) = 0.3 \quad \Pr(B) = 0.7 \quad \Pr(C) = \ 0.8 \quad \Pr(A \cap C) = 0.1.$
(e) $\Pr(A) = 0.8 \quad \Pr(B) = 0.4 \quad \Pr(C) = \ 0.5 \quad \Pr(A \cup C) = 0.7.$

*3.8 What is the probability that a six-digit telephone number has no repeated digits? Do not allow the number to start with a zero.

*3.9 A motor car manufacturer produces four different models, each with three levels of luxury, and with five colour options. One example of each combination of model, luxury level and colour is on display in a parking lot.

(a) How many cars are on display?
(b) An interested client parks his vehicle in the parking lot. A rock from an explosion at a nearby construction site lands on one of the cars. What is the probability that it lands on the client's car?
(c) What assumptions did you make in order to do part (b)?

*3.10 A pack of cards like the one described in Example 13C is being used by four players for a game of bridge, so each is dealt a hand of 13 cards. The king, queen and jack are referred to as "picture cards". Find the probability that a bridge hand (13 cards) contains

(a) 3 spades, 4 diamonds, 1 heart and 5 clubs
(b) 3 aces and 4 picture cards.

*3.11 If $\Pr(A) = 0.6, \Pr(B) = 0.15$, and $\Pr(B \mid \overline{A}) = 0.25$ find the following probabilities

(a) $\Pr(B \mid A)$ (c) $\Pr(A \cup B)$
(b) $\Pr(A \mid B)$ (d) $\Pr((A \cap \overline{B}) \cup (\overline{A} \cap B))$.

3.12 If A and B are mutually exclusive, what is $\Pr(A \mid B)$?

3.13 If $A \cap B = \emptyset$, show that $\Pr(A \mid A \cup B) = \Pr(A)/(\Pr(A) + \Pr(B))$.

3.14 The probability that a student passes Statistics is 0.8 if he studies for the exam and 0.3 if he does not study. If 60% of the class studied for the exam, and a student chosen at random from the class passes, what is the probability that he studied?

*3.15 The probability that a cancer test will detect the disease in a person who **has** cancer is 0.98. The probability that a person who does not have cancer will give a positive reading on the test is 0.1 (i.e. the test says he has the disease even though he has not). If 0.1 per cent of the population has cancer, what is the probability that a person selected at random will in fact have cancer, given that he shows a positive reading on the cancer test? Comment on your answer.

*3.16 The probability that twins are identical is 0.7. Identical twins are always of the same sex, while non-identical twins are of the same sex with probability 0.5. What is the probability that twin boys are identical twins?

3.17 The sample space for the response of a voter's attitude towards a political issue has three elementary events: $A_1 = \{\text{in favour}\}$, $A_2 = \{\text{opposed}\}$, $A_3 = \{\text{undecided}\}$. Let B be the event that a voter is under 25 years of age. Given the following table of probabilities, compute the probability that a voter is opposed to the issue, given that he is under 25.

Event	A_1	A_2	A_3	$B \mid A_1$	$B \mid A_2$	$B \mid A_3$
Probability of Event	0.4	0.5	0.1	0.8	0.2	0.5

3.18 A and B are events such that $\Pr(A) = 0.6$, $\Pr(B \mid A) = 0.3$, and $\Pr(A \cup B) = 0.72$. Are A and B independent, mutually exclusive, or both?

*3.19 If the probability is 0.001 that a 20-watt bulb will fail a 10-hour test, what is the probability that a sign constructed of 1000 bulbs will burn for 10 hours

(a) with no bulb failure?
(b) with one bulb failure?
(c) with k bulb failures?

3.20 Show that if events A and B are independent, then the following pairs of events are also independent.

(a) A and $\overline{B}$
(b) $\overline{A}$ and $\overline{B}$.

3.21 The events A, B and C are such that A and B are independent and B and C are mutually exclusive. Their probabilities are $\Pr(A) = 0.3$, $\Pr(B) = 0.4$, and $\Pr(C) = 0.2$. Calculate the probabilities of the following events.

(a) Both B and C occur.
(b) At least one of A and B occurs.
(c) B does not occur.
(d) All three events occur.
(e) $(A \cap B) \cup C$.

*3.22 A satellite is to have a number of solar panels, which will function independently of each other, and each will fail during the mission with probability 0.05. For success, at least one must function until the end of the mission. How many solar panels are necessary, if the probability of success must exceed 0.999?

Further exercises ...

*3.23 (a) In how many ways can the batting order for a cricket team (11 players) be arranged?
(b) In how many ways can the team be arranged, given that three specific players have definite positions in the batting order?
(c) In how many ways can the team be divided up into 2 teams of 5 players each, and one player left out?
(d) What is the probability that the player left out is the captain of the team?

3.24 If seven diplomats were asked to line up for a group picture with the senior diplomat in the centre, how many distinguishable arrangements are possible?

3.25 The telephone numbers for the University of Cape Town's Rondebosch campus all start with 650 followed by a four digit number.
(a) How many different telephone numbers can be accommodated on this campus?
(b) What is the probability that a randomly selected number has its last three digits (i) 000 (three zeros) (ii) all the same?

3.26 Suppose A and B are events in a sample space.
(a) If $A \cap B = B$, what is the numerical value of $\Pr(A \mid B)$?
(b) If $A \cap B = \emptyset$, what is $\Pr(A \mid B)$?
(c) If $A \cup B = A$, what is $\Pr(A \mid B)$?

3.27 Let A and B be two events defined on a sample space S.
(a) Write down an expression for each of the following events in terms of unions, intersections and complements, and express their probabilities in terms of $\Pr(A)$, $\Pr(B)$ and $\Pr(A \cap B)$.
(i) Both A and B occur.
(ii) At least one of A and B occur.
(iii) Either A occurs, or B occurs, but not both.
(iv) A occurs but B does not occur.
(v) A occurs, or B does not occur.
(b) Now suppose that $\Pr(A) = \frac{1}{4}$, $\Pr(B) = \frac{1}{3}$ and that A and B are independent events. Evaluate the probabilities in part (a).

*3.28 (a) What is the probability of drawing exactly 1 spade in a bridge hand (as defined in Exercise 3.10)?
(b) What is the probability of drawing at least 1 spade?
(c) What is the probability that a bridge hand contains 3 spades, 7 diamonds, 2 hearts and 1 club?
(d) What is the probability that a bridge hand contains both the ace and the king of spades?

3.29 If A and B are events in S, and if $\Pr(A) = \frac{1}{3}$, $\Pr(B) = \frac{3}{4}$, and $\Pr(A \cup B) = \frac{11}{12}$ find
(a) $\Pr(A \cap B)$
(b) $\Pr(A \mid B)$
(c) $\Pr(B \mid A)$.

*3.30 Let A and B be two events in a sample space. Suppose $\Pr(A) = 0.4$ and $\Pr(A \cup B) = 0.7$. Let $\Pr(B) = p$.

(a) For what value of p are A and B mutually exclusive?
(b) For what value of p are A and B independent?

3.31 There are n people in a room.

(a) What is the probability that at least two are born in the same sign of the Zodiac? (Assume 12 signs, each of equal duration.)
(b) What is this probability if $n = 13$?

*3.32 In how many ways can a student answer an eight-question, true-false examination if

(a) he marks half the questions true and half the questions false?
(b) he marks no two consecutive answers the same?

3.33 A card is drawn from an ordinary pack of 52, looked at and replaced, and the pack shuffled. How many times should this be done in order to have a 90% chance of seeing the ace of spades at least once?

*3.34 A class of 75 students is to be divided into three tutorial groups of sizes 24, 31 and 20 respectively.

(a) In how many ways can this be done?
(b) There are two brothers in the class. What is the probability that they are in the same tutorial class?

*3.35 (a) (i) How many arrangements of the letters s t a t i s t i c s are possible if repeated letters such as s are indistinguishable from one another?

(ii) If the letters are randomly arranged, what is the probability that the first and last letters are both i ?

(b) (i) How many different words (including nonsense words) can be made from the letters a s t r o n o m e r s ?

(ii) If the letters are randomly arranged, what is the probability that both the letters m o o n and the letters s t a r appear in sequence in the word?

*3.36 For safety reasons, each of 1000 parts in a spacecraft is duplicated. The spacecraft will fail in its mission if any component and its safety duplicate both fail. Each component fails (independently of any other component) with probability 0.01. What is the probability that the mission fails?

3.37 The probability that a B.Sc. student takes neither Statistics nor Chemistry is 0.3 and the probability that he takes Statistics but not Chemistry is 0.2. If B.Sc. students take Statistics and Chemistry independently of each other, what is

(a) the probability that a B.Sc. student takes Statistics?
(b) the probability that a B.Sc. student takes Chemistry but not Statistics?

3.38 Is it possible for events A and B to be defined on a sample space with the following probabilities?

(a) $\Pr(A) = 0.5$, $\Pr(B) = 0.8$ and $\Pr(A \mid B) = 0.2$

(b) $\Pr(A) = 0.5$, $\Pr(A \mid B) = 0.7$, $\Pr(A \cap B) = 0.3$ and $\Pr(A \cup B) = 0.6$.

3.39 (a) If A, B and D are three events in a sample space where $A \cap B = \emptyset$ and $A \cup B = S$, show that

(i) $\Pr(D) = \Pr(D \mid A)\Pr(A) + \Pr(D \mid B)\Pr(B)$

(ii) $\Pr(A \mid D) = \dfrac{\Pr(A)\Pr(D \mid A)}{\Pr(D)}$.

(b) Two machines are producing the same item. Last week, Machine A produced 40% of the total output, and Machine B the remainder. On average, 10% of the items produced by Machine A were defective, and 4% of the items produced by Machine B were defective.

(i) What proportion of last week's entire production was defective?

(ii) If an item selected at random from the combined output produced last week is found to be defective, what is the probability it came from Machine A?

3.40 The probability of passing Statistics without doing these exercises is 0.1 and 0.8 if they are done. If 60% of students do these exercises, what is the probability that a student has not done the exercises if he passes Statistics?

*3.41 Which of the following pairs of events would you expect to be independent, which mutually exclusive and which neither?

(a) studying Economics and being left-handed,
(b) owning a dog and paying vet's bills,
(c) the prices of shares in Anglovaal and Gold Fields (both in the mining house sector of the Johannesburg Stock Exchange) both rising today,
(d) being a member of the Canoe Club and studying for a B.A.,
(e) buying sugar-free cooldrink and buying a cream doughnut for yourself.

*3.42 An X-ray test is used to detect a disease that occurs, initially without any obvious symptoms, in 3% of the population. The test has the following error rates: 7% of people who are disease free have a positive reaction and 2% of the people who have the disease have a negative reaction. A large number of people are screened at random using the test, and those with a positive reaction are examined further.

(a) What proportion of people who have the disease are correctly diagnosed?
(b) What proportion of people with a positive reaction actually have the disease?
(c) What proportion of people with a negative reaction actually have the disease?
(d) What proportion of the tests conducted give the incorrect diagnosis?

Proofs of the counting rules ...

3.43 Prove Counting rule 5; i.e. prove that the number of combinations (selections) of r objects, selected from n objects, allowing repetitions, is $\binom{n+r-1}{r}$. [Hint: consider r objects and $n-1$ marker objects, and apply Counting rule 6. Let the markers divide the objects up into the n types of objects.]

3.44 Prove Counting rule 6; i.e. prove that the number of distinguishable arrangements of n objects, of which n_1 are of type 1 and n_2 of type 2, is given by $\binom{n}{n_1}$.

3.45 Prove Counting rule 7; i.e. prove that the number of combinations of sizes $n_1, n_2, \ldots, n_k$ choosen from a set of n items is given by $\binom{n}{n_1\ n_2 \ldots\ n_k}$.

3.46 Prove Counting rule 8; i.e. prove that the number of distinguishable arrangements of n objects, of which n_1 are of type 1, n_2 of type $2, \ldots, n_k$ of type k is given by $\binom{n}{n_1\ n_2\ \ldots\ n_k}$.

Solutions to exercises ...

3.1 $S = \{$aa, ab, ... , de, ee$\}$, 25 elementary events. 15 if one does not distinguish between, for example, *ab* and *ba*

3.2 $S = \{1111, 1112, 1121, \ \ldots\ , 3333\}$, 81 elementary events.
$A = \{1111\}$. $B = \{1122, 1212, 1221, 2211, 2121, 2112\}$
$C = \{1231, 1232, 1233, 1321, \ \ldots\ , 3321\}$, 36 elementary events.

3.3 (a) $A \cap (\overline{B \cup C}) = A \cap \overline{B} \cap \overline{C}$ (b) $(A \cap B) \cap \overline{C}$.
(c) $A \cap B \cap C$. (d) $A \cup B \cup C$.
(e) $(A \cap B) \cup (A \cap C) \cup (B \cap C)$.
(f) $(A \cap (\overline{B \cup C}) \cup (B \cap (\overline{A \cup C})) \cup (C \cap (\overline{A \cup B}))$.
(g) $(A \cap B \cap \overline{C}) \cup (A \cap C \cap \overline{B}) \cup (B \cap C \cap \overline{A})$.
(h) $(\overline{A \cap B \cap C})$ (i) $(\overline{A \cup B \cup C})$.

3.4 (a) 0.5 (b) 0.2 (c) 0.7 (d) 0.1 (e) 0.5 (f) 0.3.

3.7 (a) impossible, $\Pr(A \cup B) > 1$ (b) impossible, $\Pr(A) < \Pr(A \cap B)$.
(c) impossible, $\Pr(C) < 0$ (d) possible (e) impossible, $\Pr(A) > \Pr(A \cup C)$.

3.8 $9 \times (9)_5/(9 \times 10^5) = 0.1512$.

3.9 (a) 60 (b) 1/61 (c) An equal probability of hitting any of the 61 cars!

3.10 (a) $\binom{13}{3}\binom{13}{4}\binom{13}{1}\binom{13}{5}\Big/\binom{52}{13}$ (b) $\binom{4}{3}\binom{12}{4}\binom{36}{6}\Big/\binom{52}{13}$.

3.11 (a) 0.083 (b) 0.33 (c) 0.7 (d) 0.65.

3.12 0.

3.14 0.8.

3.15 0.00971.

3.16 0.8235.

3.17 0.2128.

3.18 Events A and B are independent.

3.19 (a) 0.3677 (b) 0.3681 (c) $\binom{1000}{k}\ 0.001^k \times 0.999^{1000-k}$.

3.21 (a) 0 (b) 0.58 (c) 0.6 (d) 0 (e) 0.32.

3.22 3

3.23 (a) 11! (b) 8! (c) $\binom{11}{5\ 5\ 1}$ (d) $\binom{10}{5\ 5}\Big/\binom{11}{5\ 5\ 1} = 0.0909$.

3.24 6!

3.25 (a) $10^4 = 10\,000$ (b) (i) $10/10\,000 = 0.001$ (ii) $(10 \times 10)/10\,000 = 0.01$

3.26 (a) 1.0 (b) 0.0 (c) 1.0.

3.27 (a) (i) $A \cap B$, $\Pr(A \cap B)$ (ii) $A \cup B$, $\Pr(A) + \Pr(B) - \Pr(A \cap B)$
(iii) $(A \cap \overline{B}) \cup (\overline{A} \cap B)$, $\Pr(A) + \Pr(B) - 2\Pr(A \cap B)$
(iv) $A \cap \overline{B}$, $\Pr(A) - \Pr(A \cap B)$
(v) $A \cup \overline{B}$, $1 + \Pr(A \cap B) - \Pr(B)$.
(b) (i) 1/12 (ii) 1/2 (iii) 5/12 (iv) 1/6 (v) 3/4.

3.28 (a) $\binom{13}{1}\binom{39}{12}/\binom{52}{13}$, (b) $1 - \binom{39}{13}/\binom{52}{13}$
(c) $\binom{13}{3}\binom{13}{7}\binom{13}{2}\binom{13}{1}/\binom{52}{13}$ (d) $\binom{2}{2}\binom{50}{11}/\binom{52}{13}$.

3.29 (a) 1/6 (b) 2/9 (c) 1/2.

3.30 (a) $p = 0.3$ (b) $p = 0.5$.

3.31 (a) $1 - (12)_n/12^n$ (b) 1.0 (certain event).

3.32 (a) $\binom{8}{4}$ (b) 2.

3.33 $(51/52)^n < 0.1$ for $n > 119$.

3.34 (a) $\binom{75}{24\ 31\ 20}$
(b) $\left(\binom{73}{22\ 31\ 20} + \binom{73}{24\ 29\ 20} + \binom{73}{24\ 31\ 18}\right)/\binom{75}{24\ 31\ 20}$

3.35 (a) (i) $\binom{10}{3\ 3\ 2} = 50400$ (ii) $\binom{8}{3\ 3}/50400 = 0.022$
(b) (i) $\binom{11}{2\ 2\ 2} = 4989600$ (ii) $\binom{5}{2\ 2}/4989600$

3.36 0.095

3.37 (a) 0.4 (b) 0.3

3.38 (a) No, $\Pr[A \cup B] = 1.14 > 1$ (!)
(b) No. Value of $\Pr[A \cup B]$ from first three statements is inconsistent with fourth.

3.39 (b) (i) 6.4% (ii) 0.625

3.40 0.0769

3.41 (a) and (d) are independent, (b) and (c) are neither, and (e) is mutually exclusive if you argue that a diet-conscious person won't buy a cream doughnut!

3.42 Let D be the event having disease, let N be the event testing negative.
(a) $\Pr(\overline{N} \mid D) = 1 - \Pr(N \mid D) = 1 - 0.02 = 0.98$ (98% of those with the disease have a positive reaction)
(b) $\Pr(D \mid \overline{N}) = 0.3022$ (30.22%)
(c) $\Pr(D \mid N) = 0.0007$ (0.07%)
(d) Misdiagnosed is the event $(N \cap D) \cup (\overline{N} \cap \overline{D})$, the union of two mutually exclusive events.

$$\begin{aligned}\Pr((N \cap D) \cup (\overline{N} \cap \overline{D})) &= \Pr(N \mid D)\Pr(D) + \Pr(\overline{N} \mid \overline{D})\Pr(\overline{D}) \\ &= 0.02 \times 0.03 + 0.07 \times 0.97 = 0.0685 \qquad (6.85\%)\end{aligned}$$

Chapter 4

RANDOM VARIABLES

KEYWORDS: Random variables, discrete and continuous random variables, probability mass functions and probability density functions.

Words or numbers...

In Chapter 3, we defined a sample space as the set consisting of all the elementary events that are possible outcomes of a random experiment. Sometimes, we expressed these elementary events quantitatively (the length of time for which a light bulb lasts, the number of items purchased by a customer, the proportion of voters who support a particular proposal), and sometimes we used verbal, qualitative descriptions of the elementary events (for random experiments such as the state of the economy, the sex of an applicant for a job, the colour of a vehicle ordered by a purchaser).

In order to manipulate the events defined on a sample space mathematically, it is necessary to attach a numerical value to each elementary event. Frequently, the elementary events are quantitative, and there is a natural and obvious way to assign numbers to them — the "survival time" (in hours) of the light bulb, the count of items purchased, the number of girls in families of four children.

However, if the elementary events are expressed qualitatively, we have to assign a number to each elementary event. For example, the economy might be classified as being "in recession", "stable" or "booming"; we could asign a "1" to the event "recession", "2" to the event "stable" and "3" to the event "booming". An applicant for a advertised post could be male or female, and we could assign "0" to the event "male applicant" and "1" to the event "female applicant". To repeat the motivation for assigning numbers to elementary events — it clears the way for us to develop a general mathematical theory for handling the probabilities of events in a sample space.

Once all the elementary events in a sample space have numerical values assigned to them, we follow the classic algebraic tradition and let X "stand for" the numerical values of the elementary events. We then call X a **random variable**. X is a **variable** because it can "take on" (or assume) different values. X is a **random** variable because the particular value it takes on depends on the outcome of a random experiment.

By convention, statisticians use the capital letters near the end of the alphabet to denote random variables. Their favourite choice is the letter X.

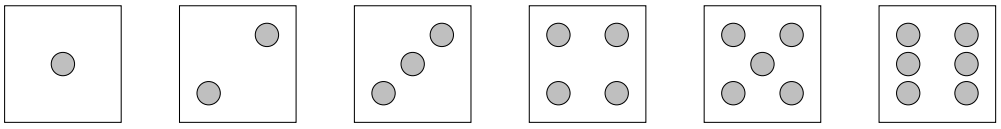
Definition of a random variable...

A **random variable** X is a numerical variable whose value is determined by the outcome of a random experiment. Expressed somewhat differently, a random variable is a function whose domain is the sample space, and whose range is the real line.

Once we are dealing with a random variable, events (which we defined in Chapter 3 as subsets of the sample space) become subsets of the real line, usually a set of points or an interval. Thus "$X = 1$", "$X < 4$" and "$5 \leq X \leq 10$" are events. And because they are events, they have probabilities — we write $\Pr[X = 1]$, $\Pr[X < 4]$ and $\Pr[5 \leq X \leq 10]$, and read them as "the probability that the random variable X takes on the value 1", "the probability that X is less than 4", and "the probability that X lies in the interval from 5 to 10, inclusive of the end points."

Statisticians have also adopted the convention that small letters (e.g. x, a, b) are used to denote particular values that a random variable may assume. Thus $\Pr[X = x]$ means the probability that the random variable X takes on some particular value x. If there is only one random variable under consideration, and hence no ambiguity, we abbreviate $\Pr[X = x]$ to $\Pr(x)$.

Example 1A: Consider tossing a die. We attach numerical values to the elementary events in the sample space in the obvious way:



1 2 3 4 5 6

If the die is unbiased, then $\Pr[X = 1] = 1/6$, i.e. $\Pr(1) = 1/6$. We can write $\Pr[X = x] = 1/6$, or $\Pr(x) = 1/6$, for $x = 1,\ 2,\ 3,\ 4,\ 5$ and 6.

It is important to realize that the definition of a random variable does not imply that a **different** numerical value needs to be assigned to each elementary event. In fact, we often want random variables in which the same value is assigned to different elementary events. The following four examples illustrate this.

Example 2A: Consider the random experiment that consists of tossing an unbiased coin 3 times (see also Example 3C of Chapter 3). If the random variable of interest is the number of heads that occur, then we attach numerical values to the elementary events as follows:

$S =$	{	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT	}
$X =$		3	2	2	2	1	1	1	0	

The event "$X = 1$" corresponds to the subset {HTT, THT, TTH} of S, and thus $\Pr[X = 1] = 3/8$. Also, clearly, $\Pr(0) = 1/8$, $\Pr(2) = 3/8$ and $\Pr(3) = 1/8$.

Example 3B: Suppose you have 5 coins in your pocket — two 5c coins, two 10c coins and a 50c coin — and you pull out two coins at random for a tip. Let the random variable X be the amount of the tip. What are the possible values for X, and the probabilities that X takes on these values?

We denote the coins 5_1 5_2 10_1 10_2 and 50. The sample space S, and the numerical values assigned to each elementary event, can be represented as:

$S =$	{	5_15_2	5_110_1	5_110_2	5_210_1	5_210_2	10_110_2	5_150	5_250	10_150	10_250	}
$X =$		10	15	15	15	15	20	55	55	60	60	

If we assume that each of the ten pairs of coins is equally likely, then $\Pr[X = 10] = 0.1$, $\Pr[X = 15] = 0.4$, $\Pr[X = 20] = 0.1$, $\Pr[X = 55] = 0.2$, and $\Pr[X = 60] = 0.2$.

Example 4B: A shocking snooker player hits the balls around at random until he gets one into a pocket. There are 15 red balls (valued at 1 point) and one each of the colours yellow, green, brown, blue, pink and black (valued from 2 to 7 respectively). What is the sample space for this "random experiment"? Let the random variable X be the "score". What values can X take on and with what probabilities?

Denoting the red balls red 1, red 2,..., red 15, the elementary events in the sample space, and the X values assigned to them, are

$S =$	{	red 1,	...,	red 15,	yellow,	green,	brown,	blue,	pink,	black }
$X =$		1	...	1	2	3	4	5	6	7

and, assuming that each ball is equally likely to be pocketed, $\Pr[X = 1] = 15/21$ and $\Pr[X = 2] = \Pr[X = 3] = \ldots = \Pr[X = 7] = 1/21$.

Example 5C: A car salesperson is scheduled to see two clients today. She sells only two models of cars, an "executive" (E) and a "basic" (B) model. Each executive model sold earns the salesperson a commission of R2000, while each basic model sold earns her only R1000. If the sale is lost (L), no commission is earned. Suppose $\Pr(E) = 0.2$, $\Pr(B) = 0.3$, and $\Pr(L) = 0.5$, and that sales are independent of each other. Let the random variable X be the total commission earned by the salesperson today. What values can X take on, and with what probabilities?

Discrete and continuous random variables...

Random variables fall into two categories — **discrete** and **continuous**. The mathematical treatment of these two types of random variables is very different — as you will learn from the remainder of this chapter.

Discrete random variables take on isolated values along the real line, usually (but by no means always) integer values. Examples of integer-valued discrete random variables are:

- the number of customers entering a store between 09h00 and 10h00
- the number of occupied tables at a restaurant
- the number of clients visited by a salesperson during a day
- the number of applicants who respond to an job advertisement

Discrete random variables with values that are not integers do also exist! This happens when the random variable consists of the ratio of two counts: for example, we might measure the effectiveness of a television advertisement for a luxury car as the number of cars sold divided by the number of times the advertisement was shown on TV. Both the numerator and the denominator are then integers, so the random variable must be a rational number. Detailed consideration of this type of random variable is beyond the scope of this book, but there are a couple of simple examples!

In contrast to discrete random variables, a **continuous random variable** can (conceptually, at least) be measured to any degree of accuracy; i.e. between every two possible values x_1 and x_2 that the random variable can assume, there is another possible value x_3, between x_1 and x_2. The set of all possible values of a continuous random variable is usually an interval of the real line. Examples of continuous random variables are:

- the distance a car travels on one litre of petrol

- the proportion of gold in a sample of ore
- the volume of milk that actually goes into a nominally one litre carton
- the time that a customer waits in the queue at a fast food outlet
- the direction of the wind at midday.

Example 6C: Which of the following are random variables? Which of the random variables are continuous and which are discrete? Write down the set of values that each random variable can take on.

(a) The number of customers arriving at a supermarket during the morning.
(b) The number of letters in the Greek alphabet.
(c) The opening price of gold in New York on Monday next week.
(d) The number of seats that will be sold for a performance of a play in a theatre with a capacity of 328.
(e) The length of time you have to wait at an autobank.
(f) The ratio between the circumference and the diameter of a circle.
(g) The last digit of a randomly selected telephone number.

The distinction between discrete and continuous random variables is critical because we develop different mathematical approaches for the two types of random variable. (Interestingly, though, in advanced treatments of random variables, the mathematical approach for both types is again unified!) We describe discrete random variables mathematically using **probability mass functions**. Continuous random variables are described by **probability density functions**. We adopt the convention of using $p(x)$ to denote a probability mass function and $f(x)$ for a probability density function.

Probability mass functions...

A function $p(x)$ is called a **probability mass function** (frequently abbreviated to p.m.f.) if it satisfies the conditions PMF1, PMF2 and PMF3.

PMF1: $p(x)$ is defined for all values of x, but $p(x) \neq 0$ only at a finite or "countably infinite" set of points.

PMF2: all values of $p(x)$ lie in the unit interval $[0, 1]$, that is $0 \leq p(x) \leq 1$.

PMF3: $\Sigma p(x) = 1$, where the sum is taken over all values of x for which $p(x) \neq 0$.

We now consider several examples of probability mass functions.

Example 7A: An unbiased die is rolled and the random variable X consists of the number of dots appearing on the upturned face. Find the probability mass function for this random variable.

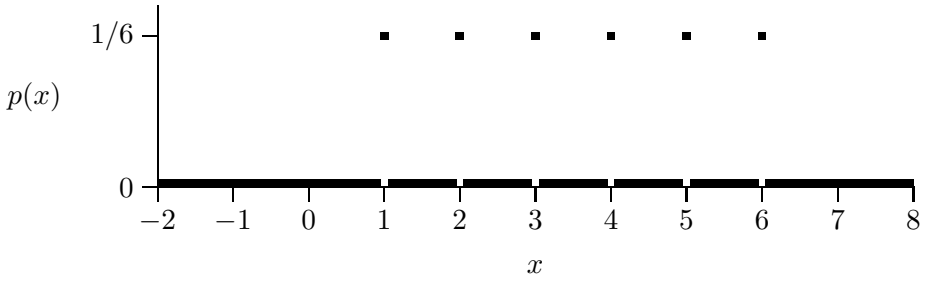
Simply letting $\Pr[X = x] = p(x)$ gives us the required probability mass function. Because the die is unbiased $\Pr[X = 1] = p(1) = 1/6$, and similarly $p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$. All other values of X represent impossible events; for example, $\Pr[X = -1] = p(-1) = 0$ (you cannot get -1 when you toss a die!), $p(0) = 0$, $p(113) = 0$, $p(187) = 0$, and so on. Defined in this way $p(x)$ is non-zero at only six isolated points, and zero for all other values of x, thus satisfying condition PMF1. PMF2 is satisfied

because $p(x)$ is either 0 or 1/6 (which both lie in the closed interval $[0, 1]$, and for PMF3 we note that

$$\sum_{x=1}^{6} p(x) = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1.$$

The probability mass function that describes tossing a single die can therefore be written as

$$\begin{aligned} p(x) &= 1/6 \quad x = 1, 2, 3, 4, 5, 6 \\ &= 0 \quad \text{all other values of } x. \end{aligned}$$



Example 8B: *"Heavily-backed favourite Enforce came through along the inside, but was overwhelmed by Susan's Dream, quoted at 9–1."* Express the anticipated performance of the filly Susan's Dream in this horse race as a probability mass function.

Let $X = 0$ describe the event that Susan's Dream loses the race and $X = 1$ the event that she wins. The quoted odds of 9–1 means that the probability of losing is estimated by the bookmaker as 9 times the probability of winning. Thus $\Pr[X = 0] = 9/10$ and $\Pr[X = 1] = 1/10$, so that

$$\begin{aligned} p(x) &= 9/10 \quad x = 0 \\ &= 1/10 \quad x = 1 \\ &= 0 \quad \text{all other values of } x. \end{aligned}$$

PMF1 is satisfied, because $p(x)$ is non-zero at only two points. Both values of $p(x)$ lie in the unit interval, so PMF2 is satisfied. The two values of $p(x)$ add to one, so PMF3 is satisfied.

Example 9C: The Minister of Environment Affairs has to decide on a fishing quota for the forthcoming season. Currently, the biomass of fish is estimated to be 20 m tonnes. The fish may have a good breeding season (with probability 0.3) and produce 10 m tonnes of young, or have a bad breeding season and produce only 1 m tonnes. A so-called "warm-water event" may occur with probability 0.1, and kill 15 m tonnes of fish, otherwise 1 m tonnes of fish will die. Find the probability mass function for X, the biomass of fish before setting the quota (assuming all events are independent). If the minister bases his decision using a policy that the biomass must remain 10 m tonnes or more with probability 0.8, what should his decision be?

Example 10C: The hostile merger bid by Minorco on Consgold in 1989 was, at one point, considered highly likely to fail by the financial media. They quoted a 12–1 chance of failure. Express the anticipated outcome of the merger as a probability mass function.

For a discrete random variable X, the probability of the event $a \leq X \leq b$ is found by summing the relevant values of the probability mass function:

$$\Pr[a \leq X \leq b] = \sum_{x=a}^{b} p(x).$$

Be careful in your handling of "$\leq$" and "$<$", and "$\geq$" and "$>$": if X assumes only integer values, then

$$\Pr[a < X < b] = \sum_{x=a+1}^{b-1} p(x).$$

Also, if you have to find $\Pr[X \geq b]$, the lower limit of the summation is b, but the upper limit is the largest value of x for which $p(x)$ is defined. You need this information for the following examples.

Example 11A:

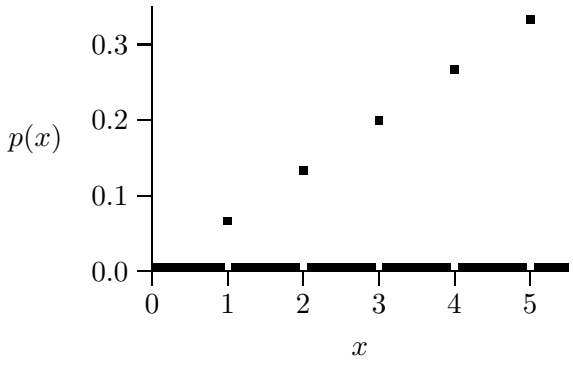
(a) Check that the function

$$\begin{array}{lll} p(x) & = x/15 & x = 1, 2, 3, 4, 5 \\ & = 0 & \text{otherwise} \end{array}$$

satisfies the conditions for being the probability mass function of some random variable X. Sketch $p(x)$.

(b) Find $\Pr[2 \leq X \leq 4]$.

(c) Find $\Pr[X \geq 3]$.



(a) PMF1: $p(x) \neq 0$ for only five values of x and $p(x)$ is defined for all values of x.

PMF2: all values of $p(x)$ lie in the interval $[0, 1]$.

PMF3: $\sum_{x=1}^{5} p(x) = \sum_{x=1}^{5} \frac{x}{15} = \frac{1+2+3+4+5}{15} = 1.$

(b) $\Pr[2 \leq X \leq 4] = \sum_{x=2}^{4} p(x) = \sum_{x=2}^{4} \frac{x}{15} = \frac{2+3+4}{15} = \frac{3}{5}.$

(c) $\Pr[X \geq 3] = \sum_{x=3}^{5} p(x) = \sum_{x=3}^{5} \frac{x}{15} = \frac{3+4+5}{15} = \frac{4}{5}$

Example 12B: Show that the function

$$\begin{aligned} p(x) &= (\frac{1}{2})^x && x = 1, 2, 3, \ldots \\ &= 0 && \text{otherwise} \end{aligned}$$

is a probability mass function.
PMF1: $p(x)$ is defined for all x, and is non-zero on the set of positive integers $\{1, 2, 3, \ldots\}$, a "countably infinite" set[1].
PMF2: $p(x)$ takes on the values 0, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, ... all of which lie in the unit interval.
PMF3:

$$\begin{aligned} \sum_{x=1}^{\infty} p(x) = \sum_{x=1}^{\infty} \frac{1}{2}^x &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \\ &= \frac{1}{2} \Big/ (1 - \frac{1}{2}) = 1. \end{aligned}$$

(Recall that sum of infinity of a geometric progression is given by $a/(1-r)$. Here $a = \frac{1}{2}$ and $r = \frac{1}{2}$.)

Example 13C: Find the sample space for the random experiment which consists of rolling a pair of dice. Find the probability mass function for the random variables X defined to be sum of the values on the dice and Y defined to be the product of the values. Find $\Pr[X \geq 10]$ and $\Pr[Y \geq 13]$.

Example 14C:

(a) Show that the function

$$\begin{aligned} p(x) &= \tbinom{3}{x}\ 0.5^3 && x = 0, 1, 2, 3 \\ &= 0 && \text{otherwise} \end{aligned}$$

satisfied the conditions for being a probability mass function.

(b) Show that it is the probability mass function for the random variable X, the number of heads obtained when three coins are flipped.

(c) Find $P[X \leq 2]$.

Example 15C: Which of the following functions satisfy the conditions for being a probability mass function?

(a)

$$\begin{aligned} p(x) &= 0.25\ (0.75)^x && x = 0, 1, 2, \ldots \\ &= 0 && \text{elsewhere} \end{aligned}$$

(b)

$$\begin{aligned} p(x) &= \tfrac{1}{5}(2x - 3) && x = 1, 2, 3, 4, 5 \\ &= 0 && \text{elsewhere} \end{aligned}$$

[1] A set is said to be countably infinite if there is an orderly way of setting about counting its members. The set of integers is a countably infinite set. However, the set $\{x | 0 \leq x \leq 1\}$, the unit interval, is non-countable — no matter how what system you use to count the numbers, you always leave out infinitely many!

(c)

$$\begin{array}{lll} p(x) & = 0.3 & x = 1 \\ & = 0.4 & x = 2 \\ & = 0.2 & x = 3 \\ & = 0.1 & x = 4 \\ & = 0.0 & \text{otherwise} \end{array}$$

(d)

$$\begin{array}{lll} p(x) & = \binom{4}{x}\binom{3}{2-x}/\binom{7}{2} & x = 0, 1, 2 \\ & = 0 & \text{otherwise} \end{array}$$

Example 16C:

(a) For what value of k will

$$\begin{array}{lll} p(x) & = \frac{k}{x!} & x = 0, 1, 2, 3, 4 \\ & = 0 & \text{otherwise} \end{array}$$

be a probability mass function?

(b) Find $\Pr[X < 2]$.

Example 17C: 10% of the customers entering a supermarket purchase a particular brand of margarine. A market researcher wishes to interview a sample of these customers. As people exit the supermarket she asks them if they have purchased this margarine. Let the random variable X be the number of people she approaches **before** she finds her first customer who has purchased the margarine. Find the probability mass function for X.

Bar graphs...

A probability mass function is conveniently plotted by means of a bar graph. This gives an easily interpretable visual impression of the shape of the distribution of probabilities associated with the random variable. The example demonstrates the method.

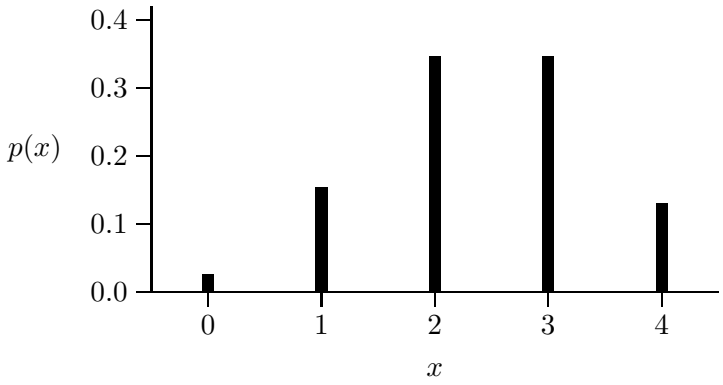
Example 18A: Plot a bar graph for the probability mass function

$$\begin{array}{lll} p(x) & = \binom{4}{x}\, 0.6^x\, 0.4^{4-x} & x = 0, 1, 2, 3, 4 \\ & = 0 & \text{otherwise} \end{array}$$

of the random variable X.
We compute the following probabilities:

x	0	1	2	3	4
$p(x)$	0.026	0.154	0.346	0.346	0.130

and plot them as a **bar graph**. The heights of the lines are equal to the probabilities of the events $X = 0,\ X = 1,\ X = 2,\ X = 3$ and $X = 4$. Naturally, the sum of the heights of the bars must be equal to one.



Example 19C: A random variable X has probability mass function

$$\begin{aligned} p(x) &= \frac{12}{25x} \quad x = 1, 2, 3, 4 \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

Plot the bar graph.

Probability density functions...

Continuous random variables are represented by **probability density functions.** The mathematical treatment of probability density functions is very different to that of probability mass functions: having separate notations for them reminds us to keep the mathematical differences in view. We use $p(x)$ for probability mass functions and $f(x)$ for probability density functions.

A function $f(x)$ is called a **probability density function** (sometimes abbreviated to p.d.f.) if it satisfies the conditions PDF1, PDF2 and PDF3.

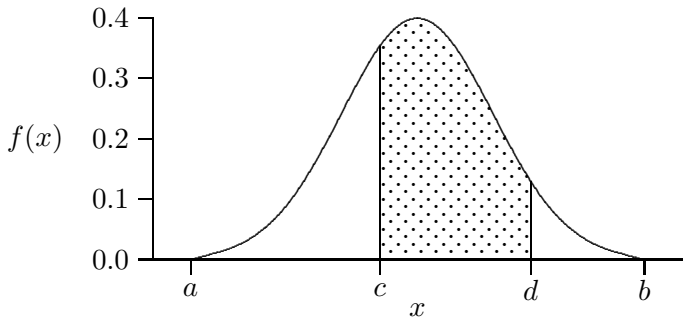
PDF1: $f(x)$ is defined for all values of x.

PDF2: all values of $f(x)$ lie in the interval $[0, \infty)$; that is $0 \leq f(x) < \infty$.

PDF3: $\int_{-\infty}^{\infty} f(x)\,dx = 1$, i.e. the "area under the curve" of a probability density function is one.

Frequently, the function $f(x)$ is non-zero only on some interval, say (a, b) (this interval may also be closed, or one of the limits may be infinity). It is then only necessary to check PDF3 on this interval: $\int_a^b f(x)\,dx = 1$. This is obvious, because then $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{a} 0\,dx + \int_a^b f(x)\,dx + \int_b^{\infty} 0\,dx = \int_a^b f(x)\,dx$ because $f(x) = 0$ outside the interval (a, b).

We have seen that probabilities for **discrete** random variables are found by calculating the values of the probability mass function $p(x)$ at the points of interest and summing them. However, for **continuous** random variables, the probability density function $f(x)$ is constructed in such a way that probabilities of events are found by integration: the area under the graph between the numbers c and d represents the probability of the event "the random variable X lies between c and d" i.e. $\Pr(c \leq X \leq d) = \int_c^d f(x)\,dx$. This is illustrated below:



This, in fact, motivates condition PDF3 of our definition of a probability density function. The random variable X **must** lie between $-\infty$ and ∞. Thus the event "$-\infty < X < \infty$" is a **certain event**, and therefore its probability must be equal to 1:

$$\text{i.e.} \qquad \Pr(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

We now consider several examples of probability density functions.

Example 20A: Show that

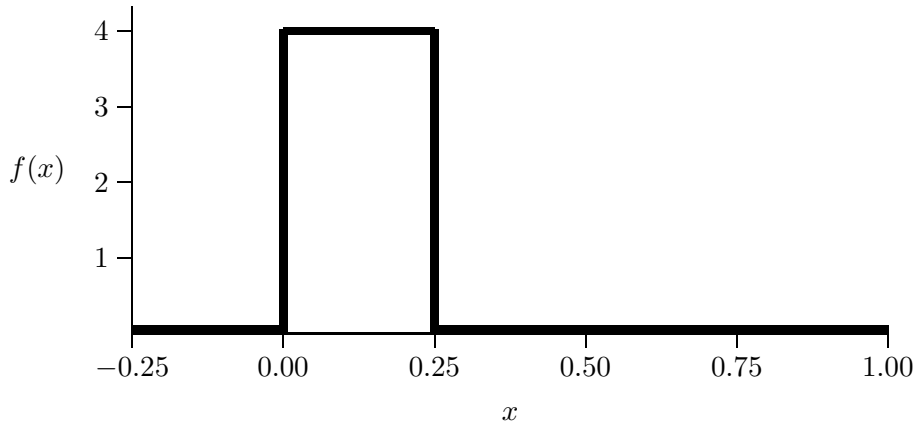
$$\begin{aligned} f(x) &= 4 \quad 0 \le x \le 0.25 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

is a probability density function. Sketch the probability density function.

PDF1: $f(x)$ is defined for all x.

PDF2: $f(x)$ takes on only the values 0 and 4, both of which are positive.

PDF3: $\int_0^{0.25} 4\,dx = [4x]_0^{0.25} = 1$.



Example 21B: In a certain risky sector of the share market, the proportion of companies that survive (i.e. are not delisted) a year is a continuous random variable lying in the interval from zero to one. A statistician examines the data collected over past years and suggests that the function

$$\begin{aligned} f(x) &= 20x^3(1-x) \quad 0 \le x \le 1 \\ &= 0 \quad\quad\quad\quad\quad\quad \text{otherwise} \end{aligned}$$

might be useful in modelling X, the annual proportion of companies that survive.

(a) Check that $f(x)$ is a probability density function.
(b) What is the probability that between 30% and 50% of the companies survive a year?
(c) What is the probability that less than 10% of the companies survive a year?

(a) PDF1 ($f(x)$ is defined for all x), and PDF2 ($f(x) > 0$) are satisfied. To check PDF3,

$$\int_0^1 20x^3(1-x)\,dx = \int_0^1 20x^3 - 20x^4\,dx = [5x^4 - 4x^5]_0^1 = 1,$$

as required.

(b) The probability that between 30% and 50% survive is

$$\Pr[0.3 < X < 0.5] = \int_{0.3}^{0.5} 20x^3(1-x)\,dx = [5x^4 - 4x^5]_{0.3}^{0.5} = 0.157$$

(c) The probability that less than 10% survive is

$$\Pr[X < 0.1] = \int_0^{0.1} 20x^3(1-x)\,dx = [5x^4 - 4x^5]_0^{0.1} = 0.00046$$

Example 22B: A remote country service station is supplied with petrol once a week. The weekly demand for petrol (measured in 1000's of litres) is a random variable with probability density function

$$\begin{array}{lll} f(x) & = \frac{1}{2500}(10-x)^3 & 0 \le x \le 10 \\ & = 0 & \text{otherwise} \end{array}$$

(a) Show that $f(x)$ is a probability density function
(b) What is the probability that between 3 and 5 thousand litres of petrol are sold in a week?
(c) If less than 2 thousand litres are sold in a week, the petrol company does not bother to deliver a supply. What is the probability of this event?
(d) If the service station has a 7 thousand litre tank, what is the probability that it runs out of petrol in a week, assuming that it started the week full?
(e) What size tank is required in order to be 98% certain that weekly demand can be met?
(f) What is the probability of selling exactly 5 thousand litres in a week?

(a) Checking the three conditions:

PDF1: $f(x)$ is defined for all x.

PDF2: $10 - x$ is positive for x in the interval $[0, 10]$, hence $f(x) \ge 0$.

PDF3:

$$\frac{1}{2500}\int_0^{10} (10-x)^3\,dx = \frac{1}{2500}\left[-\frac{1}{4}(10-x)^4\right]_0^{10} = \frac{1}{2500}\left(\frac{1}{4}\times 10^4\right)$$
$$= 1$$

(b) The probability that sales lie between 3000 and 5000 litres is

$$\Pr[3 \le X \le 5] = \frac{1}{2500}\int_3^5 (10-x)^3\,dx = \frac{1}{2500}\left[-\frac{1}{4}(10-x)^4\right]_3^5$$
$$= \frac{1}{2500}\left(-\frac{1}{4}\right)(5^4 - 7^4) = 0.1776$$

(c) The probability that sales are less than 2000 litres is

$$\Pr[0 \le X \le 2] = \frac{1}{2500}\left[-\frac{1}{4}(10-x)^4\right]_0^2$$
$$= \frac{1}{2500}\left(-\frac{1}{4}\right)(8^4 - 10^4) = 0.5904$$

(d) The probability that sales exceed 7000 litres is

$$\Pr[x \ge 7] = \frac{1}{2500}\int_7^{10}(10-x)^3\,dx = \frac{1}{2500}\left(\frac{1}{4}\times 3^4\right) = 0.0081$$

Demand exceeds capacity with probability 0.0081; on average, the tank is emptied once in every 123 weeks.

(e) We want $\Pr[0 \le X \le c] = 0.98$, or equivalently $\Pr[c \le X \le 10] = 0.02$: that is

$$\frac{1}{2500}\int_c^{10}(10-x)^3\,dx = \frac{1}{2500}\left[\frac{1}{4}(10-c)^4\right] = 0.02$$

Thus $c = 10 - (0.02 \times 10\ 000)^{\frac{1}{4}} = 10 - 3.76 = 6.24$. We need a 6240 litre tank.

(f) Probability of selling exactly 5000 litres, may be expressed as

$$\Pr[5 \le X \le 5] = \int_5^5 f(x)\,dx = 0.$$

This is a general principle for **continuous variables — the probability of the random variable taking on a particular value exactly is zero**. This seems counterintuitive, but is due to the fact that our ability to measure is always discrete (for example, digital petrol pumps measure to the "nearest" tenth of a litre). Continuous random variables are essentially unobservable, an abstract mathematical concept that is useful only because it is convenient.

As a corollary of the above, note that for **continuous** random variables

$$\Pr[c \le X \le d] = \Pr[c < X \le d] = \Pr[c \le X < d] = \Pr[c < X < d] = \int_c^d f(x)\,dx.$$

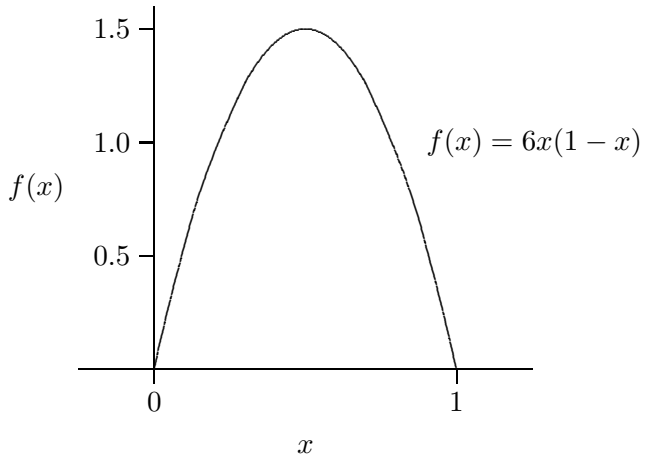
(This is not true of discrete random variables, where one has to be alert to the type of inequality.)

Examples 21B and 22B showed how a random variable and a "probability density function" could be used to "model" a practical problem. The particular probability density functions that we used were chosen to make the integration trivial, and would certainly be poor representations of reality in both situations. In the next chapter we will be considering various probability mass functions and probability density functions which have proved themselves useful in practice as models of real-world phenomena.

Example 23B:

(a) Could the following function serve as a probability density function for some random variable X?

$$\begin{aligned} f(x) &= 6x(1-x) && 0 \le x \le 1 \\ &= 0 && \text{otherwise} \end{aligned}$$



(b) What is the probability that

(i) $0 \leq X \leq 0.2$?

(ii) $0.4 \leq X \leq 0.6$?

(a) We must check that the three conditions of our definition are satisfied

PDF1: $f(x)$ is defined for all x.

PDF2: All values $f(x)$ lie in the set $\{y|0 \leq y \leq 1.5\}$ which contains no negative numbers.

PDF3: We need to check that the area under the curve between 0 and 1 is equal to 1:

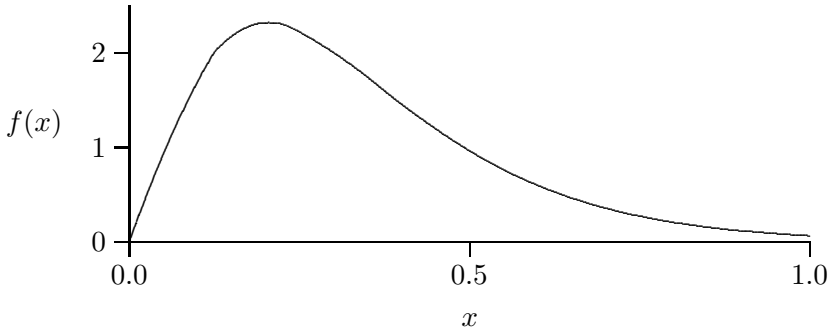
$$\begin{aligned}\int_0^1 f(x)\,dx &= \int_0^1 6x(1-x)\,dx = 6\int_0^1 (x-x^2)\,dx \\ &= 6\left[\frac{1}{2}x^2 - \frac{1}{3}x^3\right]_0^1 \\ &= 6\left(\frac{1}{2} - \frac{1}{3}\right) = 1.\end{aligned}$$

The conditions are satisfied.

(i) $\Pr[0 \leq X \leq 0.2] = 6\int_0^{0.2} (x-x^2)\,dx = 0.104$

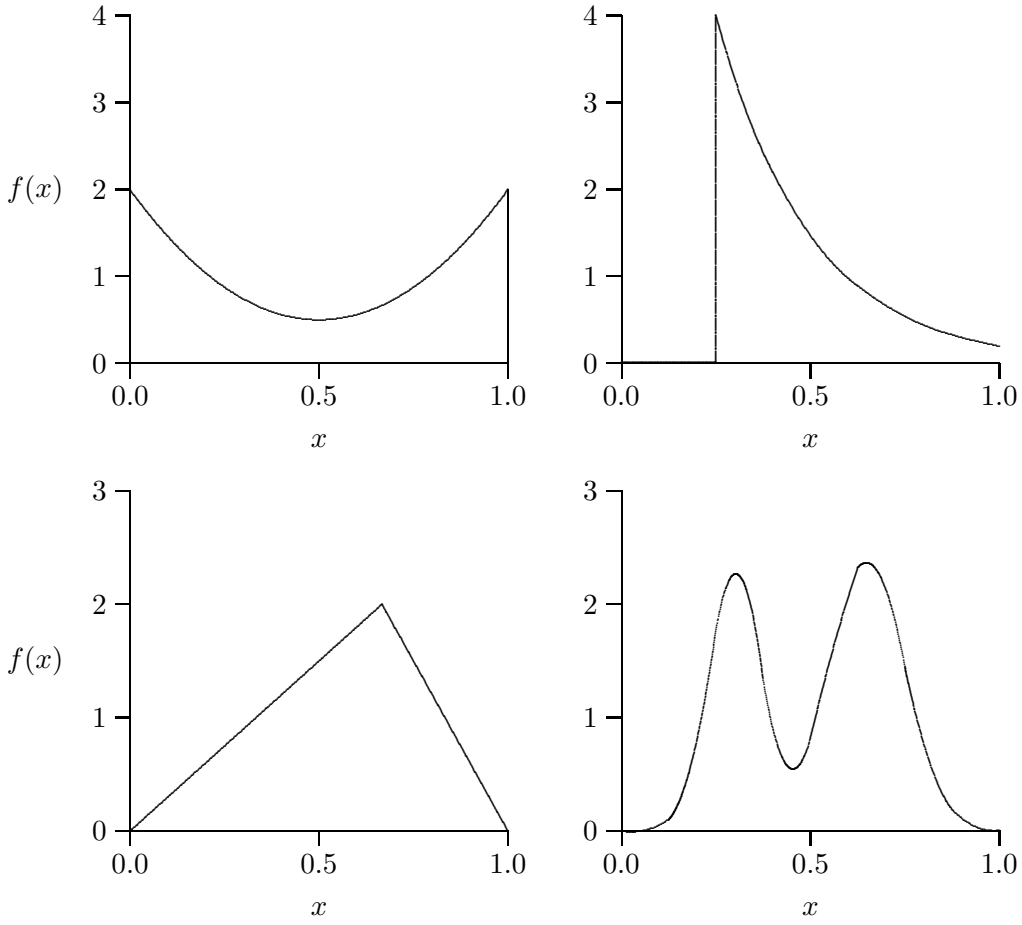
(ii) $\Pr[0.4 \leq X \leq 0.6] = 6\int_{0.4}^{0.6} (x-x^2)\,dx = 0.296$

Note that there is no requirement for the graph of a probability density function $f(x)$ to be a "smooth" curve such as this one:



In fact, a great variety of shapes are possible. The only restrictions are that $f(x)$ must be non-negative, and that the area under the curve must be equal to one. It is

important to grasp that the actual values of $f(x)$ (the "height of the curve at value x") cannot be interpreted as being the probability that the random variable X is equal to x. This interpretation was possible with graphs of the probability mass functions of discrete random variables. For continuous random variables, probabilities are computed by integrating the probability density function.



Example 24B: Let X be a random variable with probability density function

$$\begin{aligned} f(x) &= ke^{-\frac{1}{2}x} && 0 \le x \le \infty \\ &= 0 && \text{otherwise} \end{aligned}$$

What value must k assume?

To make $f(x)$ a density function, k must be chosen so that

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

i.e.

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,dx &= \int_{0}^{\infty} ke^{-\frac{1}{2}x}\,dx \\ &= \left[-2ke^{-\frac{1}{2}x}\right]_0^{\infty} = 2k = 1 \end{aligned}$$

Thus $k = \frac{1}{2}$.

A selection of examples...

Example 25C:

(a) Find the value of k, so that the function

$$\begin{aligned} f(x) &= k(x^2-1) && 1 \le x \le 3 \\ &= 0 && \text{otherwise} \end{aligned}$$

may serve as a probability density function.

(b) Find the probability that X lies between 2 and 3.

Example 26C: Verify that each of the following functions satisfies the conditions for being either a probability mass function or probability density function.

(a)
$$\begin{aligned} p(x) &= x/6 && x = 1, 2, 3 \\ &= 0 && \text{otherwise} \end{aligned}$$

(b)
$$\begin{aligned} p(x) &= \tbinom{4}{x}\tfrac{1}{2}^4 && x = 0, 1, 2, 3, 4 \\ &= 0 && \text{otherwise} \end{aligned}$$

(c)
$$\begin{aligned} f(x) &= 1 && 0 \le x \le 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

(d)
$$\begin{aligned} f(x) &= |x| && -1 \le x \le 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

(e)
$$\begin{aligned} f(x) &= -\log x && 0 < x \le 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

(f)
$$\begin{aligned} p(x) &= e^{-1}/x! && x = 0, 1, 2, \ldots \\ &= 0 && \text{otherwise} \end{aligned}$$

(g)
$$\begin{aligned} p(x) &= 1/n && x = 1, 2, 3, \ldots, n \\ &= 0 && \text{otherwise} \end{aligned}$$

(h)
$$\begin{aligned} f(x) &= \tfrac{1}{2}\sin x && 0 \le x \le \pi \\ &= 0 && \text{otherwise} \end{aligned}$$

Example 27C: The probability density function of a random variable X is given by

$$\begin{aligned} f(x) &= kx(1-x^2) && 0 \le x \le 1 \\ &= 0 && \text{elsewhere} \end{aligned}$$

(a) Show that the value of k must be 4

(b) Calculate $\Pr[0 < X < \frac{1}{2}]$

(c) Find the value of A so that $\Pr[0 < X < A] = \frac{1}{2}$.

Example 28C: For what values of A can $p(x)$ be a probability mass function?

$$\begin{array}{lll} p(x) & = (1-A)/4 & x = 0 \\ & = (1+A)/2 & x = 1 \\ & = (1-A)/4 & x = 2 \\ & = 0 & \text{otherwise} \end{array}$$

Example 29C: A small pool building company is equally likely to be able to complete 2 or 3 pool contracts each month. The company receives between 1 and 4 contracts to build pools each month, with probabilities $\Pr(1) = 0.1$, $\Pr(2) = 0.2$, $\Pr(3) = 0.5$, $\Pr(4) = 0.2$. At the beginning of this month the company has two contracts carried forward from the previous month. The random variable X of interest is the number of contracts to be carried forward to next month. Find the probability mass function of X. In particular, what is the probability that no contracts will be carried forward to next month? Assume that the number of contracts is independent of the number of pools completed. Also, to simplify the problem, assume that the contracts for a month are made at the beginning of the month.

Solutions to examples...

5C The sample space, numerical values for the elementary events and their associated probabilities are:

S	=	{	EE	EB	EL	BB	BE	BL	LL	LE	LB	}
X	=		4000	3000	2000	2000	3000	1000	0	2000	1000	
Pr	=		0.04	0.06	0.10	0.09	0.06	0.15	0.25	0.10	0.15	

The probability mass function is therefore given by

$$\begin{array}{lll} p(x) & = 0.25 & x = 0 \\ & = 0.30 & x = 1000 \\ & = 0.29 & x = 2000 \\ & = 0.12 & x = 3000 \\ & = 0.04 & x = 4000 \\ & = 0 & \text{otherwise} \end{array}$$

6C (b) & (f) are not random variables, (a), (d) & (g) are discrete, (c) & (e) are continuous.

9C

$$\begin{array}{lll} p(x) & = 0.07 & x = 6 \\ & = 0.03 & x = 15 \\ & = 0.63 & x = 20 \\ & = 0.27 & x = 29 \\ & = 0 & \text{otherwise} \end{array}$$

Set the quota at 10 m tonnes.

10C Let $X = 0$ be fail, $X = 1$ be succeed.

$$\begin{array}{lll} p(x) & = 12/13 & x = 0 \\ & = 1/13 & x = 1 \\ & = 0 & \text{otherwise} \end{array}$$

13C The sample space, numerical values for the elementary events and their associated probabilities are:

X	=	2	3	4	5	6	7	8	9	10	11	12
Probability	=	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Y =	1	2	3	4	5	6	8	9	10	12	15	16	18	20	24	25	30	36
$\Pr(Y) =$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$\Pr[x \geq 10] = 0.1667$, $\Pr[Y \geq 13] = \frac{13}{36}$.

14C (c) 0.875

15C (a), (c) and (d) are probability mass functions, but (b) is not, because $p(1) = -0.2 < 0$.

16C (a) $k = 24/65$ (b) 48/65

17C The probability mass function is

$$\begin{aligned} p(x) &= 0.1 \times 0.9^x && x = 0, 1, 2, \ldots \\ &= 0 && \text{otherwise} \end{aligned}$$

25C (a) $k = 3/20$ (b) $\Pr[2 < X < 3] = 4/5$

26C (a), (b), (f) and (g) are probability mass functions, (c), (d), (e) and (h) are probability density functions.

27C (b) 7/16 (c) $A = 0.5412$

28C $-1 \leq A \leq 1$ (otherwise probabilities are either negative or greater than one).

29C

$$\begin{aligned} p(x) &= 0.1 \times 0.5 = 0.05 && x = 0 \\ &= 0.2 \times 0.5 + 0.1 \times 0.5 = 0.15 && x = 1 \\ &= 0.2 \times 0.5 + 0.5 \times 0.5 = 0.35 && x = 2 \\ &= 0.5 \times 0.5 + 0.2 \times 0.5 = 0.35 && x = 3 \\ &= 0.2 \times 0.5 = 0.1 && x = 4 \\ &= 0 && \text{otherwise} \end{aligned}$$

EXERCISES...

*4.1 Which of the following random variables are discrete, and which are continuous?

(a) the time required to answer this question
(b) the number of words in a book chosen at random from the library
(c) the number of "heads" in 6 flips of a coin
(d) the number of goals scored in a soccer match
(e) the maximum temperature recorded at Cape Town International Airport today
(f) the volume of air breathed in by an individual when asked to "take a deep breath"
(g) the annual income to the nearest cent of a randomly chosen wage-earner
(h) the population of a randomly chosen town in the Free State
(i) the length of time you have to wait for a bus

(j) the amount of rain that falls in a day.

*4.2 Check which of the following functions can serve as probability mass functions or probability density functions.

(a)
$$\begin{aligned} p(x) &= x/6 \quad x = 1, 2, 3 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

(b)
$$\begin{aligned} f(x) &= \tfrac{3}{10}(2 - x^2) \quad -1 \le x \le 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

(c)
$$p(x) = x \qquad x = \frac{1}{16}, \frac{3}{16}, \frac{1}{4}, \frac{1}{2}$$

(d)
$$\begin{aligned} f(x) &= 2x/3 \quad -1 < x < 2 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

(e)
$$\begin{aligned} f(x) &= \tfrac{1}{4} \quad 3 < x < 7 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

(f)
$$\begin{aligned} p(x) &= \tfrac{x}{\frac{1}{2}n(n+1)} \quad x = 1, 2, \ldots, n \\ &= 0 \quad \text{otherwise} \end{aligned}$$

4.3 Show that the following functions are probability mass functions.

(a)
$$\begin{aligned} p(x) &= e^{-2}\, 2^x \,/\, x! \quad x = 0, 1, 2, \ldots \\ &= 0 \quad \text{otherwise} \end{aligned}$$

(b)
$$\begin{aligned} p(x) &= \tbinom{5}{x} \tfrac{1}{4}^x \tfrac{3}{4}^{5-x} \quad x = 0, 1, 2, 3, 4, 5 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

4.4 Show that the following functions are probability density functions.

(a)
$$\begin{aligned} f(x) &= (2\sqrt{x})^{-1} \quad 0 < x < 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

(b)
$$\begin{aligned} f(x) &= \tfrac{1}{4}\, x e^{-\frac{1}{2}x} \quad 0 \le x < \infty \\ &= 0 \quad \text{otherwise} \end{aligned}$$

*4.5 What must the value of k be so that the following functions are probability density functions?

(a)
$$\begin{aligned} f(x) &= kx^2(1 - x) \quad 0 < x < 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

(b)
$$\begin{aligned} f(x) &= ke^{-4x} \quad 0 \le x < \infty \\ &= 0 \quad \text{otherwise} \end{aligned}$$

4.6 A random variable X has probability density function given by

$$\begin{aligned} f(x) &= Ax^3 \quad 0 \le x \le 10 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Find A. What is the probability that X lies between 2 and 5, and what is the probability that X is less than 3? Sketch the density function.

*4.7 A random variable X has probability density function

$$\begin{aligned} f(x) &= e^{-x} \quad 0 \le x < \infty \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Find the number t such that $\Pr[X < t] = \frac{1}{2}$.

4.8 For the probability density function

$$\begin{aligned} f(x) &= 2x \quad 0 \le x \le 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

find the number a such that the probability that $X < a$ is three times the probability that $X \ge a$.

4.9 If $f(x) = 3x^2$ for $0 < x < 1$, and zero elsewhere, find the number b, such that X is equally likely to be greater than, or less than b.

*4.10 The probability density function of the life in hours X of a certain kind of radio tube is found to be

$$\begin{aligned} f(x) &= 100/x^2 \quad x > 100 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Three such tubes are bought for a radio set. What is the probability that none will have to be replaced during the first 150 hours of operation?

4.11 A batch of small-calibre ammunition is accepted as satisfactory if none of a sample of five shots fall more than 8 cm from the target at a given range. If X, the distance from the centre of the target to an impact point, has probability density function

$$\begin{aligned} f(x) &= xe^{-x} \quad 0 \le x < \infty \\ &= 0 \quad \text{otherwise} \end{aligned}$$

find the probability that a batch is accepted.

Further exercises...

4.12 A continuous random variable X has probability density function

$$\begin{aligned} f(x) &= k \quad a \le x \le b \\ &= 0 \quad \text{otherwise} \end{aligned}$$

for arbitrary constants a and b. Find the value of k.

4.13 Find values for c so that the following functions may serve as probability density functions:

(a)

$$\begin{array}{lll} f(x) & = c + e^{-x} & 0 \le x \le 1 \\ & = 0 & \text{otherwise} \end{array}$$

(b)

$$\begin{array}{lll} f(x) & = x + \frac{1}{2} & 0 \le x \le c \\ & = 0 & \text{otherwise} \end{array}$$

*4.14 The density function for a random variable X is given by

$$\begin{array}{lll} f(x) & = \frac{3}{4}(kx - x^2) & 0 \le x \le 2 \\ & = 0 & \text{otherwise} \end{array}$$

(a) Determine the value of k.
(b) Calculate $\Pr[0 < X < 1]$.

*4.15 (a) Check whether

$$\begin{array}{lll} p(x) & = \binom{5}{x}\binom{4}{2-x}/36 & x = 0, 1, 2 \\ & = 0 & \text{elsewhere} \end{array}$$

is a probability mass function.
(b) Calculate $\Pr[X = 0]$, $\Pr[X = 1]$ and $\Pr[X = 2]$.

Solutions to exercises...

4.1 (b) (c) (d) (g) and (h) are discrete
(a) (e) (f) and (i) are continuous.
(j) is an unusual example of a mixed continuous and discrete random variable: although the random variable is, at face value, continuous, it cannot be modelled by a conventional probability density function because the probability of no rain in a day is not zero but positive. The probability function for X needs to be something like

$$\begin{array}{lll} p(x) & = p & x = 0 \\ & = f(x) & x > 0 \\ & = 0 & \text{otherwise} \end{array}$$

with the "probability density function" $f(x)$ integrating to $1 - p$.

4.2 (a) (b) (e) and (f) satisfy conditions.
For (c), $p(x)$ is not defined for all X.
For (d), $f(x) < 0$ for $-1 < x < 0$.

4.5 (a) $k = 12$ (b) $k = 4$

4.6 $A = 1/2500$, $\Pr[2 < X < 5] = 0.0609$, $\Pr[X < 3] = 0.0081$.

4.7 0.6931

4.8 0.8660

4.9 0.7937

4.10 8/27

4.11 0.9850

4.12 $k = 1/(b-a)$

4.13 (a) $c = e^{-1}$ (b) $c = 1$

4.14 (a) $k = 2$ (b) $\frac{1}{2}$

4.15 $\Pr[X = 0] = 6/36$, $\Pr[X = 1] = 20/36$, $\Pr[X = 2] = 10/36$

Chapter 5

PROBABILITY DISTRIBUTIONS I: THE BINOMIAL, POISSON, EXPONENTIAL AND NORMAL DISTRIBUTIONS

KEYWORDS: Binomial, Poisson, exponential and normal distributions.

A number of probability mass and density functions have proved themselves useful as "models" for a large variety of practical problems in business and elsewhere. We consider four of the most frequently encountered probability distributions in this chapter — the Binomial, Poisson, Exponential and Normal Distributions.

The binomial distribution ...

The **binomial distribution** may be used as a probability model in situations in which the following conditions are satisfied:

1. We have a random experiment which has a sample space with exactly two outcomes, one of which we can label "success", and the other "failure": i.e. $S = \{\text{success, failure}\}$.

   e.g. A door-to-door salesperson calls on a prospective client — the client either purchases the product (success) or does not purchase (failure).

2. The random experiment is repeated n times, $n \geq 1$. The outcome on any one repetition is not influenced by the outcome on any other repetition. We say "**we have n independent trials of the random experiment**".

   e.g. The salesperson calls on $n = 6$ prospective clients — the clients make their purchasing decisions independently (there is no communication between them!).

3. The probability of success remains constant from trial to trial. We assume that each client is equally likely to purchase the product. Let $\Pr(\text{success}) = p$; thus $\Pr(\text{failure}) = 1-p$. It is sometimes convenient to let $q = 1-p$, so that $\Pr(\text{failure}) = q$.

Our random variable X is the number of successes we observe in n trials. If the conditions above are satisfied, then we say that we have a **binomial process**, and that

the random variable X has a binomial distribution. In the above example, X is the number of calls that resulted in sales. Because 6 calls were made, X must assume one of the values $0, 1, \ldots, 6$, and X is therefore an example of a discrete random variable.

Binomial processes occur in many contexts. From an industrial or commercial perspective, one of the most important binomial processes occurs in the field of quality. The quality of a product or service, whether it is a tomato, a nail, a personal computer, a car, an insurance policy or the punctuality of a train, can be classified as "satisfactory" or "defective". In particular, the binomial probability distribution provides the basis for deciding whether or not a consignment of goods meets the desired specifications.

BINOMIAL DISTRIBUTION

In a binomial process, we have n independent trials, each trial has two outcomes, success of failure, and $\Pr[\text{success}] = p$ for all trials. Let the random variable X be the number of successes in n trials.

Then X has the **binomial distribution**, and $\Pr[X = x]$ is given by the probability mass function

$$\begin{array}{lll} p(x) & = \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ & = 0 & \text{otherwise} \end{array}$$

Once we give values to n and p, $(n \geq 1, 0 < p < 1)$, a particular binomial distribution is specified. n and p are examples of what we call the **parameters** of the distribution. Once the parameters of a distribution have values, a particular distribution is specified.

We have a neat abbreviated notation which saves us writing "the random variable X is distributed binomially with parameters n and p". We compress all this information into the symbols $X \sim B(n, p)$.

Example 1A: A door-to-door salesperson calls on 6 clients per session. Each client makes their purchasing decision independently of the others, with probability 0.2 of purchasing the product. What are the probabilities that 0, 1, 2, 3, 4, 5 or 6 clients purchase the product?

Clearly, the three conditions for the binomial process are satisfied, and X, the number of clients who purchase the product, has a binomial distribution with $n = 6$ and $p = 0.2$: thus $X \sim B(6, 0.2)$.

Instead of simply using the formula given in the box, let us compute **from first principles** the probability of, say, 2 clients purchasing the product, i.e. $\Pr[X = 2]$:

Firstly, 2 clients out of 6 can purchase the product in many different permutations. Let A_1 be the event that the first 2 clients purchase (these are the "successes" that we count) and that clients 3 to 6 refuse to purchase (i.e. are "failures"). Then, using our usual conventions, we can write

$$A_1 = S \cap S \cap F \cap F \cap F \cap F.$$

Let the events A_2, A_3 represent other permutations of 2 successes and 4 failures, e.g.

$$A_2 = F \cap S \cap S \cap F \cap F \cap F$$
$$A_3 = F \cap F \cap S \cap S \cap F \cap F$$

How many permutations of 2 successes and 4 failures are there? Counting rule 6 tells us there are $\binom{6}{2} = 15$ such permutations, so we could write down events from A_1 to A_{15}.

Secondly, we compute $\Pr(A_1)$. Because the clients act **independently** of each other,

$$\begin{aligned}\Pr(A_1) &= \Pr(S \cap S \cap F \cap F \cap F \cap F) \\ &= \Pr(S) \times \Pr(S) \times \Pr(F) \times \Pr(F) \times \Pr(F) \times \Pr(F) \\ &= p^2(1-p)^4 = 0.2^2 \times 0.8^4.\end{aligned}$$

Recall that the probability of the intersection of independent events is the product of the individual probabilities, so that

$$\Pr(A_1) = \Pr(A_2) = \ldots = \Pr(A_{15}) = 0.8^4 \times 0.2^2$$

Thirdly, the events $A_1, A_2, \ldots, A_{15}$ are **mutually exclusive** — no client can simultaneously both purchase and refuse to purchase! Thus

$$\begin{aligned}\Pr[X = 2] &= p[A_1 \cup A_2 \cup \cdots \cup A_{15}] \\ &= \Pr(A_1) + \Pr(A_2) + \cdots + \Pr(A_{15}) \\ &= \binom{6}{2} 0.2^2\ 0.8^4 = 15 \times 0.2^2 \times 0.8^4 = 0.2458.\end{aligned}$$

Stop a while and convince yourself that the answer $\binom{6}{2} 0.2^2\ 0.8^4$ obtained from first principles is the same as that obtained by substituting $n = 6,\ p = 0.2$ and $x = 2$ into the formula for the binomial probability mass function.

Try computing the remaining probabilities from first principles, and compare them with the results obtained from the formula. The probabilities are given in the table below:

x	$p(x) = \Pr[X = x]$
0	$\binom{6}{0} \times 0.8^6 = 0.2621$
1	$\binom{6}{1} \times 0.2^1 \times 0.8^5 = 0.3932$
2	$\binom{6}{2} \times 0.2^2 \times 0.8^4 = 0.2458$
3	$\binom{6}{3} \times 0.2^3 \times 0.8^3 = 0.0819$
4	$\binom{6}{4} \times 0.2^4 \times 0.8^2 = 0.0154$
5	$\binom{6}{5} \times 0.2^5 \times 0.8^1 = 0.0015$
6	$\binom{6}{6} \times 0.2^6 = 0.0001$

The probability that all six clients purchase the product is very small (0.0001) but will occasionally occur (we expect it roughly once in every 10 000 times that a session of 6 calls are made!). The probability that none of the 6 clients purchase is 0.2621, so that in approximately a quarter of sessions of 6 calls no purchases are made. The probability that two or more purchases are made is $\Pr[X \geq 2] = 0.2458 + 0.0819 + 0.0154 + 0.0015 + 0.0001 = 0.3417$, so that in approximately one-third of sessions of 6 calls the salesperson achieves two or more sales.

Example 2B: What is the probability of a contractor being awarded only one out of five contracts? Assume that the probability of being awarded a contract is 0.5.

Let "success" = "awarded a contract". Pr (success) $= p = \frac{1}{2}$. So $q = 1 - p = \frac{1}{2}$. We have $n = 5$ trials. Let X be the number of successes in 5 trials. Then $X \sim B(5, \frac{1}{2})$.

$$P[X = x] = p(x) = \binom{5}{x} \frac{1}{2}^x \frac{1}{2}^{5-x} \qquad x = 0, 1, \ldots, 5$$

So

$$\Pr[X = 1] = p(1) = \binom{5}{1} \frac{1}{2}^5 = 5/32.$$

Example 3B: Check that the binomial distribution

$$\begin{array}{lll} p(x) & = \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \ldots, n \\ & = 0 & \text{otherwise} \end{array}$$

is a probability mass function.

(i) It is defined everywhere and $p(x) \neq 0$ on the finite set $\{0, 1, \ldots, n\}$.

(ii) $p(x)$ has no negative terms.

(iii) For convenience, let $q = 1 - p$.

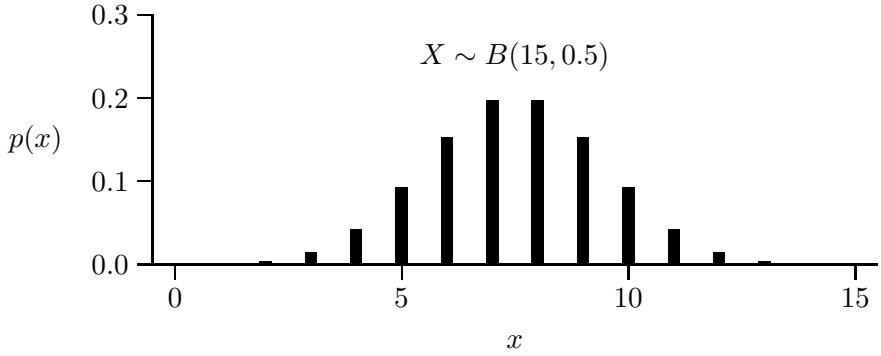
$$\begin{aligned} \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} &= \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} \\ &= \binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1} + \cdots + \binom{n}{x} p^x q^{n-x} + \cdots + \binom{n}{n} p^n q^0 \\ &= (p+q)^n \end{aligned}$$

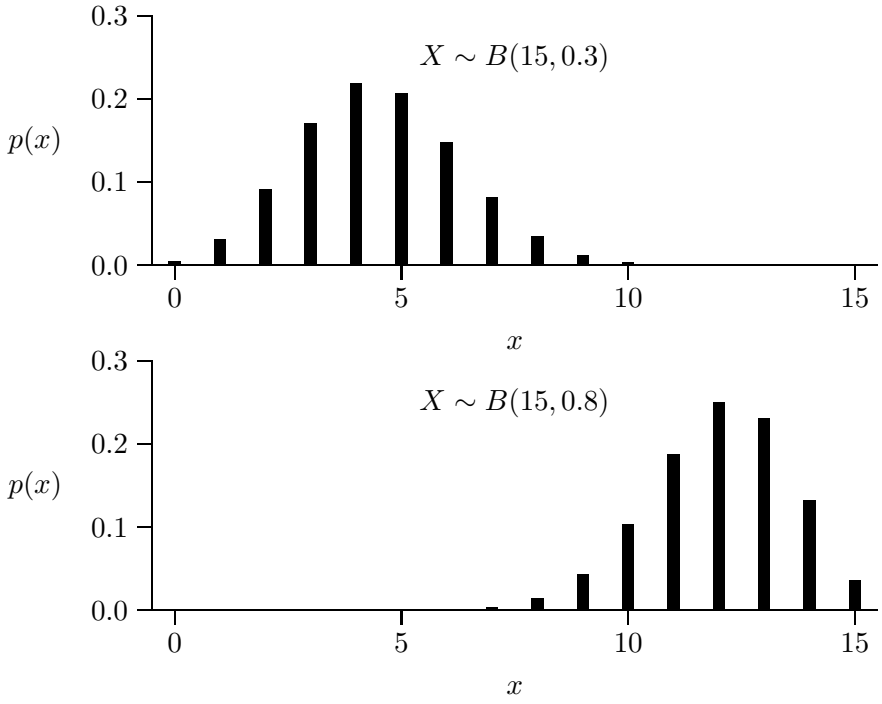
(from the binomial theorem — hence the name "binomial "distribution)

$$\begin{aligned} &= 1^n \quad (\text{because } q = 1 - p) \\ &= 1. \end{aligned}$$

Bar graphs of the binomial distribution ...

To gain some feeling for the shape of the binomial distribution, consider the following three bar graphs, for which n is fixed at 15, and p is varied.





Further examples on the binomial distribution...

Example 4B: A certain type of pill is packed in bottles of 12 pills each. 10% of the pills are chipped in the manufacturing process.

(a) Explain why the binomial distribution can provide a reasonable model for the random variable X, the number of chipped pills found in a bottle of 12 pills. What are the appropriate parameters?

(b) What is the probability that a bottle of pills contains x chipped pills, i.e. what is $\Pr[X = x]$?

(c) What are the probabilities of

(i) 2 chipped pills?

(ii) no chipped pills?

(iii) at least 2 chipped pills?

(a) We check that the three conditions are satisfied.

1. The random experiment consists of examining a pill, and deciding whether it is chipped or unchipped. Thus there are two possible outcomes, as required. Because "chipped pills" are the things we are looking for and counting, we will let "chipped pill" = "success" and "unchipped pill" = "failure".
2. There are 12 pills in the bottle. We repeat the experiment 12 times, examining each pill. It seems reasonable to assume that the pills are chipped independently of each other.
3. It also seems reasonable that the probability of a pill being chipped is the same for each pill.

Thus the binomial distribution with parameters $n = 12$ and $p = 0.10$ may be used to model the phenomenon of the number of chipped pills in a bottle of pills.

(b) Because $X \sim B(12, 0.10)$

$$\begin{array}{lll} p(x) & = \binom{12}{x} 0.10^x \ 0.90^{12-x} & x = 0, 1, \ldots, 12 \\ & = 0 & \text{otherwise} \end{array}$$

(c) (i) $\Pr[X = 2] = p(2) = \binom{12}{2} 0.10^2 \ 0.90^{10} = 0.2301$

(ii) $\Pr[X = 0] = p(0) = \binom{12}{0} 0.10^0 \ 0.90^{12} = 0.2824$

(iii) $\Pr[X \geq 2] = 1 - \Pr[X = 0] - \Pr[X = 1] = 1 - 0.2824 - 0.3766 = 0.3410.$

Example 5C: A TV manufacturer is supplied with a certain component by a specialist producer. Each incoming consignment of components is subjected to the following quality control procedure. A random sample of 10 components is individually tested. If there are one or more defective components among the 10 tested, the entire consignment is rejected. If there are no defective components in the sample, the consignment is accepted.

(a) What are the probabilities of a consignment being rejected if the true proportions of defective components are
(i) 1% (ii) 10% (iii) 30%

(b) If a sample of 20 components (instead of 10) were tested, and the consignment rejected if two or more proved defective, calculate the probabilities of rejecting a consignment for the same proportions of defective components.

(c) Which quality control procedure do you think is the better?

The Poisson distribution...

Many phenomena in physics obey the Poisson probability law named in honour of the French mathematician Simeon D. Poisson (1781–1840). The classic example is the decomposition of radio-active nuclei. In management science, the **number of demands for service** in a given period of time (e.g. on tellers in a bank, the stock pile of a factory, the runways of an airport, the lines of a telephone exchange) often obeys (either exactly or approximately) a Poisson distribution. This applies also to the occurrence of accidents, errors, breakdowns and other calamities — the number that occurs within a specified time period has a Poisson distribution under certain circumstances.

In broad terms, the condition for a "Poisson process" is that the events occur in time "at random". Loosely, this means that an event is equally likely to occur at any instant in time. If a phenomenon obeys the Poisson process, then the Poisson distribution may be used to model **the number of occurrences of the event during a fixed time period**. We can also use the Poisson distribution when we count the occurrences of an event in a fixed amount of "space". For example, the number of faults in 100 m of computer cable, the number of misprints on an A4 page, and the number of diamonds in a cubic metre of ore are all Poisson processes (if the "events" occur at random in space) and can be modelled using the Poisson distribution.

The Poisson distribution has only one parameter, namely the average rate λ at which events are occurring per time period. Because the number of events that occur in the interval must be an integer, the Poisson distribution is discrete. The probability mass function is given in the box:

POISSON DISTRIBUTION

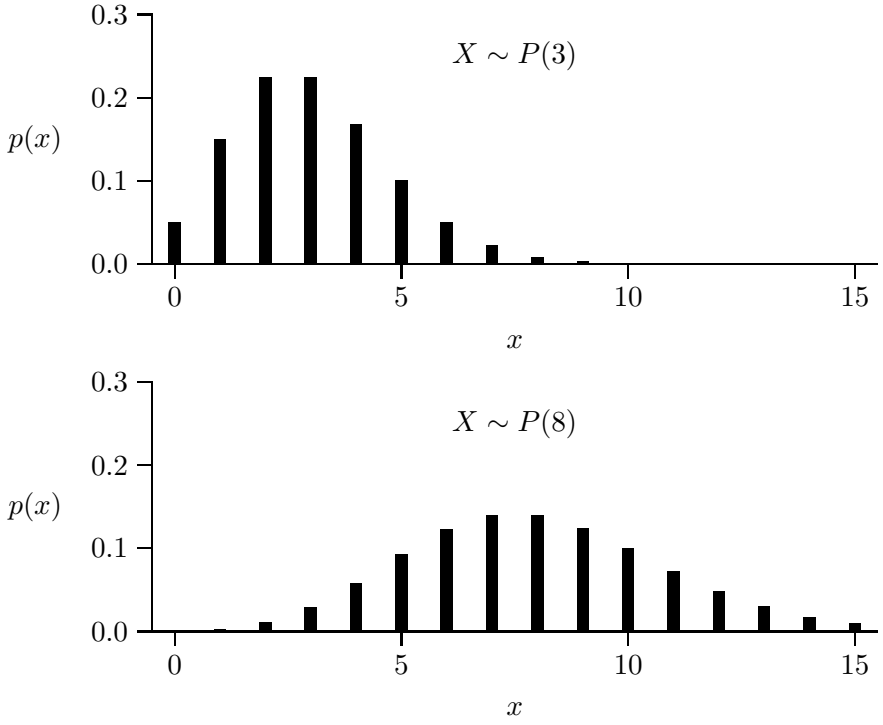
We are given a period of time during which events occur at random. The average rate at which events occur is λ events per time period.

It is critical that the time period referred to in the rate **must be the same** as the time period during which the events are counted. Let the random variable X be the number of events occurring during the time period.

Then X has the Poisson distribution with parameter λ, i.e. $X \sim P(\lambda)$, and has probability mass function

$$\begin{aligned} p(x) &= \frac{e^{-\lambda}\lambda^x}{x!} && x = 0, 1, 2, \ldots \\ &= 0 && \text{otherwise} \end{aligned}$$

The bar graphs below show the shape of Poisson distribution for two values of λ.



Example 6A: We have a large fleet of delivery trucks. On average we have 12 breakdowns per 5-day working week. Each day we keep two trucks on standby. What is the probability that on any day

(a) no standby trucks are needed?
(b) the number of standby trucks is inadequate?

Let the random variable X be the number of trucks that break down in a given day. Because we are dealing with breakdowns, it is reasonable to assume that they occur at random and that the Poisson distribution is a realistic model.

Because we are interested in breakdowns per day, we need to convert the given weekly rate into a daily rate. 12 breakdowns per 5 days is equivalent to $12/5 = 2.4$ breakdowns per day. Thus we assume that X has the Poisson distribution with parameter $\lambda = 2.4$, i.e. $X \sim P(2.4)$. Hence

$$\Pr[X = x] = p(x) = \frac{e^{-2.4}2.4^x}{x!}$$

(a) $\Pr(\text{no breakdowns}) = \Pr[X = 0] = p(0) = \dfrac{e^{-2.4}2,4^0}{0!} = 0.0907$

(b)

$$\begin{aligned}\Pr(\text{inadequate standby trucks}) &= \Pr[X > 2] \\ &= 1 - \Pr[X \leq 2] \\ &= 1 - (p(0) + p(1) + p(2)) \\ &= 1 - (0.0907 + 0.2177 + 0.2613) \\ &= 0.4303.\end{aligned}$$

This means that 9% of days we will not use our standby trucks at all, but that on 43% of days we will run out of standby trucks. We should investigate the financial implications of putting more trucks on standby.

Example 7B: Bank tellers make errors in entering figures in their ledgers at the rate of 0.75 errors per page of entries. What is the probability that in a random sample of 4 pages there will be 2 or more errors?

Because we are dealing with errors, we assume a Poisson distribution. If errors occur at 0.75 errors per page, then the error rate per 4 pages is 3. So we choose $\lambda = 3$.
Hence

$$\Pr[X = x] = \frac{e^{-3}\,3^x}{x!}$$

Then:

$$\begin{aligned}\Pr[X \geq 2] &= 1 - \Pr[X < 2] \\ &= 1 - \Pr[X = 0] - \Pr[X = 1] \\ &= 1 - \frac{e^{-3}\,3^0}{0!} - \frac{e^{-3}\,3^1}{1!} \\ &= 1 - 0.0498 - 0.1494 = 0.8008\ .\end{aligned}$$

Example 8C: Show that the function

$$\begin{aligned} p(x) &= \frac{e^{-\lambda}\lambda^x}{x!} && x = 0, 1, 2, \ldots \\ &= 0 && \text{otherwise}\end{aligned}$$

is in fact a probability mass function.
[You need the mathematical result:

$$e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \quad \Big]$$

Example 9C: Beercans are randomly tossed alongside the national road, with an average frequency 3.2 per km.

(a) What is the probability of seeing no beercans over a 5 km stretch?
(b) What is the probability of seeing at least one beercan in 200 m?
(c) Determine the values of x and y in the following statement: "40% of 1 km sections have x or fewer beercans, while 5% have more than y."

For (c), the following information is useful:

x	0	1	2	3	4	5	6	7	8
$p(x) = \dfrac{e^{-3.2}3.2^x}{x!}$	0.0408	0.1304	0.2087	0.2226	0.1781	0.1140	0.0608	0.0278	0.0111
$\Pr[X \le x] = \sum_{t=0}^{x} p(t)$	0.0408	0.1712	0.3799	0.6025	0.7806	0.8946	0.9554	0.9832	0.9943

A more formal derivation of the Poisson distribution ...

We remind ourselves of two mathematical results:

1. $\lim_{n\to\infty}(1-\lambda/n)^n = e^{-\lambda}$
2. $\lim_{n\to\infty} a/n = 0$ for any finite value a.

Suppose events are occurring "at random" with rate λ per time period. Divide the time period into n extremely short intervals of time, each of length $1/n$. These time intervals are regarded as being so short that it is assumed impossible for two or more events to occur in the same interval. With this assumption it is true to say that the probability that an event occurs in the short interval is λ/n, and thus the probability that an event does not occur is $1-\lambda/n$.

We now stand back and look at the problem from a new angle. We think of each interval as a "trial". There are exactly two outcomes of each trial — either an event occurs ("success") or does not occur ("failure"). The probability of success, p, is λ/n. There are n intervals, thus we have n trials. Let X be the number of events that occur in the time period. Clearly, X is a random variable satisfying the conditions for a binomial distribution, $X \sim B(n, \lambda/n)$ and thus

$$\begin{aligned}\Pr[x \text{ events in the time period}] &= \Pr[x \text{ successes in } n \text{ trials}] \\ &= \binom{n}{x}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}\end{aligned}$$

We now let n get so large that the assumption of two or more events in one interval being impossible becomes realistic. Ultimately, we use the two mathematical results above to see what happens when we let n tend to infinity:

$$\begin{aligned}p(x) &= \lim_{n\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \lim_{n\to\infty}\frac{n!}{x!(n-x)!}\,\frac{\lambda^x}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\ &= \frac{\lambda^x}{x!}\lim_{n\to\infty}\frac{n!}{(n-x)!\,n^x}\times\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\times\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} \\ &= \frac{\lambda^x}{x!}\lim_{n\to\infty}\left\{\frac{n}{n}\times\frac{n-1}{n}\times\frac{n-2}{n}\times\ldots\times\frac{n-x+1}{n}\right\}\times e^{-\lambda}\times 1,\end{aligned}$$

using the first of the mathematical results above. A simple re-expression of each term within brackets yields

$$\begin{aligned} p(x) &= \frac{\lambda^x}{x!} \lim_{n\to\infty} \left\{ 1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \ldots \times \left(1 - \frac{x-1}{n}\right) \right\} e^{-\lambda} \\ &= \frac{\lambda^x}{x!} \times 1 \times 1 \times 1 \times \ldots \times 1 \times e^{-\lambda} \end{aligned}$$

using the second of the mathematical results x times. Therefore, we have the result we require,

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!},$$

the probability mass function for the Poisson distribution.

The exponential distribution ...

The exponential distribution arises out of the same underlying scenario as the Poisson distribution, the Poisson process in which events occur "at random" in time or space. For the Poisson distribution, we counted the number of events that occurred in a fixed period of time. For the exponential distribution we consider **the interval of time between events**. We let the random variable X be the time between events. Obviously, X is a **continuous** random variable (it can take on **any** random variable (it can take on **any** value in the sample space $S = \{x | x \geq 0\}$), and must therefore have a probability **density** function. We motivate the formula for the density function in Example 13C.

EXPONENTIAL DISTRIBUTION

If events are occurring at random with average rate λ per unit of time, then the probability density function for the random variable X, the length of time between events is given by

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} && x \geq 0 \\ &= 0 && \text{otherwise} \end{aligned}$$

X is said to have the exponential distribution with parameter λ, and we write $X \sim E(\lambda)$.

The exponential distribution can also be used to model the "distance" between "events" in space, so long as the space is one-dimensional! For example, the exponential distribution can be used for the distance between flaws in cable, but not for the distance between flaws on an A4 page, because the page is two-dimensional!

Example 10A: A computer that operates continuously breaks down at random on average 1.5 times per week.

This tells us $\lambda = 1.5$ **per week**, and that the random variable X, the time between breakdowns, has density function

$$\begin{aligned} f(x) &= 1.5e^{-1.5x} && x \geq 0 \\ &= 0 && \text{otherwise} \end{aligned}$$

What is the probability of no breakdowns for 2 weeks?

This implies that X must be greater than 2 (think through this statement carefully) and that we want $\Pr[X > 2]$. Because the exponential distribution is continuous, we evaluate this probability by integration:

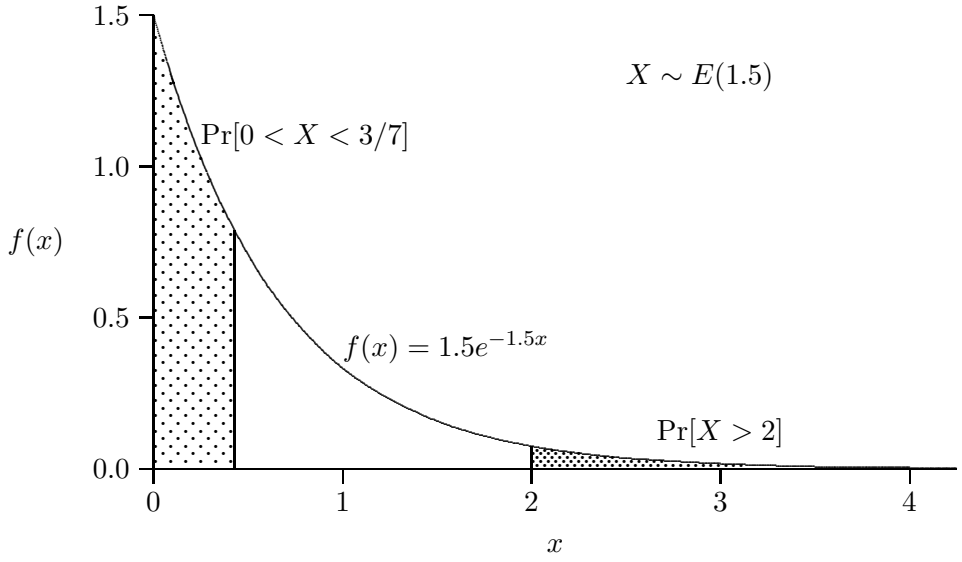
$$\begin{aligned} P[X > 2] &= \int_2^\infty 1.5e^{-1.5x}\,dx = \left[-e^{-1.5x}\right]_2^\infty \\ &= -e^{-\infty} + e^{-3} = 0 + e^{-3} \\ &= 0.0498. \end{aligned}$$

What is the probability of a breakdown within 3 days?

We first make our units of time compatible: 3 days = 3/7 week. We want the probability of a breakdown **before** 3/7 week:

$$\begin{aligned} P[0 < X < 3/7] &= \int_0^{3/7} 1.5e^{-1.5x}dx = \left[-e^{-1.5x}\right]_0^{3/7} \\ &= -e^{-0.6429} + e^{-0} = -0.5258 + 1 \\ &= 0.4742 \end{aligned}$$

These probabilities are depicted in the figure below, which also shows the general shape of the exponential distribution.



Example 11B: Show that the exponential distribution

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} && x \geq 0 \\ &= 0 && \text{otherwise} \end{aligned}$$

is a probability density function.

We check that the three condition for a probability density function are satisfied.

(i) $f(x)$ is defined everywhere. It is non-zero on the interval $[0, \infty)$, i.e. the set $\{x | 0 \leq x < \infty\}$.

(ii) $f(x)$ is never negative. Because λ is a rate, it must be positive, and $e^{-\lambda x}$ is positive.

(iii)

$$\begin{aligned}\int_0^\infty f(x)\,dx &= \int_0^\infty \lambda\, e^{-\lambda x}\,dx = [-e^{-\lambda x}]_0^\infty \\ &= -e^{-\infty} + e^0 = 0 + 1 \\ &= 1,\end{aligned}$$

as required for the area under the curve of a probability density function.

Example 12C: Let the random variable X be the time in hours for which a light bulb burns from the time it is put into service. The probability density function of X is given by

$$\begin{aligned} f(x) &= \tfrac{1}{1000} e^{-\frac{1}{1000}x} && x \geq 0 \\ &= 0 && \text{otherwise}\end{aligned}$$

(a) What is the probability that the bulb burns for between 100 and 1000 hours?
(b) What is the probability that the bulb burns for more than 1000 hours?
(c) What is the probability that the bulb burns for a further 1000 hours, given that it has already burned for 500? (Use conditional probabilities!)

Example 13C: Events occur according to a Poisson process with "intensity" λ (i.e. at rate λ per unit of time).

(a) Use the Poisson distribution to determine the probability of no events in t units of time.
(b) Now use the exponential distribution to determine the probability that the time between events is greater than t.
(c) Compare the answers to (a) and (b) and explain these results.

Example 14C: Flaws occur in telephone cable at the average rate of 4.4 flaws per km of cable. Calculate the following probabilities. (Make use of binomial, Poisson and exponential distributions.)

(a) What is the probability of 1 flaw in 100 m of cable?
(b) What is the probability of more than 3 flaws in 250 m of cable?
(c) What is the probability that the distance between flaws exceeds 500m?
(d) In ten 200 m lengths of cable, what is the probability that 8 or more are free of flaws?

The normal distribution...

The normal distribution is often referred to as the Gaussian distribution, in honour of Carl Friedrich Gauss (1777–1855), a famous German mathematician, who, for more than a century, was credited with its discovery. The same result was published at about the same time by the equally famous French mathematician the Marquis de Laplace (1749–1827). But the normal distribution had actually been discovered nearly a century earlier by Abraham de Moivre (1667–1754). In 1733 he published a mathematical pamphlet that was not widely circulated and was quickly forgotten. A copy of de Moivre's pamphlet was found in 1924, and the English statistician Karl Pearson found that it contained the formula for the normal distribution. De Moivre's precedence in discovering the normal

distribution is contained in a paper published in 1924 by Pearson (**Historical note on the origin of the normal curve of errors**) in the journal *Biometrika* volume 16 pages 402–404, an important statistical journal which is still publishing major discoveries in statistics.

The normal distribution is the most important distribution in statistics. Part of the reason for this is a result called the "**central limit theorem**", which states that if a random variable X is the sum of a large number of random increments, then X has the normal distribution.

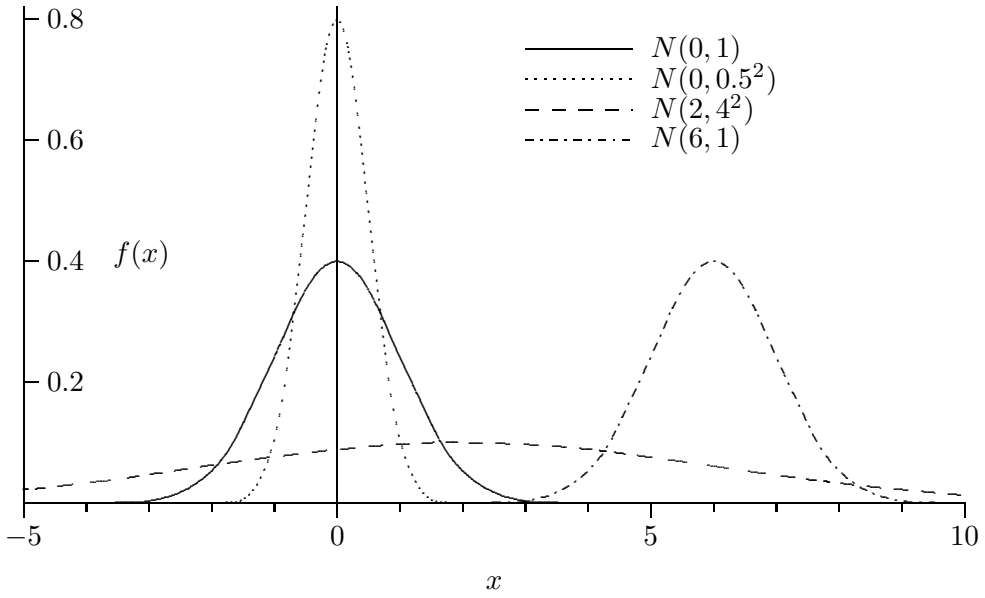
The daily turnover of a large store is the sum of the purchases of all the individual customers. The height of a 50-year old pinetree can be thought of as the sum of each year's growth — which itself is a variable affected by sunshine, temperature, rainfall, etc. So one expects the heights of 50-year old pinetrees to obey a normal distribution. Similarly, an examination mark is the sum of the scores in a large number of questions. Thus, by the central limit theorem, one expects daily turnover, the heights of trees and examination marks (approximately, at least) to be normally distributed.

The normal distribution is continuous, and has probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

There are two parameters, μ ("mu", the Greek letter m for **M**ean) and σ ("sigma", the Greek letter s for **S**tandard deviation).

The constant μ tells us where the graph is located (it can take on any real value); the constant σ (which is always positive) tells how spread out the distribution is. The graphs, depicting $f(x)$ for a few values of μ and σ, make this clear: The most striking feature of the normal distribution is that it is **bell-shaped**. Notice also that the centre of the bell is located at the value μ, and that the distribution gets flatter as σ gets larger. The plot also illustrates the fact that the area under the curve for a probability density function is one; to accommodate this, notice that as the curve gets "flatter", its maximum value has to become smaller.



If X has the normal distribution with parameters μ and σ, we abbreviate this to $X \sim N(\mu, \sigma^2)$, reading this as "the random variable X has the normal distribution with parameters μ and σ^2. When we use this notation, our convention is to write σ^2 for

the second parameter, not plain σ. The parameter σ^2 is known as the variance of the distribution. As in Chapter 1, the variance is the square of the standard deviation.

Unfortunately it is impossible to determine probabilities by integrating the normal probability density function. However (and this makes life very easy), the integration can be done by computer, and we are supplied with a table of probabilities for the normal distribution.

It should come as a surprise to you that a single table is all we need. After all, there are infinitely many combinations of μ and σ, and it seems that we ought to have a massive book of normal tables. We are luckier than we deserve to be, and there is a connecting link between all normal distributions which makes it possible to get away with a single table! We will learn how to use this amazing table by means of an example.

Example 15A: If the amount of margarine, X, in a 250 g tub is normally distributed with $\mu = 251$ and $\sigma = 3$, what is the probability that the tub will contain

(a) between 251 g and 253 g of margarine?
(b) less than 250 g of margarine?

For part (a) we want $\Pr[251 \leq X \leq 253]$

$$= \int_{251}^{253} \frac{1}{\sqrt{2\pi 9}} e^{-\frac{1}{2}(\frac{x-251}{3})^2}\, dx$$

It is impossible to evaluate this integral by finding an indefinite integral and then substituting. So we resort to our tables. We are given tables for only one set of values of the two parameters — this is all we need: when $\mu = 0$ and $\sigma = 1$ we have the **standard normal distribution** which has density function

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2} \qquad -\infty < x < \infty$$

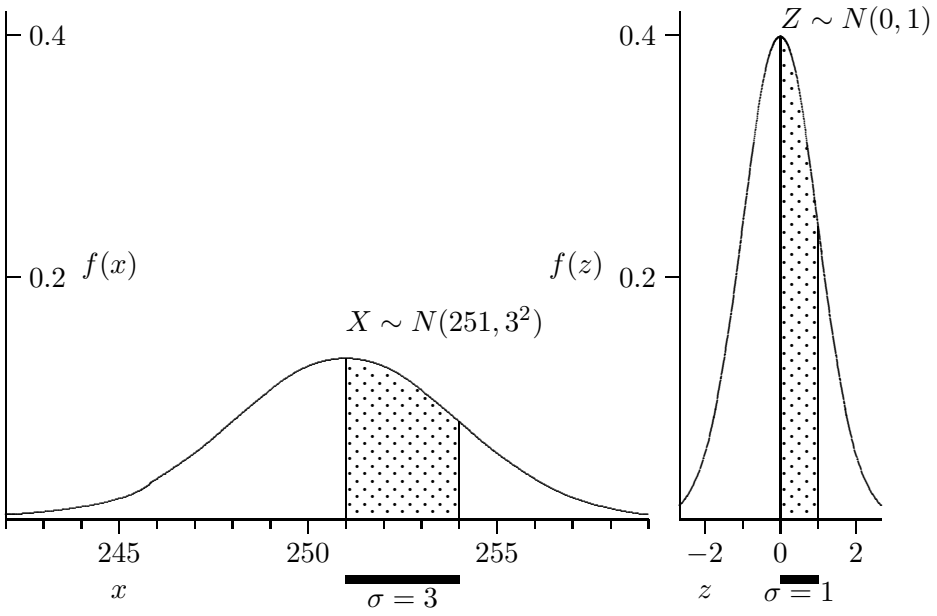
How do we make do with tables for only the standard normal distribution? Because we have an easily proved result that **the proportion of the density function that lies between the mean and a specified number of standard deviations away from the mean is always constant** regardless of the numerical values of the mean and standard deviation.

Translated into mathematical symbols, this important result can be written as

$$\int_{\mu}^{\mu+z\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\, dx = \int_0^z \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}\, dz$$

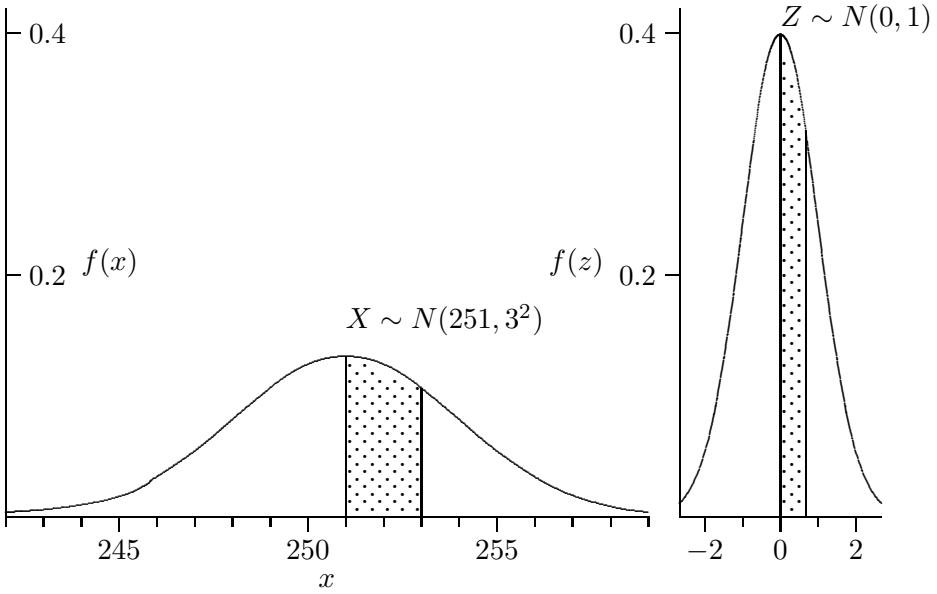
Use integration by substitution to prove this by putting $z = (x - \mu)/\sigma$.

As an example of this, the areas depicted below are equal. The shading, in both cases, shows the area under the curve between the mean and one standard deviation above the mean. Both plots have the same scale on both axes — so you can count the dots for a numerical "proof"!



254 is one standard deviation (i.e. 3 units) above 251, the mean. Thus the area between 251 and 254 in $N(251, 3^2)$ is the same as that between 0 and 1 in $N(0, 1)$.

Returning to part (a) of our margarine example, we need the area between 251 and 253 of $N(251, 3^2)$. 253 is two-thirds of a standard deviation above the mean of 251, because $(253 - 251)/3 = 2/3$. Thus $\Pr[251 < X < 253] = \Pr[0 < Z < 2/3]$, as depicted below:



Some numerical results from the normal tables help to give a "feel" for the normal distribution. The area from one standard deviation below the mean to one standard deviation above the mean is 0.683 (close to 2/3rds); i.e.

$$\Pr[\mu - \sigma < X < \mu + \sigma] = 0.683.$$

The corresponding probabilities for two, three and four standard deviations are:

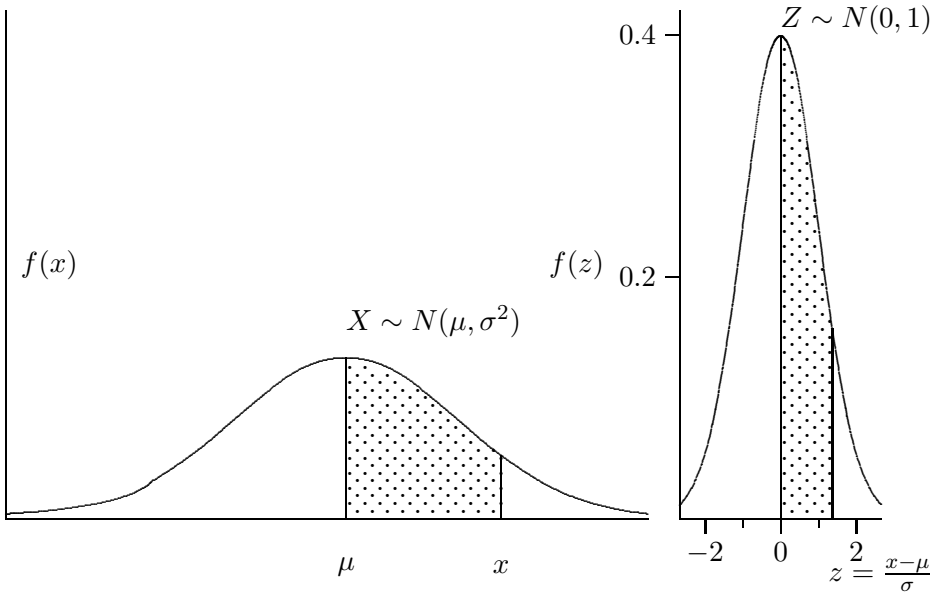
$$\Pr[\mu - 2\sigma < X < \mu + 2\sigma] = 0.954$$

$$\Pr[\mu - 3\sigma < X < \mu + 3\sigma] = 0.997$$

$$\Pr[\mu - 4\sigma < X < \mu + 4\sigma] = 0.999999$$

These results are true for all combinations of μ and σ! In general terms, two-thirds (68%) of a normal distribution is within one standard deviation of its mean, 95% is within two standard deviations, and virtually all of it is within three standard deviations,

The general result is that the area between μ and some point x for $N(\mu, \sigma^2)$ is the same as the area between 0 and $z = \frac{x-\mu}{\sigma}$ for $N(0, 1)$. The formula $z = \frac{x-\mu}{\sigma}$ tells us how many standard deviations the point x is away from the mean μ. Once again, you can count the dots in the plot below of the normal distribution with arbitrary parameters μ and σ and in the standard normal distribution $N(0, 1)$:



In our margarine example, we use $z = (x - \mu)/\sigma$ for $x = 251$ and $x = 253$ to get

$$\Pr[251 < X < 253] = \Pr\left[\frac{251 - 251}{3} < Z < \frac{253 - 251}{3}\right] = \Pr[0 < Z < 0.67]$$

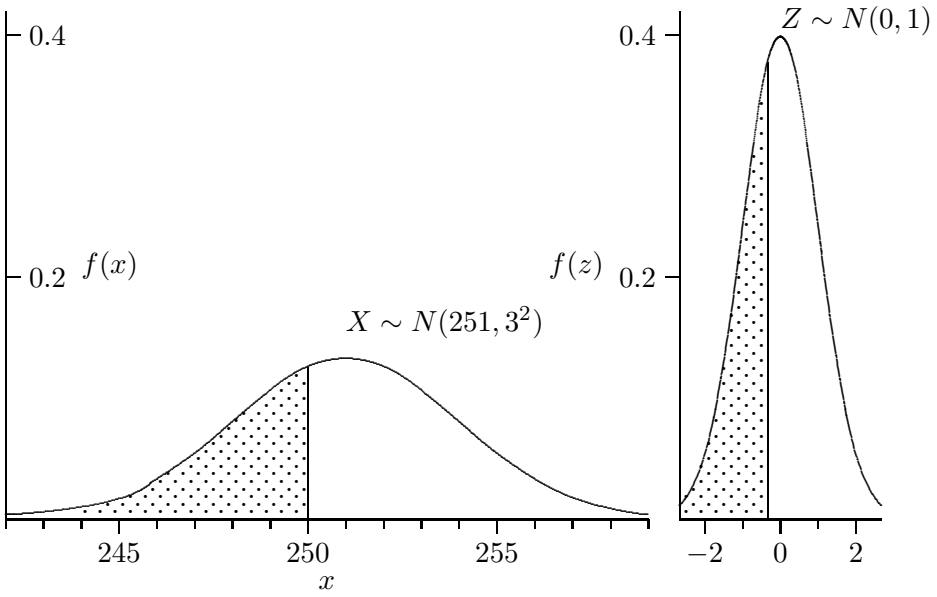
From the table for the standard normal distribution (Table 1) we read off this probability as 0.2486. Thus

$$\Pr[251 < X < 253] = 0.2486,$$

almost a quarter of margarine tubs contain between 251 g and 253 g of margarine.

Part (b) of our question asked for the probability that a tub of margarine was underweight, i.e. the probability that $X < 250$. The area between $-\infty$ and 250 in $N(251, 3^2)$ is the same as the area between $-\infty$ and $(250 - 251)/3 = -1/3$ in $N(0, 1)$:

$$\Pr[X < 250] = \Pr\left[Z < \frac{250 - 251}{3}\right] = P[Z < -1/3].$$



Because our tables give us areas between 0 and a point z, we have to go through the steps depicted in the diagrams below to find this probability. We make use of the facts that the normal distribution is symmetric, and that the area from 0 to ∞ is 0.5.

Alternatively, we can write:

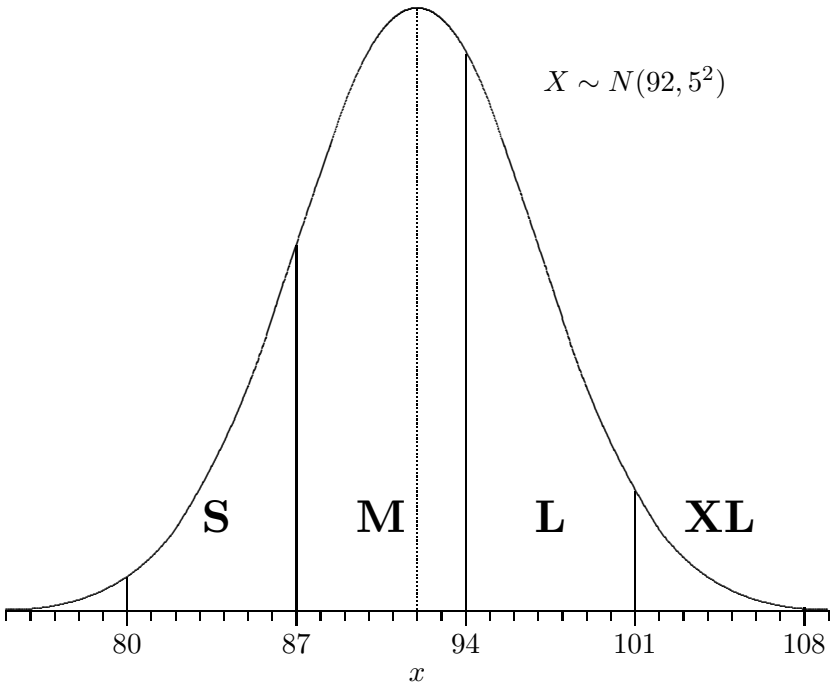
$$\begin{aligned}\Pr[X < 250] &= \Pr[Z < (250 - 251)/3] = \Pr[Z < -1/3] = \Pr[Z > 1/3] \\ &= 0.5 - \Pr[0 < Z < 1/3] = 0.5 - 0.1293 = 0.3707.\end{aligned}$$

The value 0.1293 is looked up in Table 1. Thus 37% of the tubs will contain less margarine than stated. Notice that because the normal distribution is **symmetric** we only need tables for "half" of the distribution.

Example 16B: If $\mu = 4$ and $\sigma = 8$ what is the probability that a normally distributed random variable X lies between 2 and 18?

$$\begin{aligned}\Pr[2 < X < 18] &= \left[\frac{2-4}{8} < \frac{X-\mu}{\sigma} < \frac{18-4}{8}\right] \quad (\text{letting } \; z = (x-\mu)/\sigma) \\ &= \Pr\left[-0.25 < Z < 1.75\right] \\ &= 0.0987 + 0.4599 = 0.5586\end{aligned}$$

Example 17B: A t-shirt manufacturer knows that the chest measurements of his customers are normally distributed with mean 92 cm and standard deviation 5 cm. He makes his t-shirts in four sizes — S (fit size range 80–87 cm), M (to fit 87–94), L (to fit 94–101) and XL (to fit 101–108). What proportion of customers fit into each size t-shirt?



We need to find the z-values for each of the boundary points, by using the formula $z = (x - \mu)/\sigma$.
Then, from our normal tables, we find the area between each of these points and the mean. This gives

x	$z = (x - 92)/5$	Area between x and μ
80	-2.4	0.4918
87	-1.0	0.3413
94	0.4	0.1554
101	1.8	0.4641
108	3.2	0.4993

The proportions for each size are then found by subtraction (or addition in the case of size M), as follows:

Size	Proportion
S	$0.4918 - 0.3413 = 0.1505$ (15.05%)
M	$0.3413 + 0.1554 = 0.4967$ (49.67%)
L	$0.4641 - 0.1554 = 0.3087$ (30.87)%
XL	$0.4993 - 0.4641 = 0.0352$ (3.52%)

Check for yourself that 0.89% of customers don't fit into any size t-shirt.

Example 18C: The mean inside diameter of washers produced by a machine is 0.403 cm and the standard deviation is 0.005 cm.
Washers with an internal diameter less than 0.397 cm or greater than 0.406 cm are considered defective. What percentage of the washers produced are defective, assuming the diameters are normally distributed?

Example 19C: In a large group of men 4% are under 160 cm tall and 52% are between 160 cm and 175 cm tall. Assuming that heights of men are normally distributed, what are the mean and standard deviation of the distribution?

Example 20C: A soft-drink vending machine is set to discharge an average of 215 ml of cooldrink per cup. The amount discharged is normally distributed with standard deviation 10 ml.

(a) If 225 ml cups are used, what proportion of cups overflow?
(b) What is the probability that a cup contains at least 200 mℓ of cooldrink?
(c) What size cups ought to be used if it is desirable that only 2% of cups overflow?

Sums and differences of independent normal random variables...

Suppose we have a number of tasks that have to be completed in sequence e.g. when a building is constructed. Suppose the time taken for each task obeys a normal distribution, each having a given mean and variance and is independent of the time taken for the other tasks. Obviously the **total time taken** will also be a random variable. What will its distribution be and what will its mean and variance be? Without proof, we state that the total time taken will be normally distributed with mean total time equal to the sum of the means for each task, and **variance** equal to the sum of the **variances** (not the standard deviations). Mathematically, we write this as follows. If the time X_i taken for the i th task is such that

$$X_i \sim N(\mu_i, \sigma_i^2)$$

and if it is independent of the time taken for other tasks, then the distribution of the random variable $Y = \sum_{i-1}^{n} X_i$ is

$$Y \sim N(\mu, \sigma^2)$$

where $\mu = \sum_{i=1}^{n} \mu_i$, and $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$.

Sometimes we need to consider the **difference** of two independent normally distributed random variables. Suppose

$$X_1 \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad X_2 \sim N(\mu_2, \sigma_2^2)$$

then, letting $Z = X_1 - X_2$, we state, without proof, that

$$Z \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

The mean of the random variable Z is found by **subtraction**, but the variance is still found by **addition**.

Example 21B: You have 4 chores to perform before getting to Statistics lectures by 08h00. The time (in minutes) to perform each chore is normally distributed with mean and standard deviation as given below:

	mean (μ)	std. dev. (σ)
1. Shower	5	0.5
2. Get dressed	4	1.0
3. Eat breakfast	10	3.5
4. Drive to university	15	5.0

(a) If you get up at 07h20, what is the probability of being late?

(b) (i) At what time should you get up so as to be 99% sure you will not be more than 3 minutes late?

(ii) What is the probability, getting up at this time, that you will be there after 08h10?

(a) The total time taken to get to university is a normally distributed random variable X with mean

$$\mu = 5 + 4 + 10 + 15 = 34 \text{ minutes}$$

and variance

$$\sigma^2 = 0.5^2 + 1.0^2 + 3.5^2 + 5.0^2 = 38.5$$

and therefore standard deviation $\sigma = 6.205$.
The probability that you take more than the allowed 40 minutes is

$$\Pr[X > 40] = \Pr\left[Z > \frac{40 - 34}{6.205}\right] = \Pr[Z > 0.97] = 0.1660$$

On average, you will be late one day in six, because $1/0.1660 \approx 6$.

(b) (i) We must choose x so that $\Pr[X < x] = 0.99$. From tables $\Pr[Z < z] = 0.99$ implies $z = 2.33$. Thus, using the formula $z = (x - \mu)/\sigma$,

$$2.33 = \frac{x - 34}{6.205},$$

which has solution $x = 48.5$ minutes.
48.5 minutes before 08h03 is 07h14.5.

(ii) Probability of taking a total of more than 45.5 minutes is

$$\Pr[X > 45.5] = \Pr\left[Z > \frac{45.5 - 34}{6.205}\right] = \Pr[Z > 1.85] = 0.0322.$$

You'll be late about one day in 31, on average, but (by part (b)) more than three minutes late only once in every 100 days.

Example 22C: Plastic caps seal the ends of the tube into which your degree certificate is placed when you graduate. Suppose the tubes have a mean diameter of 24.0mm and a standard deviation of 0.15mm, and that the plastic caps have a mean diameter of 23.8 mm and a standard deviation of 0.11mm. If the diameter of the cap is 0.10 mm or more larger than that of the tube, the cap cannot be squashed into the tube, and if the diamater of the cap is 0.45 mm or more smaller than that of the tube, it will not seal the tube, but will just keep falling out. If a tube and and plastic cap are selected at random, what are the probabilities of (a) the cap being too large for the tube, and (b) the cap falling out of the tube?

Multiplying a normal random variable by a constant

Suppose that an American textbook says that the heights of students have a normal distribution with mean 67 inches and standard deviation 4 inches. How do we convert this information to a normal distribution with heights in centimetres?

A result, which we do not prove, comes to our aid. It says, if the random variable $X \sim N(\mu, \sigma^2)$, and if a and b are constants, then the random variable $Y = aX + b$ also has a normal distribution, with

$$Y \sim N(a\mu + b, a^2\sigma^2).$$

To solve the inches and centimetres problem, we note that the conversion factor, a, from inches to centimetres is 2.54 (one inch = 2.54 cm) and $b = 0$. So if $X \sim N(67, 16)$, where X is measured in inches, then $Y = 2.54X + 0$ will be in centimetres, and

$$Y \sim N(2.54 \times 67 + 0, 2.54^2 \times 16) = N(170.2, 103.2).$$

Also consider the case where you'd like to convert from degrees Celsius (X) to degrees Fahrenheit (Y). It is known that the the relationship is given by $Y = \frac{9}{5}X + 32$. So in general, if $X \sim N(\mu, \sigma^2)$ then $Y \sim N(\frac{9}{5}\mu + 32, \frac{81}{25}\sigma^2)$. If it was know that the temperature in degrees Fahrenheit was normally distributed, what would the corresponding distribution in degrees Celsius be?

Example 23C: Another textbook says that the mass of an Ostrich is normally distributed with mean 68745 g and variance 13201000g^2.

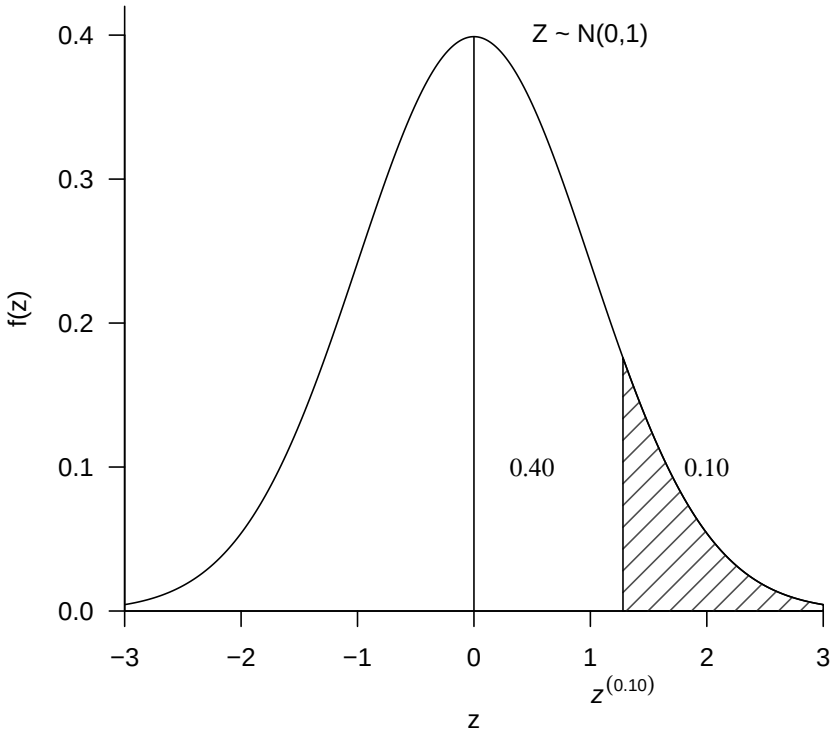
(a) Convert this information to a random variable with mass measured, more sensibly, in kilograms.
(b) What is the probability that three ostriches weigh more than 225 kg?

Percentage points of the standard normal distribution

Often, instead of wanting to find $\Pr[Z > z]$ for some given value of z, we are given a probability p and need to find the value of z that makes the equation $\Pr[Z > z] = p$ true. It is convenient to use the notation $z^{(p)}$ to denote the value of z which provides the solution to this equation, and to describe it as the $100p\%$ point of the distribution.

Example 24A: Find $z^{(0.10)}$, the 10% point of the standard normal distribution.

In other words, we are asked to find the point along the standard normal distribution such that 10% of the distribution lies to the right of it:



Remember that Table 1 is constructed to give probabilities between 0 and z. Therefore, to find $z^{(0.10)}$, we search in the body of Table 1 until we find the closest value we can to 0.40. We find 0.3997 when $z = 1.28$. Thus $z^{(0.10)} = 1.28$; we say tat 1.28 is the **10% point of the standard normal distribution**. Sometimes, we need to be more precise, and say that 1.28 is the upper 10% point of the standard normal distribution. Clearly, because of the symmetry of the standard normal distribution, -1.28 is the lower 10% point of the standard normal distribution. The **lower** 10% point is also, in as perverted way, the **upper** 90% point, so that we can even write $z^{(0.90)} = -1.28$!

Example 25C: Find (a) $z^{(0.05)}$, (b) $z^{(0.025)}$, (c) $z^{(0.01)}$, (d) $z^{(0.005)}$, (e) $z^{(0.25)}$, (f) $z^{(0.5)}$, (g) $z^{(0.95)}$ and (h) $z^{(0.99)}$.

Solutions to examples

5C (a) $1 - \binom{10}{0}q^{10}$ (i) 0.0956 (ii) 0.6513 (iii) 0.9718
(b) $1 - \binom{20}{0}q^{20} - \binom{20}{1}pq^{19}$
(i) 0.0169 (ii) 0.6083 (iii) 0.9924
(c) The second procedure is less likely to reject relatively satisfactory consignments, and is more likely to reject very poor consignments. However, it costs twice as much to do the checking, so there is a trade-off.

9C (a) 0.1125×10^{-6} (b) 0.4727 (c) $x = 2 \quad y = 6$

12C (a) 0.5369 (b) 0.3679 (c) 0.3679

13C (a) and (b) $e^{-\lambda t}$ (c) The events are the same.

14C (a) 0.2834 (Poisson) (b) 0.0257 (Poisson) (c) 0.1108 (Exponential) (d) 0.0158 (Binomial)

18C 38.94%

19C $\mu = 173.83,\ \sigma = 7.895$

20C (a) 0.1587 (b) 0.9332 (c) 235.5 ml (The machine is pretty useless!)

22C (a) 0.0537 (b) 0.0901

23C (a) The conversion factor is to divide by 1000. $Y \sim N(68.745, 13.201)$, where Y is measured in kilograms. (b) If one ostrich weighs Y kg, then three weigh $V = Y_1 + Y_2 + Y_3$ kg, and $V \sim N(3 \times 68.745, 3 \times 13.201) = N(206.235, 39.603)$. $\Pr[V > 225] = \Pr[Z > 2.98] = 0.00144$.

25C (a) $z^{(0.05)} = 1.64$, (b) $z^{(0.025)} = 1.96$, (c) $z^{(0.01)} = 2.33$, (d) $z^{(0.005)} = 2.58$, (e) $z^{(0.25)} = 0.67$, (f) $z^{(0.5)} = 0$, (g) $z^{(0.95)} = -1.64$ and (h) $z^{(0.99)} = -2.33$.

Exercises on the binomial distribution ...

*5.1 Suppose that 25% of the people entering a supermarket are aged between 18 and 30 years, classified as young adults. A market researcher has to fulfil a quota of 10 interviews. What is the probability her quota of interviews contains

(a) exactly x young adults?
(b) no young adults?
(c) between four and six (inclusive) young adults?

*5.2 (a) A true-false test is given with five questions. To pass you need at least four right. You guess each answer. What is the probability that you pass?
(b) The true-false test is replaced with a multiple-choice test with four alternative answers, only one of which is correct. If you guess, what is the probability that you pass?

*5.3 A shopper has a choice between a supermarket and a hypermarket. She chooses the supermarket 60% of the time, and the hypermarket 40% of the time. What are the probabilities that, on her next seven shopping trips,

(a) she shops at the hypermarket three times?
(b) she shops at least twice at the supermarket?
(c) she shops at only one of the stores?

*5.4 An anti-aircraft battery in England during World War II had on the average 3 out of 10 successes in shooting down flying bombs that came within range. What was the chance that, if eight bombs came within range, two or more were shot?

Poisson distribution...

*5.5 What is the probability of finding 12 errors in a 200-page book if the printers have an error rate of 0.075 errors per page? Assume that a Poisson distribution may be used to model the occurrence of errors.

*5.6 A pump fails, on the average, once in every 500 hours of operation.

(a) Find the probability that the pump has more than one failure during a 500-hour period.
(b) What is the probability of exactly 3 failures in 2000 hours of operation?

*5.7 The average demand on a factory store for a certain electric motor is 8 per week. When the storeman places an order for these motors, delivery takes one week. If the demand for motors has a Poisson distribution, how low can the storeman allow his stock to fall before ordering a new supply if he wants to be at least 95% sure of meeting all requirements while waiting for his new supply to arrive?

5.8 The average number of accidental drownings per year is 3.5 per 100 000 population.

(a) Find the probability that in a city with a population of 200 000 there will be between 4 and 8 (inclusive) accidental drownings per year.
(b) What are the probabilities that in towns of 15 000, 20 000 and 50 000 there will be no drownings in a year?

Exponential distribution...

*5.9 If the average drowning rate is 3.5 per 100 000 population per year what is the probability that the time interval between drownings in a city of 200 000 will be less than one month?

*5.10 Customers arrive at a restaurant at the rate of 90 per hour during lunch time.

(a) If a customer has just arrived, what is the probability that it will be at least another minute before the next customer arrives?
(b) If a minute has already passed by since the last customer arrived, what is the probability that it is at least another minute before the next customer arrives?
(Hint: use conditional probabilities.)

*5.11 The life of an electronic device is known to have the exponential distribution with parameter $\lambda = 1/1000$.

(a) What is the probability that the device lasts less than 1000 hours?
(b) What is the probability it will last more than 1200 hours?
(c) If three such devices are taken at random, what is the probability that one will last less than 800 hours, another between 500 and 1200 hours, and the third between 1200 and 2000 hours?

*5.12 The duration (in minutes) of showers on a tropical island is approximately exponentially distributed with $\lambda = 1/5$.

(a) Out of 3 showers, what is the probability that not more than 2 will last for 10 minutes or more?
(b) What is the probability that a shower will last at least 2 minutes more, given that it has already lasted 5 minutes?

NORMAL DISTRIBUTION...

5.13 If the random variable Z has the standard normal distribution (i.e. $Z \sim N(0;1)$ find the following probabilities:

(a)	$P[0 < Z < 1]$	(b)	$P[0 < Z < 1.96]$
(c)	$P[-1.64 < Z < 1.64]$	(d)	$P[Z \geq 2]$
(e)	$P[Z < -1.38]$	(f)	$P[Z < 2.1]$
(g)	$P[-2.3 < Z < 1.6]$	(h)	$P[1 < Z < 2]$
(i)	$P[-1.74 < Z < -0.86]$	(j)	$P[-3 \leq Z \leq 3]$

5.14 Given that Z has standard normal distribution, what values must z^* have in order to make each of the following statements true?

(a) $P[0 < Z < z^*] = 0.475$
(b) $P[-z^* < Z < z^*] = 0.95$
(c) $P[Z < z^*] = 0.05$
(d) $P[Z > z^*] = 0.005$
(e) $P[Z < z^*] = 0.99$

*5.15 If X is distributed as a normal variable with mean 3 and variance 4, find

(a) $P[X < 4]$
(b) $P[|X| < 6]$
(c) $P[3.5 < X < 6.5]$

*5.16 If $X \sim N(0; \frac{1}{4})$ find

(a) $P[X > 2]$
(b) $P[0 < X < 1]$

5.17 If $X \sim N(1;4)$ find numbers x_0 and x_1 such that

(a) $P[X > x_0] = 0.10$
(b) $P[X > x_1] = 0.80$

*5.18 If X has a normal distribution, and if $P[X < 10] = 0.8413$, what is the value of the mean if the distribution is known to have variance $\sigma^2 = 16$?

*5.19 Sports shirts are frequently classified as S, M, L and XL for small, medium, large and extra large neck sizes. S fits a neck circumference of less than 37 cm, M fits between 37 and 40.5 cm and L fits between 40.5 and 44 cm while XL fits necks over 44 cm in circumference. The neck circumference of adult males has a normal distribution with $\mu = 40$ and $\sigma = 2$.

(a) What proportion of shirts should be manufactured in each category?
(b) If you wanted to define categories S, M, L, XL so that each category contained 25% of the total population of adult males, what neck sizes must you assign to each of these categories.

*5.20 Suppose that the profit (or loss) per day of a shopkeeper dealing in a perishable item is approximately normally distributed with mean R10 and standard deviation R5. What is the probability that

(a) he makes a loss on any one day?
(b) his profit exceeds R14?
(c) he makes exactly R10 profit?

5.21 Consider an I.Q. test for which the scores of adult Americans are known to have a normal distribution with expected value 100 and variance 324, and a second I.Q. test for which the scores of adult Americans are known to have a normal distribution with expected value 50 and variance 100. Under the assumption that both tests measure the same phenomenon ("intelligence"), what score on the second test is comparable to a score of 127 on the first test? Explain your answer.

FURTHER EXERCISES, WITH THE DISTRIBUTIONS MIXED UP!...

5.22 Which of the four probability distributions we have considered might serve as models for

(a) the intervals of time between breakdowns of a computer?
(b) the number of times a total of 6 occurs when 2 dice are thrown 5 times?
(c) the precise masses of packets of 36 biscuits?
(d) the number of telephone calls received each day by a telephone counselling service?

*5.23 The average number of oil tankers arriving each day at a Persian Gulf port is known to be 7. The facilities at the port can handle at most 10 tankers per day. If tankers arrive at random, what is the probability that on a given day tankers have to be turned away?

*5.24 Airplane engines operate independently in flight and fail with probability 1/10. A plane makes a successful flight if at most half of its engines fail. Determine the probability of a successful flight for two-engined and four-engined planes.

*5.25 An ice-cream vendor's sales follow a Poisson distribution with an average rate of 10 per hour.

(a) What is the probability that he sells at least one ice-cream in his first hour of operation?
(b) How much should his stock of ice-cream be at any point in time if he wants to be at least 95% sure that he does not run out of ice-cream in the following hour?
(c) Given that he has made no sale during the past 15 minutes, what is the probability that he makes a sale within the next 20 minutes?

5.26 A die is thrown 10 times. What is the probability of obtaining at least three even numbers?

*5.27 The annual income of residents of Bishopscourt is normally distributed with mean R25 000 and standard deviation R5000. What is the highest income of the lowest 20% of income earners in Bishopscourt?

5.28 A liquid culture medium contains on the average m bacteria per ml. A large number of samples is taken, each of 1 ml, and bacteria are found to be present in 90% of the samples. Estimate m.

5.29 The strength of a plastic produced by a certain process is known to be normally distributed. If 10% of the plastic has a strength of at least 4000 kg, and 70% has a strength exceeding 3000 kg, what are the mean and standard deviation of the distribution?

*5.30 A bank has 175 000 credit card holders. During one month the average amount spent by each card holder totalled R192,50 with a standard deviation of R60,20. Assuming a normal distribution, determine the number of card holders who spent more than R250.

*5.31 The maximum (stated) load of a passenger lift is 8 passengers or 600 kg. If the masses of people using the lift can be considered to be normally distributed with mean 70 kg and standard deviation 15 kg, how often will the combined mass of 8 passengers exceed the 600 kg limit?

5.32 A road is constructed so that the right-turn lane at an intersection has a capacity of 3 cars. Suppose that 30% of cars approaching the intersection want to turn right. If a string of 15 cars approaches the intersection, what is the probability that the lane will be insufficiently large to hold all the cars wanting to turn right?

*5.33 The weekly demand for sulphuric acid from the store of a chemical factory is normally distributed with mean 246 litres and standard deviation 50 litres. After placing an order with the sulphuric acid manufacturers, delivery to the store takes one week.

(a) How low can the stock of sulphuric acid be allowed to fall before ordering a new supply in order to be 95% sure of meeting all requirements while waiting for the new supply to arrive?

(b) What volume of sulphuric acid should then be ordered so that, with 95% certainty, it will not be necessary to reorder within 6 weeks?

THE KEEN MAY LIKE TO TRY THESE EXERCISES...

5.34 Suppose the number of eggs a bird lays in its nest has a Poisson distribution with parameter λ. Suppose each egg hatches with probability p. Show that the number of eggs that hatch in a nest has a Poisson distribution with parameter λp.

5.35 (a) Suppose we have n trials of a random experiment with three possible outcomes which have probabilities p_1, p_2 and p_3, $(p_1 + p_2 + p_3 = 1)$. Show that the probability that x_1 of the n trials have the first outcome, x_2 the second outcome, x_3 the third outcome $(x_1 + x_2 + x_3 = n)$ is given by the **trinomial distribution**

$$p(x_1, x_2, x_3) = \binom{n}{x_1\ x_2\ x_3} p_1^{x_1}\ p_2^{x_2}\ p_3^{x_3}$$

(b) Extend this result to the **multinomial distribution:** there are now m possible outcomes with probabilities p_1, $p_2, \ldots, p_m$, $(\sum_{i=1}^{m} p_i = 1)$, and the

probability that x_1 trials have the first outcome, x_2 the second, ..., x_m the m th outcome ($\sum_{i=1}^{m} x_i = n$), is given by

$$\binom{n}{x_1\ x_2\ \ldots\ x_m} p_1^{x_1}\ p_2^{x_2} \ldots p_m^{x_m}$$

Solutions to exercises...

5.1 (a) $\binom{10}{c}\ 0.25^x\ 0.75^{10-x}$ (b) 0.0563 (c) 0.2206

5.2 (a) 0.1875 (b) 0.0156

5.3 (a) 0.2903 (b) 0.9812 (c) 0.0296

5.4 0.7447

5.5 0.0829

5.6 (a) 0.2642 (b) 0.1954

5.7 13

5.8 (a) 0.6473 (b) 0.5916, 0.4966 and 0.1738

5.9 0.442

5.10 (a) 0.2231 (b) 0.2231

5.11 (a) 0.632 (b) 0.301 (c) $0.551 \times 0.305 \times 0.166 = 0.0279$

5.12 (a) Pr(a shower lasts longer than 10 mins) = 0.135, $\Pr(X \leq 2) = 0.9975$
(b) 0.6703

5.13 (a) 0.3414 (b) 0.4750 (c) 0.8990 (d) 0.0228
(e) 0.0838 (f) 0.09821 (g) 0.9345 (h) 0.1359
(i) 0.1540 (j) 0.99730

5.14 (a) 1.964 (b) 1.96 (c) -1.64 (d) 2.58 (e) 2.33

5.15 (a) 0.6915 (b) 0.9332 (c) 0.3612

5.16 (a) 0.0000 (b) 0.4773

5.17 (a) 3.56 (b) -0.68

5.18 $\mu = 6$

5.19 (a) 7% S, 53% M, 38% L, 2% XL (b) S $<$ 38.65 $<$ M $<$ 40 $<$ L $<$ 41.35 $<$ XL

5.20 (a) 0.0228 (b) 0.2119 (c) 0

5.21 65

5.22 (a) Exponential (b) Binomial (c) Normal (d) Poisson

5.23 0.0985

5.24 0.9900 and 0.9963 for 2- and 4-engined planes, respectively

5.25 (a) 0.999995 (b) 15 (c) 0.964

5.26 0.9453

5.27 R20 800

5.28 $m = 2.3026$

5.29 $\mu = 3288.89,\ \sigma = 555.56$

5.30 29 488

5.31 0.1736

5.32 0.7031

5.33 (a) 328.0 litres (b) 1676.9 litres

Chapter 6

MORE ABOUT RANDOM VARIABLES

KEYWORDS: Mean, variance, standard deviation, and coefficient of variation of a random variable; the distribution function and the median of a random variable; the normal approximations to the binomial and Poisson distributions.

Mean and variance of a random variable...

In chapter 1 we learnt about measures of location and spread for samples of data. We now develop equivalent concepts for random variables. The most important measures of location and spread for random variables are given the same names, the **mean** and the **variance**, as were used for samples of data. However, the formulae defining the **mean of a random variable** and the **variance of a random variable** are completely different from the formula which defined the **mean of a sample**, which we denoted $\bar{x}$, and the **variance of a sample**, denoted s^2. The justification for using the names **mean** and **variance** in both contexts is fairly subtle and will be made clear in chapter 8.

As you might expect, discrete and continuous random variables are handled separately, but the notation is the same in both cases. The **mean of a random variable** X is denoted by μ or $\mathrm{E}[X]$ ("the expected value of X") and the **variance of a random variable** X is denoted by σ^2 or $\mathrm{Var}[X]$ ("the variance of X").

A discrete random variable X has a probability mass function $p(x)$; the values of $p(x)$ are non-zero for a countable set of x-values. For a discrete random variable X, the mean is defined to be

$$\mu = \mathrm{E}[X] = \sum_x x\, p(x).$$

In words, this says that the mean of a discrete random variable is equal to the sum of a set of terms, each term being one of the values that the random variable can take on (x), multiplied by the probability of taking that value ($p(x)$). The variance makes immediate use of μ as defined above. The variance of a discrete random variable is defined to be

$$\sigma^2 = \mathrm{Var}[X] = \sum_x (x-\mu)^2\, p(x).$$

The sum is taken over the set of values for which the probability mass function $p(x)$ is positive.

A continuous random variable X has a probability density function $f(x)$; usually, the values of $f(x)$ are non-zero on some interval (a, b). For a continuous random variable X, the mean is defined to be

$$\mu = \mathrm{E}[X] = \int_a^b x\, f(x)\, dx$$

and the variance is

$$\sigma^2 = \mathrm{Var}[X] = \int_a^b (x-\mu)^2\, f(x)\, dx.$$

The limits of integration are taken over the interval for which the probability density function $f(x)$ is non-zero.

As was the case with the variance of a sample in chapter 1, there is an alternative formula for the variance of a random variable, which also provides a "short-cut" for many problems. If X is discrete, use

$$\sigma^2 = \mathrm{Var}[X] = \big(\sum_x x^2\, p(x)\big) - \mu^2.$$

For X continuous, use

$$\sigma^2 = \mathrm{Var}[X] = \int_a^b x^2\, f(x)\, dx - \mu^2.$$

You should easily be able to prove that the pairs of formulae for the variance of discrete and continuous random variables are equivalent.

For both discrete and continuous random variables, the **standard deviation of the random variable** X is defined to be the square root of the variance:

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

Note: Let X be a random variable with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Also define the random variable Y as a linear function of X, $Y = aX + b$ where a and b are constants. Then the mean and variance of Y is given by:

$$\mathrm{E}[Y] = a\mathrm{E}[X] + b = a\mu + b$$

$$\mathrm{Var}(Y) = a^2\mathrm{Var}(X) = a^2\sigma^2$$

The **coefficient of variation of the random variable** X is defined to be the ratio of the standard deviation and the mean:

$$\mathrm{CV} = \frac{\sqrt{\mathrm{Var(X)}}}{\mathrm{E(X)}} = \sigma/\mu.$$

The coefficient of variation has uses in sampling theory. It is frequently multiplied by 100 and then expressed as a percentage. The coefficient of variation is only sensibly defined if the lower limit of the random variable X is zero.

Example 1A: Suppose the random variable X has probability density function

$$\begin{aligned} f(x) &= 6x(1-x) && 0 \le x \le 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

Find (a) the mean, (b) the variance and (c) the coefficient of variation of the random variable X.

(a) To find the mean, we use the definition for a continuous random variable:

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x\, f(x)\, dx \\ &= \int_0^1 x \times 6x(1-x)\, dx = 6\int_0^1 (x^2 - x^3)\, dx \\ &= 6\left[\frac{1}{3}\, x^3 - \frac{1}{4}\, x^4\right]_0^1 = 6\left(\frac{1}{3} - \frac{1}{4})\right) = \frac{1}{2} \end{aligned}$$

(b) We use the alternative formula for finding the variance:

$$\begin{aligned} \mathrm{Var}[X] &= \int_{-\infty}^{\infty} x^2\, f(x)\, dx - \mu^2 \\ &= \int_0^1 x^2 \times 6x(1-x)\, dx - \left(\frac{1}{2}\right)^2 \\ &= 6\int_0^1 (x^3 - x^4)\, dx - \frac{1}{4} \\ &= 6\left[\frac{1}{4}x^4 - \frac{1}{5}x^5\right]_0^1 - \frac{1}{4} = 6\left(\frac{1}{4} - \frac{1}{5}\right) - \frac{1}{4} = \frac{1}{20} \end{aligned}$$

(c) The coefficient of variation is given by

$$\mathrm{CV} = \frac{\sqrt{\mathrm{Var}[X]}}{\mathrm{E}[X]} = \frac{\sqrt{1/20}}{\frac{1}{2}} = 0.4472$$

Example 2A: The discrete random variable X has probability mass function given by

$$\begin{aligned} p(x) &= 1/6 && x = 1,\ 2,\ 3,\ 4,\ 5,\ 6 \\ &= 0 && \text{otherwise} \end{aligned}$$

Find (a) the mean, (b) the variance and (c) the coefficient of variation of the random variable.

(a) We use the definition of the mean of a discrete random variable:

$$\begin{aligned} \mathrm{E}(X) &= \sum_x x\ p(x) \\ &= \sum_{x=1}^{6} x \times 1/6 \\ &= 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 \\ &= 3.5 \end{aligned}$$

(b) We use the alternative formula for finding the variance of a discrete random variable:

$$\begin{aligned}\text{Var}(X) &= \big(\sum_x x^2\, p(x)\big) - \mu^2 \\ &= \big(1/6 + 4/6 + 9/6 + 16/6 + 25/6 + 36/6\big) - 3.5^2 \\ &= 2.917\end{aligned}$$

(c) Coefficient of variation is

$$\text{CV} = \frac{\sqrt{\text{Var(X)}}}{\text{E(X)}} = \frac{\sqrt{2.917}}{3.5} = 0.488$$

In the next two examples, it will help you to be reminded that the mean is the sum of the values that the random variable takes on multiplied by the probabilities of taking on these values.

Example 3C: You are contemplating whether it is worth your while driving out to the Northern Suburbs to call on a client. You estimate that there is a 40% chance that the client will purchase your product. Commission on the sale is R50, but the petrol will cost you R15.

(a) Should you call on the client?
(b) What probability of purchase would lead to an expected net gain of zero for the call?

Example 4C: You are considering investing in one of two shares listed on the stock exchange. You estimate that the probability is 0.3 that share A will decline by 15% and a probability of 0.7 that it will rise by 30%. Correspondingly, for share B, you estimate that the probability is 0.4 that it will decline by 15% and that the probability that it will rise by 30% is 0.6. The "return" on a share is defined as the percentage price change.

(a) Calculate the expected return for each share.
(b) Calculate the standard deviations of the returns for each share.
(c) Which of the two shares would you say is the more "risky"? Why?
(d) Calculate the coefficient of variation for each share.
(e) Which of the two shares would you buy? Why?

Example 5B: Suppose the random variable X has the Poisson distribution with parameter λ. Find E[X].

The probability mass function for the Poisson distribution is

$$\begin{aligned} p(x) &= \frac{e^{-\lambda}\lambda^x}{x!} && x = 0,\ 1,\ 2,\ \ldots \\ &= 0 && \text{otherwise}\end{aligned}$$

Using the definition of the mean of a discrete random variable, we have

$$\begin{aligned}
\mathrm{E}[X] &= \sum x \times p(x) \\
&= \sum_{x=0}^{\infty} x \times \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} \frac{x}{x} \times \frac{e^{-\lambda}\lambda^x}{(x-1)!} \qquad \text{because } x! = x \times (x-1)! \\
&= \lambda\, e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda\, e^{-\lambda} \left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots\right) \\
&= \lambda\, e^{-\lambda} \times e^{\lambda} \\
&= \lambda
\end{aligned}$$

The mean of a Poisson distribution is its parameter, λ.

Example 6B: Suppose $X \sim B(n, p)$. Find the expected value of X.

The probability mass function for the binomial distribution is given by

$$\begin{array}{lll} p(x) & = \binom{n}{x} p^x\, q^{n-x} & x = 0,\ 1,\ \ldots n \\ & = 0 & \text{otherwise} \end{array}$$

Once again, we use the definition of the mean of a discrete random variable:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{x=0}^{n} x \times \binom{n}{x} p^x\, q^{n-x} \\
&= \sum_{x=1}^{n} x \times \binom{n}{x} p^x\, q^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!}\, p^x\, q^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!}\, p^{x-1}\, q^{n-x}.
\end{aligned}$$

Now substitute $y = x - 1$, and substitute $m = n - 1$. With this substitution, note that when $x = 1$, $y = 0$, and that when $x = n$, $y = n - 1$, which in terms of the second substitution is $y = m$. We use these to adjust the lower and upper limits of the summation.

$$\begin{aligned}
\mathrm{E}[X] &= np \sum_{y=0}^{m} \frac{m!}{y!(m-y)!}\, p^y\, q^{m-y} \\
&= np \sum_{y=0}^{m} \binom{m}{y}\, p^y\, q^{m-y} \\
&= np \times 1,
\end{aligned}$$

because

$$\sum_{y=0}^{m} \binom{m}{y}\, p^y\, q^{m-y} = 1;$$

it is the sum of all possible values of the probability mass function of the binomial distribution $B(m, p)$. Therefore, as required

$$\mathrm{E}[X] = np,$$

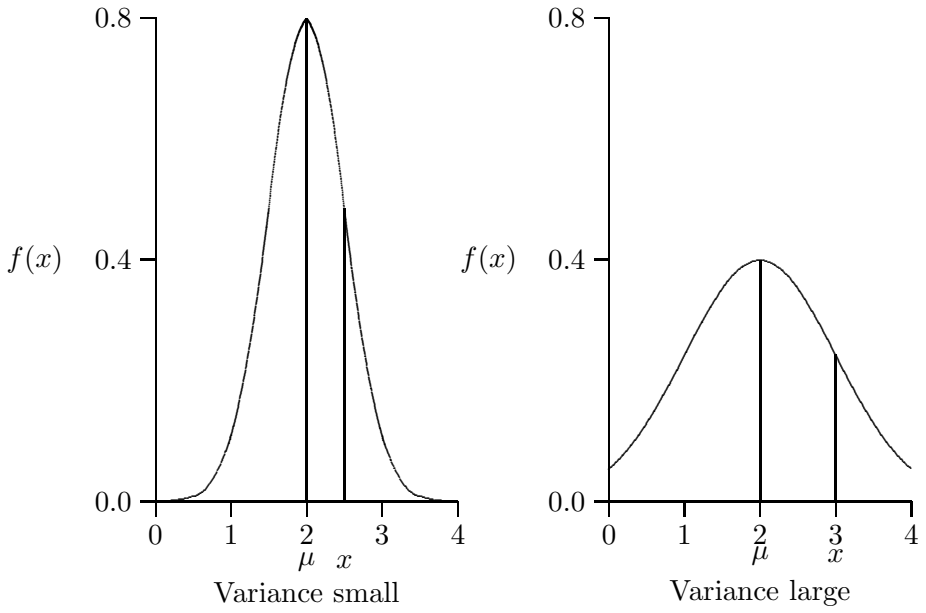
which says that the mean of the binomial distribution $B(n, p)$ is the product of the two parameters of the distribution, n and p.

A geometrical interpretation of the mean and the variance...

The formulae for finding the mean are mathematically equivalent to performing the following operations:

$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx$ is equivalent to cutting the shape of the graph of $f(x)$ out of a piece of tin or cardboard of uniform thickness and finding the point along the x-axis on which it balances.

$\mathrm{E}[X] = \sum_x x\, p(x)$ is equivalent to hanging masses corresponding to $p(x)$ at the point x along a ruler, and finding the point at which the ruler balances.



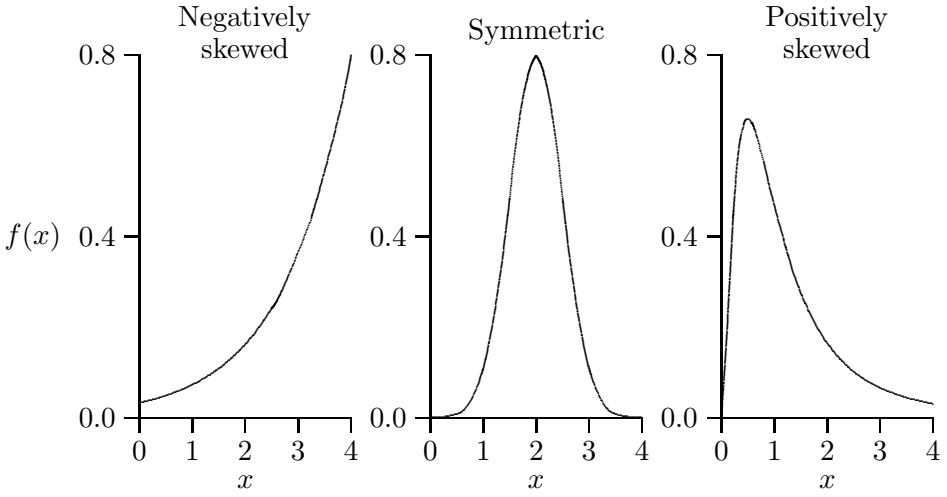
The variance may be thought of as a measure of the average distance of the random variable X from its mean. If the p.d.f. or p.m.f. is very "flat", the variance will be large. If the probability function is very "peaked", the variance will be small. This is illustrated above. In the plot on the left, $f(x)$ is peaked, and the terms $(x - \mu)^2$ are on average small, leading to a small variance. On the other hand, if as in the plot on the right, $f(x)$ is flat, then the terms $(x - \mu)^2$ tend to be large, leading to a large variance.

Applied Mathematics students will see the relationship between "means" and "centres of gravity" and between "variances" and "moments of inertia".

Skewness ...

Just as a histogram can be **skew**, so can the distribution of a random variable. We use the same terminology here as in chapter 1. A random variable is said to be **positively skewed** if it has a long tail on the right-hand side. Similarly, a random variable has **negative skewness** if it has a long tail on the left-hand side. **Symmetric distributions** are just that: symmetric, so that the tail on the left is a mirror image of the tail on the right.

Statisticians sometimes need to describe the shape of the tails of a probability distribution. Even if two distributions may have the same mean and variance, the shapes of the tails may be quite different. Statisticians distinguish between **heavy-tailed** distributions, in which the probability of observations far from the mean is relatively large, and **light-tailed** distributions, in which observations far from the mean are unlikely.



The distribution function $F(x)$ ...

Let X be a random variable with probability density function $f(x)$ or probability mass function $p(x)$. The function that gives the probability that X takes on a value less than or equal to x is called the **distribution function** and is denoted by $F(x)$:

$$F(x) = \Pr[X \leq x].$$

If X is continuous,

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

and if X is discrete

$$F(x) = \sum_{t \leq x} p(t).$$

Note that for X continuous, $F'(x) = f(x)$, i.e. the derivative of the distribution function is the probability density function.

Example 7A: Find the distribution function $F(x)$ for the exponential distribution.

$$\begin{aligned} F(x) = \Pr[X \le x] &= \int_{-\infty}^{x} f(x)\ dx = \int_{0}^{x} \lambda e^{-\lambda x}\ dx \\ &= \left[-e^{-\lambda x}\right]_0^x = -e^{-\lambda x} + e^0 \\ &= 1 - e^{-\lambda x} \end{aligned}$$

The distribution function should be defined for the domain $(-\infty,\ \infty)$: thus we write

$$\begin{aligned} F(x) &= 0 & x < 0 \\ &= 1 - e^{-\lambda x} & x \ge 0 \end{aligned}$$

as the distribution function for the exponential distribution. Differentiating $F(x)$ yields

$$\begin{aligned} \frac{dF(x)}{dx} = f(x) &= 0 & x < 0 \\ &= \lambda e^{-\lambda x} & x \ge 0 \end{aligned}$$

the probability density function of the exponential distribution.

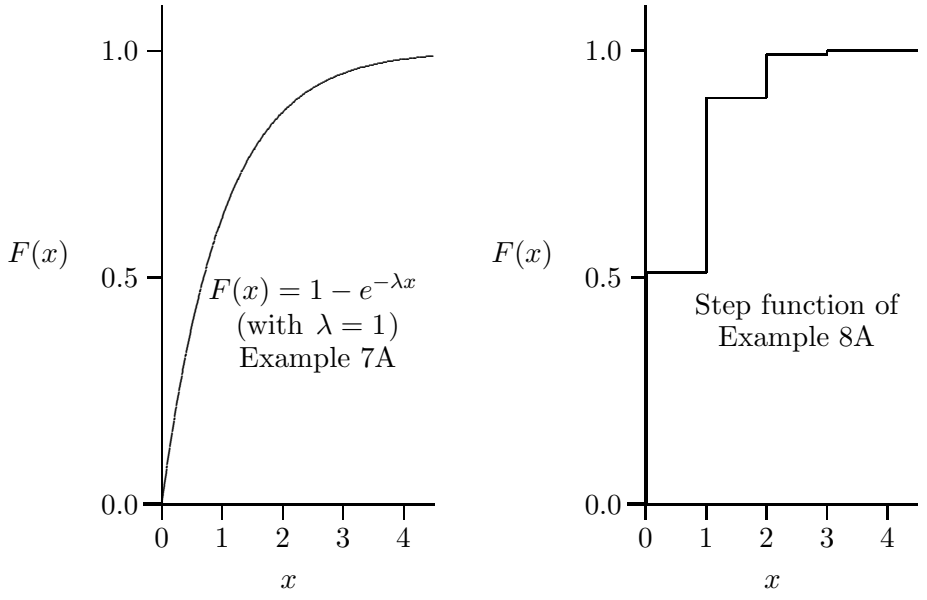
Example 8A: Find the distribution function for the random variable X with the binomial distribution $X \sim B(3,\ 0.2)$.

$$\Pr[X = x] = \binom{3}{x} 0.2^x\ 0.8^{3-x} \qquad x = 0,\ 1,\ 2,\ 3$$

which yields $p(0) = 0.512,\ p(1) = 0.384,\ p(2) = 0.096,\ p(3) = 0.008$. Thus $\Pr[X \le 0] = 0.512,\ \Pr[X \le 1] = 0.512 + 0.384 = 0.896$, etc, so that

$$\begin{aligned} F(x) &= 0 & x < 0 \\ &= 0.512 & 0 \le x < 1 \\ &= 0.896 & 1 \le x < 2 \\ &= 0.992 & 2 \le x < 3 \\ &= 1 & x \ge 3 \end{aligned}$$

The graphs of the distribution function $F(x)$ for Examples 7A and 8A are shown below. The graph of $F(x)$ is always an increasing function, between 0 and 1. If X is a discrete random variable, then $F(x)$ will be a "**step function**".



The distribution function gives us an expression which can be used **directly** to compute probabilities. For a continuous random variable, the distribution function can frequently be expressed as a formula, and this does away with the need to integrate to compute probabilities. For a discrete random variable, the distribution function is always a step function (why?), and is therefore less useful for computing probabilities than for a continuous random variable. Notice that the domain of the distribution function $F(x)$ is **always** the entire real line for both discrete and continuous random variables.

Example 9C: Find the distribution function for the random variable X with density function

$$\begin{aligned} f(x) &= \tfrac{1}{2}/\sqrt{x} \quad 0 \le x \le 1 \\ &= 0 \qquad\quad \text{otherwise} \end{aligned}$$

Example 10B: Suppose events are occurring at random with average rate λ per unit of time. What is the probability density function of the random variable X, the waiting time to the second event?
We first find $F(x)$.

$$\begin{aligned} F(x) &= \Pr[X \le x] = \Pr\left[\text{waiting less than } x \text{ units of time for the 2nd event}\right] \\ &= 1 - \Pr\left[\text{waiting more than } x \text{ units of time for the 2nd event}\right] \\ &= 1 - \Pr\left[\text{less than 2 events take place in } x \text{ units of time}\right] \\ &= 1 - \Pr\left[0 \text{ events in } x \text{ units}\right] - \Pr[1 \text{ event in } x \text{ units}] \\ &= 1 - e^{-\lambda x} - \lambda x e^{-\lambda x}, \end{aligned}$$

using the Poisson distribution. We now differentiate $F(x)$ to find the density function $f(x)$.

$$\begin{aligned} f(x) = \frac{dF(x)}{dx} &= \lambda e^{-\lambda x} - \lambda e^{-\lambda x} + \lambda^2 x e^{-\lambda x} \\ &= \lambda^2 x e^{-\lambda x}. \end{aligned}$$

Thus the density function of the random variable X, the waiting time to the second event, is given by

$$\begin{aligned} f(x) &= \lambda^2 x e^{-\lambda x} \quad x \ge 0 \\ &= 0 \qquad\qquad\quad \text{otherwise} \end{aligned}$$

This is an extension of the exponential distribution which is the density function of the waiting time to the first event. See Exercises 6.23 and 6.24.

Example 11C: A company which sells expensive woodworking machinery has established that the time between sales can be modelled by an exponential distribution, with parameter $\lambda = 2$ per five-day working week. The company has had no sales over the last week, and the manager fears that if no sales are made in the next few days, the company will be in serious financial difficulty. Throw some light on the situation by computing an expression for the probability that the number of days between two sales will be x or fewer days. Plot this function. Can you allay the manager's fears?

Example 12C: An estate agent has five houses to sell. She believes her situation can be modelled by a binomial process, and that her probability of selling each house within a month is 0.4. Compute and plot the distribution function of the number of sales in the month. How would this plot help the estate agent?

Example 13C: A student believes that she is equally likely to obtain a mark between 45% and 60% for an examination.

(a) Find the distribution function for the random variable giving the student's examination mark. Assume that the random variable is continuous!

(b) Use the distribution function to determine the probability that the student obtains less than 50% for her examination.

The median of a random variable...

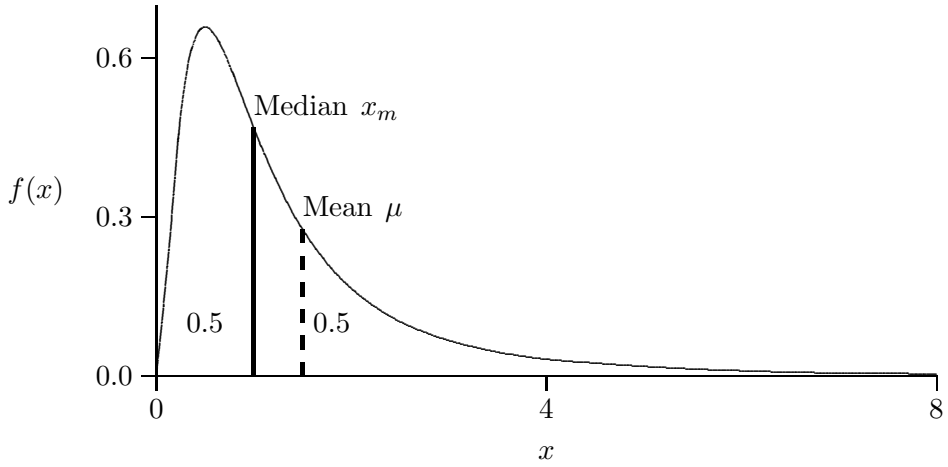
The distribution function gives us a way of defining the **median of a random variable** (which is not the same as the median of a sample, considered in chapter 1). If X is a random variable with distribution function $F(x)$, then the median of X, denoted x_m, satisfies the equation

$$F(x_m) = \frac{1}{2}.$$

If X is continuous, the median is clearly the value x_m such that

$$\int_{-\infty}^{x_m} f(x)\ dx = \frac{1}{2}.$$

Half the density function lies below the median, and half lies above it: the picture makes this clear. The mean and the median of a random variable only coincide when the density function is symmetric.



If X is discrete, $F(x)$ is a step function, and the median is taken to be the lowest value of x for which $F(x) \geq \frac{1}{2}$.

Example 14A: Find the median of the exponential distribution.

In Example 7A we showed that the distribution function of the exponential distribution is

$$F(x) = 1 - e^{-\lambda x}.$$

The median x_m is therefore the solution to the equation

$$\frac{1}{2} = 1 - e^{-\lambda x_m}$$

Rearranging, this yields

$$e^{-\lambda x_m} = \frac{1}{2}.$$

Now take natural logarithms to obtain

$$-\lambda x_m = \log_e \frac{1}{2},$$

so that, finally, the median is given by

$$x_m = 0.6931/\lambda.$$

Example 15C: Find the distribution functions and medians of the random variables having the following probability density/mass functions. Compare the medians with the means.

(a)

$$\begin{array}{lll} p(x) & = \binom{4}{x}0.5^4 & x = 0,\ 1,\ 2,\ 3,\ 4 \\ & = 0 & \text{otherwise} \end{array}$$

(b)

$$\begin{array}{lll} p(x) & = \binom{4}{x}0.4^x\ 0.6^{4-x} & x = 0,\ 1,\ 2,\ 3,\ 4 \\ & = 0 & \text{otherwise} \end{array}$$

(c)

$$\begin{array}{lll} f(x) & = 1/5 & 3 < x < 8 \\ & = 0 & \text{otherwise} \end{array}$$

(d)

$$\begin{array}{lll} f(x) & = 1/x & 1 \leq x \leq e \\ & = 0 & \text{otherwise} \end{array}$$

Using the normal distribution to approximate the binomial and Poisson distributions...

As surprising as it might appear, it is possible (under certain conditions) to compute approximate probabilities for both the binomial and Poisson distributions, which are both discrete, using our tables for the normal distribution, which is continuous.

So far we have shown (Examples 6B and 5B, respectively) that if $X \sim B(n,p)$ then $\mathrm{E}[X] = np$ and if $X \sim P(\lambda)$, then $\mathrm{E}[X] = \lambda$. We now need $\mathrm{Var}[X]$ for both distributions. Finding these variances is a fairly stiff exercise which is left for you to do (Exercise 6.22). We simply state that for the binomial distribution, $\mathrm{Var}[X] = npq = np(1-p)$ and for the Poisson distribution, $\mathrm{Var}[X] = \lambda$, the same as the mean.

NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

If $X \sim B(n,p)$ and both np and $n(1-p)$ are greater than 5 and $0.1 < p < 0.9$ then

$$X \approx N\big(np,\ np(1-p)\big),$$

where $\approx$ means "is approximately distributed".

The normal distribution used to approximate the binomial distribution is the one with the same mean and variance as the binomial distribution being approximated.

The same principle applies for the normal approximation to the Poisson distribution; the approximating normal distribution has the same mean and variance as the Poisson distribution being approximated.

NORMAL APPROXIMATION TO THE POISSON DISTRIBUTION

If $X \sim P(\lambda)$ and $\lambda > 10$ then

$$X \dot{\sim} N(\lambda,\ \lambda).$$

The examples show how to use the approximation to compute binomial and Poisson probabilities.

Example 16A: If you toss an unbiased coin 20 times what is the probability of exactly 15 heads?

X, the number of heads, has a binomial distribution $B(20,\ \frac{1}{2})$ with mean $np = 10$ and variance $np(1-p) = 5$. Also $n(1-p) = 10$. Because $np > 5$ and $n(1-p) > 5$ and $0.1 < p < 0.9$, we can approximate the distribution of X by means of a normal distribution which we will denote Y. The appropriate normal distribution Y to use to appproximate the binomial distribution X is the one with the same mean and variance as X. Thus we take $\mu = 10$ and $\sigma^2 = 5$ so that $Y \sim N(10,\ 5)$. We write $X \dot{\sim} N(10,\ 5)$ and say "X is distributed approximately normally with mean 10 and variance 5".

We now have to get around the problem of using a **continuous** distribution for a **discrete** random variable. The probability that the normal distribution takes on the value 15 is zero. If we are going to obtain a positive probability, we need an **interval** over which to evaluate the area under the curve. How do we choose this interval? The approximation has been designed in such a way that to obtain the probability that $X = 15$ for the binomial distribution, we calculate $\Pr[14.5 < Y < 15.5]$ for the normal distribution $(10,\ 5)$.

We learnt in Chapter 5 how to find probabilities for any normal distribution; we transform it to the standard normal distribution, using the formula $z = (x-\mu)/\sigma$ with $\mu = 10$ and $\sigma = \sqrt{5}$. Thus

$$\begin{aligned}\Pr(14.5 < Y < 15.5) &= \Pr\Big(\frac{14.5-10}{\sqrt{5}} < Z < \frac{15.5-10}{\sqrt{5}}\Big)\\ &= \Pr(2.01 < Z < 2.46)\\ &= 0.49305 - 0.4778 = 0.01525\end{aligned}$$

from the table of the standard normal distribution. Thus $\Pr(14.5 < Y < 15.5) = 0.0153$.

By way of comparison, the exact answer obtained from the binomial probability distribution is computed as

$$\Pr(X = 15) = \binom{20}{15}\frac{1}{2}^{20} = 0.0148.$$

Our approximate answer is within $3\frac{1}{2}\%$ of the true value.

Example 17A: If sales of tractors by a dealer occur in accord with a Poisson distribution with rate $\lambda = 25$ sales per month, what is the probability of 30 or more sales in a month?

Because λ exceeds 10, X, the number of tractors sold, can be closely approximated by a normal distribution Y with $\mu = 25$ and $\sigma^2 = 25$. To find the probability that the discrete distribution is 30 or more, we obtain the probability that the continuous distribution exceeds 29.5.

$$\begin{aligned}\Pr[X \geq 30] = \Pr[Y > 29.5] &= \Pr[Z > (29.5 - 25)/5)] \\ &= \Pr[Z > 0.9] \\ &= 0.1841\end{aligned}$$

The method we have demonstrated to compute the approximate probabilities is summarized in the block below:

PROCEDURE FOR USING THE APPROXIMATIONS

If the random variable X has a binomial or Poisson distribution and satisfies the conditions for being approximated by a normally distributed random variable Y, then

$$\Pr[a \leq X \leq b] = \Pr[a - \frac{1}{2} < Y < b + \frac{1}{2}]$$

Example 18B: An actuarial lifetable states that the probability that a 40-year old man will die before age 60 years is 0.17. An insurance company insures 300 men aged 40. What is the probability that the number of insured men who will die before age 60 lies between 50 and 60 (inclusive)?

Let the random variable X be the number who will die between age 50 and age 60. Clearly, $X \sim B(300,\ 0.17)$.

We calculate $np,\ n(1-p),\ np(1-p)$:

$$np = 300 \times 0.17 = 51, \qquad n(1-p) = 249, \qquad np(1-p) = 42.3.$$

The conditions for using the normal approximation are therefore satisfied. Thus $X \sim B(300,\ 0.17)$ can be approximated by $Y \sim N(51,\ 42.3)$.

$$\begin{aligned}\Pr[50 \leq X \leq 60] &= \Pr[49.5 < Y < 60.5] \\ &= \Pr[(49.5 - 51)/\sqrt{42.3} < Z < (60.5 - 51)/\sqrt{42.3}] \\ &= \Pr[-0.23 < Z < 1.46] \\ &= 0.0910 + 0.4279 \\ &= 0.5189\end{aligned}$$

Example 19C: The average number of newspapers sold at a busy intersection during rush hour is 5 per minute. What is the probability that in a 15-minute period during rush hour more than 85 newspapers are sold at the intersection?

Example 20C: A car ferry can accommodate 298 cars. Because bookings are not always taken up, the operators accept 335 bookings for each ferry crossing, hoping that no more than 298 cars arrive. If individual bookings are taken up independently with probability 0.85, what is the probability, when 335 bookings have been made, that more cars will arrive than can be accommodated for a particular crossing?

Solutions to examples...

3C (a) $\mathrm{E}[X] = 5 > 0$, so you should make the call.
(b) $\mathrm{E}[X] = 0$ implies $35p + (-15)(1-p) = 0$, so that $p = 0.3$.

4C (a) $\mathrm{E}[X_A] = 16.5\%$, $\mathrm{E}[X_B] = 12.0\%$,
(b) $SD[X_A] = 20.62\%$, $SD[X_B] = 22.05\%$,
(c) Share B is more risky, because it has the larger standard deviation.
(d) Coefficients of variation. A: 1.25, B: 1.83.
(e) Buy share A. It has both a higher expected return and a lower variance.

9C The distribution function is

$$\begin{array}{lll} F(x) & = 0 & x < 0 \\ & = \sqrt{x} & 0 \le x \le 1 \\ & = 1 & x > 1 \end{array}$$

11C $X \sim E(\frac{2}{5})$. Compute the distribution function $F(x) = 1 - e^{-\frac{2}{5}x}$ for $x \ge 0$ (and $F(x) = 0$ for $x < 0$). Some values are $F(x)$: $\Pr[X \le 5] = 0.865$ is the probability of making a sale within 5 days; $\Pr[X \le 4] = 0.798$, $\Pr[X \le 3] = 0.699$, $\Pr[X \le 2] = 0.55$.

12C The distribution function is

$$\begin{array}{lll} F(x) & = 0 & x < 0 \\ & = 0.0777 & 0 \le x < 1 \\ & = 0.3369 & 1 \le x < 2 \\ & = 0.6825 & 2 \le x < 3 \\ & = 0.9129 & 3 \le x < 4 \\ & = 0.9897 & 4 \le x < 5 \\ & = 1 & x \ge 5 \end{array}$$

13C (a) $f(x) = 1/15$ for $45 < x < 60$ (and zero otherwise).

$$\begin{array}{lll} F(x) & = 0 & x < 45 \\ & = (x-45)/15 & 45 \le x \le 60 \\ & = 1 & x > 60 \end{array}$$

(b) 0.333

15C (a) The distribution function is

$$\begin{array}{lll} F(x) & = 0 & x < 0 \\ & = 0.0625 & 0 \le x < 1 \\ & = 0.3125 & 1 \le x < 2 \\ & = 0.6875 & 2 \le x < 3 \\ & = 0.9375 & 3 \le x < 4 \\ & = 1 & x \ge 4 \\ x_m & = 2 & \mu = 2 \end{array}$$

(b) The distribution function is

$$\begin{array}{lll} F(x) & = 0 & x < 0 \\ & = 0.1296 & 0 \le x < 1 \\ & = 0.4752 & 1 \le x < 2 \\ & = 0.8208 & 2 \le x < 3 \\ & = 0.9744 & 3 \le x < 4 \\ & = 1 & x \ge 4 \\ x_m & = 2 & \mu = 1.6 \end{array}$$

(c) The distribution function is

$$\begin{array}{lll} F(x) & = 0 & x \le 3 \\ & = (x-3)/5 & 3 < x < 8 \\ & = 1 & x \ge 8 \\ x_m & = 5.5 & \mu = 5.5 \end{array}$$

(d) The distribution function is

$$\begin{array}{lll} F(x) & = 0 & x < 1 \\ & = \log x & 1 \le x \le e \\ & = 1 & x \ge e \\ x_m & = 1.6487 & \mu = 1.7183 \end{array}$$

For the symmetric distributions, (a) and (c), the mean and median coincide.

19C 0.1131

20C 0.0179

Exercises...

*6.1 Find the mean, the variance, the distribution function and the median of the random variables having the following probability density/mass functions.

(a)

$$\begin{array}{lll} f(x) & = x/50 & 0 < x < 10 \\ & = 0 & \text{otherwise} \end{array}$$

(b)

$$\begin{array}{lll} p(x) & = (x-1)/15 & x = 1,\ 2,\ \ldots,\ 6 \\ & = 0 & \text{otherwise} \end{array}$$

(c)

$$\begin{array}{lll} p(x) & = 0.1 & x = 10 \quad \text{or} \quad x = 40 \\ & = 0.3 & x = 20 \\ & = 0.5 & x = 30 \\ & = 0 & \text{otherwise} \end{array}$$

(d)

$$\begin{array}{lll} f(x) & = 3x^2/125 & 0 < x < 5 \\ & = 0 & \text{otherwise} \end{array}$$

(e)

$$\begin{aligned} f(x) &= 10x^9 & 0 < x < 1 \\ &= 0 & \text{otherwise} \end{aligned}$$

*6.2 Let X be a random variable with probability function

$$\begin{aligned} f(x) &= kx(\tfrac{1}{2} - x) & 0 \leq x \leq \tfrac{1}{2} \\ &= 0 & \text{otherwise} \end{aligned}$$

(a) Determine k so that $f(x)$ is a density function.
(b) Find the mean and variance of X.
(c) Find the distribution function $F(x)$ and the median.

*6.3 A random variable X has probability density function

$$\begin{aligned} f(x) &= Ax + B & 0 \leq X \leq 1 \\ &= 0 & \text{otherwise} \end{aligned}$$

If the mean of X is 0.5, find A and B.

*6.4 For the opportunity to roll a die you pay n cents. If you score a six, you get a reward of R2.00, and get your n cents returned. How much should you pay in order to make this a fair game? (Note: a game is defined to be "fair" if the expected gain is zero.)

*6.5 A charitable institution wishes to raise funds by holding either a braaivleis, or a dinner in a hotel. If they choose a braaivleis they will lose R1200 if it rains, or make R3000 if it does not rain. If they hold a dinner they will make R2400 if it rains and lose R300 if it does not rain. Which should they choose (a) if the probability of rain is 1/3, (b) if the probability of rain is 1/2?

6.6 Find the mean and the variance of the exponential distribution.

6.7 Find the mean of the normal distribution.

*6.8 What is the probability of obtaining more than 300 sixes in 1620 tosses of a fair die?

*6.9 Experience has shown that 15% of travellers reserving flights with Wildebeest Airlines do not take up their seats. If each plane has 50 seats and 58 bookings are accepted beforehand, what is the probability that everyone wishing to make the flight can be accommodated?

*6.10 The university lost-property office receives, on average, 64 articles per weekday. What is the probability that on one day

(a) between 70 and 80 articles (inclusive) are received?
(b) exactly 64 articles are received?
(c) less than 50 articles are received?

6.11 Each kilogram of grass seeds contains an average of 200 weed seeds. What is the probability that a kilogram of seeds contains more than 225 weed seeds? (Assume that the number of weed seeds has a Poisson distribution.)

*6.12 A holiday farm has 110 bungalows. During the winter holidays, the farm has an average occupancy of 60 bungalows per night. Assuming that the random variable giving the number of bungalows let per night has a binomial distribution, compute the probabilities

(a) fewer than 70 bungalows are let for a night, and
(b) between 65 and 75 bungalows (inclusive) are let for a night.

6.13 Prove that the mean and the median of a continuous random variable having a symmetric probability distribution function are equal.

*6.14 A random variable X has probability density function

$$\begin{aligned} f(x) &= k(1-x^2) && -1 < x < 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

(a) Determine the value of k.
(b) Find the mean and variance of the random variable X.
(c) Determine the probability that the random variable X lies within an interval of one standard deviation on either side of the mean.
(d) Find the distribution function and the median.

6.15 The probability density function of a random variable X is given by

$$\begin{aligned} f(x) &= a && 0 \le x \le 1 \\ &= b && 1 \le x \le 3 \\ &= 0 && \text{otherwise} \end{aligned}$$

where a and b are constants.

(a) Given that the mean of the random variable X is 5/4, prove that $a = \frac{1}{2}$ and $b = \frac{1}{4}$.
(b) Find the variance of X.
(c) Find $F(x)$, the distribution function, and x_m the median.
(d) What is the probability that 3 out of 5 observations on the variable X are less than the median?

*6.16 An automatic teller machine (ATM) is installed in a busy shopping area. The number of customers using the ATM per hour can be modelled by a Poisson random variable with a mean of 70. Find, using an approximation, the probability that between 15 and 35 customers (inclusive) use the ATM in any half-hour period.

6.17 A fair coin is tossed 250 times. Find the probability that the number of heads will not differ from 125 by

(a) more than 10
(b) less than 7.

6.18 On average, a newspaper is delivered late twice per month. Calculate the approximate probability that the paper will be late more than 30 times in a year.

6.19 A random variable X has probability density function

$$f(x) = Ax + B \qquad 0 \le x \le 3$$

We are told that the mean of X is $\mathrm{E}(X) = 1$.

(a) Find the values of the constants A and B.
(b) Find the variance of X.
(c) Find the distribution function of X.
(d) Find the median of X.

*6.20 Consider the following game. You pay an amount x, and then toss a coin until a head appears. If a head is obtained on the first or second throw, you lose. If a head is obtained on the third or fourth throws, you win R1. If a head appears on the fifth or subsequent throw, you win R5.

(a) If $x = 75$c, what are your expected winnings (or losses) per game?
(b) What should x be to make the game a "fair game"?

6.21 The random variable X has the probability mass function (known as the "truncated Poisson distribution")

$$\begin{aligned} p(x) &= \frac{e^{-\lambda}\lambda^x}{x!\,(1-e^{-\lambda})} && x = 1,\ 2,\ \ldots \\ &= 0 && \text{otherwise} \end{aligned}$$

(a) Check that the conditions for a probability mass function are satisfied.
(b) Find the mean of the random variable X.

Some more difficult exercises ...

6.22 (a) Show that the variance of the random variable X can be expressed as

$$\mathrm{Var}[X] = \mathrm{E}[X(X-1)] + \mathrm{E}[X] - (\mathrm{E}[X])^2,$$

where $\mathrm{E}[X(X-1)]$ means $\sum_x x(x-1)p(x)$.
(b) Use this result to find the variance of the binomial and Poisson distributions:
(i) if $X \sim B(n,p)$, then $\mathrm{Var}(X) = np(1-p)$
(ii) if $X \sim P(\lambda)$, then $\mathrm{Var}(X) = \lambda$.

6.23 (a) Show that the probability density function derived in Example 10B

$$\begin{aligned} f(x) &= \lambda^2 x e^{-\lambda x} && x \geq 0 \\ &= 0 && \text{otherwise} \end{aligned}$$

satisfies the conditions for being a probability density function.
(b) Find the mean and variance of the random variable X, the waiting time to the second event.

6.24 Generalize the results of Example 10B to determine the probability density function of the waiting time to the n-th event.

Solutions to exercises...

6.1 (a) $\mu = 20/3$, $\sigma^2 = 50/9$, and $x_m = 7.07$
(b) $\mu = 14/3$, $\sigma^2 = 14/9$, and $x_m = 5$
(c) $\mu = 26$, $\sigma^2 = 64$, and $x_m = 30$
(d) $\mu = 15/4$, $\sigma^2 = 15/16$, and $x_m = 3.97$
(e) $\mu = 10/11$, $\sigma^2 = 5/726$, and $x_m = 0.7943$

6.2 (a) $k = 48$
(b) $\mu = \frac{1}{4}$, and $\sigma^2 = 1/80$
(c) The distribution function is

$$\begin{array}{rll} F(x) & = 0 & x < 0 \\ & = 12x^2 - 16x^3 & 0 \leq x \leq \frac{1}{2} \\ & = 1 & x > \frac{1}{2} \end{array}$$

The median $x_m = \frac{1}{4}$ because $f(x)$ is symmetric.

6.3 $A = 0 \qquad B = 1$

6.4 40 cents

6.5 (a) Choose a braai, then E(gain) = 1600
(b) Choose a dinner, then E(gain) = 1050

6.6 $\mu = 1/\lambda$, and $\sigma^2 = 1/\lambda^2$

6.7 $\mathrm{E}[X] = \mu$

6.8 0.0212

6.9 0.6700

6.10 (a) 0.2254 (b) 0.0478 (c) 0.0351

6.11 0.0359

6.12 (a) 0.966 (b) 0.193

6.14 (a) $k = 3/4$ (b) $\mu = 0 \quad \sigma^2 = 1/5$
(c) $\Pr[-0.4472 < X < 0.4472] = 0.6261$
(d) The distribution function is

$$\begin{array}{rll} F(x) & = 0 & x \leq 1 \\ & = \frac{1}{4}(3x - x^3 + 2) & -1 \leq x \leq 1 \\ & = 1 & x > 1 \end{array}$$

The median $x_m = 0$ because $f(x)$ is symmetric.

6.15 (b) $\sigma^2 = 37/48$
(c) The distribution function is

$$\begin{array}{rll} F(x) & = 0 & x < 0 \\ & = \frac{1}{2}x & 0 \leq x \leq 1 \\ & = \frac{1}{2} + \frac{1}{4}(x-1) & 1 < x \leq 3 \\ & = 1 & x \geq 3 \end{array}$$

Median is $x_m = 1$

(d) 0.3125

6.16 0.5316

6.17 0.8164 0.4122

6.18 0.0918

6.19 (a) $A = -2/9 \quad B = 2/3$ (b) $\sigma^2 = 0.5$
(c) The distribution function is given by

$$\begin{array}{lll} F(x) & = 0 & x < 0 \\ & = -x^2/9 + 2x/3 & 0 \le x < 3 \\ & = 1 & x \ge 3 \end{array}$$

(d) $x_m = 0.8787$

6.20 (a) Expected loss of 25c per game.
(b) 50c

6.21 (b) $\mathrm{E}(X) = \lambda/(1 - e^{-\lambda})$

6.23 (b) $\mathrm{E}(X) = 2/\lambda \qquad \mathrm{Var}(X) = 2/\lambda^2$

6.24 $F_n(x) = 1 - \sum_{k=0}^{n-1} e^{-\lambda x}(\lambda x)^k/k!$, where $F_n(x)$ is the distribution function of the waiting time to the nth event. The density function is found by differentiation:

$$\begin{array}{lll} f_n(x) & = \lambda^n x^{n-1} e^{-\lambda x}/(n-1)! & x \ge 0 \\ & = 0 & \text{otherwise} \end{array}$$

Chapter 7

PROBABILITY DISTRIBUTIONS II: THE NEGATIVE BINOMIAL, GEOMETRIC, HYPERGEOMETRIC AND UNIFORM DISTRIBUTIONS

KEYWORDS: Negative binomial, geometric, hypergeometric and uniform distributions.

Recap ...

To date we have learnt about four important probability distributions, two of which were discrete, the binomial and Poisson distributions, and two of which were continuous, the exponential and normal distributions. There are very many other useful probability distributions and in this chapter we extend our repertoire.

The negative binomial distribution...

The underlying scenario for the negative binomial distribution is identical to that of the binomial distribution. But there is a twist in the tail! The binomial distribution is applicable if we have a random experiment with two outcomes, "success" and "failure", and if the probability of success, p, did not vary between trials of the experiment. These two conditions must also be satisfied for the negative binomial distribution to be applicable. For the **binomial** distribution, we fixed n, **the number of trials**, and we counted the number of successes we observed on these trials, letting the random variable X be the observed number of successes. But for the **negative binomial** distribution, however, we fix r, **the number of successes**, and count the number of failures until we have the rth success. Thus we let the random variable X be the number of failures before we obtain r successes. **We count failures, not trials!** The probability mass function for the negative binomial distribution is easily derived from first principles.

If the random variable X takes on the value x, so that there are x failures before r successes, there must have been a total of $x + r$ trials and the rth success must have occurred on the very last trial, the $(x + r)$th trial.[1] This in turn implies that there were

[1]Have you ever wondered why, when you are looking for something, you always find it in the last place that you look for it? Answer! Because once you have found it, you (hopefully!) stop looking for it. Similarly for a negative binomial process.

$r-1$ successes and x failures in the first $x+r-1$ trials. Thus

$$\begin{aligned}\Pr[X=x] &= \Pr[x \text{ failures in } x+r-1 \text{ trials}] \\ &\quad \times \Pr[\text{success in } (x+r)\text{th trial}] \\ &= \binom{x+r-1}{x} p^{r-1}\, q^x \times p \\ &= \binom{x+r-1}{x} p^r\, q^x\end{aligned}$$

We clinch the above discussion into the box:

NEGATIVE BINOMIAL DISTRIBUTION

We have a series of independent trials, each of which has two outcomes, success or failure. $\Pr[\text{success}] = p$ for each trial. Let $q = 1-p$.

Let r be a fixed number of successes, and let the random variable X be the number of failures obtained before we have r successes.

Then X has the **negative binomial distribution** with parameters r and p, i.e. $X \sim NB(r,\ p)$, and has probability mass function

$$\begin{aligned}p(x) &= \binom{x+r-1}{x} p^r\, q^x \quad x = 0, 1, 2, \ldots \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise}\end{aligned}$$

The name **negative** binomial distribution is not a good one! The mathematicians have a result called the **negative binomial theorem** which states that, under certain conditions,

$$(1-q)^{-r} = \sum_{x=0}^{\infty} \binom{r+x-1}{x} q^x.$$

This result, with its **negative** exponent, is used to prove the condition PMF3, that $\sum_{x=0}^{\infty} p(x) = 1$. In the same way that the binomial theorem gave its name to the binomial distribution, this mathematical result has transferred its name to the negative binomial distribution. There is nothing else about the negative binomial distribution that is negative!

Example 1A: A market research company requires each of its fieldworkers to conduct 10 interviews per day. Not everybody approached by a fieldworker agrees to participate in an interview. In fact, only 60% of approaches lead to an interview. What is the probability that the 10th interview is obtained from the 15th person approached?

Let the random variable X be the number of failures before 10 successes. Here $r = 10,\ p = 0.6$, so that $X \sim NB(10, 0.6)$. We want to find $\Pr[X = 5]$:

$$\Pr[X=5] = \binom{5+10-1}{5} 0.6^{10}\ 0.4^5 = 0.1240.$$

The special case of the negative binomial distribution when $r = 1$ is called the geometric distribution. Check for yourself that the mass function simplifies to the function given below:

GEOMETRIC DISTRIBUTION

Under the same conditions as for the negative binomial distribution, let the random variable X be the number of trials **before** the first success. Then X has the geometric distribution with parameter p, $X \sim G(p)$, and has probability mass function

$$\begin{aligned} p(x) &= pq^x \quad x = 0,\ 1,\ 2,\ \ldots \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

Example 2A: Suppose that you interview job applicants in succession until you find a person that satisfies the job description. Suppose that, at each interview, the probability of finding the right person is 0.3.

(a) What is the probability that you appoint the third person you interview?
(b) What is the probability that you will need to do five or more interviews?

(a) Let X be the number of trials **before** you succeed. Clearly $X \sim G(0.3)$. We want $\Pr[X = 2]$:

$$\Pr[X = 2] = 0.3 \times 0.7^2 = 0.147$$

(b) The probability of needing at least five interviews implies that the number of unsuccessful interviews must be four or more.

$$\begin{aligned} \Pr[X \geq 4] &= 1 - \Pr[X \leq 3] \\ &= 1 - \Pr[X = 0] - \Pr[X = 1] - \Pr[X = 2] - \Pr[X = 3] \\ &= 1 - 0.3 - 0.3 \times 0.7 - 0.3 \times 0.7^2 - 0.3 \times 0.7^3 \\ &= 0.2401 \end{aligned}$$

Example 3B: Some notebook computers make use of an "active" colour display on their screens. One reason why these computers are expensive is that the manufacturing process is so delicate that many of the screens produced are defective, and have to be discarded when tested during the assembly process. Only 58% of all screens produced are free from defects and can be used. If an order is placed for five notebook computers, what is the probability that

(a) the fifth non-defective screen is the eighth screen tested?
(b) no more than nine screens are required?

Let the random variable X be the number of defective screens tested before the fifth non-defective screen is found. Thus X has the negative binomial distribution:

$$X \sim NB(5, 0.58).$$

(a) We want $\Pr[X = 3] = \binom{3+5-1}{3} 0.58^5 \times 0.42^3 = 0.170$.
(b) Here we need

$$\begin{aligned} \Pr[X \leq 4] &= \sum_{x=0}^{4} \binom{x+4}{x} 0.58^5 \times 0.42^x \\ &= 0.066 + 0.138 + 0.174 + 0.170 + 0.143 \\ &= 0.691 \end{aligned}$$

Example 4C: Show that the geometric distribution satisfies the conditions for a probability mass function. (The sum you have to evaluate is a geometric progression). Show, assuming the result for the negative binomial theorem given earlier, that the negative binomial distribution satisfies the conditions for a probability mass function.

Example 5C: Suppose that the probability of passing the "board" examination is 0.45, that this probability does not vary with time, and that each attempt is independent of previous attempts. What is the probability that you pass the examination on your fifth attempt?

Example 6C: A breakfast cereal manufacturer places 1 of 16 picture cards of famous soccer players in each packet of cereal. Each picture is equally likely to be contained in a packet. You have collected 15 of the 16 cards. What is the probability that you will have to buy x more packets to complete the set?

Example 7C: A computer salesman has established that if he takes an interested customer out to lunch, the probability of making a sale increases from 0.4 to 0.65. The salesman needs to make only 7 more sales to reach his target, and decides to continue taking all interested customers out to lunch until he reaches his target. What is the probability that he will need to take customers out to lunch

(a) on 10 occasions?
(b) on more than 10 occasions?
(c) What are these probabilities if he does not take the customers out to lunch?

The hypergeometric distribution ...

The binomial and negative binomial distributions require that the probability of success remains the same from trial to trial. In many practical situations this is unrealistic — in particular it is unrealistic in sampling problems when the sampling is done "without replacement".

Consider, for example, a population of N articles, M of which are defective. We draw a sample of size n. Let the random variable X be the number of defective articles in n articles sampled without replacement.

For the event $X = x$ to occur, we must draw x articles from the M defective articles, and $n-x$ from the $N-M$ non-defective articles. Counting rule 3 of chapter 3 tells us that we can choose x articles from M in $\binom{M}{x}$ ways, and that we can choose $n-x$ articles from $N-M$ in $\binom{N-M}{n-x}$ ways. Thus the total number of ways in which the event $X = x$ can occur is

$$\binom{M}{x}\ \binom{N-M}{n-x}$$

The total number of ways in which a sample of size n can be drawn from N articles is $\binom{N}{n}$. Thus using the rule for computing probabilities when the elementary events are equiprobable, we have

$$\Pr[X = x] = \binom{M}{x}\ \binom{N-M}{n-x} \Big/ \binom{N}{n}.$$

HYPERGEOMETRIC DISTRIBUTION

Given a population of size N, of which M are defective, a sample of size n $(n \leq N)$ is drawn. Let the random variable X be the number of defectives in the sample. Then X has the hypergeometric distribution with parameters N, M and n, $X \sim H(N,\ M,\ n)$ and X has probability mass function

$$\begin{aligned} p(x) &= \binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n} && x = 0,\ 1,\ \ldots, n \\ &= 0 && \text{otherwise} \end{aligned}$$

Example 8A: A fisherman caught 10 lobsters, 3 of which were undersized. An inspector of the Sea Fisheries Branch measured a random sample of 4 lobsters. What is the probability that the sample contains no undersized lobsters?

Here $N = 10, M = 3$ and $n = 4$. If X is the number of undersized lobsters in the sample of 4, then

$$\Pr[X = 0] = \binom{3}{0}\binom{10-3}{4-0}\Big/\binom{10}{4} = 0.1667$$

Conversely, the probability that the inspector finds at least one undersized lobster is $1 - 0.1667 = 0.8333$.

Example 9B: A team of 15 people is chosen from a class of 65 MBA students to play a social rugby match. The class contains 25 engineers. What is the probability that the team contains

(a) four engineers?
(b) at least four engineers?

(a) Let X be the number of engineers in the sample. Then $N = 25+40 = 65,\ M = 25$ and $n = 15$, so that $X \sim H(65,\ 25,\ 15)$. Thus

$$\Pr[X = 4] = \binom{25}{4}\binom{40}{11}\Big/\binom{65}{15} = 0.1410.$$

(b) The probability that X is 4 or more is

$$\begin{aligned} &\Pr[X \geq 4] = 1 - \Pr[X \leq 3] = 1 - \{p(0) + p(1) + p(2) + p(3)\} \\ &= 1 - \left\{\left[\binom{25}{0}\binom{40}{15} + \binom{25}{1}\binom{40}{14} + \binom{25}{2}\binom{40}{13} + \binom{25}{3}\binom{40}{12}\right]\Big/\binom{65}{15}\right\} \\ &= 0.9176 \end{aligned}$$

Example 10C: There are addition errors in 3 out of a total of 32 invoices. An auditor checks a random sample of 10 invoices. What are the probabilities of finding (a) 0, (b) 1, (c) 2 and (d) all 3 errors in the sample?

If N is much larger than n, then the difference between using the binomial distribution (which assumes that sampling is done **with replacement**) and using the hypergeometric distribution is small. If $n/N < 0.1$, so that the size of the sample is less than 10% of the total population size, then the binomial distribution may be used to give satisfactory approximations to hypergeometric probabilities:

BINOMIAL APPROXIMATION TO THE HYPERGEOMETRIC DISTRIBUTION

If $X \sim H(N, M, n)$ and if $n/N < 0.1$, then $X \approx B(n, p)$ with $p = M/N$

Example 11B: A company is established in two cities, Johannesburg and Cape Town. The total staff complement is 240, of which only 32 are based in Cape Town. If 10 of the total of 240 staff are randomly selected to attend a course on VAT, what is the probability that two of the 10 are from Cape Town. Calculate both exact and approximate probabilities.

Let X be the number of staff members selected from Cape Town.

$$X \sim H(240,\ 32,\ 10)$$

The exact probability is given by

$$\Pr[X = 2] = \binom{32}{2}\binom{208}{8} \Big/ \binom{240}{10} = 0.2604$$

Using the binomial approximation to the hypergeometric distribution, we put $p = 32/240 = 0.1333$ so that $X \approx B(10,\ 0.1333)$, and

$$\Pr[X = 2] = \binom{10}{2}\ 0.1333^2 \times 0.8667^8 = 0.2546.$$

The error in the approximation is

$$\frac{0.2604 - 0.2546}{0.2604} = 0.022 \text{ or } 2.2\%.$$

The uniform distribution...

This is the simplest possible continuous distribution. It is used to model the situation where all values in some interval (a, b) are equally likely to occur. At first glance the uniform distribution looks uninspiringly simple; but it is this very simplicity that gives it its importance. It is possible, but not trivial, to programme a computer to produce a series of numbers that look like a random sample from the uniform distribution.

UNIFORM DISTRIBUTION

If the continuous random variable X is equally likely to take on any value in the interval (a, b), then X has the uniform distribution, $X \sim U(a, b)$, with probability density function

$$\begin{aligned} f(x) &= \frac{1}{b-a} && a \le x \le b \\ &= 0 && \text{otherwise} \end{aligned}$$

Example 12A: Suppose the mass of a nominally 500 g tub of margarine is equally likely to take on any value in the interval $(495, 510)$. What is the probability that a randomly chosen tub will have a mass less than 500 g?

Let the random variable X be the mass of a tub of margarine. Because $X \sim U(495,\ 510)$

$$\begin{aligned} f(x) &= \tfrac{1}{510-495} = 1/15 && 495 < x < 510 \\ &= 0 && \text{otherwise} \end{aligned}$$

$$\begin{aligned} \Pr(X < 500) &= \int_{495}^{500} \frac{1}{15}\,dx = \Big[\frac{x}{15}\Big]_{495}^{500} \\ &= (500 - 495)/15 = 1/3 \end{aligned}$$

Example 13C: If $X \sim U(a, b)$, investigate the following properties of X:

(a) Show that the function given for the uniform distribution satisfies the conditions for a probability density function.
(b) Show that $E[X] = \frac{1}{2}(b + a)$ and that $\text{Var}[X] = (b - a)^2/12$
(c) Show that the distribution function is given by

$$\begin{aligned} F(x) &= 0 && x < a \\ &= (x - a)/(b - a) && a \le x \le b \\ &= 1 && x > b \end{aligned}$$

Example 14C: An investor knows that his share portfolio is equally likely to yield an annual return anywhere in the interval between 5% and 35%. The fixed deposit rate is 13.5%. What is the probability that he would be better off investing his funds in a fixed deposit account (rather than in his share portfolio) over the forthcoming year?

Example 15C: The final mark (Y) for a particular statistics course comprises of a 30% weighting for the class record and a 70% weighting for the examination. A student believes that she is equally likely to obtain a mark anywhere between 45% and 65% for her examination (X).

(a) Find the probability density function for her final mark (Y) if she has a class record of 50%.
(b) Find the distribution function of her final mark.
(c) Find the probability she gets a third-class pass (between 50% and 60%) as a final mark.
(d) Find the probability that she gets a lower second (between 60% and 70%).
(e) What is the probability that she fails (below 50%)?

Solutions to examples...

5C $\Pr[\text{passing on 5th attempt}] = \Pr[X = 4] = 0.55^4 \times 0.45 = 0.0412$

6C $\Pr[x \text{ packets}] = \Pr[x - 1 \text{ failures}] = \frac{1}{16}\left(\frac{15}{16}\right)^{x-1}$

7C $X \sim NB(7, 0.65)$ (a) $\Pr[X = 3] = 0.1766$
(b) $\Pr[X > 3] = 1 - \Pr[X \le 3] = 0.4862$ (c) 0.0297 and 0.9452

10C $\Pr[x \text{ errors}] = p(x) = \binom{3}{x}\binom{29}{10 - x} \Big/ \binom{32}{10}$

(a) $p(0) = \binom{3}{0}\binom{29}{10}\Big/\binom{32}{10} = 0.3105$

(b) $p(1) = \binom{3}{1}\binom{29}{9}\Big/\binom{32}{10} = 0.4657$

(c) $p(2) = 0.1996$ (d) $p(3) = 0.0242$

(Note that these probabilities sum to 1.)

14C $\Pr(5 < X < 13.5) = 0.283$

15C (a) The random variable $Y \sim U(46.5, 60.5)$.

$$\begin{array}{lll} f(y) & = \frac{1}{60.5-46.5} = \frac{1}{14} & 46.5 < y < 60.5 \\ & = 0 & \text{otherwise} \end{array}$$

(b) The distribution function is

$$\begin{array}{lll} F(y) & = 0 & y < 46.5 \\ & = (y - 46.5)/14 & 46.5 \leq y < 60.5 \\ & = 1 & y \geq 60.5 \end{array}$$

(c) 0.714

(d) 0.036

(e) 0.250, although if the examiners have compassion, and pass her on a mark between 49 and 50, the probability of failing is 0.179!

Exercises...

*7.1 What is the probability that a fair die is tossed z times before a 6 appears

(a) for the first time

(b) for the second time

(c) for the rth time?

*7.2 A parliamentary candidate needs to collect 300 signatures before he can be nominated. If the probability that a voter approached at random will give a signature is 0.15, what is the probability that 1300 voters need to be approached before the 300th signature is collected?

7.3 By assuming that the negative binomial distribution is a probability mass function show that

$$\sum_{x=0}^{\infty} \binom{x+r-1}{x} q^x = (1-q)^{-r}.$$

*7.4 Show that if $X \sim NB(r,\ p)$, then $E[X] = rq/p$ and (more difficult) that $\mathrm{Var}[X] = rq/p^2$.

(Hint: the same procedure as was used for finding the mean and variance of the binomial and Poisson distributions is applied here.)

7.5 In exercise 7.2, what number of voters can be expected to refuse to sign before 300 signatures have been collected?

7.6 The Blood Transfusion Service knows that 6.3% of the population belong to the A-negative blood group.

(a) If people donate blood at random, what is the probability that x people will not belong to the A-negative group before
   (i) the first A-negative donor
   (ii) the fourth A-negative donor?

(b) What is the expected number of donors not having A-negative blood before
   (i) the first A-negative donor
   (ii) the fourth A-negative donor?

*7.7 A company that specializes in the breeding of fish needs to estimate the number of fish in one of the dams on their fish farm. In order to estimate the number of fish N in the dam, M are caught, marked and released. After sufficient time has elapsed for the marked fish to mix thoroughly, fish are caught one by one until r marked fish have been caught. The fish are released as soon as they have been examined for marks.

(a) What is the probability that x unmarked fish are examined before r marked fish are caught?
(b) What is the expected number of unmarked fish that need to be examined before r marked fish are caught?
(c) What, therefore, would you suggest as an estimate of N, the total population?
(d) If 300 fish are marked, and 189 unmarked fish are caught before 50 marked fish are caught, what is your estimate of the total population?

7.8 (a) Plot the bar graph and the distribution function of the negative binomial distribution with parameters
   (i) $r = 4 \qquad p = 0.8$
   (ii) $r = 4 \qquad p = 0.5$.

(b) Thus determine the medians of these two negative binomial distributions.

7.9 A small shop has 10 cartons of milk left, of which three are sour. Unaware of this, you ask for four cartons of milk.

(a) If the four cartons are selected at random, what is the probability that you get x cartons of sour milk?
(b) Evaluate these probabilities, and plot them as a bar graph.

7.10 If $X \sim H(N,\ M,\ 2)$ write down the probability mass function, and show that it sums to one.

7.11 Show that the mean of the random variable X having the hypergeometric distribution $H(N,\ M,\ n)$ is nM/N. It is much more difficult to show that the variance is given by $n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$.

*7.12 An engineer has 60 fuses of which 7 are defective. He selects 5 fuses (without replacement) for a particular job.

(a) What is the exact probability of getting 1 defective fuse in the 5 fuses selected?
(b) Use the binomial approximation to the hypergeometric distribution to estimate this probability.
(c) What is the percentage error of the approximation?

7.13 A company employs five male and three female computer programmers. Four of the eight programmers are selected at random to serve on a committee. One of

the four is chosen from the committee to report to the manager. If this person is female, find the conditional probability that the committee consists of two males and two females.

(Hint: use Bayes' theorem and the hypergeometric distribution.)

7.14 A manufacturer of light bulbs reports that among a consignment of 10 000 sent to a supermarket, 2500 were faulty. A shopper selects 10 of these bulbs at random. What is the approximate probability that more than 2 are faulty?

7.15 A child plays with a pair of scissors and a piece of string 10 cm long. He cuts the string into two at a randomly chosen place.

(a) What is the probability that the piece of string to the left of the pair of scissors is less than 4 cm long?

(b) What is the probability that the shorter piece of string is less than 2.5 cm long?

*7.16 A taxi travels between two cities A and B which are 100 km apart. There are service stations at A and B and at the midpoint of the route. If the taxi breaks down, it does so at random at any point along the route between the cities. If a tow truck is dispatched from the nearest service station, what is the probability that it has to travel more than 15 km to reach the taxi?

7.17 A radio station announces the time every 15 minutes between midnight and 06h00. If you wake up at random in the early hours of the morning and switch on the radio, what is the probability that you have to wait less than 5 minutes to find out the time?

*7.18 Between 08h00 and 09h00 buses leave the residence for the university at the following numbers of minutes past 08h00:

00 03 05 07 10 12 15 30 37 55 60

(a) Calculate the probability of having to wait less than 2 minutes for a bus, if you arrive at the residence at a time uniformly distributed over the interval

(i) 08h00 to 09h00

(ii) 08h00 to 08h20

(iii) 08h02 to 08h30

(b) Calculate the probability of having to wait less than 5 minutes for a bus if you arrive between 08h00 and 08h35.

Solutions to exercises...

7.1 (a) $(\frac{1}{6})(\frac{5}{6})^{z-1}$ (b) $\binom{z-1}{z-2}(\frac{1}{6})^2(\frac{5}{6})^{z-2}$
(c) $\binom{z-1}{z-r}(\frac{1}{6})^r(\frac{5}{6})^{z-r}$. Remember that the negative binomial distribution counts the number of **failures** before r **successes**. Here $z = x + r$ **trials**.

7.2 The number of refusals $X \sim NB(300,\ 0.15)$. Therefore $p(1000) = \binom{1299}{1000} 0.15^{300} 0.85^{1000}$.

7.3 Note that this result is analogous to the binomial theorem expansion of $(x+y)^n$ for negative values of n.

7.5 If $X \sim NB(300,\ 0.15)$, then $E(X) = 300 \times 0.85/0.15 = 1700$.

7.6 (a) (i) 0.063×0.937^x (ii) $\binom{x+3}{x} 0.063^4\ 0.937^x$.
(b) (i) $q/p = 14.9$ (ii) $qr/p = 59.5$.

7.7 (a) The number of unmarked fish $X \sim NB(r, M/N)$.

Then $p(x) = \binom{x+r-1}{x}(M/N)^r\ (1-M/N)^x$.
(b) $E(X) = r(N-M)/M$ (c) $N = M(\mathrm{E}(X)+r)/r$
(d) We have to hope that our observed value of X, namely 189, is close to $E(X)$:

We then estimate $N = 300(189+50)/50 = 1434$.

7.8 (b) (i) $x_m = 1$ (ii) $x_m = 3$.

7.9 (a) The number of sour milk cartons $X \sim H(10,\ 3,\ 4)$.

Thus $p(x) = \binom{3}{x}\binom{7}{4-x}/\binom{10}{4}$.
(b) $p(0) = 0.167$ $p(1) = 0.500$ $p(2) = 0.300$ $p(3) = 0.033$.

7.12 (a) The number of defective fuses has a hypergeometric distribution: $X \sim H(60,\ 7,\ 5)$.
Thus $p(1) = \binom{7}{1}\binom{53}{4}/\binom{60}{5} = 0.3753$.
(b) Because $n/N < 0.1$, $X \dot{\sim} B(5,\ 7/60)$. Thus

$$p(1) \approx \binom{5}{1}\left(\frac{7}{60}\right)^1\left(\frac{53}{60}\right)^4 = 0.3552.$$

(c) Percentage error $= \left(\frac{0.3753-0.3552}{0.3752}\right) 100 = 5.4\%$

7.13 Let A_i be the event "the committee contains i females". Let B be the event "the computer programmer is female". Then

$$\Pr(B|A_0) = 0,\ \Pr(B|A_1) = \frac{1}{4},\ \Pr(B|A_2) = \frac{1}{2},\ \Pr(B|A_3) = \frac{3}{4}.$$

Also:

$$Pr(A_0) = \binom{3}{0}\binom{5}{4}\Big/\binom{8}{4} = 5/10$$
$$\Pr(A_1) = \binom{3}{1}\binom{5}{3}\Big/\binom{8}{4} = 30/70,\ \text{etc.}$$

By Bayes' theorem,

$$\begin{aligned}\Pr(A_2|B) &= \Pr(B|A_2)\Pr(A_2)\Big/\left(\sum_{i=0}^{3}\Pr(B|A_i)(\Pr(A_i)\right)\\ &= \frac{1}{2}\times\frac{30}{70}\Big/\left(0\times\frac{5}{70}+\frac{1}{4}\times\frac{30}{70}+\frac{1}{2}\times\frac{30}{70}+\frac{3}{4}\times\frac{5}{70}\right)\\ &= 0.5614\end{aligned}$$

7.14 Let X be the number of faulty bulbs in 10.

$X \sim H(10\ 000,\ 2500,\ 10)$. $X \approx B(10,\ \frac{1}{4})$.

$\Pr[X > 2] = 1 - 0.5256 = 0.4744$.

7.15 Let X be length of string to the left of the pair of scissors. $X \sim U(0,\ 10)$.
(a) $\Pr[X < 4] = 0.4$ (b) 0.5

7.16 0.4

7.17 0.33

7.18 (a) (i) 0.333 (ii) 0.600 (iii) 0.464
(b) 0.657.

Chapter 8

MORE ABOUT MEANS

KEYWORDS: Population mean and variance, statistic, sampling distribution, central limit theorem, the sample mean is a random variable, confidence interval, tests of hypotheses, null and alternative hypotheses, significance level, rejection region, test statistic, one-sided and two-sided alternatives, Type I and Type II errors

The unobtainable carrot...

We introduce now two new concepts — the mean and variance of a population — and then immediately acknowledge that they are virtually unobtainable. Consider this problem. We want to make global statements about the travelling times to work of all people employed in a country. To be precise, we wish to determine the mean and variance of these travelling times. Now let your imagination run riot considering the logistics of such an operation (observers, stop-watches, data collection and processing, ...). There is no doubt that the mean and variance of the travelling times of all employed people are numbers that exist — it is just too expensive and too time consuming to obtain them. So what do we do? We take the travelling times of a **sample** of employed people and use the sample mean and variance to **estimate** the mean and variance of the travelling times of the **population** of employed people. The operation of taking a sample from a population is not a trivial one, and we shall discuss it further in Chapter 11.

We now have three related concepts, each called a **mean**: the mean of a sample (chapter 1), the mean of a probability distribution (chapter 6) and now the mean of a population. The sample mean is used to estimate the population mean. When a probability distribution is chosen as a **statistical model** for a population, one of the criteria for determining the parameters of the probability distribution is that the mean of the probability distribution should be equal to the population mean. This paragraph so far is also true when we replace the word **mean** with the word **variance**. It is a universal convention to use the symbols μ and σ^2 for the population mean and population variance respectively, and the fact that these symbols are also used for the mean and the variance of a probability distribution causes no confusion.

Because these notions of the mean and variance are important, we risk saying them again. The population mean and variance are quantities that belong to the population as a whole. If you could examine the entire population of interest then you could determine the **one true value** for the population mean and the **one true value** for the population variance. Usually, it is impracticable to do a census of every member of a population

to determine the population mean. The standard procedure is to take a random sample from the population of interest and **estimate** μ, the population mean, by means of $\bar{x}$, the sample mean.

Statistics...

We remind you again of the special definition for the noun statistic, within the subject Statistics, of a "statistic". By definition a **statistic** is any value computed from the elements of a random sample. Thus $\bar{x}$, s^2 are examples of statistics.

The random variable, $\overline{X}$ ...

We argued above that the population parameter μ has a fixed value. But the sample mean $\bar{x}$, the statistic which estimates μ, depends on the particular random sample drawn, and therefore varies from sample to sample. Thus the sample mean $\bar{x}$ is a random variable — it takes on different values for different random samples. In accordance with our custom of using capital letters for random variables and small letters for particular values of random variables, we will now start referring to the sample mean as the random variable $\overline{X}$.

Because the statistic $\overline{X}$ is a random variable it must have a probability distribution. We have a special name for the probability distributions of statistics. They are called **sampling distributions**. This name is motivated by the fact that statistics depend on samples.

In order to find the sampling distribution of $\overline{X}$, consider the following. Suppose that we take a sample of size n from a population which has a normal distribution with known mean μ and known variance σ^2. Apart from dividing by n, which is a fixed number, we can then think of $\overline{X}$ as the sum of n values, let us call them X_1, X_2, X_3, ..., X_n, each of which has a normal distribution, with mean μ and variance σ^2; i.e. $X_i \sim N(\mu, \sigma^2)$. In Chapter 5, we stated that the sum of independent normal distributions also has a normal distribution; the mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances. Thus

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

because there are n equal means and n equal variances. But we do not want the distribution of $\sum_{i=1}^{n} X_i$; we want the distribution of

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We also saw in Chapter 5 that if the random variable $X \sim N(\mu, \sigma^2)$, then the distribution of $aX \sim N(a\mu, a^2\sigma^2)$. Applying this result with $a = 1/n$, we have

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}).$$

This statement about the distribution of $\overline{X}$ is true for all values of n. The bottom line is that if a sample of any specified size n is taken from a population having a normal

distribution with mean μ and variance σ^2, then the mean of that sample will also have a normal distribution, with the same mean μ, but variance σ^2/n.

But what happens if we take a sample from a population that **does not have a normal distribution**? Suppose that we do however know that this distribution has mean μ and variance σ^2. Suppose we take a sample of size n, and compute the sample mean $\overline{X}$. As mentioned above, apart from division by the sample size, the sample mean $\overline{X}$ consists of $\sum_{i=1}^{n} X_i$, the sum of n random variables. In Chapter 5, we mentioned the "central limit theorem," which states that **the sum of a large number of independent identically distributed random variables always has a normal distribution**. Thus, by the central limit theorem, the sample mean has a normal distribution if the sample size is "large enough". A sample size of 30 or more is large enough so that the distribution of the sample mean can be assumed to have a normal distribution; the approximation to a normal distribution is often good even for much smaller samples. It can be shown that if we draw a sample of n observations from a population which has population mean μ and variance σ^2, then $\overline{X}$ can be modelled by a normal distribution with mean μ and variance σ^2/n, i.e.

$$\overline{X} \approx N(\mu,\ \sigma^2/n)$$

("the sample mean is approximately distributed normally, with mean μ and variance σ^2/n"). The approximation is invariably very good for $n \geq 30$. But if the population which is being sampled has a normal distibution, then

$$\overline{X} \sim N(\mu,\ \sigma^2/n)$$

for all values of n, including small values.

The central limit theorem is a powerful result. Firstly, it tells us about the sampling distribution of $\overline{X}$, **regardless of the distribution of the population from which we sample**. The approximation becomes very good as the sample size increases. If the population which we sampled had a distribution which was similar to a normal distribution, the approximation is good even for small samples. If the sampled population had an exact normal distribution, then $\overline{X}$ too has a normal distribution for all sample sizes. But even when the population from which the samples were taken looks nothing like a normal distribution, the approximation of the sampling distribution to a normal distribution shape gets better and better as the sample size n increases.

Secondly, (and remember that we are thinking of $\overline{X}$ as a random variable and that therefore it has a mean) it tells us that the mean of the sample mean $\overline{X}$ is the same as the mean of the population from which we sampled. The sample mean is therefore likely to be "close" to the true population mean μ. In simple terms, the sample mean is a good statistic to use to estimate the population mean μ.

Thirdly, the inverse relation between the sample size and variance of the sample mean has an important practical consequence. With a large sample size, the sample mean is likely, on average, to be closer to the true population mean than with a small sample size. In crude terms, sample means based on large samples are better than sample means based on small samples.

The rest of this chapter, and all of the next, builds on our understanding of the distribution of $\overline{X}$. This idea is a fairly deep concept, and many students take a while before they really understand it. The above section should be revisited from time to time. Each time, you should peel off a new layer of the onion of understanding.

Let us now consider various problems associated with the estimation of the mean. For the remainder of this chapter we will make the (usually unrealistic) assumption that

even though the population mean μ is unknown, we do in fact know the value of the population variance σ^2. In the next chapter, we will learn how to deal with the more realistic situation in which both the population mean and variance are unknown.

Estimating an unknown population mean when the population variance is assumed known...

We motivate some theory by means of an example.

Example 1A: We wish to estimate the population mean μ of travelling times between home and university. For the duration of this chapter we have to assume the population variance σ^2 known, so that σ is also known and let us suppose $\sigma = 1.4$ minutes. On a sample of 40 days, we use a stopwatch to measure our travelling time, with the following results (time in minutes)

17.2	18.3	16.8	15.9	17.4	16.8	16.3	18.4	19.1	29.3
18.1	16.9	17.2	17.6	15.8	18.4	17.9	16.5	17.9	16.8
16.4	19.3	17.5	18.1	17.2	17.7	17.1	16.0	18.1	22.3
19.1	17.0	20.5	16.1	18.7	19.0	17.3	17.6	18.2	16.5

Adding these numbers and dividing by 40, gives the sample mean $\overline{X} = 17.76$.

Fine, $\overline{X}$ estimates μ, so we have what we wanted. But how good is this estimate? How much confidence can we place in it? To answer this question we need to make use of the sampling distribution of $\overline{X}$. We shall see that we can use this distribution to form an interval of numbers, called a **confidence interval**. We will also be able to make a statement about the probability that the confidence interval method yields an interval which includes the population mean.

Confidence intervals...

In many situations, in business and elsewhere, a sample mean by itself is not very useful. Such a single value is called a **point estimate**. For example, suppose that the breakeven point for a potential project is R20 million in revenues, and that we are given a point estimate for revenues of R22 million. Our dilemma is that this point estimate of revenue is subject to variation — the true unknown value may well be below R20 million. Clearly, it will be far more helpful in taking a decision if we could be given a method of calculating an interval of values along with a probability statement about the likelihood that the interval includes the true value. If we could be told that the true value for revenue is likely to lie in the interval R21 million to R23 million, we would decide to go ahead with the project. But if we were told it was likely that the true value lay in the interval R12 million to R32 million, we would most certainly want to get more or better information before investing in the project. Notice that, for both of these intervals, the point estimate of revenues, R22 million, lies at the midpoint of the interval. Our hesitation to invest, given the second scenario, is not due to the position of the midpoint, but to the width of the interval.

The most commonly used probability associated with confidence interval methods is 0.95, and we then talk of a 95% confidence interval. This phrase means that the probability is 0.95 that the confidence interval from a random sample will include the true population value. Conversely, the probability that the confidence interval from a

random sample does not include the population mean is 0.05. Put another way, the confidence interval will not include the population mean 1 time in 20, on average. Let us develop the method for setting up such a confidence interval for the mean (assuming, as usual for this chapter, that σ^2 is known).

We make use of the fact that (for large samples)

$$\overline{X} \sim N(\mu,\ \sigma^2/n).$$

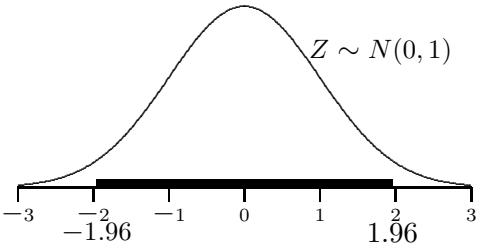
We make the usual transformation to obtain a standard normal distribution:

$$Z = (\overline{X} - \mu)/\frac{\sigma}{\sqrt{n}} \sim N(0,\ 1)$$

From our normal tables we know that

$$\Pr[-1.96 < Z < 1.96] = 0.95$$

i.e. the probability that the standard normal distribution lies between -1.96 and $+1.96$ is 0.95.



We can substitute $(\overline{X} - \mu)/\frac{\sigma}{\sqrt{n}}$ in place of Z in the square brackets because it has a standard normal distribution:

$$\Pr[-1.96 < (\overline{X} - \mu)/\frac{\sigma}{\sqrt{n}} < 1.96] = 0.95.$$

Rearranging, we obtain:

$$\Pr[\overline{X} - 1.96\ \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}}] = 0.95.$$

Look at this probability statement carefully: in words, it says, the probability is 0.95 that an interval formed by the method $(\overline{X} - 1.96\ \frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}})$ includes μ, the population mean. The method supplies a 95% confidence interval for μ. We apply this result to our introductory example.

Example 1A, continued: Look back and check that $\overline{X} = 17.76,\ \sigma = 1.4$ and $n = 40$. Substitute these values into the formula for the 95% confidence interval:

$$(17.76 - 1.96 \times 1.4/\sqrt{40}\ ,\ 17.76 + 1.96 \times 1.4/\sqrt{40})$$

which reduces to $(17.33\ ,\ 18.19)$. Thus, finally, we can state that we are 95% certain that our method covers the population mean of the travelling time from home to university. This interval method specifies a set of most likely values. We now not only have the **point estimate** given by $\overline{X}$, we also have some insight into the accuracy of our estimate.

To obtain confidence intervals with different probability levels, all that needs to be changed is the z-value obtained from the normal tables. The box below gives the appropriate z-values for the most frequently used confidence intervals.

CONFIDENCE INTERVAL METHOD FOR μ, σ^2 KNOWN

If we have a random sample of size n with sample mean $\overline{X}$, then A% confidence intervals for μ are given by

$$\left(\overline{X} - z^* \frac{\sigma}{\sqrt{n}} \ , \ \overline{X} + z^* \frac{\sigma}{\sqrt{n}}\right)$$

where the appropriate values of z^* are given by:

A%	z^*
90%	1.64
95%	1.96
98%	2.33
99%	2.58

Example 2B: An estimate of the mean fuel consumption (litres/100 km) of a car is required. A sample of 47 drivers each drive the car under a variety of conditions for 100 km, and the fuel consumed is measured. The sample mean turns out to be 6.73 litres/100 km. The value of σ is known to be 1.7ℓ/100 km. Determine 95% and 99% confidence intervals for μ.

We have $\overline{X} = 6.73$, $\sigma = 1.7$ and $n = 47$. Thus a 95% confidence interval for μ is given by

$$(6.73 - 1.96 \times 1.7/\sqrt{47} \ , \ 6.73 + 1.96 \times 1.7/\sqrt{47})$$

which is $(6.24 \ , \ 7.22)$. We are 95% sure that the method will include true fuel consumption. The 99% confidence interval is found by replacing 1.96 by 2.58:

$$(6.73 - 2.58 \times 1.7/\sqrt{47} \ , \ 6.73 + 2.58 \times 1.7/\sqrt{47})$$

which is $(6.09 \ , \ 7.37)$. We are 99% sure that the method will include true fuel consumption. Note the penalty we have to pay for **increasing** the level of confidence: the interval is considerably **wider**. Making a wider interval involves accepting margins of variability in sample means.

Suppose now that there was a block in the communication between the experimenter and the statistician and that the actual sample size was not 47, but 147. What now are the 95% and 99% confidence intervals?

The appropriate confidence interval methods give intervals of

$$\begin{aligned}(6.73 - 1.96 \times 1.7/\sqrt{147} \ , \ 6.73 + 1.96 \times 1.7/\sqrt{147} \\ = (6.46 \ , \ 7.00)\end{aligned}$$

and

$$\begin{aligned}(6.73 - 2.58 \times 1.7/\sqrt{147} \ , \ 6.73 + 2.58 \times 1.7/\sqrt{147}) \\ = (6.37 \ , \ 7.09)\end{aligned}$$

respectively. Note that if everything else remains unchanged, the method provides confidence intervals that become **shorter** with an **increase** in the sample size n. Resist the temptation to conclude that by increasing the sample size it is more likely that the confidence interval method covers the true mean. The increase in sample size results in the method supplying narrower confidence intervals, but the probability of intervals that include the true mean μ is unchanged.

Example 3C: A chain of stores is interested in the expenditure on sporting equipment by high school pupils during a winter season. A random sample of 58 high school pupils yielded a mean winter expenditure on sporting equipment of R168.15. Assuming that the population standard deviation is known to be R37.60, find the 95% confidence interval for the true mean winter expenditure on sporting equipment.

Determining the sample size to achieve a desired accuracy ...

Example 4A: A rugby coach of an under-19 team is interested in the average mass of 18-year-old males. The population standard deviation of mass for this male age class is known to be 6.58 kg. What size sample is needed in order to be 95% sure to obtain an interval of the average mass of 18-year-old males within 2 kg of the true value?

Clearly, we are referring to a confidence interval of the form $(\overline{X}-2\ ,\ \overline{X}+2)$. Because the 95% confidence interval method yields $(\overline{X}-1.96\ \sigma/\sqrt{n}\ ,\ \overline{X}+1.96\ \sigma/\sqrt{n})$ it follows that, with $\sigma = 6.58$

$$2 = 1.96 \times 6.58/\sqrt{n},$$

and thus $n = (1.96 \times 6.58/2)^2 = 42$, rounding **upwards** to the next integer.

The general method for determining sample sizes is given in the box:

SAMPLE SIZE n REQUIRED TO ACHIEVE DESIRED ACCURACY L, WHEN σ^2 IS KNOWN

To obtain a sample mean $\overline{X}$ which is within L units of the population mean, with confidence level A%, the required sample size is

$$n = (z^*\,\sigma/L)^2$$

where the appropriate values of z^* are

A%	z^*
90%	1.64
95%	1.96
98%	2.33
99%	2.58

Example 5C: The population variance of the amount of cooldrink supplied by a vending machine is known to be $\sigma^2 = 115$ ml^2.

(a) The machine was activated 61 times, and the mean amount of cooldrink supplied on each occasion was 185 ml. Find a 99% confidence interval for the mean.

(b) What size samples are required if the estimated mean is required to be within (i) 1 ml (ii) 0.5 ml of the true value, with probability 0.99?

Testing whether the mean is a specified value when the population variance is assumed known...

We again use an example to motivate our problem.

Example 6A: A manufacturer claims that, on average, his batteries last for 100 hours before they go flat. The population standard deviation of battery life is known to be 12 hours. We take a sample of 50 batteries and find the sample mean $\overline{X}$ is 95.5 hours. Does this sample mean undermine the manufacturer's claim?

As a first approach to this problem, let us assume that the manufacturer's claim is true, and that the true mean μ is in fact 100 hours. We now ask how likely it is that a random sample mean will be more extreme than the $\overline{X} = 95.5$ we have obtained. Which is the same as asking: "What is the probability that $\overline{X} < 95.5$, given that $\mu = 100$?" We know that

$$\overline{X} \sim N(\mu,\ \sigma^2/n)$$

and that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,\ 1).$$

In this example $Z = \dfrac{\overline{X} - 100}{12/\sqrt{50}} \sim N(0,\ 1)$.

Here $\overline{X} = 95.5$, and the corresponding z-value is therefore

$$z = \frac{95.5 - 100}{12/\sqrt{50}} = -2.65.$$

Thus

$$\Pr[\overline{X} < 95.5] = \Pr[Z < -2.65] = 0.0040,\ \text{from tables.}$$

This equation says, if the manufacturer's claim is true (i.e. $\mu = 100$) then the probability of getting a sample mean of 95.5 or less is 0.004, a very small probability.

Now we have to take a decision. Either

(a) the manufacturer is correct, and a very unlikely event has occurred, one that will occur on average 4 times in every 1000 samples, or
(b) the manufacturer's claim is not true, and the true population mean is less than 100, making a population mean of 95.5 or less a more likely event.

The statistician here would go for alternative (b). He would reason that alternative (a) is so unlikely that he can safely reject it, and he would conclude that the manufacturer's claim is exaggerated.

Tests of hypotheses...

The problem posed in example 6A introduces us to the concept of **statistical inference** — how we infer or draw conclusions from data. **Tests of hypotheses**, also called **significance tests**, are the foundation of statistical inference.

Whenever a claim or assumption needs to be examined by means of a significance test, we have a step-by-step procedure, as outlined below. We will use the Example 6A to illustrate the steps. A modified procedure to perform hypothesis tests will be discussed later.

1. Set up a **null hypothesis**. This statement almost always suggests the value of a population parameter. Here the null hypothesis is that the true mean μ is equal to 100. For our null hypothesis we usually take any claim that is made (and usually we are hoping to be able to reject it! If the evidence against it in our random sample is strong enough to do so).

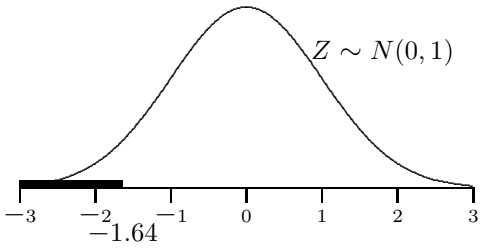
   We abbreviate the above null hypothesis to

   $$H_0 : \mu = 100.$$

2. An **alternative hypothesis** H_1 is specified. H_1 is accepted if the test enables us to reject H_0. Here the alternative hypothesis is $H_1 : \mu < 100$. We shall see that this inequality is a "**one-sided alternative**" and gives rise to a "**one-tailed significance test**".

3. A **significance level** is chosen. The significance level expresses the probability of rejecting the null hypothesis when it is in fact true. We usually work with a 5% or 0.05 significance level. We will then make the mistake of rejecting a true null hypothesis in 5% of all random samples, i.e. one time in twenty. On the other hand if H_0 is true we will accept H_0 in 95% of random samples. If the consequences of wrongly rejecting the null hypothesis are serious, a 1% or 0.01 level may be used. Here a 5% significance level would suffice.

4. We determine, from tables, the set of values that will lead to the rejection of the null hypothesis. We call this the **rejection region**. Our test statistic will have a standard normal distribution — thus we use normal tables. Because here H_1 is one-sided and contains a "$<$" sign, the rejection region is in the lower (left hand) end of the standard normal distribution. We reject H_0 if the sample mean $\overline{X}$ is too far below the hypothesized population mean μ, that is, if $\overline{X} - \mu$ is too negative. Because the significance level is 5% we must therefore find the lower 5% point of the standard normal distribution; this is -1.64.

   Thus the rejection region ties up with the distribution of the test statistic, the form of the alternative hypothesis, and the "size" of the significance level. The value we look up in the table is frequently called the **critical value** of the test statistic.

5. We calculate the **test statistic**. We know that $\overline{X} \sim N(\mu,\ \sigma^2/n)$. If H_0 is true then $\overline{X} \sim N(100,\ 12^2/50)$ and

   $$Z = \frac{\overline{X} - 100}{12/\sqrt{50}} \sim N(0,\ 1).$$

   In our example, $\overline{X} = 95.5$ and the observed value of the test statistic z is

   $$z = (95.5 - 100)/(12/\sqrt{50}) = -2.65.$$

6. We state our conclusions. We determine whether the test statistic value we observed falls into the rejection region. If it does, then we reject the null hypothesis H_0 and accept the alternative H_1. The result of the sampling is then said to be **statistically significant**. This phrase means that the signal from the random sample suggests we can reject the null hypothesis.

The feeling you should have by now is that if $\overline{X}$, the mean from our sample, is "too far" from 100, then we will reject H_0. But how far is "too far"?

We will reject H_0 if the calculated value of $Z = (\overline{X} - 100)/(12/\sqrt{50})$ is less than -1.64. Because -2.65, the observed value of Z, is less than -1.64 we reject H_0 at the 5% significance level.

In essence what we are doing is to say that it is an unlikely event (with probability less than 0.05) to get a sample mean of 95.5, or smaller, when the population mean is 100. We infer the true mean is unlikely to be 100 and we conclude that it must be less than 100. Another possible perspective is to remember that we are testing the position of μ, which is a measure of location. We are opting for the conclusion that the entire distribution is located (or centred) on a point to the left of the hypothesized location; this conclusion is consistent with the sample mean we obtained ($\overline{X} = 95.5$) being lower than $\mu = 100$

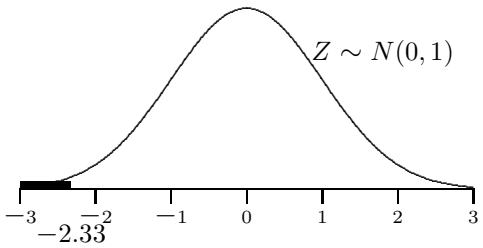
Example 7B: The mean output of typists in a typing pool is known to be 30 letters per day with a standard deviation of 10 letters. A new typist in the pool types 795 letters in her first 30 days, i.e. $795/30 = 26.5$ letters per day. Is her performance so far below average that we should fire her? Use a 1% significance level.

1. We set up the null hypothesis by giving the typist the benefit of the doubt: $H_0 : \mu = 30$.

2. The alternative hypothesis is $H_1 : \mu < 30$.

3. We use a 1% significance level — the probability of rejecting a true null hypothesis is then specified as 0.01. Because the decision to fire a person is an important one, we do not want to take this decision incorrectly. Using a 1% significance level means that we will make this incorrect decision only one time in a hundred.

4. An observed z-value of less than -2.33 will lead us to reject the null hypothesis:



5. The test statistic is
$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{26.5 - 30}{10/\sqrt{30}} = -1.92.$$

6. Because $-1.92 > -2.33$ we do not reject H_0. We thus decide to keep our new typist on. Note that if we had chosen a 5% significance level, the critical value would have been -1.64, and on that criteria we would have decided that our new typist was below standard.

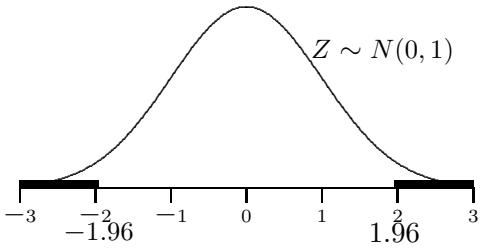
Example 8C: From past records it is known that the checkout times at supermarket tills have a standard deviation of 1.3 minutes. Past records also reveal that the average checkout time at a particular type of till is 4.1 minutes. A new type of till is monitored and 64 randomly sampled customers had an average checkout time of 3.8 minutes. Does the new till result in a significant reduction in checkout times? Use a 1% level of significance.

One-sided and two-sided alternative hypothesis ...

The following example illustrates the use of a two-sided alternative hypothesis. Guidelines on the choice of one-sided and two-sided alternative hypotheses is given after the example.

Example 9B: A farmer who has used a specific fertilizer for many years knows that his average yield of tomatoes is 2.5 tons/ha with a standard deviation of 0.53 tons/ha. The fertilizer is discontinued and he has to use a new brand. He suspects that the new fertilizer might alter the yield, but he has no idea whether the change will be an increase or a decrease. At the end of the season he finds that the average yield in 35 plots has been 2.65 tons/ha. At the 5% level of significance, test whether this sample data is statistically significant.

1. $H_0 : \mu = 2.5$
2. $H_1 : \mu \neq 2.5$. The farmer does not know in advance in which direction the change in yield might go. He wants to be able to reject H_0 if the yield either increases or decreases significantly.
3. Significance level : 5%.
4. Rejection region. Because of the form of the alternative hypothesis, we need to have two rejection regions, so that we can reject the null hypothesis either if the change of fertilizer decreases the yield or if it increases the yield. The two rejection regions are constructed to have the same size, but with their total probability set at 0.05. We reject the null hypothesis if $|z| > 1.96$. The diagram illustrates this rejection region.



5. We calculate the test statistic:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \ : \ z = \frac{2.65 - 2.5}{0.53/\sqrt{35}} = 1.67.$$

6. The observed z-value, 1.67, does not lie within either part of the rejection region; thus we cannot reject H_0. On the available statistical evidence the farmer concludes that the new fertilizer is not significantly different to the old.

Some guidelines...

The null hypothesis, the alternative hypothesis and the level of significance should all be determined before the data is gathered. The guideline for the choice of alternative hypotheses is: always use a two-sided alternative unless there are good **theoretical** reasons for using a one-sided alternative. The use of a one-sided alternative is never justified by claiming "the data point that way". In the hypothesis testing procedure we have considered above, the significance level ought also to be predetermined.

Type I and type II errors...

By now you are probably saying: it is better to use a 1% significance level than a 5% significance level, because with a 1% level we make the mistake of rejecting a true null hypothesis only 1 time in 100, whereas we make this mistake 1 time in 20 with a 5% significance level. If you have thought of that, then you have done well, but you may have overlooked something. There is another type of error one can make: the error of accepting the null hypothesis when it is false. The higher the significance of the test (i.e. a smaller α), the more difficult it is to reject the null hypothesis, and the more likely we are to accept the null hypothesis when it is false. Statisticians call the two classes of errors **type I** and **type II errors**, respectively.

		True situation	
		H_0 true	H_0 false
Our decision	Do not reject H_0	correct decision	**type II error**
	Reject H_0	**type I error**	correct decision

The probability of committing a type I error is the significance level of the test, sometimes also referred to as **the size of the test** $\Pr(\text{Type I error}) = \alpha$. The probability of committing a type II error varies depending on how close H_0 is to the true situation, and is difficult to control. The tradition of using 5% and 1% significance levels is based on the experience that, at these levels, the frequency of type II errors is acceptable.

Comparing two sample means (with known population variances)...

We recall a result we first used in chapter 5. If X_1 and X_2 are random variables with normal distributions,

$$X_1 \sim N(\mu_1\ ,\ \sigma_1^2), \quad X_2 \sim N(\mu_2\ ,\ \sigma_2^2)$$

and X_1 and X_2 are independent, then the distribution of the random variable $Y = X_1 - X_2$ is given by

$$Y = X_1 - X_2 \sim N(\mu_1 - \mu_2\ ,\ \sigma_1^2 + \sigma_2^2).$$

Example 10A: A cross-country athlete runs an 8 km time trial nearly every Wednesday as part of his weekly training programme. Last year, he ran on 49 occasions, and the mean of his times was 30 minutes 25.4 seconds (30.42 minutes). So far this year his mean time has been 30 minutes 15.7 seconds (30.26 minutes) over 35 runs. Assuming that the standard deviation last year was 0.78 minutes, and this year is 0.65 minutes, do these data establish whether, at the 5% significance level, there has been a reduction in the athlete's time over 8 km?

We work our way through our six-point plan:

1. Let the population means last year and this year be μ_1 and μ_2 respectively. Our null hypothesis specifies that there is no change in the athlete's times between last year and this year:

$$H_0 \ : \ \mu_1 = \mu_2 \ , \quad \text{or equivalently,} \quad H_0 \ : \ \mu_1 - \mu_2 = 0.$$

   Notice once again how the null hypothesis expresses the concept we are hoping to disprove. The null hypothesis can helpfully be thought of as the **hypothesis of no change**, or **of no difference**.

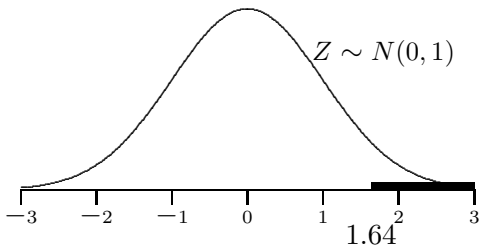
2. The alternative hypothesis involves the statement the athlete hopes is true:

$$H_1 \ : \ \mu_1 > \mu_2 \ , \quad \text{or equivalently,} \quad H_1 \ : \ \mu_1 - \mu_2 > 0.$$

   The alternative hypothesis does not specify the amount of the change; it simply states that there has been a decrease in average time between years 1 and 2.

3. Significance level : we specified that we would perform the test at the 5% level.

4. Rejection region. Because the test statistic has a standard normal distribution (we will show this in the next paragraph), because the significance level is 5%, and because we have a one-sided "greater than" alternative hypothesis, we will reject H_0 if the observed value of the test statistic is greater than 1.64.



5. Test statistic. In general, let us suppose that we have a random sample of size n_1 from one population and a random sample of size n_2 from a second population. Suppose the sample means are $\overline{X}_1$ and $\overline{X}_2$, respectively, and that the population means and variances are μ_1, μ_2, σ_1^2 and σ_2^2. Then

$$\overline{X}_1 \sim N(\mu_1, \ \sigma_1^2/n_1) \quad \text{and} \quad \overline{X}_2 \sim N(\mu_2, \ \sigma_2^2/n_2).$$

   Now, using the result about the distribution of differences

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

We now transform the difference variable to the standard normal distribution by subtracting its mean, and dividing by its standard deviation:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,\ 1).$$

This is our test statistic, and we can substitute for each variable: from the problem description we have

$$\begin{array}{lll} \overline{X}_1 = 30.42 & n_1 = 49 & \sigma_1 = 0.78 \\ \overline{X}_2 = 30.26 & n_2 = 35 & \sigma_2 = 0.65\ , \end{array}$$

and from the null hypothesis we have

$$\mu_1 - \mu_2 = 0$$

Thus, substituting, we compute our observed z-value as

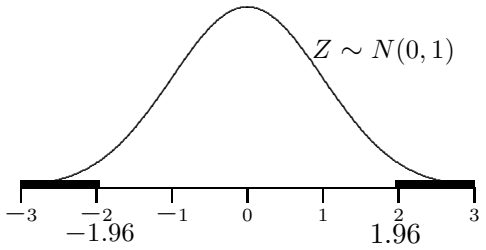
$$z = \frac{30.42 - 30.26 - (0.0)}{\sqrt{\frac{0.78^2}{49} + \frac{0.65^2}{35}}} = 1.02.$$

6. Conclusion. The observed value of the test statistic does not lie in the rejection region. We have to disappoint our athlete and tell him to keep trying.

Example 11B: A retail shop has two drivers that transport goods between the shop and a warehouse 15 km away. They argue continuously about the choice of route between the shop and warehouse, each claiming that his route is the quicker. To settle the argument, you wish to decide (5% significance level) which driver's route is the quicker. Over a period of months you time the two drivers, and the data collected are summarized below.

	Driver 1	Driver 2
Number of observations	$n_1 = 38$	$n_2 = 43$
Average time (minutes)	$\overline{X}_1 = 20.3$	$\overline{X}_2 = 22.5$
Standard deviation (minutes)	$\sigma_1 = 3.7$	$\sigma_2 = 4.1$

1. H_0 : $\mu_1 - \mu_2 = 0$. Both drivers take equally long.
2. H_1 : $\mu_1 - \mu_2 \neq 0$. We do not know in advance which driver is quicker.
3. Significance level : 5%.
4. We have a two-sided alternative at the 5% level. Thus we reject H_0 if the test statistic exceeds 1.96 or is less than -1.96; i.e. we reject H_0 if $|z| > 1.96$.



5. The test statistic is calculated from the data and the null hypothesis

$$z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{20.3 - 22.5 - 0}{\sqrt{\frac{3.7^2}{38} + \frac{4.1^2}{43}}}$$
$$= -2.54$$

6. Because $|-2.54| > 1.96$ we reject H_0 and conclude that the driving times are not equal. By inspection, it is clear that the route used by driver 1 is the quicker.

Example 12C: Two speedreading courses are available. Students enrol independently for these courses. After completing their respective courses, the group of 27 students who took course A had an average reading speed of 620 words/minute, while the group of 38 students who took course B had an average speed of 684 words/minute. If it is known that reading speed has a standard deviation of 25 words/minute, test (at the 5% significance level) whether there is any difference between the two courses.

Example 13C: The scientists of the Fuel Improvement Centre think they have found a new petrol additive which they hope will reduce a car's fuel consumption by 0.5 ℓ/100 km. Two series of trials are conducted, one without and the other with the additive. 40 trials without the additive show an average consumption of 9.8 ℓ/100 km. 50 trials with the additive show an average consumption of 9.1 ℓ/100 km. Do these data establish evidence that the additive reduces fuel consumption by 0.5 ℓ/100 km? Use a 5% significance level, and assume that the population standard deviations without and with the addition of the additive are 0.8 and 0.7 ℓ/100 km respectively.

Example 14C: Show that a 95% confidence interval for $\mu_1 - \mu_2$, the difference between two independent means, with variances assumed known, is given by

$$\left(\overline{X}_1 - \overline{X}_2 - 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \ , \ \overline{X}_1 - \overline{X}_2 + 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right).$$

Find a 95% confidence interval for the difference between mean travelling times in example 15B.

A modified hypothesis testing procedure...

These methods can be used *only* when we have random sample data, or when we explicitly report that we are assuming the data set is from a random sample.

The six-point plan for hypothesis testing, with steps 1 to 3 completed **before** the data are considered, represents the classical approach to hypothesis testing:

1. Null hypothesis H_0
2. Alternative hypothesis H_1
3. Significance level α
4. Rejection region for test statistic
5. Observed test-statistic, calculated from data under the assumption that H_0 is true
6. Conclusion

However, in practice, and particularly in the presentation of statistical conclusions in the journal literature, an alternative approach has been widely adopted:

1. Null hypothesis H_0
2. Alternative hypothesis H_1
3. Observed test-statistic, calculated from data under the assumption that H_0 is true
4. p-value
5. Conclusion

In this second approach, instead of explicitly using the fixed significance level α specified beforehand, we report the p-value derived from the data set under the assumption that the null hypothesis is true. The p-value is defined to be the **probability of obtaining a test-statistic value equal to or more extreme than the observed test-statistic** if we were to take other random samples of the same size n from the same population used in the first sample. In other words, the p-value is the probability of seeing what we have seen (in the data via the test statistic) or something more extreme under the assumption that the null hypothesis is true. So, a small p-value provides evidence against the null hypothesis, and we'll reject the null hypothesis when we have a sufficiently small p-value. We say we have a statistically significant result, when we observe a sufficiently small p-value.

To determine if a p-value is sufficiently small we need a (internal) boundary. It has become convention to use 0.05, so that p-values smaller than this level would result in a rejection of the null hypothesis. We will adopt the convention that if the significance level α is not explicitly stated, then the modified hypothesis testing procedure is to be followed.

Since the only test you have encountered thus far is a z-test, we'll illustrate how p-values are calculated in this context, but please note that the concept is applicable in all hypothesis testing.

We said that a p-value is defined to be the probability of obtaining a test-statistic value equal to or more **extreme** than the observed test-statistic. The test statistic is computed from the data, but how "extreme" it is depends on the alternative hypothesis. In an upper tail test (greater than alternative), one would expect to observe relatively

large test statistic values if the null hypothesis was false. This is why the p-value would be calculated as follows:

$$p\text{-value } = Pr\left(Z \geq \text{test stat}|H_0 \text{ is true}\right)$$

On the other hand, in a lower tailed test you'd expect to observe relative small test statistic values if the null hypothesis was false. The p-value would therefore be calculated as:

$$p\text{-value } = Pr\left(Z \leq \text{test stat}|H_0 \text{ is true}\right)$$

In the case of a two-tailed test, we don't know in which tail the test statistic would lie beforehand (but it can lie in only one!). To account for this, the area associated with the test statistic is calculated (depending on the tail in which it lies) and this area is then multiplied by 2 to obtain the relevant p-value.

The p-value is conventionally reported as a probability (rather than as a percentage). Statistical computer packages will calculate an exact p-value. When performing a z-test we can calculate these (exact) p-values from our standard normal tables. In some hypothesis tests, it will be come necessary to approximate p-values due to the nature of the tables. We'll handle this approximation then!

Example 15A: The weekly wage of semi-skilled workers in an industry is assumed to have population mean R286.50 and standard deviation R29.47. If a company in this industry pays a sample of 40 of its workers an average of R271.78 per week, can the company be accused of paying inferior wages?

1. $H_0 : \mu = 286.50$.
2. $H_1 : \mu < 286.50$. Inferior wages on average
3. The test statistic is z

$$\begin{aligned} z &= \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{271.78 - 286.50}{29.47/\sqrt{40}} \\ &= -3.16. \end{aligned}$$

4. Examining the column in the table for a one-sided alternative, we see that $z = -3.16$ is significant at the 5% level (because $z < -1.64$), the 1% level ($z < -2.33$), the 0.5% level ($z < -2.58$), the 0.1% level ($z < -3.09$), but not at the 0.05% level (because $z > -3.29$). The "most extreme" level at which the observed value of the test statistic $z = -3.16$ is significant is the 0.1% level.
5. We illustrate the conventional form of writing the conclusion.

   > "We have tested the sample mean against the population mean, and have found a statisticallysignificant difference ($z = -3.16,\ P < 0.001$). We conclude that the company is paying inferior wages."

   Notice carefully the shorthand method for presenting the results. The value of the test statistic is given, and an approximate p-value.

Example 16B: The specifications for extra-large eggs are that they have a mean mass of 125 g and standard deviation 6 g. A sample of 24 reputedly extra-large eggs had an average mass of 123.2 g. Are the sampled eggs smaller than the specifications permit?

1. $H_0 : \mu = 125$ g.
2. $H_1 : \mu < 125$ g.
3. Test statistic $z = \dfrac{123.2 - 125}{6/\sqrt{24}} = -1.47$.
4. From the table, $z = -1.47$ is statistically significant at the 10% level (but not at 5%).
5. The test shows that the mass of the sample of 24 eggs was close to being significantly lower than permitted by the specifications ($z = -1.47,\ P < 0.10$). Further investigation is recommended to clarify the situation.

Example 17C: The standard deviation of salaries in the private sector is known to be R4500, while in the public sector it is R3000. A random sample of 80 people employed in the private sector has a mean income of R8210, and a sample of 65 public sector employees has a mean income of R7460. Test the hypothesis that incomes in the private sector are significantly higher than those in the public sector.

Example 18C: A tellurometer is an electronic distance-measuring device used by surveyors. A new model has been developed, and its calibration is tested by repeatedly and independently measuring a distance of exactly 500 m. After 100 such measurements, the mean is 499.993 m with a standard deviation of 0.023 m. Does the new tellurometer have a systematic bias? Assume that the sample is sufficiently large that $\sigma \approx s$.

Example 19C: In the course of a year a statistician performed 60 hypothesis tests on different independent sets of data, each at the 5% significance level. Suppose that in every case the null hypothesis was true. What is the probability that he made no incorrect decisions? What is the expected number of incorrect decisions?

Hypothesis testing using the normal approximation to the binomial and Poisson distributions...

The normal distribution can also be used, under specified conditions, to test whether the parameters p and λ of the binomial and Poisson distributions have particular values. The examples illustrate the procedure. Note that these are not tests on a mean μ of a random variable, but on the parameters of binomial and Poisson distributions.

Strictly speaking a continuity correction is needed since a discrete distribution is being approximated by a continuous one. When n is large, the continuity correction is usually omitted, however for borderline cases it can be important (refer to using the normal distribution to approximate the binomial and Poisson distributions in Chapter 6). Please note that these corrections have been omitted in the examples and exercises of this chapter.

Example 20A: A new coin is produced. It is necessary to test whether such coins are unbiased. A random sample of 1000 coins are tossed independently and 478 heads appear. Is the coin unbiased? Use a 1% level of significance.

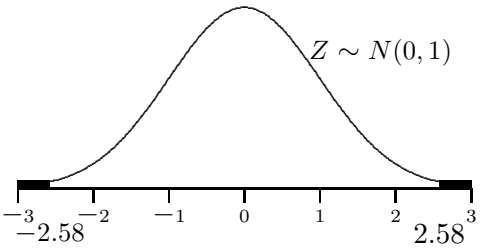
Let p be the probability of getting a head. The random variable of interest X, the number of heads in 1000 trials, has a binomial distribution with parameters n and p.

1. $H_0 \ : \ p = \frac{1}{2}$. This equation states that the coin is unbiased.

2. $H_1 \ : \ p \neq \frac{1}{2}$. If p differs from $\frac{1}{2}$ then the bias can go either way — remember that the null and alternative hypotheses ought to be formulated before you conduct the experiment. This is an example of a "**two-sided alternative**" and gives rise to a "**two-tailed test**".

3. Significance level : $\alpha = 1\%$.

4. Because we will be using a test statistic z with the normal distribution, we will reject H_0 if the observed z-value is greater than 2.58 or less than -2.58.



Our rejection region will be $|z| > 2.58$

5. The random variable X, the number of heads in 1000 trials, has a binomial distribution with mean np and variance npq. If H_0 is true, then $p = \frac{1}{2}$ and $np = 500$ and $npq = 250$. Since $n = 1000$, X satisfies the conditions for it to be approximated by the normal distribution (see Chapter 6 for "Normal approximation to the binomial distribution") with mean 500 and variance 250 (i.e. standard deviation $\sqrt{250}$) i.e. $X \sim N(500,\ 250)$ and the transformation $Z = (X - \mu)/\sigma$ to the standard normal distribution yields

$$z = \frac{478 - 500}{\sqrt{250}} = -1.39.$$

6. This value does not lie in the rejection region : we therefore cannot reject H_0 and instead we conclude that the coin is unbiased.

In general, the procedure for testing the null hypothesis that the parameter p of the binomial distribution is some particular value, is summarized as follows. If $X \sim B(n,p)$, if np and $n(1-p)$ are both greater than 5 and if $0.1 < p < 0.9$, then the distribution of the binomial random variable can be approximated by normal distribution $N(np,\ npq)$, so that the formula for calculating the test statistic z is

$$Z = \frac{X - np}{\sqrt{npq}}, z = \frac{x - np}{\sqrt{npq}}$$

The observed value of X is x the number of successes in n trials, and the value for p is taken from the null hypothesis.

Likewise, the procedure for testing the null hypothesis that the parameter λ of the Poisson distribution is some particular value, is summarized as follows. The Poisson distribution $P(\lambda)$ can be approximated by the normal distribution $N(\lambda,\ \lambda)$ provided $\lambda > 10$. So the formula for calculating the test statistic is thus

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}, z = \frac{x - \lambda}{\sqrt{\lambda}}$$

where the observed value of X is x the Poisson count and the value for λ is taken from the null hypothesis.

Example 21B: The number of aircraft landing at an airport has a Poisson distribution. Last year the parameter was taken to be 120 per week. During a week in March this year an aviation department official recorded 143 landings. Does this datum suggest that the parameter has increased? Use a 5% significance level.

1. $H_0\ :\ \lambda = 120$. The null hypothesis states that the rate is unchanged.
2. $H_1\ :\ \lambda > 120$. The value has increased.
3. Significance level : 5%.
4. Reject H_0 if the observed z-value exceeds 1.64.
5. Using the normal approximation to the Poisson distribution, we compute the statistic
$$z = \frac{X - \lambda}{\sqrt{\lambda}} = \frac{143 - 120}{\sqrt{120}} = 2.10.$$
6. Because $2.10 > 1.64$, we reject H_0, and conclude that the rate of landings has increased.

   In other words, it is an unlikely event to observe 143 landings in a week, if the true rate is 120. We therefore prefer the alternative hypothesis still.

Example 22C: A large motor car manufacturer enjoys a 21% share of the South African market. Last month, out of 2517 new vehicles sold, 473 were produced by this manufacturer. Does management have cause for alarm? Quote the p-value.

Example 23C: A die is rolled 300 times, and 34 sixes are observed. Is the die biased?

Example 24C: The complaints department of a large department store deals on average with 20 complaints per day. On Monday, 9 May 2010, there were 33 complaints.

(a) How frequently should there be 33 or more complaints?
(b) At the 1% significance level, would you advise management to investigate this day's complaints further.
(c) What would be plausible numbers of complaints per day at the 5% and 1% significance levels?

Example 25C: In a large company with a gender initiative, the proportion of female staff in middle-management positions was 0.23 in 1990. Currently, in one of the regional offices of the company, there are 28 women and 63 men in middle-management positions. If this regional office is assumed to be representative of the company as a whole, has there been a significant increase in the proportion of women in middle management? Perform the test at the 5% significance level.

Solutions to examples...

3C $(158.47,\ 177.83)$.

5C (a) $(181.46,\ 188.54)$.
(b) (i) 766 (ii) 3062.
Note: to improve the precision of the sample mean by a factor $\frac{1}{2}$ required a sample 4 times as large.

8C Reject H_0 if $z < -2.33$. Observed $z = -1.85$. Cannot reject H_0.

12C $z = -10.17$. Reject H_0.

13C $H_0 : \mu_1 - \mu_2 = 0.5$. $z = 1.25$. Cannot reject H_0.

14C 95% confidence interval : $(-3.90,\ -0.50)$.

17C Using modified procedure, $z = 1.20,\ P > 0.1$, insignificant.

18C $z = -3.043$, p-value $= 0.0012 \times 2 = 0.0024\ \ (P < 0.005)$, significant.

19C $\Pr(\text{all correct decisions}) = 0.95^{60} = 0.046$.
$\text{E(incorrect decisions)} = 60 \times 0.05 = 3$.

22C $z = -2.72$, p-value $= 0.0033\ \ (P < 0.005)$, significant.

23C $H_0\ :\ p = 1/6,\ z = -2.47$, p-value $= 0.0136\ \ (P < 0.05)$, significant.

24C (a) Use normal approximation to Poisson: $\Pr[x \geq 33] = 0.0026$, or once in 385 days.
(b) $H_0 : \lambda = 20,\ H_1 : \lambda > 20$. Reject H_0 if $z > 2.33$, observed $z = 2.91$. Reject H_0.
(c) Largest acceptable number of complaints is 27 at the 5% level and 30 at the 1% level.

25C $H_0 : p = 0.23$. Observed $z = 1.606$. Cannot reject H_0.

Exercises using the central limit theorem...

*8.1 A manufacturer of light bulbs prints "Average life 2100 hours" on the package of its bulbs. If the true distribution of lifetimes has a mean of 2130 and standard deviation 200, what is the probability that the average lifetime of a random sample of 50 bulbs will exceed 2100 hours? (Hint: use the result $\overline{X} \sim N(\mu,\ \sigma^2/n)$.)

8.2 Tubes produced by a company have a mean lifetime of 1000 hours and a standard deviation of 160 hours. The tubes are packed in lots of 100. What is the probability that the mean lifetime for a randomly selected pack will exceed 1020 hours?

Exercises on confidence intervals and sample sizes...

*8.3 (a) A sample survey was conducted in a city suburb to determine the mean family income for the area. A random sample of 200 households yielded a mean of R6578. The standard deviation of incomes in the area is known to be R1000. Construct a 95% confidence interval for μ.

(b) Suppose now that the investigator wants to be within R50 of the true value with 99% confidence. What size sample is required?

(c) If the investigator wants to be within R50 of the true value with 90% confidence, what is the required sample size?

8.4 We want to test the strength of lift cables. We know that the standard deviation of the maximum load a cable can carry is 0.73 tons. We test 60 cables and find the mean of the maximum loads these cables can support to be 11.09 tons. Find intervals that will include the true mean of the maximum load with probabilities (a) 0.95 and (b) 0.99.

*8.5 A random variable has standard deviation $\sigma = 4$. The 99% confidence interval for the mean of the random variable was $(19.274\ ,\ 20.726)$. What was the sample mean and what was the size of the sample?

*8.6 During a student survey, a random sample of 250 first year students were asked to record the amount of time per day spent studying. The sample yielded a mean of 85 minutes, with a standard deviation of 30 minutes. Construct a 90% confidence interval for the population mean.

Exercises on hypothesis testing (use the six-point plan) ...

*8.7 The mean height of 10-year-old males is known to be 150 cm and the standard deviation is 10 cm. An investigator has selected a sample of 80 males of this age who are known to have been raised on a protein-deficient diet. The sample mean is 147 cm. At the 5% level of significance, decide whether diet has an effect on height.

8.8 A machine is supposed to produce nuts with a mean diameter of 20 mm and a standard deviation of 0.2 mm. A sample of 40 nuts had a mean of 20.05 mm. Make a decision whether or not the machine needs adjustment, using a 5% significance level.

8.9 Suppose that the time taken by a suburban train to get from Cape Town to Wynberg is a random variable with a $N(\mu,\ 4)$ distribution. 25 journeys take an average of 27.7 minutes. Test the hypothesis $H_0 : \mu = 27$ against the alternatives (a) $H_1 : \mu \neq 27$ and (b) $H_1 : \mu > 27$ at the 5% level of significance.

*8.10 The lifetime of electric bulbs is assumed to be exponentially distributed with a mean of 40 days. A sample of 100 bulbs lasted on average 37.5 days. Using a 1% level of significance, decide whether the true mean is in fact less than 40 days. (Hint: remember that the mean and variance of the exponential distribution are $1/\lambda$ and $1/\lambda^2$ respectively.)

8.11 A mechanized production line operation is supposed to fill tins of coffee with a mean of 500.5 g of coffee with a standard deviation of 0.6 g. A quality control specialist is concerned that wear and tear has resulted in a reduction in the mean. A sample of 42 tins had a mean content of 500.1 g. Use a 1% significance level to perform the appropriate test.

*8.12 (a) A coin is tossed 10 000 times, and it turns up heads 5095 times. Is it reasonable to think that the coin is unbiased? Use 5% significance level.
(b) A coin is tossed 400 times, and it turns up heads 220 times. Is the coin unbiased? Use 5% level.

8.13 The "beta coefficient" is a measure of risk widely used by financial analysts. Larger values for beta represent higher risk. A particular analyst would like to determine whether gold shares are more risky than industrial shares. From past records, it is known that the standard deviation of betas for gold shares is 0.313 and for industrial shares is 0.507. A sample of 40 gold shares had a mean beta coefficient of 1.24 while a sample of 30 industrial shares had a mean of 0.72. Using a 1% level of significance, conduct the appropriate statistical test for the financial analyst.

*8.14 The standard deviation of scores in an I.Q. test is known to be 12. The I.Q.'s of random samples of 20 girls and 20 boys yield averages of 110 and 105 respectively. Use a 1% significance level to test the hypothesis that the I.Q.'s of boys and girls are different.

Exercises on hypothesis testing (use the modified procedure)...

8.15 In an assembly process, it is known from past records that it takes an average of 4.32 hours with a standard deviation of 0.6 hours for a computer part to be assembled. A new procedure is adopted. A random sample of 100 items using the new procedure took, on average, 4.13 hours. Assuming that the standard has remained unaltered, test whether the new procedure is effective in reducing assembly time.

*8.16 It is accepted the standard deviation of the setting time of an adhesive is 1.25 hours under all conditions. It is, however, well known that the adhesive sets more quickly in warm conditions. A new additive is developed that reputedly accelerates setting under freezing conditions. A sample of 50 joints glued with the adhesive, plus additive, had a mean setting time of 4.78 hours, whereas a sample of 30 joints glued without the additive in the adhesive took, on average, 5.10 hours to set. The entire experiment was conducted at 0°C. Is the additive effective in accelerating setting time?

*8.17 A reporter claims that at least 60% of voters are concerned about conservation issues. Doubting this claim, a politician samples 480 voters and of them 275 expressed concern about conservation. Test the reporter's claim.

*8.18 If the average rate of computer breakdowns is 0.05 per hour, or less, the computer is deemed to be operating satisfactorily. However, in the past 240 hours the computer has broken down 18 times. Test the null hypothesis that $\lambda = 0.05$ against the alternative that λ exceeds 0.05.

8.19 The mean grade of ore at a gold mine is known to be 4.40 g with a standard deviation of 0.60 g per ton of ore milled. A geologist has randomly selected 30 samples of ore in a new section of the mine, and determined their mean grade to be 4.12 g per ton of ore. Test whether the ore in the new section of the mine is of inferior quality.

Solutions to exercises...

8.1 0.8554.

8.2 0.1056.

8.3 (a) (6439.4, 6716.6) (b) 2663 (c) 1076.

8.4 (a) (10.91, 11.27) (b) (10.85, 11.33).

8.5 Mean was 20.0, $n = 202$.

8.6 (81.9, 88.1).

8.7 Observed value of test statistic $z = -2.68$ critical value $= -1.64$, reject H_0.

8.8 $z = 1.58 < 1.96$, cannot reject H_0.

8.9 (a) $z = 1.75 < 1.96$, cannot reject H_0. (b) $z = 1.75 > 1.64$, reject H_0.

8.10 $z = -0.63 > -2.33$, cannot reject H_0.

8.11 $z = -4.32 < -2.33$, reject H_0.

8.12 (a) Yes, $z = 1.90 < 1.96$, cannot reject H_0: coin is unbiased.
(b) No, $z = 2.00 > 1.96$, reject H_0.

8.13 $z = 4.95 > 2.33$, reject H_0.

8.14 $z = 1.32 < 2.58$, cannot reject H_0.

8.15 Conclusion: the new procedure reduced assembly time significantly ($z = -3.17$, p-value $= 0.0008$ $(P < 0.001)$).

8.16 Conclusion: the adhesive with the additive did not accelerate setting time significantly ($z = -1.11$, p-value $= 0.1335$ $(P < 0.20)$). Further investigation recommended.

8.17 Use normal approximation to $B(480, 0.6)$. Conclusion: accept the reporter's claim ($z = 1.21$, p-value $= 0.1131 \times 2 = 0.2262$ $(P > 0.20)$) that the proportion does **not** differ significantly from 60%.

8.18 Use normal approximation to $P(12)$. Conclusion: the computer is having breakdowns at a rate significantly higher than 0.05/hr. ($z = 1.73$, p-value $= 0.0418$ $(P < 0.05)$).

8.19 Conclusion: The new section of the mine does involve inferior grade ore ($z = -2.56$, p-value $= 0.0052$, $P < 0.01$).

Chapter 9

THE t- AND F-DISTRIBUTIONS

KEYWORDS: t-distribution, degrees of freedom, pooled sample variance, F-distribution

Population variance unknown...

So far, for confidence intervals for the mean, and for tests of hypotheses about means, we have always had to make the very restrictive assumption that the population variance (or standard deviation) was known. Until the early 1900's, whenever the **population** variance σ^2 was unknown, the **sample** variance s^2 was substituted in its place. This is reasonably satisfactory when the samples are "large enough", and consequently the sample variance lies reliably close to the population variance. We shall see in this chapter that when the sample size exceeds about 30 then it is reasonable to make this substitution. Problems arise with small samples, because the sample variance s^2 is then very variable and can be far removed from the population variance σ^2, which we think of as being the one "true" value.

At the turn of the century, W.S. Gosset, who worked for an Irish brewery, needed to do statistical tests using small samples. This motivated him to tackle the mathematical statistical problem of how to use s^2, the estimate of σ^2, in these tests. He solved the problem. Because of his professional connections, he published the theory he developed under the pen-name "Student". Gosset in 1908 published the actual distribution of

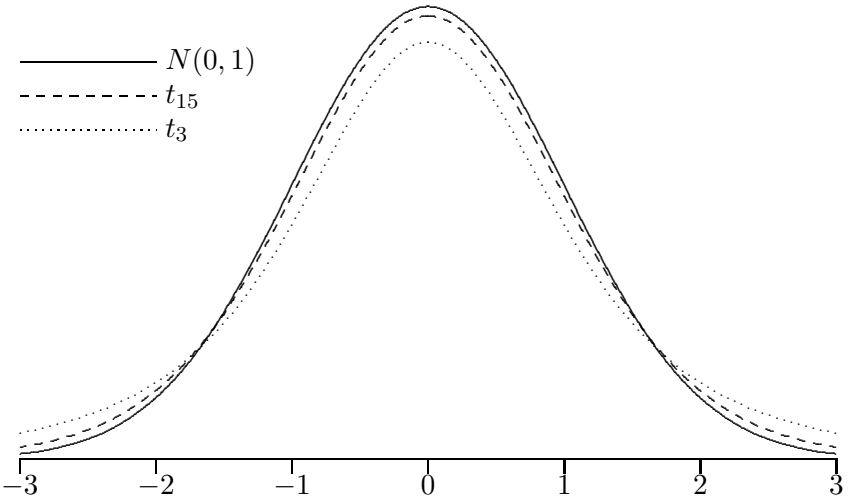
$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

We know, from chapter 8, that $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has the standard normal distribution. But when σ, the single "true" value for the standard deviation in the population, is replaced by s, the estimate of σ, this is no longer true, although, as the sample size n increases, it rapidly becomes an excellent approximation. But for small samples it is far from the truth. This is because the sample variance s^2 (and also s) is itself a random variable — it varies from sample to sample. We will discuss the sampling distribution of s^2 in the next chapter.

We know that the size of the sample influences the accuracy of our estimates. The larger the sample the closer the estimate is likely to be to the true value. Student's t-distribution takes account of the size of the sample from which s is calculated.

The shape of the t-distribution is similar to that of the normal distribution. **However, the shape of the distribution varies with the sample size**. It is longer-

or heavier-tailed than the normal distribution when the sample size is small. As the sample size increases, the t-distribution and normal distribution become progressively closer, and, ultimately, they are identical. The standard normal distribution, and two t-distributions are plotted.



Degrees of freedom...

In order to gain some insight into the notion of "degrees of freedom", consider again the definition of the sample variance:

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

The terms $x_i - \bar{x}$ are the deviations of each of the x_i from the sample mean. To achieve a given sample mean for, say, six numbers, five of these can be chosen at will, but the last is then fully determined. Suppose we are given the information that the mean of six numbers is 5 and that the first five of the six numbers are 4, 9, 5, 7 and 3. The sixth number **must** be 2, otherwise the sample mean would not be 5. It is fixed, it has no "freedom". In general if we are told that the mean of n numbers is $\bar{x}$, and that the first $n-1$ numbers are x_1, x_2, ..., x_{n-1}, then it is easy to see that x_n must be given by

$$x_n = n\bar{x} - x_1 - x_2 \cdots - x_{n-1}.$$

In other words, once we are given $\bar{x}$ and the first $n-1$ of the x_i we have enough information to compute the sample variance. Thus, only $n-1$ of the deviation terms $x_i - \bar{x}$ in s^2 contain real information; the last term is just a formality, but it must be included! We say that s^2, based on a sample of size n, has $n-1$ **degrees of freedom**. (This is part of the reason why the formula for the sample variance s^2 calls for division by $n-1$, and not n.)

We will be encountering the concept of degrees of freedom regularly. We have a simple rule which helps in making decisions about degrees of freedom.

THE DEGREES OF FREEDOM RULE

For each parameter we need to estimate prior to evaluating the current parameter of interest, we lose one degree of freedom.

For example, when we use s^2 to estimate σ^2, we first need to use $\bar{x}$ to estimate μ. So we lose one degree of freedom. There will be further examples later on in which we will "lose" two or more degrees of freedom!

Because the shape of the t-distribution varies with the sample size, there is a whole family of t-distributions. It is therefore necessary to have some means of indicating which t-distribution is being used in a specific situation. We do this by using a subscript. Intuition suggests that the subscript should be the sample size, but it turns out that the sensible subscript is the degrees of freedom of the standard deviation, which is one less than the sample size. So we use the notation

$$t_{n-1} = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

and say that the expression on the right hand side has the t-distribution with $n-1$ degrees of freedom, or simply the t_{n-1}-distribution.

As for the normal distribution, we need tables for looking up values for the t-distribution. The shapes of the t-distributions are dependent on the degree of freedom; thus we cannot get away with a single table for all t-distributions as we did with the normal distribution. We really do appear to need a separate table for each number of degrees of freedom. But we take a short cut, and we only present a selection of key values from each t-distribution in a single table (Table 2). If you think about it, you can now begin to understand why we can do this; even for the normal distribution, we repeatedly only use a handful of values; $z^{(0.05)} = 1.64$, $z^{(0.025)} = 1.96$, $z^{(0.01)} = 2.33$ and $z^{(0.005)} = 2.58$ are by far the most frequently used percentage points of the standard normal distribution. In Table 2, there is one line for each t-distribution; on that line, we present 11 percentage points.

As mentioned previously, exact p-values are usually computed using statistical software. For the z-test we could calculate these exactly since our tables gave us probabilities associated with z-scores. But due to the fact that our t-tables report critical values rather than probabilities, p-values can no longer be computed exactly. To obtain approximate p-values: (1) identify the appropriate degrees of freedom, (2) highlight the relevant line (corresponding to the df) in the t-tables, (3) identify where the test statistic would lie along that line and (4) determine the approximate probability (range) associated with the test statistic by looking at the probabilities corresponding with the two values on either side of the test statistic.

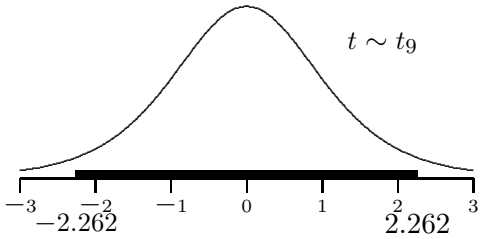
To illustrate, let's assume we observed a t-test statistic of 3.2156 with 12 degrees of freedom. If we look in the line (on the t-table) corresponding to 12 degrees of freedom, we see that our value lies between 3.055 and 3.428. Now the tail probabilities associated with these two values respectively are 0.005 and 0.0025. This implies that the area in the tail associated with the test statistic would lie somewhere between 0.0025 and 0.005. So we have that $0.0025 <$ p-value < 0.005 in this instance (for a upper tail test)! For a two-tailed test, these values need to be multiplied by 2 and the p-value would lie between 0.005 and 0.01.

Confidence intervals (using the sample variance)...

Example 1A: Ten direct flights to Johannesburg from Cape Town took, on average, 103 minutes. The **sample** standard deviation was 5 minutes. Determine a 95% confidence interval for the true mean flight duration.

We have $\bar{x} = 103$, $s = 5$ and $n = 10$. We know that the statistic $(\overline{X} - \mu)/\frac{s}{\sqrt{n}}$ has the t-distribution with 9 degrees of freedom, denoted by t_9.

In Table 2, we see that, for t_9, the points between which 95% of the distribution lies are -2.262 and $+2.262$.



Because $2\frac{1}{2}\%$ (or 0.025) of the t_9 distribution lies to the right of 2.262 we write

$$t_9^{(0.025)} = 2.262$$

and speak of the "$2\frac{1}{2}\%$ point of the t_9 distribution".

We can write

$$\Pr(-2.262 < t_9 < 2.262) = 0.95.$$

But

$$\frac{\overline{X} - \mu}{s/\sqrt{10}} \sim t_9.$$

Therefore

$$\Pr\left(-2.262 < \frac{\overline{X} - \mu}{s/\sqrt{10}} < 2.262\right) = 0.95.$$

Manipulation of the inequalities, as done in the same context in Chapter 8, yields

$$\Pr\left(\overline{X} - 2.262\ \frac{s}{\sqrt{10}} < \mu < \overline{X} + 2.262\ \frac{s}{\sqrt{10}}\right) = 0.95.$$

Substituting $\overline{X} = 103$ and $s = 5$ yields

$$\overline{X} - 2.262\ \frac{s}{\sqrt{10}} = 103 - 2.262\,\frac{5}{\sqrt{10}} = 99.4$$

for the lower limit of the confidence interval and

$$\overline{X} + 2.262\ \frac{s}{\sqrt{10}} = 103 + 2.262\,\frac{5}{\sqrt{10}} = 106.6$$

for the upper limit. Thus a 95% confidence interval for the mean flying time is given by $(99.4,\ 106.6)$.

CONFIDENCE INTERVAL FOR μ, WHEN σ^2 IS ESTIMATED BY s^2

If we have a random sample of size n from a population with a normal distribution, and the sample mean is $\overline{x}$ and the sample variance is s^2, then confidence intervals for μ are given by

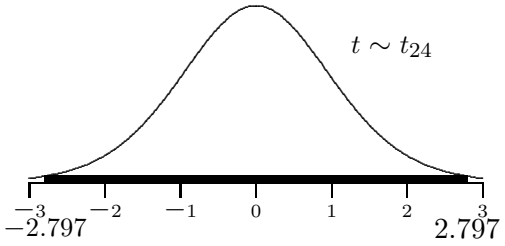
$$\left(\overline{X} - t^*_{n-1}\ \frac{s}{\sqrt{n}}\ ,\ \overline{X} + t^*_{n-1}\ \frac{s}{\sqrt{n}}\right)$$

where the t^* values are obtained from the t-tables. For 95% confidence intervals, use the column in the tables headed 0.025. For 99% confidence intervals, use the column headed 0.005.

Example 2B: A random sample of 25 loaves of bread had a mean mass of 696 g and a standard deviation of 7 g. Calculate the 99% confidence interval for the mean.

We have $\overline{x} = 696$, and $s = 7$ has 24 degrees of freedom. Because we want 99% confidence intervals, we must look up $t_{24}^{(0.005)}$ in the t-tables:

$$t_{24}^{(0.005)} = 2.797.$$



Thus the 99% confidence interval for the mean is

$$(696 - 2.797 \times 7/\sqrt{25}\ ,\ 696 + 2.797 \times 7/\sqrt{25})$$

which reduces to $(692.08, 699.92)$.

Example 3C: An estimate of the mean fuel consumption (litres/100 km) of a new car is required. A sample of 18 drivers each drive the car under a variety of conditions for 100 km, and the fuel consumption is measured for each of the drivers. The sample mean was calculated to be $\bar{x} = 6.73$ litres/100 km. The sample standard deviation was $s = 0.35$ litres/100 km. Find the 99% confidence interval for the population mean.

Testing whether the mean is a specified value (population variance estimated from the sample)...

Example 4A: A poultry farmer is investigating ways of improving the profitability of his operation. Using a standard diet, turkeys grow to a mean mass of 4.5 kg at age 4 months. A sample of 20 turkeys, which were given a special enriched diet, had an average mass of 4.8 kg after 4 months. The sample standard deviation was 0.5 kg. Using the 5% significance level test whether the new diet is effectively increasing the mass of the turkeys.

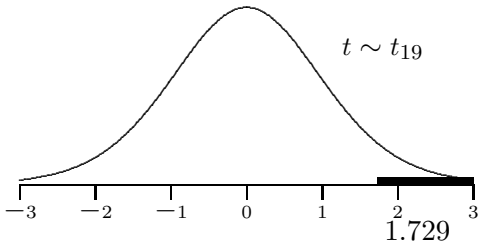
We follow the standard hypothesis testing procedure.

1. $H_0 : \mu = 4.5$. As usual, for our null hypothesis we assume no change has taken place.

2. $H_1 : \mu > 4.5$. A one-sided alternative is appropriate, because an enriched diet should not cause a loss of mass.

3. Significance level: 5%

4. The rejection region is found by reasoning as follows. Because the population variance is unknown, we need to use the t-distribution. The degrees of freedom for t will be 19, because s is based on a sample of 20 observations. We will thus reject H_0 if the "observed t-value" exceeds $t_{19}^{(0.05)}$. From the t-tables, $t_{19}^{(0.05)} = 1.729$.



5. The formula for the test statistic is

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}.$$

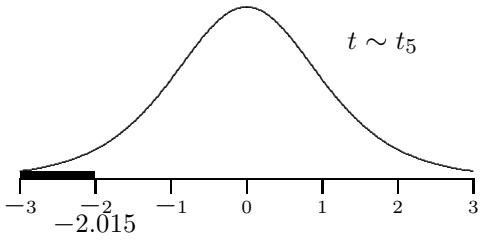
We substitute the appropriate values, and compute

$$t = \frac{4.8 - 4.5}{0.5/\sqrt{20}} = 2.68.$$

6. Because $2.68 > 1.729$ we reject H_0, and conclude that, at the 5% significance level, we have established that the new enriched diet is effective.

Example 5B: The average life of 6 car batteries is 30 months with a standard deviation of 4 months. The manufacturer claims an average life of 3 years for his batteries. We suspect that he is exaggerating. Test his claim at the 5% significance level.

1. $H_0 : \mu = 36$ months.
2. $H_1 : \mu < 36$ months.
3. Significance level : 5%.
4. Degrees of freedom is $6 - 1 = 5$. So we use the t_5-distribution. From the form of the alternative hypothesis and the significance level, we will reject H_0 if the observed t-value is less than $-t_5^{(0.05)}$. From our tables $-t_5^{(0.05)} = -2.015$.



5. Substituting into the test statistic $t_{n-1} = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$ yields

$$t_5 = \frac{30 - 36}{4/\sqrt{6}} = -3.67$$

6. This lies in the rejection region: we conclude that the true mean is significantly less than 36 months.

Example 6C: A purchaser of bricks believes that their crushing strength is deteriorating. The mean crushing strength had previously been 400 kg, but a recent sample of 81 bricks yielded a mean crushing strength of 390 kg, with a standard deviation of 20 kg. Test the purchaser's belief at the 1% significance level.

Example 7C: The specifications for a certain type of ball bearings stipulate a mean diameter of 4.38 mm. The diameters of a sample of 12 ball bearings are measured and the following summarized data computed:

$$\Sigma x_i = 53.18 \qquad \Sigma x_i^2 = 235.7403.$$

Is the sample consistent with the specifications?

Comparing two sample means: matched pairs or paired samples...

We'll introduce the idea of a paired t-test via an example.

Example 8A: Consider a medicine designed to reduce dizziness. We want to know if it is effective. We have 10 patients who complain of dizziness and we examine the reduction in dizzy spells for each patient when taking the medicine. In other words, we ask how many dizzy spells each patient had per month before taking the medication and again while on (after) the medication. Since a specific "before" score can be linked to a specific "after" score, taking a difference in the dizzy spells is a sensible measure of the reduction in dizzy spells. Note the data below, where the difference in dizzy spells have been calculated.

	1	2	3	4	5	6	7	8	9	10
Before (B):	19	18	9	8	7	12	16	22	19	18
After (A):	17	24	12	4	7	15	19	25	16	24
$d = B - A$	2	-6	-3	4	0	-3	-3	-3	3	-6

The two samples have effectively been reduced to one, by taking the difference in scores. Therefore, testing if the two populations ("before" and "after") have the same mean, is equivalent to testing if the (population) mean of the difference scores is 0. So, the two paired samples have been reduced to one sample on which we can perform a one-sample t-test.

The relevant summary statistics obtained from the is $\bar{d} = -1.5$ and $s_d = 3.57$. Performing the hypothesis test assuming a 5% level of significance (using the six-point plan), we have the following:

1. $H_0 : \mu_d = 0$
2. $H_1 : \mu_d > 0$
3. $\alpha = 0.05$
4. The critical value is 1.833
5. If the number of dizzy spells have reduced while on the treatment, the before score should be greater than the after score. And since differences were defined by, $d = B - A$, we'd expect them to be positive. Note that the sample information is not consistent with this hypothesis; hypotheses are determined by what you are interested in finding out, NOT by the data!

$$t_9 = \frac{-1.5 - 0}{3.57/\sqrt{10}} = -1.33$$

Note that n corresponds to the number of observations, and since the test was performed on the difference scores, we count 10 differences.

6. Since the test statistic $-1.33 < 1.833$ we fail to reject the null hypothesis and conclude that the medication isn't effective at reducing dizzy spells.

Note: It is assumed that the population of difference scores is normally distributed with a mean of μ_d. Also, in the case above the p-value associated with the test statistic would have been between 0.8 and 0.9 (clearly resulting in a fail to reject decision).

As already mentioned, in this example we appear to have two sets of data - the "before" data and the "after" data. But were these really two separate samples? If we had 10 randomly selected patients for "before" and then a **different** 10 randomly selected patients for "after" then the answer would be **yes** and the samples would be **independent** of one another. But, since dizziness varies from patient to patient, we looked at the **same** 10 patients and took two readings from each individual. Thus our two sets of data were **dependent** and to perform the test we used a **single sample of differences**. These dependent measures are known as **repeated measures**.

Example 9C: A study was conducted to determine if the pace of music played in a shopping centre influences the amount of time that customers spend in the shopping centre. To test the hypothesis, a sample of 8 customers was observed on two different days. On the one day, the customers did their shopping while slow music was playing. On the same day the next week, they did their shopping while fast music was playing. Other than the pace of the music, circumstances were similar. The following observations were made:

Type of music	Time spent in shopping centre (minutes)							
Slow music (A):	67	52	98	24	55	43	48	70
Fast music (B):	58	44	80	30	48	42	40	63
$d = A - B$	9	8	18	-6	7	1	8	7

The following sample means and standard deviations were observed:

$$\bar{a} = 57.125 \quad s_a = 21.86$$

$$\bar{b} = 50.625 \quad s_b = 15.74$$

$$\bar{d} = 6.5 \quad s_d = 6.87$$

Do the data above provide sufficient evidence that the pace of the music has an influence on the amount of time that shoppers spend in a shopping centre?

Another way for getting dependent samples is when individuals are paired on some criteria to reduce spurious variation in measurements. As a consequence of the latter, the test we used is often called the paired t-test.

Example 10A: Twenty individuals were paired on their initial rate of reading. One of each pair was randomly assigned to method I for speed reading and the other to method II. After the courses the speed of reading was measured. Is there a difference in the effectiveness of the two methods?

Pair	1	2	...	9	10
Method I:	1114	996	...	996	894
Method II:	1032	1148	...	1032	1012
$d = I - II$	82	-152	...	-36	-118

The relevant summary statistics are

$$\bar{d} = -40 \quad s_d = 71.6 \text{ and } n = 10 \text{ pairs}$$

Once again we assume that this is a random representative sample from the population of differences and that the differences in reading speed are normally distributed with mean μ_d.

1. $H_0 : \mu_d = 0$
2. $H_1 : \mu_d \neq 0$
3. $t_9 = \frac{-40-0}{71.6/\sqrt{10}} = -1.77$
4. The p-value for this statistic lies between 0.1 and 0.2 and
5. Therefore we do not to reject the null hypothesis (p-value is too large) and conclude that there isn't a difference in reading speed between the two methods.

Confidence intervals...

As before, it is also possible to construct confidence intervals for these differences. Following the same logic as before and using the fact that

$$\frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \sim t_{n-1},$$

the 95% confidence interval can be constructed by:

$$\Pr\left(-t_{n-1}^{0.025} < \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} < t_{n-1}^{0.025}\right) = 0.95$$

So

$$\Pr\left(\bar{d} - t_{n-1}^{0.025}\frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1}^{0.025}\frac{s_d}{\sqrt{n}}\right) = 0.95$$

Therefore the 95% confidence interval is:

$$\left(\bar{d} - t_{n-1}^{0.025}\frac{s_d}{\sqrt{n}} \;;\; \bar{d} + t_{n-1}^{0.025}\frac{s_d}{\sqrt{n}}\right)$$

Example 10A continued: To compute the 95% confidence interval:

$$\left(-40 - 2.626\frac{71.6}{\sqrt{10}} \;;\; -40 + 2.626\frac{71.6}{\sqrt{10}}\right)$$

Which is equivalent to

$$(-40 - 51.216; -40 + 51.216)$$

And this results in a 95% confidence interval for difference in speed reading methods of -91.22 to 11.22. Note that zero lies in this interval.

Example 8A continued: The 95% confidence interval is:

$$\left(-1.5 - 2.626\frac{3.57}{\sqrt{10}}\ ;\ -1.5 + 2.626\frac{3.57}{\sqrt{10}}\right)$$

Which is equivalent to

$$(-1.5 - 2.554; -1.5 + 2.554)$$

The corresponding 95% confidence interval for difference in dizzy spells per month is -4.05 to 1.05. Note again that zero lies in this interval.

Comparing two independent sample means (variance estimated from the sample)...

When we have small samples from two populations and want to compare their means, the procedure is a little more complex than you might have expected. In Chapter 8, the test statistic for comparing two means was

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

You would anticipate that the test statistic now might be

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

But, unfortunately, mathematical statisticians can show that this quantity does not have the t-distribution. In order to find a test statistic which does have the t-distribution, an additional assumption needs to be made. This assumption is that the population variances in the two populations from which the samples were drawn are equal. The examples below highlight this difference between comparing the means of two populations when the variances are known and when the variances are estimated from the samples.

Example 11A: Two varieties of wheat are being tested in a developing country. Twelve test plots are given identical preparatory treatment. Six plots are sown with Variety 1 and the other six plots with Variety 2 in an experiment in which the crop scientists hope to determine whether there is a significant difference between yields, using a 5% significance level.

The results were:

Variety 1 :	1.5	1.9	1.2	1.4	2.3	and	1.3	tons per plot
Variety 2 :	1.6	1.8	2.0	1.8	2.3			tons per plot

One of the plots planted with Variety 2 was accidently given an extra dose of fertilizer, so the result was discarded. The means and standard deviation are calculated. They are

$$\bar{x}_1 = 1.60 \qquad s_1 = 0.42 \qquad n_1 = 6$$
$$\bar{x}_2 = 1.90 \qquad s_2 = 0.27 \qquad n_2 = 5.$$

We follow the standard hypothesis testing procedure.

1. $H_0 : \mu_1 - \mu_2 = 0$.
2. $H_1 : \mu_1 - \mu_2 \neq 0$ (a two-tailed test).
3. Significance level : 5%.

4. & 5. Before we can find out the rejection region we need to know the "degrees of freedom". The procedure is rather different from that in the test when the population variances were known. Instead of working with the individual variances σ_1^2 and σ_2^2 we **assume that both populations have the same variance** and we **pool** the two individual sample variances s_1^2 and s_2^2 to form a joint estimate of the variance, s^2. This assumption of equal variances is required by the mathematical theory underlying the t-distribution, into which we will not delve. Of course, we do need to check this assumption of equal variances, which we shall discuss later in the chapter (the F-distribution).

The general formula for the **pooled variance** is

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where s_1^2 is based on a sample of size n_1 and s_2^2 is based on a sample of size n_2. In the above example, $n_1 = 6$ and $n_2 = 5$. Therefore

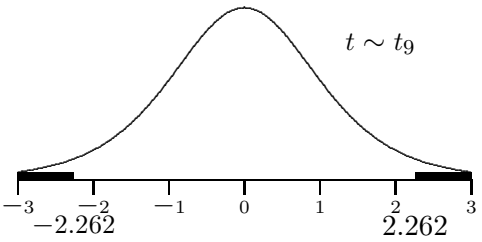
$$\begin{aligned} s^2 &= \frac{5 \times (0.42)^2 + 4 \times (0.27)^2}{6 + 5 - 2} \\ &= 0.13 \end{aligned}$$

Therefore, the pooled standard deviation

$$s = \sqrt{0.13} = 0.361.$$

How many "degrees of freedom" does s have? s_1^2 has 5 and s_2^2 has 4. Therefore s^2 has $5+4 = 9$ degrees of freedom. In general, s^2 has $(n_1-1)+(n_2-1) = n_2+n_2-2$ degrees of freedom. We lose two degrees of freedom because we estimated the two parameters μ_1 (by $\overline{X}_1$) and μ_2 (by $\overline{X}_2$) before estimating s^2.

Thus we use the t-distribution with 9 degrees of freedom, and because we have a two-sided alternative and a 5% significance level we need the value of $t_9^{(0.025)}$, which from the tables is 2.262. So we will reject H_0 if the observed t_9-value is less than -2.262 or greater than 2.262.



The formula for calculating the test statistic in this hypothesis-testing situation, the so-called **two-sample t-test**, is

$$t_{n_1+n_2-2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the sample means, the value of $\mu_1 - \mu_2$ is determined by the null hypothesis, s is the pooled sample standard deviation, and n_1 and n_2 are the sample sizes.

Substituting our data into this formula:

$$t_{6+5-2} = \frac{1.60 - 1.90 - 0}{0.361\sqrt{\frac{1}{6} + \frac{1}{5}}}.$$

Thus

$$t_9 = -1.372.$$

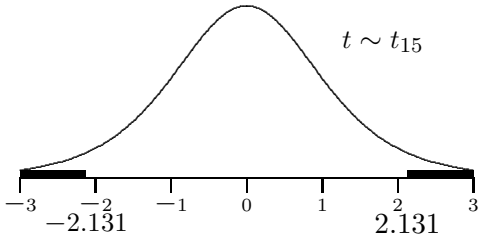
6. Because -1.372 does not lie in the rejection region we conclude that the difference between the varieties is not significant.

Example 12B: A marketing specialist considers two promotions in order to increase sales of do-it-yourself hardware in a supermarket. During a trial period the promotions are run on alternative days. In the first promotion, a free set of drill bits is given if the customer purchases an electric drill. In the second, a substantial discount is given on the drill. The marketing specialist is particularly interested in the average amount spent on do-it-yourself hardware by customers who took advantage of the promotion. On the basis of randomly selected samples, the following data were obtained of the amount spent on items of do-it-yourself hardware.

Promotion	Free gift	Discount
n	8	9
$\overline{x}$	R490	R420
s	R104	R92

Test, at the 5% level of significance, whether there is any difference in the effectiveness of the two promotions.

1. $H_0 : \mu_1 = \mu_2$.
2. $H_1 : \mu_1 \neq \mu_2$.
3. Significance level : 5%.
4. Degrees of freedom : $n_1 + n_2 - 2 = 15$. From t-tables, if the observed t_{15} value exceeds 2.131, we reject H_0.



5. We need first to calculate s^2, the pooled variance:

$$\begin{aligned} s^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{7 \times 104^2 + 8 \times 92^2}{15} \\ &= 9561.60 \end{aligned}$$

and so the pooled standard deviation is

$$s = 97.78.$$

We now calculate the observed test statistic

$$t_{n_1+n_2-2} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t_{15} = \frac{490 - 420 - 0}{97.78\sqrt{\frac{1}{8} + \frac{1}{9}}} = 1.47.$$

6. Because $1.47 < 2.131$, we cannot reject H_0, and we conclude that we cannot detect a difference between the effectiveness of the two promotions.

Example 13C: Two methods of assembling a new television component are under consideration by management. Because of more expensive machinery requirements, method B will only be adopted if it is significantly shorter than method A by more than a minute. In order to determine which method to adopt a skilled worker becomes proficient in both methods, and is then timed with a stopwatch while assembling the component by both methods. The following data were obtained:

Method A	$\overline{x}_1 = 7.72$ minutes	$s_1 = 0.67$ minutes	$n_1 = 17$
Method B	$\overline{x}_2 = 6.21$ minutes	$s_2 = 0.51$ minutes	$n_2 = 25$

What decision should be taken?

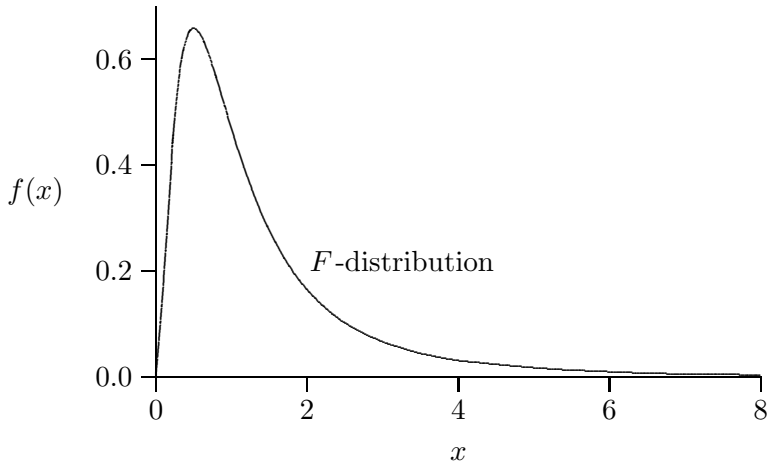
Testing variances for equality...

The t-tests for comparing means of two populations, as introduced in the previous section, required us to assume that the two populations had equal variances.

This assumption can be tested easily. Using the statistical jargon we have developed, what we need is a test of $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$. (Situations do arise when we need one-sided alternatives and the test we will develop can handle both one-sided and two-sided alternatives.)

Because s_1^2 and s_2^2 are our estimates of σ_1^2 and σ_2^2, our intuitive feeling is that we would like to reject our null hypothesis when s_1^2 and s_2^2 are "too far apart". This might suggest the quantity $s_1^2 - s_2^2$ as our **test statistic**. However, we cannot easily find the **sampling distribution** of $s_1^2 - s_2^2$. It turns out that the test statistic which is mathematically convenient to use is the ratio $F = s_1^2/s_2^2$. The sampling distribution of the statistic F is known as the F-distribution, or Fisher distribution. Sir Ronald Fisher was a British statistician who was one of the founding fathers of the discipline of Statistics.

Because variances are by definition positive, the statistic F is always positive. When $H_0 : \sigma_1^2 = \sigma_2^2$ is true, we expect the sample variances to be nearly equal, so that F will be close to one. When H_0 is false, and the population variances are unequal, then F will tend to be either large or small, where in this context small means close to zero. Thus, we accept H_0 for F-values close to one, and we reject H_0 when the F-value is too large or too small. The rejection region is obtained from F-tables, but the shape of probability density function for a typical F-distribution is shown here.



The most striking feature of the probability density function of the F-distribution is that it is **not** symmetric. It is positively skewed, having a long tail to the right. The mode (the x-value associated with the maximum value of the probability density function) is less than one, but the mean is greater than one, the long tail pulling the mean to the right. The lack of symmetry makes it seem that we will need separate tables for the upper and lower percentage points. However, by means of a simple trick (to be explained later), we can get away without having tables for the lower percentage points of the F-distribution.

Because the F-statistic is the ratio of two sample variances, it should come as no surprise to you that there are two degrees of freedom numbers attached to F — the degrees of freedom for the variance in the numerator, and the degrees of freedom for the denominator variance. It would therefore appear that we need an encyclopaedia of tables for the F-distribution! To avoid this, it is usual to only present the four most important values for each F-distribution; the 5%, 2.5%, 1% and 0.5% points. The conventional way of presenting F-tables is to have one table for each of these percentage points; in this book Table 4.1 gives the 5% points, Table 4.2 the 2.5% points, Table 4.3 the 1% points and Table 4.4 the 0.5% points. Within each table, the rows and columns are used for the degrees of freedom in the denominator and numerator, respectively.

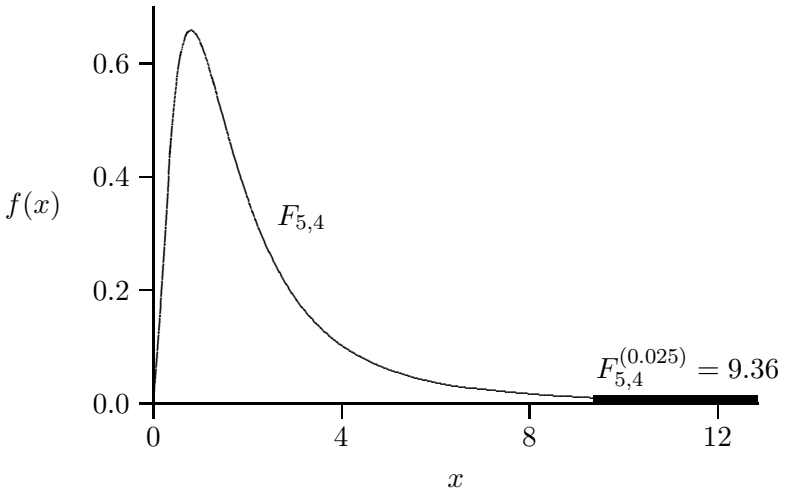
Example 14A: In example 11A, the sample standard deviations were $s_1 = 0.42$ and $s_2 = 0.27$. Let us test at the 5% level to see if the assumption of equal variances was reasonable.

1. $H_0 : \sigma_1^2 = \sigma_2^2$.
2. $H_1 : \sigma_1^2 \neq \sigma_2^2$.
3. Significance level : 5%.
4. The rejection region. If s_1^2, based on a sample of size n_1 (and therefore having $n_1 - 1$ degrees of freedom), is the numerator, and s_2^2 (sample size n_2, degrees of freedom $n_2 - 1$) is the denominator, then we say that $F = s_1^2/s_2^2$ has the F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. We write

$$\frac{s_1^2}{s_2^2} \sim F_{n_1-1,\ n_2-1}.$$

In example 11A, the sample sizes for s_1^2 and s_2^2 were 6 and 5 respectively. Thus we use the F-distribution with $6 - 1$ and $5 - 1$ degrees of freedom, i.e. $F_{5,4}$.

Because we have a two-sided test at the 5% level, we need the upper and lower $2\frac{1}{2}\%$ points of $F_{5,4}$. This means that we must use Table 4.2, and go to the intersection of column 5 and row 4, where we find that the upper $2\frac{1}{2}\%$ point of $F_{5,4}$ is 9.36. We write $F_{5,4}^{(0.025)} = 9.36$. (Notice that, in F-tables, the usual matrix convention of putting rows first, then columns, is **not** adopted.)



We will reject H_0 if our observed F value exceeds 9.36. The tables do not enable us to find the lower rejection region, but for the reasons explained below, we do not in fact need it.

5. The observed F-value is

$$F = s_1^2/s_2^2 = 0.42^2/0.27^2 = 2.42.$$

6. Because $2.42 < 9.36$, we do not reject H_0. We conclude that the assumption of equal variances is tenable, and that therefore it was justified to pool the variances for the two-sample t-test in example 11A.

The trick that enables us never to need lower percentage points of the F-distribution is to adopt the convention of always putting the numerically larger variance into the numerator — so that the calculated F-statistic is always larger than one — and adjusting the degrees of freedom. Let s_1^2 and s_2^2 have $n_1 - 1$ and $n_2 - 1$ degrees of freedom respectively. Then, if $s_1^2 > s_2^2$, consider the ratio $F = s_1^2/s_2^2$ which has the F_{n_1-1,n_2-1}-distribution. If $s_2^2 > s_1^2$, use $F = s_2^2/s_1^2$ with the F_{n_2-1,n_1-1}-distribution. This trick depends on the mathematical result that $1/F_{n_1-1,n_2-1} = F_{n_2-1,n_1-1}$.

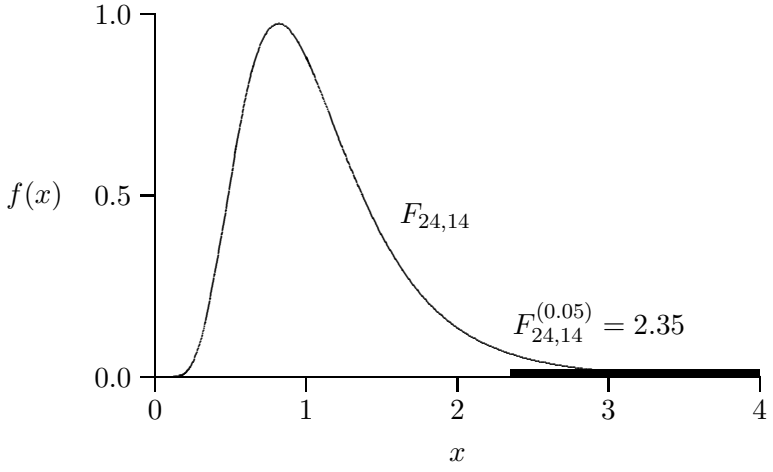
Example 15B: We have two machines that fill milk bottles. We accept that both machines are putting, on average, one litre of milk into each bottle. We suspect, however, that the second machine is considerably less consistent than the first, and that the volume of milk that it delivers is more variable. We take a random sample of 15 bottles from the first machine and 25 from the second and compute sample variances of 2.1 ml^2 and 5.9 ml^2 respectively. Are our suspicions correct? Test at the 5% significance level.

1. $H_0 : \sigma_1^2 = \sigma_2^2$.
2. $H_1 : \sigma_1^2 < \sigma_2^2$.
3. Significance level : 5% (and note that we now have a one-sided test).

4. & 5. Because $s_2^2 > s_1^2$, we compute

$$F = s_2^2/s_1^2 = 5.9/2.1 = 2.81.$$

This has the $F_{n_2-1,n_1-1} = F_{24,14}$-distribution. From our tables, $F_{24,14}^{(0.05)} = 2.35$



6. The observed F-value of $2.81 > 2.35$ and therefore lies in the rejection region.

   We reject H_0 and conclude that the second bottle filler has a significantly larger variance (variability) than the first.

Example 16C: Packing proteas for export is time consuming. A florist timed how long it took each of 12 labourers to pack 20 boxes of proteas under normal conditions, and then timed 10 labourers while they each packed 20 boxes of proteas with background music. The average time to pack 20 boxes of proteas under normal conditions was 170 minutes with a standard deviation of 20 minutes, while the average time with background music was 157 minutes, with a standard deviation of 25 minutes. At the 5% significance level, test whether background music is effective in reducing packing time. Test also the assumption of equal variances.

An approximate t-test when the variances cannot be assumed to be equal...

The F-test described above ought always to be applied before starting the t-test to compare the means. It is conventional to use a 5% significance level for this test. If we do not reject the null hypothesis of equal variances, we feel justified in computing the pooled sample variance which the two-sample t-test requires. What happens, however, if the F-test forces us to reject the assumption that the variances are equal? We resort to an **approximate** t-test, which does not pool the variances, but which makes an adjustment to the degrees of freedom.

The test statistic is

$$t^* = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}.$$

This statistic has **approximately** the t-distribution. By experimentation, researchers have determined that the degrees of freedom for t^* can be approximated by

$$n^* = \left\{ (s_1^2/n_1 + s_2^2/n_2)^2 \Big/ \left[\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1} \right] \right\} - 2.$$

This messy formula inevitably gives a value for n^* which is not an integer. It is feasible to interpolate in the t-tables, but we will simply take n^* to be the nearest integer value to that given by the formula above.

Example 17B: Personnel consultants are interested in establishing whether there is any difference in the mean age of the senior managers of two large corporations. The following data gives the ages, to the nearest year, of a random sample of 10 senior managers, sampled from each corporations:

Corporation 1	52	50	53	42	57	43	52	44	51	34
Corporation 2	44	45	39	49	43	49	45	47	42	46

Conduct the necessary test.

We compute

$$\bar{x}_1 = 47.80 \qquad s_1 = 6.86$$
$$\bar{x}_2 = 45.00 \qquad s_2 = 3.20$$

We first test for equality of variances:

1. $H_0 : \sigma_1^2 = \sigma_2^2$.
2. $H_1 : \sigma_1^2 \neq \sigma_2^2$.
3. $F = s_2^2/s_2^2 = 6.86^2/3.20^2 = 4.60$.
4. Using the $F_{9,9}$-distribution, and remembering that the test is two-sided, we see that this F-value is significant at the 5% level ($F_{9,9}^{(0.025)} = 4.03$), although not significant at the 1% level ($F_{9,9}^{(0.005)} = 6.54$).
5. The conclusion is that the population variances are not equal ($F_{9,9} = 4.60,\ P < 0.05$).

Thus pooling the sample variances is not justified, and we have to use the approximate t-test:

1. $H_0 : \mu_1 = \mu_2$.
2. $H_1 : \mu_1 \neq \mu_2$.
3. Substituting into the formula for the approximate test statistic yields

$$t^* = \frac{(47.8 - 45.0)}{\sqrt{\dfrac{6.86^2}{10} + \dfrac{3.20^2}{10}}} = 1.17.$$

Substituting into the degrees of freedom formula yields

$$n^* = \left\{ \left(\frac{6.86^2}{10} + \frac{3.20^2}{10} \right)^2 \Big/ \left[\frac{(6.86^2/10)^2}{11} + \frac{(3.20^2/10)^2}{11} \right] \right\} - 2$$
$$= 13.57 \approx 14$$

4. Because $t_{14}^{(0.100)} = 1.345$, we accept the null hypothesis even at the 20% significance level.

5. We conclude that there is no difference in the mean age of senior employees between the two corporations ($t_{14} = 1.17,\ P > 0.20$).

Example 18C: A particular business school requires a satisfactory GMAT examination score as its entrance requirement. The admissions officer believes that, on average, engineers have higher GMAT scores than applicants with an arts background. The following GMAT scores were extracted from a random sample of applicants with engineering and arts backgrounds.

Engineering	600	650	640	720	700	620	740	650
Art	550	450	700	420	750	500	520	

Investigate the admission officer's belief.

Example 19C: The dividend yield of a share is the dividend paid by the share during a year divided by the price of the share. A financial analyst wants to compare the dividend yield of gold shares with that of industrial shares listed on the Johannesburg Stock Exchange. She takes a sample of gold shares and a sample of industrial shares and computes the dividend yields. What conclusions did she come to?

Gold	3.6	4.0	3.9	5.0	2.7	3.7	4.6	3.5	4.5	3.5	3.9	3.7
shares	4.6	4.0	3.2	15.5	5.7	3.6	4.1	6.2				
Industrial	3.2	2.5	8.4	8.7	2.7	3.1	5.3	4.3	5.6	4.0	5.1	6.8
shares	2.5	8.4	5.2	4.7	3.1	5.5	6.5	3.6	4.5	1.6	3.1	

The data are summarized as:

	Gold	Industrial
n	20	23
$\overline{x}$	4.675	4.710
s	2.677	2.004

A IMPORTANT FOOTNOTE TO t-TESTS AND F-TESTS...

Whenever the t-test or F-test is applied, it is assumed that the population from which the sample was taken has a normal distribution. Can we test this assumption? And how should we proceed if the test shows that the distribution of the data is not a normal distribution?

The first of these questions can be answered using the methods of chapter 10. The answer to the second question is that methods for doing tests for non-normal data do exist, but are beyond the scope of this book. For the record, they are called **non-parametric** tests — the tests using the normal distribution and the t-distribution are known as **parametric** tests.

Another question. What should we do when there are three (or more) populations which we want to compare simultaneously? In example 18C, we might have wanted to do a comparison between students with engineering, arts and science backgrounds, and

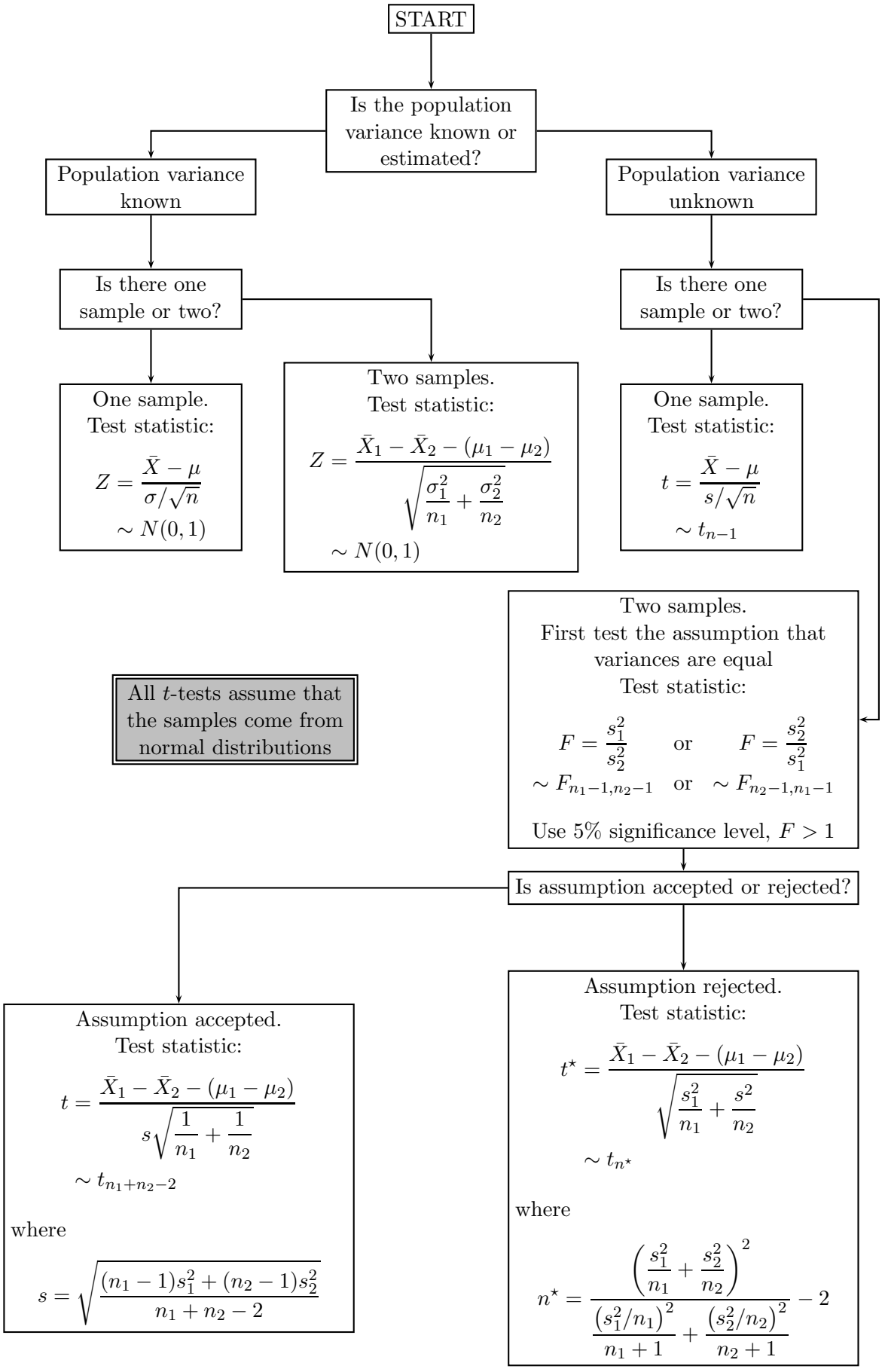


Figure 9.1: Decision Tree for Hypothesis Testing on Means

to test the null hypothesis that there is no difference between the mean GMAT scores for these three groups of students. The method to use then is called the analysis of variance, usually abbreviated to ANOVA.

A summary of hypothesis testing on means...

The **decision tree** displayed in Figure 9.1 and the comments below aim to give clear guidelines to help you decide which test to apply, and presents the formulae for all the test statistics of Chapters 8 and 9.

If the sample is large, greater than 30 say, then the central limit theorem applies and $\overline{X}$ has a normal distribution. If the sample is smaller than 30, then $\overline{X}$ will have a normal distribution if the population from which the sample is drawn has a normal distribution. If the sample is small, and the underlying population does not have a normal distribution, these techniques are not valid, and "non-parametric" statistical tests must be applied. These tests fall beyond the scope of our course.

If the sample variance is estimated from a **large** sample, then it is common practice to assume that $\sigma^2 = s^2$ and to use the normal distribution (i.e. the methods of Chapter 8, where the **population** variance was assumed known). In any case, the t-distribution is nearly identical to the normal distribution for **large** degrees of freedom, and the percentage points are almost equal. We will adopt the convention that for degrees of freedom 30 or fewer the t-distribution is used, between 30 and 100 use of both the t- and the normal distribution is acceptable, and over 100 the normal distribution is mostly used. This note only applies **if the variance is estimated**.

Example 20B: Use the decision tree to decide which test to apply. An experiment compared the abrasive wear of two different laminated materials. Twelve pieces of material 1 and 10 pieces of material 2 were tested and in each case the depth of wear was measured. The results were as follows:

Material 1	$\overline{x}_1 = 8.5$ mm	$s_1 = 4$ mm	$n_1 = 12$
Material 2	$\overline{x}_2 = 8.1$ mm	$s_2 = 5$ mm	$n_2 = 10$

Test the hypothesis that the two types of material exhibit the same mean abrasive wear at the 1% significance level.

Begin at START

The population variance is estimated. Go right. There are samples from two populations. Go right again. The assumption that the variances are equal is accepted. Check this for yourself. Go left.

Pool the variances and use the test statistic with $n_1 + n_2 - 2$ degrees of freedom. Do this example as an exercise.

Example 21C: In an assembly process, it is known from past records that it takes an average of 3.7 hours with a standard deviation of 0.3 hours to assemble a certain computer component. A new procedure is adopted. The first 100 items assembled using the new procedure took, on average, 3.5 hours each. Assuming that the new procedure did not alter the standard deviation, test whether the new procedure is effective in reducing assembly time.

Solutions to examples ...

3C $6.73 \pm 2.898 \times 0.35/\sqrt{18}$, which is $(6.49, 6.97)$

6C $t_{80} = -4.50 < -2.374$, reject H_0.

7C $t_{11} = 2.34,\ P < 0.05$, significant difference.

9C $t_7 = 2.68,\ P < 0.05$, significant difference.

13C $t_{40} = 2.80,\ P < 0.005$, significant. Adopt new method.

16C $F_{9,11} = 1.56 < 3.59$, cannot reject H_0 and pool variances. $t_{20} = 1.36 < 1.725$, cannot reject H_0.

18C $F = 6.28 > F_{6,7}^{(0.025)} = 5.12$. Therefore cannot pool variances. $t^\star = 2.18$, $n^\star = 8.2$, so use degrees of freedom 8. Using one-sided test, there is a significant difference ($P < 0.05$).

19C $F_{19,22} = 1.784$, so variances can be pooled. Two-sample t-test yields $t_{41} = -0.0244$, no significant difference. Dividend yield on gold shares not significantly different from that on industrial shares ($t_{41} = -0.0244,\ P > 0.20$).

21C Path through flow chart: start, standard deviation known, go left, one sample, go left again, and use test statistic $z = (\overline{X} - \mu)/\frac{\sigma}{\sqrt{n}} \sim N(0, 1)$.

Exercises on confidence intervals ...

9.1 Find the 95% confidence interval for the mean salary of teachers if a random sample of 16 teachers had a mean salary of R12 125 with a standard deviation of R1005.

*9.2 We want to estimate the mean number of items of advertising matter received by medical practitioners through the post per week. For a random sample of 25 doctors, the sample mean is 28.1 and the sample standard deviation is 8. Find the 95% confidence limits for the mean.

9.3 Over the past 12 months the average demand for sulphuric acid from the stores of a large chemical factory has been 206 ℓ; the sample standard deviation has been 50 ℓ. Find a 99% confidence interval for the true mean monthly demand for sulphuric acid.

*9.4 A sample of 10 measurements of the diameter of a brass sphere gave mean $\overline{x} = 4.38$ cm and standard deviation $s = 0.06$ cm. Calculate (a) 95% and (b) 99% confidence intervals for the actual diameter of the sphere. Why is the second confidence interval longer than the first?

Exercises on hypothesis testing (one sample)...

*9.5 In a textile manufacturing process, the average time taken is 6.4 hours. An innovation which, it is hoped, will streamline the process and reduce the time, is introduced. A series of 8 trials used the modified process and produced the following results:

6.1 5.9 6.3 6.5 6.2 6.0 6.4 6.2.

Using 5% significance level, decide whether the innovation has succeeded in reducing average process time.

9.6 In 1989, the Johannesburg Stock Exchange (JSE) boomed, and the Allshare Index showed an annual return of 55.5%. A sample of industrial shares yielded the following returns:

54.5	47.8	42.3	59.8	33.1	49.7	16.0	50.3
52.8	56.1	23.2	52.5	65.7	47.5	32.5	46.7

Test, at the 5% significance level, whether the performance of industrial shares lagged behind the market as a whole.

9.7 The mean score on a standardized psychology test is supposed to be 50. Believing that a group of psychologists will score higher (because they can "see through" the questions), we test a random sample of 11 psychologists. Their mean score is 55 and the standard deviation is 3. What conclusions can be made?

*9.8 The specification quoted by ABC Alloys for a particular metal alloy was a melting point of 1660° C. Fifteen samples of the alloy, selected at random, had a mean melting point of 1648° C with a standard deviation of 45° C. Is the melting point lower than specified?

Exercises using F-test for equality of variances ...

9.9 If independent random samples of size 10 from two normal populations have sample variances $s_1^2 = 12.8$ and $s_2^2 = 3.2$, what can you conclude about a claim that the two populations have the same variance? Use a 5% significance level.

*9.10 From a sample of size 13 the estimate of the standard deviation of a population was calculated as 4.47 and from a sample of size 16 from another population the standard deviation was 8.32. Can these populations be considered as having equal variances? Use a 5% level of significance.

EXERCISES ON HYPOTHESIS TESTING (TWO SAMPLES) ...

9.11 The national electricity supplier claims that switching off the hot water cylinder at night does not result in a saving of electricity. In order to test this claim a newspaper reporter obtains the co-operation of 16 house owners with similar houses and salaries. Eight of the selected owners switch their cylinders off at night. The consumption of electricity in each house over a period of 30 days is measured; the units are kWh (kilowatt-hours). The following data are collected:

OFF GROUP	ON GROUP
$n_1 = 8$	$n_2 = 8$
$\bar{x}_1 = 680$	$\overline{x}_2 = 700$
$s_1^2 = 450$	$s_2^2 = 300$

Test, at the 5% level, whether there is a significant saving in electricity if the cylinder is switched off at night. Test the assumption that the variances are equal, also at the 5% level.

*9.12 A comparison is made between two brands of toothpaste to compare their effectiveness at preventing cavities. 25 children use Hole-in-None and 30 children use Fantoothtic in an impartial test. The results are as follows:

Sample size	24	30
Average number of new cavities	1.6	2.7
Standard deviation	0.7	0.9

At the 1% level of significance, investigate whether one brand is better than the other.

9.13 A company claims that its light bulbs are superior to those of a competitor on the basis of a study which showed that a sample of 40 of its bulbs had an average lifetime of 522 hours with a standard deviation of 28 hours, while a sample of 30 bulbs made by the competitor had an average lifetime of 513 hours with a standard deviation of 24 hours. Test the null hypothesis $\mu_1 - \mu_2 = 0$ against a suitable one-sided alternative to see if the claim is justified.

*9.14 The densities of sulphuric acid in two containers were measured, four determinations being made on one and six on the other. The results were:

(1) 1.842 1.846 1.843 1.843

(2) 1.848 1.843 1.846 1.847 1.847 1.845

Do the densities differ at the 5% significance level,

(a) if there is no reason beforehand to believe that there is any difference in density between the containers?

(b) if we have good reason to suspect that, if there is any difference, the first container will be less dense than the second?

9.15 A teacher used different teaching methods in two similar statistics classes of 35 students each. Each class then wrote the same examination. In one class, the mean was $\overline{x}_1 = 82\%$ with $s_1 = 3\%$. In the other class, the results were $\overline{x}_2 = 77\%$ and $s_2 = 7\%$. Test to see if this provides the teacher with evidence that one teaching method is superior to the other. Use a 5% significance level.

9.16 Two drivers, A and B, do fuel consumption tests on a single car. The cars are refuelled every 100 km, and the number of litres required to refill the tank measured. Drivers A and B drive 2100 km and 2800 km, respectively, and so are refuelled 21 and 28 times. Driver B has recently read a pamphlet entitled *Fuel Economy Tips*, and has been putting these ideas into practice. The results are summarized in the table below:

Driver	Sample size	Sample mean (ℓ/100 km)	Sample standard deviation (ℓ/100 km)
A	21	9.03	1.73
B	28	8.57	0.89

Is Driver B more economical than Driver A?

Exercises using normal, t- and F-distributions...

9.17 The following statistics were calculated from random samples of daily sales figures for two departments of a large store:

Department	Hardware	Crockery
Sample size	15	15
Mean daily sales (rands)	1400	1250
Standard deviation (rands)	180	120

The sales manager feels that mean sales in hardware is significantly higher than in crockery. Test this idea statistically using a 1% significance level.

*9.18 Travel times by road between two towns are normally distributed; a random sample of 16 observations had a mean of 30 minutes and a standard deviation of 5 minutes.

(a) Find a 99% confidence interval for the mean travel time.

(b) Estimate how large a sample would be needed to be 95% sure that the sample mean was within half a minute of the population mean. You need to modify the formula for estimating sample sizes in Chapter 8 to take account of the fact that you are given a **sample** standard deviation based on a small sample rather than the **population** standard deviation.

9.19 An insurance company has found that the number of claims made per year on a certain type of policy obeys a Poisson distribution. Until five years ago, the rate of claims averaged 13.1 per year. New restrictions on the acceptance of this type of insurance were introduced five years ago, and since then 51 claims have been made.

(a) Test whether the restrictions have been effective in reducing the number of claims.

(b) Find an approximate 95% confidence interval for the average claim rate under the new restrictions.

9.20 A new type of battery is claimed to have two hours more life than the standard type. Random samples of new and standard batteries are tested, with the following summarized results:

	Sample size	Sample mean (hours)	Sample standard deviation (hours)
New	94	39.3	7.2
Standard	42	36.8	6.4

Test the claim at the 5% significance level.

*9.21 The mean commission of floor-wax salesmen has been R6000 per month in the past. New brands of wax are now providing stiffer competition for the salesmen, but inflation has pushed up the commission per sale. Management wishes to test whether the figure of R6000 per month still prevails, and examines a sample of 120 recent monthly commission figures. They are found to have a mean of R5850 and a standard deviation of R800. Test at a 5% significance level whether the mean commission rate has changed.

9.22 Two bus drivers, M and N, travel the same route. Over a number of journeys the times taken by each driver to travel from bus stop 5 to bus stop 19 were noted. The summarized results are presented below:

Driver	Trips	Mean (minutes)	Standard Deviation (minutes)
M	12	18.1	1.9
N	21	21.3	3.9

(a) Test whether driver M is more consistent in journey times than driver N.
(b) Test whether there is a significant difference in the times taken by each driver.

9.23 It is necessary to compare the precision of two brands of detectors for measuring mercury concentration in the air. The brand B detector is thought to be more accurate than the brand A detector. Seven measurements are made with a brand A instrument, and six with a brand B instrument one lunch hour. The results (micrograms per cubic metre) are summarized as follows:

$$\bar{x}_A = 0.87 \qquad s_A = 0.019$$
$$\bar{x}_B = 0.91 \qquad s_B = 0.008$$

At the 5% significance level, do the data provide evidence that brand B measures more precisely than brand A?

9.24 Show that if we test at the 5% significance level (using either the t- or normal distributions) the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, we will reject H_0 if and only if μ_0 lies outside the 95% confidence interval for μ.

*9.25 The health department wishes to determine if the mean bacteria count per ml of water at Zeekoeivlei exceeds the safety level of 200 per ml. Ten 1 mℓ water samples are collected. The bacteria counts are:

225 210 185 202 216 193 190 207 204 220

Do these data give cause for concern?

*9.26 A normally distributed random variable has standard deviation $\sigma = 5$. A sample was drawn, and the 95% confidence interval for the mean was calculated to be $(70.467, 73.733)$. The experimenter subsequently lost the original data. Tell him

(a) what the sample mean of the original data was, and
(b) what size sample he drew.

9.27 The following are observations made on a normal distribution with mean μ and variance σ^2:

50 53 47 51 49

(a) Find 95% and 99% confidence intervals for μ
  (i) assuming $\sigma^2 = 5$
  (ii) assuming σ^2 is unknown, and has to be estimated by s^2.
(b) Comment on the relative lengths of the confidence intervals found in (a).

Solutions to exercises ...

9.1 (11 590, 12 660)

9.2 $(24.80, 31.40)$

9.3 (161, 251)

9.4 (a) $(4.335, 4.425)$ (b) $(4.315, 4.445)$
The higher the level of confidence, the wider the interval.

9.5 $t_7 = -2.83 < -1.895$, reject H_0.

9.6 $t_{15} = -2.97 < -1.753$, reject H_0, industrial shares performed worse than the All-Share Index.

9.7 $t_{10} = 5.53$, $P < 0.0005$, very highly significant.

9.8 $t_{14} = -1.03$, $P > 0.10$, at this stage cannot reject H_0, the evidence is consistent with the specifications, but further investigation is warranted.

9.9 $F_{9,9} = 4.0 < 4.03$, cannot reject H_0.

9.10 $F_{15,12} = 3.46 > 3.18$, reject H_0.

9.11 $F_{7,7} = 1.50 < 4.99$, cannot reject H_0 and thus pooling of variances justified.
$t_{14} = -2.066 < -1.761$, reject H_0.

9.12 $F_{29,24} = 1.65 < 2.22$, cannot reject H_0, pool variances.
$t_{53} = -4.98 < -2.672$, reject H_0.

9.13 $F_{39,29} = 1.36 < 1.79$ (approx by F(40,30) at 5% level), cannot reject H_0 and pool variances.
$t_{68} = 1.41,\ 0.05 < P < 0.10$, insignificant.

9.14 $F_{3,5} = 1.07 < 7.76$, cannot reject H_0 and pool variances.
(a) $t_8 = 2.19 < 2.306$, cannot reject H_0.
(b) $t_8 = 2.19 > 1.860$, reject H_0.

9.15 $F_{34,34} = 5.44 > 2.30$ (approx by F(30,34) at 1% level), reject H_0 at 1% level. Variances cannot be pooled. Degrees of freedom
$n^* = 46.8 \approx 47.\ t^* = 3.88 > 2.01$,
reject H_0.

9.16 $F_{20,27} = 3.78,\ P < 0.01$, significant, so variances cannot be pooled. Degrees of freedom
$n^* = 28.7 \approx 29.\ t^* = 1.11,\ P < 0 > 20$,
insignificant.

9.17 $F_{14,14} = 2.25 < 2.48$, cannot reject H_0 and hence pool the variances. $t_{28} = 2.69 > 2.467$, reject H_0.

9.18 (a) (26.32 , 33.68)
(b) $n = (t^{\star}_{m-1} s/L)^2 = (2.131 \times 5/\frac{1}{2})^2 = 455$, where m is the size of the sample used for estimating s (in this case 16), so the that degrees of freedom for t is 15.

9.19 (a) $z = -1,79,\ P < 0.05$, significant reduction.
(b) (7.4 , 13.0).

9.20 $F_{93,41} = 1.27 < 1.53$ (approx by F(90,40) at 5% level), cannot reject H_0 and pool variances ($s = 6.965$).
$t_{134} = 0386 < 1.656$, cannot reject H_0.

9.21 $t_{119} = -2.05 < -1.98$, reject H_0.

9.22 (a) $F_{20,11} = 4.214,\ P < 0.05$, significant differences in variances. Pooling not justifiable.
(b) $n^* = 32,\ t^* = -3.16,\ P < 0.005$, significant difference in means.

9.23 $F_{6,5} = 5.64 > 4.95$, reject H_0.

9.25 $t_9 = 1.25,\ P < 0.20$, insignificant.

9.26 (a) $\bar{x} = 72.1$ (b) $n = 36$.

9.27 (a) The confidence intervals are:

	95%	99%
(i)	(48.04 , 51.96)	(47.42 , 52.48)
(ii)	(47.22 , 52.78)	(45.40 , 54.60)

(b) 99% confidence intervals are wider than 95% confidence intervals. If σ^2 is estimated by s^2, the confidence interval is wider.

Chapter 10

THE CHI-SQUARED DISTRIBUTION

KEYWORDS: χ^2-distribution, goodness of fit tests, observed and expected frequencies, contingency tables, tests of association, tests and confidence intervals for the variance.

Many seemingly unrelated applications...

There is a surprisingly large number of tests of hypotheses for which the sampling distribution of the test statistic has a χ^2-distribution, either exactly or approximately. In terms of numbers of applications in Statistics, the chi-squared distribution is probably in second place, after the normal distribution.

We consider here only three of these applications of the chi-squared distribution — "goodness of fit tests", "tests of association in contingency tables", and "tests and confidence intervals for the sample variance". Although the rationale behind the first two applications is completely different, the calculation of the test statistics is almost identical.

Goodness of fit tests...

In chapter 9, to do the two-sample t-test, we found that we had to make the assumption that the variances were equal, and then immediately provided a method for testing this assumption, the F-test. At the very end of the chapter, we said that another assumption underlying all t-tests was that the samples were drawn from populations which had normal distributions. There is therefore a need to be able to test this assumption. The test described in this section enables us to do just this. Not only can we test if the process that generated the data has a normal distribution, we can test whether data fits **any** specified distribution. In our first example, we test whether the process generating misprints in the pages of newspapers produces data with a Poisson distribution.

Example 1A: In the printing industry, it is generally thought that misprints occur at random — i.e. they occur independently of each other. Thus the number of misprints per page can be expected to obey a Poisson distribution with some parameter λ. To test this hypothesis, 200 pages from newspapers were examined, and the number of misprints on each page noted. The data are presented below. A 5% significance level is to be used.

Number of misprints	Observed number of pages
0	43
1	69
2	53
3	21
4	8
5	6

The table tells us that 43 of the 200 pages examined were free of misprints, 69 of the pages had one misprint, 53 pages had two misprints each (a total of 106 misprints on these 53 pages), ..., 6 pages had five misprints on each page (30 misprints on these 6 pages).

In order to test whether the Poisson distribution fits this data, the first problem is to decide what value to use for the parameter λ of the Poisson distribution. This can either be specified by the null hypothesis, or the data can be used to estimate λ. We treat these two situations separately.

Let us first consider the test of the null hypothesis that the data can be thought of as a sample from a Poisson distribution with parameter $\lambda = 1.2$ misprints per page against the alternative that the data come from some other distribution.

Thus we have

1. H_0 : Data come from a Poisson distribution with $\lambda = 1.2$
2. H_1 : The distribution is not Poisson with $\lambda = 1.2$
3. Significance level : 5%

4. & 5. We need firstly to compute the expected (theoretical) frequencies, assuming that the null hypothesis is true. If misprints are occurring in accordance with a Poisson distribution with rate $\lambda = 1.2$ then the probability that a page contains x misprints is

$$p(x) = e^{-1.2}1.2^x/x!$$

Thus the probability of no misprints is

$$p(0) = e^{-1.2} = 0.3012$$

Thus 30.12% of pages are expected to have no misprints. A sample of 200 pages can therefore be expected to have 60.24 pages with no misprints (30.12% of 200). This theoretical frequency is to be compared with an observed frequency of 43.

Similarly $p(1) = e^{-1.2} \times 1.2 = 0.3614$. Thus the expected frequency of one error is $200 \times 0.3614 = 72.28$. We compare this with an observed 69 pages with one error.

Continuing in this way we build up a table of observed and expected frequencies.

number of misprints (x)	observed number of pages (O_i)	theoretical probability $p(x)$	expected frequency (E_i)
0	43	$1.2^0 \times e^{-1.2}/0! = 0.3012$	60.24
1	69	$1.2^1 \times e^{-1.2}/1! = 0.3614$	72.28
2	53	$1.2^2 \times e^{-1.2}/2! = 0.2169$	43.38
3	21	$1.2^3 \times e^{-1.2}/3! = 0.0867$	17.34
4	8	$1.2^4 \times e^{-1.2}/4! = 0.0260$	5.20
5 or more	6	$1 - \sum_{x=0}^{4} p(x) = 0.0078$	1.56

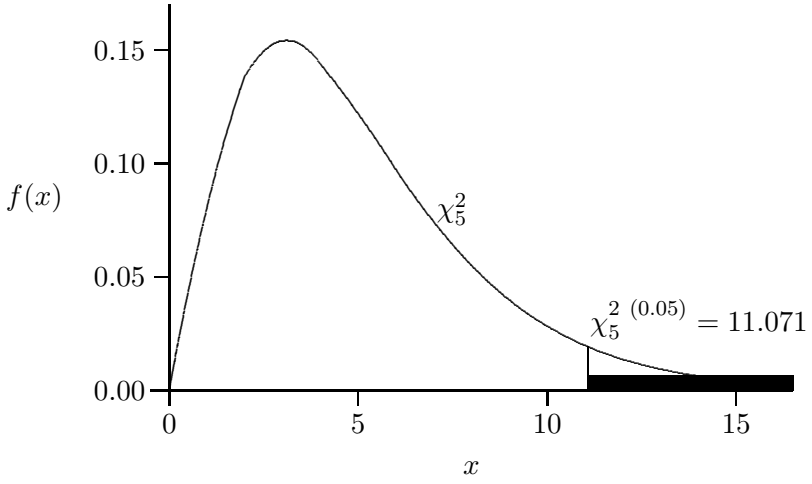
Even if H_0 is true, we anticipate that the observed and expected frequencies will not be exactly equal; this is because we expect some sampling fluctuation in our random sample of 200 pages. We would clearly like to reject H_0, however, if the difference between the observed frequencies and the expected frequencies is "too large".

We need to find a test statistic which is a function of the differences between observed and expected frequencies and which has a known sampling distribution. The sum of these differences is of no use because they sum to zero. So we square the differences and they all become positive. A difference of 3 is negligible if the observed and expected frequencies are 121 and 124, but it important if the frequencies are 6 and 9. To take account of this we divide the squared differences by their **expected frequency**. The "right" statistic to use is

$$D^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where the sum is taken over all the **cells** in the table.

The statistic D^2 has approximately a chi-squared (χ^2) distribution. Like the t- and F-distributions, the χ^2 distribution has a degrees of freedom number attached to it. In tests like the one above, the correct degrees of freedom is given by $k-1$, where k is the number of "cells" into which the data are categorized. Here $k = 6$, and therefore the degrees of freedom for χ^2 is 5. Thus we will reject H_0 if D^2 exceeds the 5% point of the χ^2_5 distribution. From our tables $\chi^{2(0.05)}_5 = 11.071$.



If D^2 *exceeds* 11.071 then the observed and expected frequencies are "too far" apart for their differences to be explained by chance sampling fluctuations alone.

Let us compute D^2 for the above data.

$$D^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(43 - 60.24)^2}{60.24} + \frac{(69 - 72.28)^2}{72.28} + \cdots + \frac{(6 - 1.56)^2}{1.56}$$
$$= 22.17$$

6. This value lies in the rejection region. We thus reject the null hypothesis that the data are sampled from a Poisson distribution with $\lambda = 1.2$.

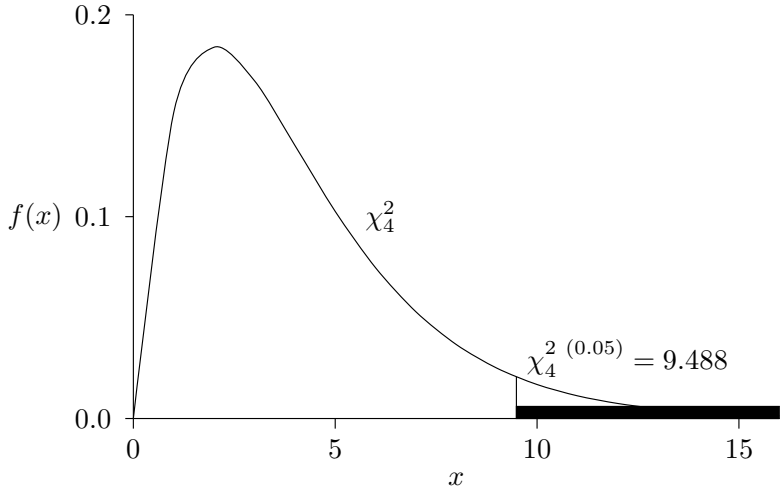
We said earlier that the statistic D^2 has approximately a χ^2 distribution. This approximation is good provided all the **expected frequencies** exceed "about" 5. Looking back at the table of expected frequencies, you will see that this condition has been violated — the expected number of pages with 5 (or more) misprints is 1.56, which is much less than 5. To get around this we amalgamate adjoining classes. This reduces the number of cells, and also the degree of freedom. Here it is necessary to amalgamate the last two cells:

number of misprints	observed number of pages (O_i)	theoretical probability	expected frequency (E_i)
0	43	0.3012	60.24
1	69	0.3614	72.28
2	53	0.2169	43.38
3	21	0.0867	17.34
4 or more	8+6 = 14	0.0338	6.76

All the expected values exceed 5, and we now correctly compute D^2 as

$$D^2 = \frac{(43 - 60.24)^2}{60.24} + \cdots + \frac{(14 - 6.76)^2}{6.76}$$
$$= 15.78$$

We now have only 5 cells, so the degrees of freedom for χ^2 is 4. The 5% point of χ^2_4 is 9.488. Because $15.78 > 9.488$, we reject H_0 and come to the same conclusion as before (but using the right method this time!).



Let us now use the data to estimate λ, and see what difference this makes to the test. Our null and alternative hypotheses are now

1. H_0 : the data fit some Poisson distribution, and
2. H_1 : the data fit a distribution other than the Poisson distribution.
3. Significance level : 5%

4. & 5. To find λ we need to estimate the rate at which misprints occurred in our sample data. The total number of misprints that occurred was $0 \times 43 + 1 \times 69 + 2 \times 53 + \cdots + 5 \times 6 = 300$ misprints. 300 misprints in 200 pages implies that the mean rate at which misprints occur is 1.5 misprints per page. We therefore try to fit a Poisson distribution with $\lambda = 1.5$.

Using the same procedure as before, we find a table of expected frequencies.

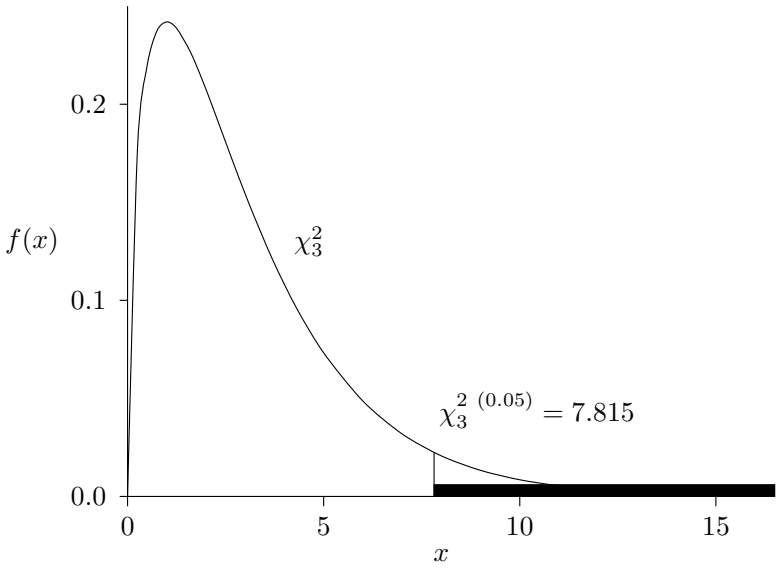
number of misprints	observed number of pages (O_i)	theoretical probability	expected frequencies (E_i)
0	43	0.2231	44.62
1	69	0.3347	66.94
2	53	0.2510	50.20
3	21	0.1255	25.10
4	8 } 14	0.0471	9.42 } 13.14
5 or more	6 }	0.0186	3.72 }

We amalgamate the last two cells, so that all expected values exceed 5. We now have five cells.

The rule for finding the degrees of freedom in this case is

> Degrees of freedom $= k - d - 1$
> where k is the number of cells, and
> d is the number of parameters estimated from the data.

Here $k = 5$ and $d = 1$, because we estimated just one parameter, λ, from the data. Thus we must use χ^2 with $5-1-1 = 3$ degrees of freedom. The 5% point of χ^2 is 7.815. We reject H_0 if D^2 exceeds 7.815. Notice that goodness of fit tests are intrinsically one-sided — we reject H_0 if D^2 is too large. If D^2 is small, it means that the distribution specified by H_0 fits the data very well.



Using the formula $D^2 = \sum \dfrac{(O_i - E_i)^2}{E_i}$ we calculate

$$\begin{aligned} D^2 &= \frac{(43 - 44.62)^2}{44.62} + \cdots + \frac{(14 - 13.14)^2}{13.14} \\ &= 1.01 \end{aligned}$$

6. Because 1.01 lies in the acceptance region, we cannot reject H_0. It is reasonable to conclude that the Poisson distribution with $\lambda = 1.5$ misprints per page fits the data well.

A SHORT-CUT FORMULA...

As in the case with calculating the sample variance there is a short-cut formula for calculating D^2.

$$\begin{aligned} D^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \sum \frac{O_i^2 - 2O_iE_i + E_i^2}{E_i} \\ &= \sum \frac{O_i^2}{E_i} - 2\sum O_i + \sum E_i \end{aligned}$$

But $\sum O_i = \sum E_i = n$, the sample size. (Each summation is over the k cells.)

Therefore

$$D^2 = \sum \frac{O_i^2}{E_i} - n.$$

This is the best formula for calculating D^2.

Example 2B: A random sample of 230 students took an I.Q. test. The scores they obtained have been summarized in the table below:

Score	Observed frequency
< 90	18
$90 - 110$	87
$110 - 130$	104
> 130	21

Test at the 5% level whether these data come from a normal distribution.

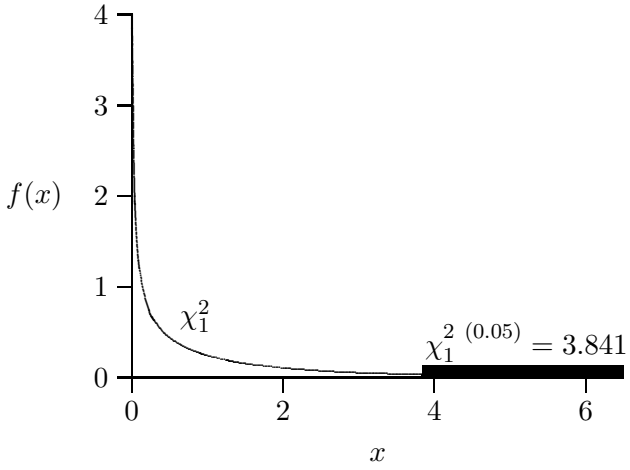
1. H_0 : The data follow a normal distribution (parameters to be estimated from the data).

2. H_1 : The distribution is not normal.

3. Significance level : 5%.

4. We use the χ^2 distribution with $k-d-1$ degrees of freedom. Here there are $k = 4$ cells and $d = 2$ parameters (μ and σ) are estimated.

   Thus we use χ^2_1, and reject H_0 if D^2 exceeds $\chi_1^{2(0.05)} = 3.843$.



5. We need to estimate μ and σ. Let us suppose that before the data was summarized into the table above, the sample mean $\bar{x}$ and the standard deviations were computed to be 111 and 16.3 respectively.

   We now use $N(111,\ 16.3^2)$ to compute the theoretical probabilities that randomly selected I.Q.'s fall into the 4 cells in our table.

   If $X \sim N(111,\ 16.3^2)$, then

$$\begin{aligned}\Pr(X < 90) &= \Pr\left(Z < \frac{90 - 111}{16.3}\right)\\ &= \Pr(Z < -1.29)\\ &= 0.0985\end{aligned}$$

   Thus the probability that a randomly selected single individual has an I.Q. less than 90 is 0.0985. In a sample of size 230 we therefore expect $230 \times 0.0985 = 22.7$ students to have I.Q.'s below 90.

We do a similar calculation for the remaining cells, to obtain the table:

score	observed frequency	theoretical probability	expected frequency
< 90	18	$\Pr[Z < -1.29] = 0.0985$	22.7
$90 - 110$	87	$\Pr[-1.29 \leq Z < -0.06] = 0.3776$	86.8
$110 - 130$	104	$\Pr[0.96 \leq Z < 1.17] = 0.4029$	92.7
≥ 130	21	$\Pr[Z \geq 1.17] = 0.1210$	27.8

Thus

$$D^2 = \sum \frac{O_i^2}{E_i} - n = \frac{18^2}{22.7} + \frac{87^2}{86.8} + \frac{104^2}{92.7} + \frac{21^2}{27.8} - 230$$
$$= 234.01 - 230 = 4.01$$

6. Because $4.01 > 3.841$, we reject H_0 and conclude that the normal distribution is not a good fit to this data.

   This example illustrates the method to use to test whether data fit a normal distribution. In practice, however, the data would be divided into many more than the 4 cells we used in this example.

   Look back at the plot of the χ_1^2-distribution in example 2B. The shape is quite unlike that of the typical chi-squared distribution with more than one degree of freedom, as depicted in example 1A. It is closely related to the exponential distribution of chapter 5. We saw that $\chi_1^{2\ (0.05)} = 3.841$, i.e. the 5% point of the chi-squared distribution with one degree of freedom, is 3.841. And it is not an accident that the square root of 3.841 is equal to 1.96, the 2.5% point of the standard normal distribution — there is a mathematical relationship between the normal and chi-squared distributions.

Example 3C: A T-shirt manufacturer makes a certain line of T-shirt in three colours: white, red and blue. 50% of the T-shirts made are white, 25% are red and 25% blue. One outlet reported sales of 47 white T-shirts, 32 red and 21 blue. Test whether sales are consistent with the manufactured proportions.

Example 4C: A popular clothing store is interested in establishing the distribution of customers arriving at the cashiers. During a 100-minute period, the number of customers arriving per minute was counted, with the following results:

Number of customers per minute	0	1	2	3	4	5 or more
observed frequency	8	25	26	21	15	5

Can a Poisson distribution be used to model the arrival times?

Tests of association in contingency tables...

Another use for the chi-squared distribution is in **contingency tables**. A contingency table is simply a table (or matrix) of counts. Each entry in the table is called a *cell*. Each member of a sample is classified according to two variables, e.g. eye colour and hair colour, and each cell in the table represents the count of the number of members of the sample who have a particular combination of the two variables, e.g. blue eyes and blond hair. The chi-squared distribution is the sampling distribution of the test statistic which tests whether there is an association (or relationship) between the two variables: e.g. Is there a tendency for eye colour to be related to hair colour? Or are eye colour and hair colour independent?

Example 5A: A financial analyst is interesting in determining whether the size of a company (as measured by its market capitalization) has any association with its performance (as categorized by the annual percentage price change of a share in the company on the stock market). The table below gives the number of companies falling into each category in a sample of 420 firms.

	Performance (% price change)				
Company size	$< 0\%$	0–20%	20–40%	$> 40\%$	Total
Small	66	90	28	39	223
Medium	24	33	17	11	85
Large	55	37	10	10	112
Total	145	160	55	60	420

There were thus 66 companies whose performances were negative (column $< 0\%$) and which were small, etc. This rectangular collection of frequencies of occurrence is a typical example of a contingency table. This is a 3×4 contingency table. It is customary to give the number of rows first, then the number of columns. The designer of the South African decimal currency had the right idea — **rands** and **cents**, **rows** and **columns**. In general we talk of an $r \times c$ contingency table with r rows and c columns.

The financial analyst wants to know if there is a significant relationship (at the 5% level) between the performance and the size of the companies. We use our standard approach to hypothesis testing.

1. We start by assuming that there is no relationship between the two variables; thus we have H_0 : the performance and the size of a company are **statistically independent**.

   You will be wise at this point to reread the last few pages of Chapter 3, where the concept of independence was developed!

2. The alternative hypothesis, which the financial analyst is trying to establish, is H_1: the performance and the size of a company are **associated**.

3. Significance level : 5%.

4. & 5. Under the assumption of independence made in the null hypothesis, we calculate theoretical expected frequencies for each cell. If the theoretical frequencies (which assume independence) and the observed frequencies are "too different", we reject

the null hypothesis of independence, and conclude that there is a dependence or a relationship or association between the two variables. A careful examination of the table then helps us to determine the nature of the association.

From the table above, we see that 145 out of the 420 companies had negative performance figures. Thus, using the relative frequencies in the sample, we estimate the probability that a randomly selected company will have negative performance figures as Pr[negative performance] = 145/420 = 0.345.

Similarly, 223 out of the 420 companies were classified as small. We estimate that the probability that a randomly selected company will be small as Pr[small] = 223/420 = 0.531.

If, as H_0 tells us, the performance of a company is **independent** of its size, then the probability that a randomly selected company will have both negative performance figures and be small will be the **product** of the individual probabilities, i.e.

$$\begin{aligned}\Pr[\text{negative performance and small}] &= \Pr[\text{negative performance}] \times \Pr[\text{small}] \\ &= \frac{145}{420} \times \frac{223}{420} = 0.345 \times 0.531 \\ &= 0.1833\end{aligned}$$

Thus, if H_0 is true, the proportion of companies that has negative performance and is small will be 0.1833, or, expressed as a percentage, 18.33%. We have 420 companies in our sample, so we expect $0.183 \times 420 = 77.0$ to have negative performance and be small. This **expected frequency** can be computed more directly as

$$\frac{145}{420} \times \frac{223}{420} \times 420 = \frac{145 \times 223}{420} = 77.0$$

Thus the calculation of the expected frequencies for this (and every other) cell of the table reduces to a very simple formula:

$$\text{expected frequency} = \frac{\text{column total} \times \text{row total}}{\text{grand total}}$$

The full set of expected frequencies is given in brackets under the observed values in the table below:

	Performance (% price change)				
Firm size	$< 0\%$	0–20%	20–40%	$> 40\%$	Total
Small	66	90	28	39	223
	(77.0)	(85.0)	(29.2)	(31.9)	
Medium	24	33	17	11	85
	(29.3)	(32.4)	(11.1)	(12.1)	
Large	55	37	10	10	112
	(38.7)	(42.7)	(14.7)	(16.0)	
Total	145	160	55	60	420

In all the above arithmetic, don't lose sight of the fact that the expected frequencies have been computed assuming that the two variables are independent. Thus, the larger the differences between the observed frequencies in the table and the expected frequencies, the less likely it is that the two variables are independent. The comparison of observed and expected frequencies makes use of the same statistic as was used for goodness of fit tests:

$$D^2 = \sum \frac{O_i^2}{E_i} - n,$$

where we sum over all the cells in the contingency table, and n is the total number of observations.

Once again we have a degrees of freedom problem. It can be shown that the following rule gives us the degrees of freedom we require:

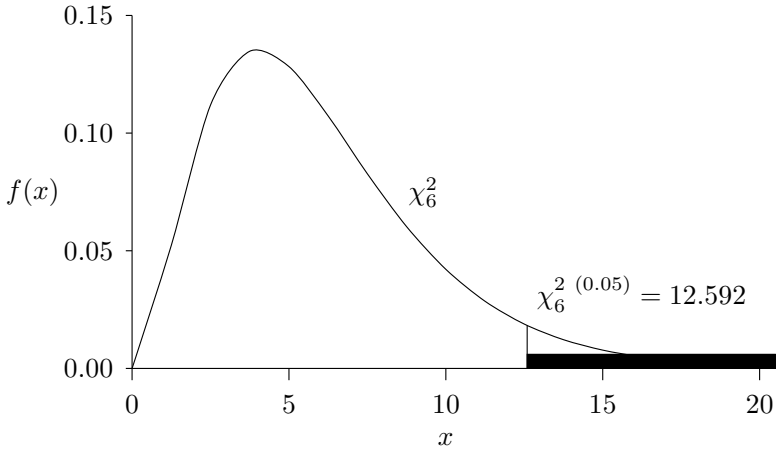
> DEGREES OF FREEDOM FOR TESTS OF ASSOCIATION
>
> For an $r \times c$ contingency table, the appropriate chi-squared distribution has degrees of freedom $= (r-1)(c-1)$

Here we have a 3×4 contingency table, hence the degrees of freedom for chi-squared is $(3-1)(4-1) = 6$.

Using the 5% significance level, we determine from tables that the 5% point of χ^2_6 is 12.592. We will reject H_0 if $D^2 > 12.592$. Notice that tests of association for contingency tables are almost invariably one-sided.

We compute D^2 using the short-cut formula:

$$\begin{aligned} D^2 &= \frac{66^2}{77.0} + \frac{90^2}{85.0} + \cdots + \frac{10^2}{14.7} + \frac{10^2}{16.0} - 420 \\ &= 19.09. \end{aligned}$$



6. Because $19.09 > 12.592$ we reject H_0. Thus our financial analyst has demonstrated that there is a significant relationship between the performance of a firm and its size. Examination of the table of observed and expected frequencies shows that expected values exceed observed values in the top left and bottom right corners of the contingency table, and vice versa in the top right and bottom left corners. This indicates an inverse relationship between these two variables, that is, as the size of a firm decreases, the performance improves.

Example 6B: A photographic company wishes to compare its present automatic film processing machines with two other machines that have recently come onto the market. It processes 194 films on the present machine (called A), and hires the other machines (B and C) for short periods and processes smaller numbers of films on them. Using very stringent criteria, the manager classifies each film as being satisfactorily processed or not. The following contingency table is obtained:

	Processing machine		
	A	B	C
Satisfactory	93	24	8
Unsatisfactory	101	6	18

Test whether the classification of the film as satisfactory or not is independent of the machine the film was processed on.

No significance level is given, and we use the modified hypothesis testing procedure.

1. H_0 : The classification is independent of the machine
2. H_1 : The classification is dependent on the machine
3. We compute the expected values, given in brackets in the table below:

	Processing machine			
	A	B	C	Total
Satisfactory	93	24	8	125
	$((194 \times 125)/250 = 97)$	(15)	(13)	
Unsatisfactory	101	6	18	125
	(97)	(15)	(13)	
Total	194	30	26	250

Next we compute D^2:

$$\begin{aligned} D^2 &= \frac{93^2}{97} + \frac{24^2}{15} + \cdots + \frac{18^2}{13} - 250 \\ &= 14.98. \end{aligned}$$

4. The degrees of freedom are $(3-1) \times (2-1) = 2$. Examining the row in the tables for χ^2_2, we see that the highest level at which the test statistic, 14.98, is significant is the 0.1% level $(P < 0.001)$.

5. Our conclusion is that the classification of the film is significantly dependent on the machine used $(\chi^2_2 = 14.98,\ P < 0.001)$. (Note that we report the value of the test statistic as a χ^2 value.) Examination of the contingency table shows that machine B tends to produce a large proportion of photographs classified as satisfactory, while machine C tends to produce a large proportion of unsatisfactory photographs. Processing machine B is the recommended choice.

Example 7C: An analysis of the tries, penalty goals and dropped goals scored by South Africa in rugby tests against the British Isles, New Zealand, Australia and France gave the following contingency table:

	Tries	Penalties	Drops
British Isles	70	31	6
New Zealand	44	38	11
Australia	79	31	7
France	43	32	3

The number of penalities scored against New Zealand and France seems unexpectedly high, and this leads you to want to test the hypothesis that the mode of scoring is dependent on the opponents.

Example 8C: An investment broking company is keen to establish whether there is an association between their clients' perception of their attitude towards risk and the type of investments they prefer. They obtained 340 responses to a questionnaire designed to capture this information. A summary of the number of clients falling into the various categories is shown below:

	Risk category		
Type of investment	Risk averse	Risk neutral	Risk lover
Fixed deposits	79	58	49
Bonds	10	8	9
Unit trusts	12	10	19
Options and futures	10	34	42

Test the hypothesis that the type of investment is dependent on the risk perception.

Confidence intervals and tests for the variance...

It can be shown that if a sample is drawn from a population having a normal distribution, then a simple transformation of the sample variance s^2 has exactly the χ^2-distribution:

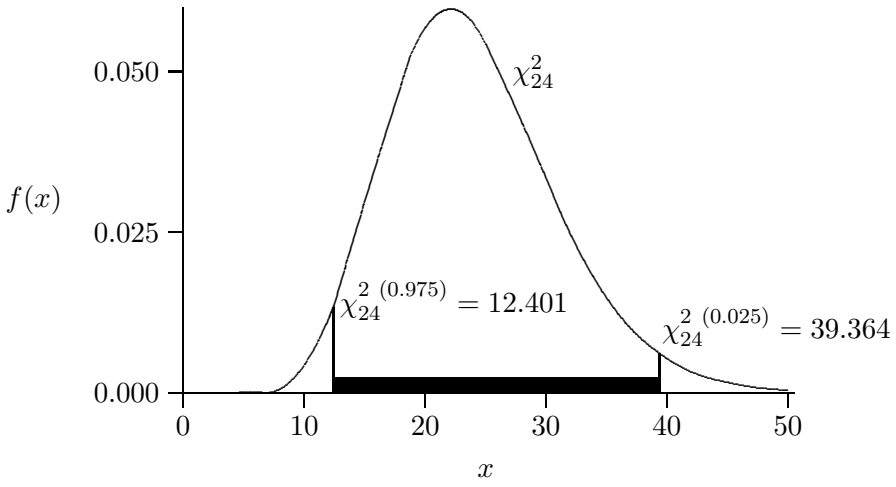
$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}.$$

This states that the sample variance s^2, multiplied by its degrees of freedom $(n-1)$, and divided by the population σ^2 has the χ^2-distribution with $n-1$ degrees of freedom. We can use this result to set up confidence intervals for the population variance, and to do hypothesis tests on the variance.

Example 9A: A dairy is concerned about the variability in the amount of milk per bottle. A sample of 25 bottles was examined, and the sample standard deviation of the contents was computed to be 3.71 ml.

(a) Find 95% confidence intervals for σ, the population standard deviation.
(b) At the 1% significance level, test whether the observed sample standard deviation is consistent with the industrial specification that the population standard deviation must not exceed 2.5 ml.

(a) Confidence Interval: The degrees of freedom of the appropriate χ^2-distribution is one less than the sample size. Here, $n-1 = 24$ degrees of freedom. To obtain a 95% confidence interval, the upper and lower $2\frac{1}{2}\%$ points of the χ^2-distribution must be obtained from tables. Because the χ^2-distribution is not symmetric, the lower $2\frac{1}{2}\%$ point must be obtained separately. The upper $2\frac{1}{2}\%$ point which we require is $\chi^{2(0.025)}_{24} = 39.364$, and the lower $2\frac{1}{2}\%$ point, which we look up, is $\chi^{2(0.975)}_{24} = 12.401$.



Thus we can write:

$$\Pr[12.401 < \chi^2_{24} < 39.364] = 0.95$$

Because $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$, and because $n-1 = 24$ and $s^2 = 3.71^2$ we have

$$\Pr[12.401 < 24 \times 3.71^2/\sigma^2 < 39.364] = 0.95$$

Rearranging, so that inequalities produce a confidence interval for σ^2, we have

$$\Pr[24 \times 3.71^2/39.364 < \sigma^2 < 24 \times 3.71^2/12.401] = 0.95.$$

This reduces to

$$\Pr[8.39 < \sigma^2 < 26.64] = 0.95.$$

The 95% confidence interval for σ^2, the population variance, is given by $(8.39, 26.64)$, and the 95% confidence interval for σ, the population standard deviation, obtained by taking square roots, is $(2.90,\ 5.16)$. Unlike the confidence intervals for the mean, the point estimate of the variance does not lie at the midpoint of the confidence interval (and the same is true for the confidence interval for the standard deviation).

CONFIDENCE INTERVAL FOR σ^2

If we have a random sample of size n from a population with a normal distribution, and the sample variance is s^2, then A% confidence intervals for σ^2 are given by

$$((n-1)\ s^2/\chi^{2\ *}_{n-1}\ ,\ (n-1)\ s^2/\chi^{2\ **}_{n-1})$$

where the value $\chi^{2\ *}_{n-1}$ is the appropriate **upper** percentage point and $\chi^{2\ **}_{n-1}$ is the appropriate **lower** percentage point, determined from tables. The confidence interval for σ is found by taking square roots.

(b) Hypothesis test.

1. Our null hypothesis specifies the maximum acceptable population standard deviation. H_0 is thus taken to be
$$H_0 : \sigma = 2.5.$$
2. The alternative hypothesis states the region of unacceptable standard deviations:
$$H_1 : \sigma > 2.5.$$
3. Significance level : 1%.
4. We will reject H_0 if our test statistic exceeds the upper 1% of the χ^2-distribution with $n-1$ degrees of freedom: i.e. if $\chi^{2(0.01)}_{24} = 42.980$ is exceeded.
5. Our test statistic is
$$\begin{aligned}\chi^2 = (n-1)s^2/\sigma^2 &= 24 \times 3.71^2/2.5^2 \\ &= 52.85\end{aligned}$$
6. Because $52.85 > 42.980$, we reject the null hypothesis, and conclude that the specification is not met, and that therefore the dairy has a problem.

Example 10B: Variability in the return of a traded security is often thought of as a measure of "total" risk of the security. A certain portfolio manager will only invest in a security if its population standard deviation of return does not exceed 10% per month. A sample of 18 monthly returns on a particular security yielded a sample deviation of 14.2% per month.

(a) Find a 90% confidence interval for the population standard deviation.
(b) Test the null hypothesis that the standard deviation does not exceed the upper limit required by the portfolio manager.

(a) From tables, $\chi^{2(0.95)}_{17} = 8.672$ and $\chi^{2(0.05)}_{17} = 27.587$. Thus the 90% confidence interval for the population **variance** σ^2 is given by

$$(17 \times 14.2^2/27.587, 17 \times 14.2^2/8.672) = (124.3, 395.3)$$

and the 90% confidence interval for the standard deviation is

$$\left(\sqrt{124.3}, \sqrt{395.3}\right) = (11.1, 19.9)$$

(b)
1. $H_0 : \sigma = 10$.
2. $H_1 : \sigma > 10$.
3. The test statistic is

$$\begin{aligned}\chi^2 &= (n-1)s^2/\sigma^2 = 17 \times 14.1^2/10^2 \\ &= 34.28.\end{aligned}$$

4. From the row in the tables for χ^2_{17}, we see that 34.28 is significant at the 1% level.
5. We conclude that the security being investigated does not meet the variability requirements of the portfolio manager ($\chi^2_{17} = 34.28,\ P < 0.01$), and should be regarded by him as too risky.

Example 11C: A certain moulded concrete product is manufactured in Johannesburg and Cape Town. Since the aggregate used in the production might differ between the two regions, management is interested in comparing the mass and variablity of the product between the regions. A sample of 43 items produced in Johannesburg had a mean mass of 53.2 kg with a standard deviation of 4.2 kg, while a sample of 23 items produced in Cape Town had a mean mass of 54.4 kg with a standard deviation of 3.3 kg. Set up 95% confidence intervals for the population means and standard deviations for products manufactured in both Johannesburg and Cape Town. Test if there is a significant difference between the means and variances of the two populations. If you find there is no significant difference, find appropriate pooled estimates of the overall mean and standard deviation, and find 95% confidence intervals for these overall estimates.

Solutions to examples...

3C D^2 (or χ^2_2) $= 2.78,\ P > 0.20$ ($\chi^{2(0.20)} = 3.219$). Not significant.

4C $\lambda = 2.25$

D^2 (or χ^2_4) $= 3.00,\ P > 0.20,$ ($\chi_4^{2(0.20)} = 5.989$). Not significant, Poisson distribution fits.

7C D^2(or χ^2_6) $= 14.4198,\ P < 0.05$ ($\chi_6^{2(0.05)} = 12.592$). Significant, mode of scoring dependent on opponents.

8C D^2 (or χ^2_6) $= 29.97,\ P < 0.001,$ ($\chi_6^{2(0.001)} = 22.45$). Significant, type of investment dependent on risk.

11C 95% confidence intervals tabulated as follows:

	μ	σ
Johannesburg	(51.9 , 54.5)	(3.5 , 5.3)
Cape Town	(53.0 , 55.8)	(2.6 , 4.7)
pooled	(52.6 , 54.6)	(3.3 , 4.7)

Test of variances: $F_{42.22} = 1.62,\ P > 0.10$, pooling of variances justified.

Pooled estimate of standard deviation: $s = 3.9$ with 64 degrees of freedom.

Test of means: $t = -1.19,\ P > 0.20$, no significant difference.

Pooled estimate of mean: $\bar{x} = 53.6,\ n = 66$.

Exercises on goodness of fit tests...

10.1 On a true-false test with 100 questions, a student gets 61 correct. The lecturer claims that the student was merely guessing, and was just lucky to get such a high mark. Test this claim at the 5% significance level.

*10.2 In 20 soccer cup finals in South Africa between 1959 and 1978, the 40 teams involved scored the following numbers of goals per match:

goals per match	0	1	2	3	4	5	6	or more
number of teams:	8	6	13	8	3	2	0	

Is it true that the number of goals scored per team per match fits a Poisson distribution?

10.3 A large furniture store also sells television sets. Their planning department is interested in the distribution of their daily sales of television sets. A sample of 210 days shows the following sales volumes:

Number of sales	Observed frequency (days)
0	60
1	71
2	53
3	20
4 or more	6

Test at the 5% significance level whether the daily sales volumes occur in accordance with a Poisson distribution.

10.4 In order to check whether a die is balanced it was rolled 240 times and the following results were obtained:

Face	1	2	3	4	5	6
No. of occurrences	33	51	49	36	32	39

Test the hypothesis that the die is balanced.

*10.5 There is an annoying traffic light at an intersection near your home. Each time you leave home, you note the colour as you pass the last lamp-post before the traffic light. During a month, your records show that there was a red light 95 times, amber 15 times, and green 30 times. Thinking this unreasonable, you contact the Traffic Department, who claim that the light is set to be red for 50%, amber for 10% and green for 40% of the time for traffic arriving at the intersection from the direction of your home. Test the Traffic Department's claim at the 1% significance level.

10.6 A vegetable processing company has frozen peas as one of its major lines. They have employed a consultant to research ways of improving their product. The consultant experiments with hybrids between two varieties of peas. He knows that, according the Mendelian inheritance theory, when the two varieties of peas are hybridized, 4 types of seeds — yellow smooth, green smooth, yellow wrinkled and green wrinkled — are produced in the ratio $9 : 3 : 3 : 1$. In the experiment, the consultant observes 102 yellow smooth, 30 green smooth, 42 yellow wrinkled and 15 green wrinkled seeds. Are the results consistent with the theory at the 5% level of significance?

*10.7 It is hypothesised that the number of bricks laid per day by a team followed a normal distribution with mean 2000 and standard deviation 300. A record of bricks laid over a typical 100 day period produces the following data:

Bricks laid per day	number of days
under 1500	5
1500–1750	14
1750–2000	30
2000–2250	28
2250–2500	17
over 2500	6

Test the above hypothesis.

*10.8 A survey of 320 families with 5 children revealed the distribution shown in the table below. Is this data consistent with the hypothesis that male and female births are equally probable?

No. of males	5	4	3	2	1	0
No. of families	18	56	110	88	40	8

Exercises on contingency tables...

10.9 A hair colour company wants to know if blondes have more boy friends. They obtained the following results from a random sample.

	No boy friends	One boy friend	More than one boy friend
blonde	4	17	30
non-blonde	19	43	62

What should the company decide?

*10.10 Records of 500 car accidents were examined to determine the degree of injury to the driver and whether or not he was wearing a safety belt at the time of the accident. The data are summarized below:

	Safety belt	No safety belt
Minor injury	128	168
Major injury	49	104
Death	11	40

Test the hypothesis that the severity of the injury to the driver was independent of whether or not the driver wears a safety belt. Use a 1% significance level. By comparing observed and expected values, interpret your result.

10.11 An insecticide company is concerned about the effectiveness of their product across a range of insect species. 200 specimens of each of four species of insects are placed in a container and a prescribed amount of an insecticide is added. After an hour the number of survivors for each species are noted:

	Species			
	A	B	C	D
Survived	27	42	68	31
Killed	173	158	132	169

Does the insecticide differ in its effectiveness according to species at the 1% level of significance?

10.12 Does the following sample indicate in the population sampled that preference for cars of certain makes is independent of sex?

	Make of car		
	A	B	C
Men	60	80	110
Women	80	70	100

*10.13 A nationwide survey is conducted to determine the public's attitude towards the abolition of capital punishment, and to compare it with that of police officers.

A sample of 300 members of the public produced the results:

In favour	Indifferent	Against
90	60	150

A sample of 100 police officers showed:

In favour	Indifferent	Against
20	10	70

Do the attitude patterns between police and public differ significantly? Use a 1% significance level.

EXERCISES ON THE SAMPLE VARIANCE...

*10.14 The amount of drug put into a tablet must remain very constant. For a certain drug the maximum acceptable standard deviation is 1 mg. Analysis of 10 tablets produced the following data (amount of drug in milligrams):

26 28 26 25 32 27 29 25 30 28

(a) Construct a 95% confidence interval for the true standard deviation.
(b) Using a 1% significance level, decide whether the fluctuation in drug content exceeds the acceptable level of 1 mg.

*10.15 The exact contents of seven similar containers of motor oil were (in millilitres)

499 501 500 498 501 501 499

(a) Calculate the mean and standard deviation.
(b) What assumption is it necessary to make in order to construct confidence intervals for the mean and the standard deviation?
(c) Calculate 95% confidence intervals for both the mean and the standard deviation.

10.16 An agricultural economist needs to estimate crop losses due to insect damage. To do this, he needs to estimate the mean number of insects per square metre. He wishes to be 95% sure of being within two insects of the true mean number/m^2. He has no idea of the population variance and thus runs a pilot survey, collecting data from 61 quadrats of 1 m^2. The sample mean is 85.1 and the sample variance $s^2 = 101.4$.

(a) Find 95% confidence intervals for μ and for σ^2.
(b) What size sample is suggested in order to be 95% sure of getting within 2 of the true mean?
(c) The researcher, however, is a cautious fellow, and thinks that his point estimate of s^2 might be underestimating σ^2. Show that he can be 95% sure that $\sigma^2 < 140.6$. Using this conservatively large estimate of σ^2, show that a sample of size 141 is required.
(d) He therefore samples a further 80 quadrats, and computes $\bar{x} = 87.3$ and $s^2 = 121.4$ from the combined sample. Find 95% confidence intervals for μ and σ^2.
(e) If the distribution of the number of insects per square metre is very skew, comment on the reliability of the confidence intervals obtained in (d).

Solutions to exercises...

10.1 $D^2 = 4.84 > 3.841 = \chi_1^{2(0.05)}$, reject H_0.

10.2 $D^2 = 3.84 < 4.642 = \chi_3^{2(0.20)}$.
Cannot reject H_0, Poisson distribution provides satisfactory fit $(P > 0.20)$.

10.3 $D^2 = 1.55 < 7.815 = \chi_3^{2(0.05)}$.
Cannot reject H_0, Poisson distribution provides good fit.

10.4 $D^2 = 8.30 > 7.289 = \chi_5^{2(0.20)}$.
p-value: $0.10 < P < 0.20$, the die is not unbalanced.

10.5 $D^2 = 21.07 > 9.210 = \chi_2^{2(0.01)}$, reject H_0.

10.6 $D^2 = 3.08 < 7.815 = \chi_3^{2(0.05)}$, cannot reject H_0.

10.7 When categories have not been combined (expected frequency almost 5): $D^2 = 0.72 < 7.289 = \chi_5^{2(0.20)}$, cannot reject H_0.
When categories have been combined: $D^2 = 0.56 < 4.642 = \chi_3^{2(0.20)}$, cannot reject H_0.

10.8 $D^2 = 11.96 > 11.071 = \chi_5^{2(0.05)}$.
Conclude that male and female births are not equally probable $(P < 0.05)$.

10.9 $D^2 = 2.10 < 3.219 = \chi_2^{2(0.20)}$.
No evidence to show that blondes have more boyfriends $(P > 0.20)$.

10.10 $D^2 = 11.63 > 9.210 = \chi_2^{2(0.01)}$, reject H_0.

10.11 $D^2 = 30.81 > 11.345 = \chi_3^{2(0.01)}$, reject H_0.

10.12 $D^2 = 4.00 > 3.219 = \chi_2^{2(0.20)}$.
p-value: $0.10 < P < 0.20$, no difference between the two sexes.

10.13 $D^2 = 12.47 > 9.210 = \chi_2^{2(0.01)}$, reject H_0.

10.14 (a) $(1.56\ ,\ 4.15)$.
(b) $D^2 = 46.4 > 21.666 = \chi_9^{2(0.01)}$, reject H_0.

10.15 (a) We need to assume that the exact contents of the containers of motor oil are normally distributed.
(b) For $\mu : (498.7, 501.0)$.
For $\sigma^2 : (0.613, 7.160)$.

10.16 (a) For $\mu : (82.52 : 87.69)$.
For $\sigma^2 : (73.03\ ,\ 150.32)$.
(b) $n = \dfrac{t_{60}^{(0.025)} s}{L} = 102$.
(c) Because $\chi_{60}^{2(0.95)} = 43.185$,
(d) $\left[\sigma^2 > \frac{60 \times 101.4}{43.185}\right] = \Pr[\sigma^2 > 140.9] = 0.95$
$n = 141$.
(e) For $\mu : (85.48\ ,\ 89.12)$.
For $\sigma^2 : (94.33\ ,\ 155.73)$.

(f) By the central limit theorem, the confidence interval for the mean is likely to be reliable (unless the distribution is so very skew that even a sample of size 141 is not large enough for the sample mean to be approximately normally distributed). On the other hand, the confidence interval for σ^2 depends on the distribution being normal, and hence is unreliable.

Chapter 11

PROPORTIONS AND SAMPLE SURVEYS

KEYWORDS: Population and sample proportions, confidence intervals and hypothesis tests for proportions, finite populations, sample surveys, sampling methods, random numbers, random number tables.

Sample surveys...

Statisticians are frequently required to estimate the **proportion** of a population having some characteristic. We are all familiar with the opinion polls that take place around election time and which purport to inform us what **proportion** of the electorate will support each party or each candidate. Market researchers run surveys to determine the **proportion** of the population who saw and absorbed the television advertisement for their client's product.

Ideally, if it were convenient, quick and affordable, one would choose to obtain data from each element of the population. Then we could estimate the population proportion exactly, using the information from the complete data set.

However each datum we obtain takes some time to observe and record, and also generates costs that must be covered. So a complete collection of all data may have large time and cost impacts. In reality there are very often severe constraints on time and money available for data collection and capture.

Estimating proportions from sample data, rather than from the complete population data, is the usual challenge that confronts us. How could such a strategy of making conclusions about the entire population, on the basis of only an **incomplete subset** of the population, ever make sense?

In general the strategy can only make sense if we have reason to believe that the two aggregated parts of the population, comprising the sampled and observed elements, on the one hand, and the unsampled and unobserved elements, are essentially similar or equivalent.

If that **belief** in equivalence is correct, then the sample can be thought of as being **representative** of the unsampled group, as well as being obviously representative of itself.

When the belief is correct, the sample will be representative not only of itself and of the unsampled group, but therefore also representative of the population in its entirety. In these circumstances, we believe the sample, the non-sample and the entire population are essentially similar, and in that sense, representative of each other.

On the other hand, if we have no reason to believe in the equivalence of the sample and non-sample subsets, or worse still, if we have reason to believe they are actually different from one another, we have a severe problem. The strategy of using the sample data to estimate population features will be incorrect and misleading.

The situation we have described thus far implies that we would still have to find a way of ensuring the **reasonableness** or validity of a belief that a particular sample of size n is representative of the population (of size $N > n$) from which it is drawn.

We cannot check the belief in the absence of information about the non-sampled part of the population. Generally there is no such information available. Thus the belief is not verifiable. If the information for verifiablity were available, then there may be no additional information value or purpose a sample could actually serve.

Instead of ensuring that a sample is representative, the statistician relies upon a less strict criterion, called **randomness**, as a device by which to eliminate any conscious or unconscious bias in the selection of sample elements from the population. In its simplest form, randomness implies that, at every stage in the sampling, each and every element in the population has the same chance of being selected into the sample as each and every other population element.

It is possible to achieve this type of random selection in practice, using various simple techniques such as listing and numerically labelling each element of a population (e.g., from 1 to N), putting labels 1 to N into a container and then drawing n numbered labels from the thoroughly shaken container.

When a sample consisting of n elements has been chosen through an appropriate randomisation method, we indicate its near-representative quality, by calling it a **random sample**. A random sample is highly likely to be either representative or close to representative of the population.

While it is possible that a random sample might turn out to be unrepresentative of its population, that outcome is very rare, if the sample size is moderate. Moreover, the likeliness, of such a technically possible outcome, decreases in probability to almost zero, as the random sample size n increases.

In practice, there are some golden rules for studying a population using statistical methods:

1. Use random selection to ensure we have random samples.
2. Moderate sample size n can help keep costs and time requirements within limits.
3. Random sample size n should be large enough for all the objectives of the study.

In this chapter we focus upon the study objective of using random sample data, and sample proportions to estimate corresponding population proportions.

The two most frequently asked questions (which are in fact interrelated) are:

1. What size sample is needed to estimate a proportion?
2. What is the reliability or margin of error of the estimate?

We will answer both questions.

Confidence intervals for proportions...

We will use π, the Greek letter p, to denote the true **population proportion** ($0 \leq \pi \leq 1$) and P to denote a **sample proportion**. P is used to estimate π, in the same way as the sample mean $\overline{X}$ is used to estimate the population mean μ, and the sample variance s^2 is used to estimate the population variance σ^2.

Note that the meaning of π here is different from its familiar use as the constant for circular dimensions $\pi = \frac{22}{7}$ or $\pi = \frac{355}{113}$. Here the use of π parallels the use of μ, Greek letters for unknown constants, namely proportion and mean respectively.

Suppose we sample 1000 voters and 500 say that they will vote for a particular candidate, then $P = 0.50$. We cannot claim this sample estimate to be the exact population proportion π. Another sample may well yield $P = 0.47$ or $P = 0.54$. Although the population proportion π is constant (at least for a short period of time!), the sample proportions will vary from sample to sample. P is yet another example of a statistic; it is a random variable and therefore has a sampling distribution.

When the sample size n is large, and P is not close to either zero or one, the sampling distribution of P is closely approximated by the normal distribution

$$P \dot{\sim} N\left(\pi,\ \frac{\pi(1-\pi)}{n}\right).$$

Aside: Strictly speaking a continuity correction is needed since a discrete distribution is being approximated by a continuous one. When n is large, the continuity correction is usually omitted, however for borderline cases it can be important (refer to using the normal distribution to approximate the binomial and Poisson distributions in Chapter 6). Also note that these corrections have been omitted in the examples and exercises of this chapter.

But π is unknown — it is the very number we are trying to estimate — therefore we do not know the variance of this approximate normal distribution. But, in this case, we can generally get away with substituting $P = \pi$ in the expression for the variance. This substitution works well because, in sample surveys for proportions, we usually have large samples, with n frequently equal to values of 1000 or larger. Making this substitution, we have that P has, to a good approximation, the normal distribution

$$P \dot{\sim} N\left(\pi,\ \frac{P(1-P)}{n}\right).$$

We now construct confidence intervals for π in much the same way that we constructed confidence intervals for μ when the standard deviation was assumed to be known:

$$(P-\pi)\Big/\sqrt{\frac{P(1-P)}{n}} \sim N(0,\ 1).$$

Thus

$$\Pr\left[-1.96 < (P-\pi)\Big/\sqrt{\frac{P(1-P)}{n}} < 1.96\right] = 0.95,$$

and, by rearrangement, 95% confidence intervals for π are given by

$$\Pr\left[P - 1.96\sqrt{\frac{P(1-P)}{n}} < \pi < P + 1.96\sqrt{\frac{P(1-P)}{n}}\right] = 0.95.$$

The 99% confidence interval is found, as usual, by replacing 1.96 by 2.58.

It is important to understand that in rearranging the terms, we are changing the meaning in our approach. We began with a probability statement about (random) sample proportions P from a particular population with proportion π.

Instead, we now adopt a **method** of using a random sample to define an interval.This method will have a chosen probability (e.g. 95%, or 99%) of providing an interval covering the unknown true value of the population proportion π, and a corresponding probability (e.g. 5%, or 1%) of failure to cover that value, regardless of the specific population to which it is applied.

The purpose of the method is to handle the uncertainty that is always our predicament when we can only access information from part of a population, rather than the entire population. The validity of the method depends upon the sample being random.

The subtlety of the change in meaning is that we have moved our attention from simply and only the single random sample of size n that constitutes our data, to a method of handling any random sample of the same size n from any population! We interpret the re-arranged probability expression to imply we have 95% confidence that the method will yield an interval that includes the unknown parameter value.

Example 1A: Suppose 500 voters in a sample of 1000 say they will vote for a candidate. Find the 95% confidence interval for π.

We have $P = 0.50$ and $n = 1000$. So the 95% confidence interval for π is given by

$$\left(0.5 - 1.96\sqrt{\frac{0.5(1-0.5)}{1000}}\ ,\ 0.5 + 1.96\sqrt{\frac{0.5(1-0.5)}{1000}}\right)$$

$$= (0.5 - 0.031,\ 0.5 + 0.031) = (0.469\ ,\ 0.531)$$

Equivalently, we can say that we are 95% sure that the interval (46.9% , 53.1%) contains the true population proportion. Note that, in all our formulae and calculations, proportions lie between 0 and 1. However, it is often convenient to communicate our results as percentages. The quantity that we add to, and subtract from, the point estimate of proportion to form the confidence intervals, is called the **reliability margin** of the estimate **at the given confidence level**, and is conventionally denoted by L, and expressed as a percentage. Thus

$$L = 100 \times 1.96\sqrt{\frac{P(1-P)}{n}}\ \%$$

at the 95% confidence level.

In this example, the plus/minus term is $100 \times 0.031 = 3.1\%$ so we say that, at the 95% confidence level, our estimate has a reliability margin of 3.1%. Note that the confidence interval for the percentage is of the form $(100P \pm L)\%$, as in $50.0\% \pm 3.1\%$.

Note that the use of the term reliability here is different from (and inverse to) the way it is used in common speech. Here reliability margin is a **margin of error** or variability that naturally arises in random samples of a given size n. Our preference is always for smaller reliability margin values, because the corresponding confidence intervals are narrower. We say our estimates should be more precise.

The only way to achieve this goal of greater precision is to reduce L. By inspection, we can see L is small when $\sqrt{n}$ is large, and hence n large. In principle we prefer sample size n as large as any time and cost constraints will permit.

We now have an answer to the second of the two questions we asked at the beginning of the chapter.

Example 2B: A market research company establishes that 28 out of 323 randomly sampled households have more than one television set. Compute a 95% confidence interval for the percentage of households having more than one television set. What is the reliability margin, at the 95% level, of the estimate?

We calculate $P = 28/323 = 0.0867$, and $n = 323$. Thus the 95% confidence interval is

$$\left(0.0867 - 1.96\sqrt{\frac{0.0867 \times 0.9133}{323}}\ ,\ 0.0867 + 0.0307\right)$$
$$= (0.0560\ ,\ 0.1174).$$

Expressing proportion in percentage terms, the percentage of households with more than one television set is within the confidence interval $(5.60\%, 11.74\%)$. The reliability margin of our estimate at the 95% confidence level is 3.07%.

Example 3C: In a questionnaire, 146 motorists out of a (random) sample of 252 stated that when they next replaced tyres on their cars, they would insist on radial ply tyres. Find a 95% confidence interval for the proportion of motorists in the population who will get radials. What is the reliability margin at this confidence level?

Finite populations...

The method used above presupposes that the population being sampled is infinite, or that the sampling is done "with replacement". If neither of these assumptions is true, we are sampling without replacement from a finite population. Then the random error in the method will be reduced.

Intuitively we might expect this result because, by sampling without replacement, no element can be selected twice, and every resulting sample of size n has a greater chance of being representative of the population. Effectively we have eliminated all the random samples which have any duplicated elements.

If the size of the population being sampled is N, and the sampling of an element is done randomly n times without replacement, then the sampling distribution of P is once again approximately normal, but with a reduced variance:

$$P \approx N\left(\pi,\ \frac{P(1-P)(N-n)}{n(N-1)}\right).$$

Confidence intervals are constructed using the same procedure as before, with the necessary modification for the diminished variance. Thus, the 95% confidence interval is given by

$$\left(P - 1.96\sqrt{\frac{P(1-P)(N-n)}{n(N-1)}}\ ,\ P + 1.96\sqrt{\frac{P(1-P)(N-n)}{n(N-1)}}\right)$$

with a reliability margin of our estimator P at the 95% confidence level given by

$$L = 100 \times 1.96\sqrt{\frac{P(1-P)(N-n)}{n(N-1)}}\ \%.$$

Example 4B: A lecturer, anxious to estimate the overall pass rate in a class of 823 students, takes a random sample of 100 scripts and finds 27 failures. Find a 95% confidence interval for the failure rate.

We have $P = 0.27,\ n = 100$, and $N = 823$. Because the sample size is more than 10% of the population size we use the modified confidence interval:

$$\left(0.27 - 1.96\sqrt{\frac{0.27(1-0.27)(823-100)}{100(823-1)}}\ ,\ 0.27 + 0.082\right) = (0.188\ ,\ 0.352).$$

At this stage, all the lecturer can say is that he has used a method for which there is a 95% probability that the interval (18.8% , 35.2%) includes the true class failure rate. The reliability margin at this confidence level is 8.2%, which is a comparatively large figure, and the interval so wide it does not provide useful information! To get a narrower confidence interval (at the 95% level), a larger sample size would be needed.

Example 5C: In a constituency of 8000 voters, 748 voters out of a sample of 1341 voters state they will support the Materialistic Party, 510 voters state they will vote for the Ecological Party and the remaining 83 voters are undecided. Find 95% confidence intervals for the population proportions which will support each party, and the population proportion of undecided voters. What are the associated reliabilities?

Sample sizes...

When you approach a statistician with our first question: "What size sample is needed?" he will reply by asking four further questions:

1. What is N, the size of the population?
2. Do you want 95% or 99% (or some other level) confidence intervals?
3. What margin of error or reliability margin (L%) can you accept?
4. Do you have a rough estimate, P, of π?

The two formulae for the reliability margin L given earlier connect these four quantities with the sample size. If the population was infinite, we had

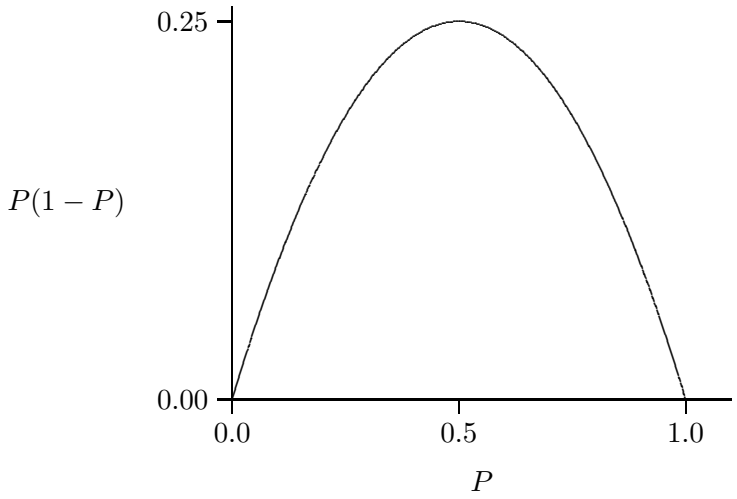
$$L = 100 \times 1.96\sqrt{\frac{P(1-P)}{n}}.$$

Making n the subject of the formula yields

$$n = (100/L)^2 \times 1.96^2 \times P(1-P).$$

This formula is appropriate if the answers to the four questions are:

1. The population is very large.
2. 95% confidence level. (For 99% confidence level use 2.58 in place of 1.96. For other levels use the appropriate value from the normal tables.)
3. Reliability margin of L% is given a value.
4. The rough estimate of π is substituted for P. If no estimate of π is available, let $P = 0.5$. To see the logic behind this choice, let us consider the function $y = P(1-P)$ further. In the region of interest, for values of P between 0 and 1, the graph looks like this:



It is easy to show that the **maximum** of $y = P(1-P)$ occurs at $P = 0.5$. Thus taking $P = 0.5$ in the sample size formula gives the largest possible sample size that might be required to achieve a margin of error $L\%$. Because it is an expensive exercise to take a sample, we like our samples to be as small as possible. Thus if an estimate of π is available, we should always use it to determine sample size, because it will yield a smaller n. If you wish to err on the side of caution, then, in the sample size formula, use a value for P which is a little closer to 0.5 than your estimate of π.

If the answer to question 1 is that the population size is finite and of size N, then we determine n from the reliability margin expression incorporating the reduced variance:

$$n = \frac{N}{1 + \dfrac{(N-1)(L/100)^2}{1.96^2\ P(1-P)}}.$$

In finite populations randomly sampled without replacement, the formula for the required minimal sample size n will always be smaller than the required size for random sampling with replacement.

Example 6A: What size sample is needed in each of the following situations? The numbers 1 to 4 refer to the four relevant questions for determining the sample size.

(a) $1.N = \infty$	2. 95% level	$3.L = 3\%$	$4.P = 0.5$
(b) $1.N = \infty$	2. 95%	$3.L = 3\%$	$4.P = 0.25$
(c) $1.N = 10\ 000$	2. 95%	$3.L = 2\%$	$4.P = 0.35$
(d) $1.N = 20\ 000$	2. 99%	$3.L = 1\%$	$4.P = 0.10$
(e) $1.N = 15\ 000$	2. 99%	$3.L = 1.5\%$	$4.P = 0.85$.

(a) $n = (100/3)^2 \times 1.96^2 \times 0.5 \times 0.5 = 1068$.

(b) $n = (100/3)^2 \times 1.96^2 \times 0.25 \times 0.75 = 801$.

(c) $$n = \frac{100\ 000}{1 + \dfrac{10\ 000(2/100)^2}{1.96^2 \times 0.35 \times 0.65}} = 1794.$$

(d) $$n = \frac{20\ 000}{1 + \dfrac{20\ 000(1/100)^2}{2.58^2 \times 0.1 \times 0.9}} = 4610.$$

(e) $$n = \frac{15\ 000}{1 + \dfrac{15\ 000(1.5/100)^2}{2.58^2 \times 0.85 \times 0.15}} = 3015.$$

Example 7B: The Student Health Service at a university with 12 000 students wishes to conduct a Health Awareness Survey by interviewing a sample of students. They desire a reliability margin of not more than 5% at a 95% confidence level in all questions that seek to estimate proportions. What size sample is required?

We are given $N = 12\,000$, we are told to find a 95% confidence interval (so $z = 1.96$) and to use reliability margin of $L = 5\%$. In the absence of any guidance about the likely value of π, we use $P = 0.5$ in the sample size formula because it gives the maximum possible sample size. Thus

$$n = \frac{12\,000}{1 + \dfrac{12\,000(5/100)^2}{1.96^2 \times 0.5 \times 0.5}} = 373.$$

Example 8C: In a town of 45 000 households, a market research organisation is conducting a survey into the use of three brands of soap powder. They want to estimate the proportion of households using each product, and at the 95% confidence level, they want a reliability margin of 2%. They provisionally estimate the proportion of users of Brand A as 15%, of Brand B as 22% and of Brand C as 17%. What size sample do they need to take?

Testing that the proportion is a specific value...

Example 9A: It is suspected that, in a lower-middle-class suburb, residents replace their cars less frequently than the national average. For example, it is known that, nationally, the proportion of cars under three years old is 27.1%. A researcher investigates and finds that 37 out of 155 cars belonging to residents of the suburb, were less than 3 years old. At the 5% significance level, test whether the proportion of cars in the suburb is less than the national average?

1. The null hypothesis states that the true proportion of residents in the lower-middle-class suburb with new cars is the same as the national proportion:

$$H_0 : \pi = 0.271.$$

2. The alternative hypothesis expresses the suspicion:

$$H_1 : \pi < 0.271.$$

3. Significance level : 5%.

4. The test statistic, we will see, has an approximate normal distribution. Thus, as in chapter 8, we will reject H_0 if the observed value of $z < -1.64$.

5. Test statistic. We asserted earlier that the random variable P, the estimate of π, has a normal distribution. If H_0 is true, then if P is based on a sample of size n, its distribution is given by

$$P \approx N\left(\pi,\ \frac{\pi(1-\pi)}{n}\right).$$

Thus the test statistic is

$$Z = \frac{P - \pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}} \approx N(0,1).$$

Note that the null hypothesis specifies the value for π. For this problem, $P = 37/155 = 0.239$ and $n = 155$, from the sample, and $\pi = 0.271$ is specified by the null hypothesis. Thus

$$\begin{aligned} z &= (0.239 - 0.271) \Big/ \sqrt{\frac{0.271(1-0.271)}{155}} \\ &= -0.896. \end{aligned}$$

6. Because $-1.64 < -0.896$, we are not able to reject H_0. Thus our data does not indicate that residents of this suburb buy new cars significantly (i.e. discernibly) less frequently than the national average.

Example 10B: A town of 3000 households is subjected to an intensive advertising campaign for Easyspread yellow margarine, including free samples. Beforehand, the proportion of Easyspread users in the town was the same as the national average, 0.132. One week later, a sample of 350 households was asked if they had bought Easyspread within the last seven days, and 64 made a positive response. Has the campaign been effective?

1. $H_0 : \pi = 0.132$, the proportion of Easyspread users has not changed.
2. $H_1 : \pi > 0.132$ (if the campaign is successful).
3. Test statistic. Because of the finite population size we use the adjusted variance:

$$P \dot{\sim} N\left(\pi,\ \frac{\pi(1-\pi)(N-n)}{n(N-1)}\right).$$

Our test statistic is thus

$$Z = (P - \pi) \Big/ \sqrt{\frac{\pi(1-\pi)(N-n)}{n(N-1)}} \dot{\sim} N(0,1).$$

For this example, $P = 64/350 = 0.183$, $n = 350$, $N = 3000$ and $\pi = 0.132$:

$$\begin{aligned} z &= (0.183 - 0.132) \Big/ \sqrt{\frac{0.132 \times 0.868 \times 2650}{350 \times 2999}} \\ &= 3.00. \end{aligned}$$

4. From the tables of the normal distribution at $z = 3.00$ we obtain we obtain $Pr = 0.49865$, so that this z-value is significant at better than the 0.5% level.
5. We conclude that the campaign has been highly effective in increasing Easyspread's share of the market ($z = 3.00$, $P < 0.005$).

Example 11C: A manufacturer claims that his market share is 60%. However, a random sample of 500 customers reveals that only 275 are users of his product. Test at the 5% significance level whether the population market share is less than that claimed by the manufacturer.

Testing for a difference between proportions...

Example 12A: At election time surveys are conducted in two suburbs which fall into the same constituency. In Broomhill, 175 voters out of a sample of 318 were in favour of a given candidate. In Crosspool, 143 voters out of a sample of 307 were in favour of this candidate. At the 5% level, is there a difference between the proportions of voters supporting the candidate in each suburb?

1. Let π_1 and π_2 are the population proportions supporting the candidate in Broomhill and Crosspool, respectively. The null hypothesis says that the proportions in each suburb are the same:
   $H_0 : \pi_1 = \pi_2$.

2. The alternative hypothesis states the population proportions are unequal:
   $H_1 : \pi_1 \neq \pi_2$.

3. Significance level : 5%.

4. Rejection region. Assuming that the test statistic will have the standard normal distribution, and because we have a two-tailed test at the 5% significance level, we will reject H_0 if $|z| > 1.96$.

5. Test statistic. We know that P_1 and P_2, the sample estimates of π_1 and π_2 based on samples of size n_1 and n_2 respectively have normal distributions

$$P_1 \dot{\sim} N\left(\pi_1\ ,\ \frac{\pi_1(1-\pi_1)}{n_1}\right) \qquad P_2 \dot{\sim} N\left(\pi_2\ ,\ \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

   In chapter 5, we saw that the difference between two random variables, each having a normal distribution, is also a normal distribution. Thus we can write

$$P_1 - P_2 \dot{\sim} N\left(\pi_1 - \pi_2\ ,\ \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

   If H_0 is true, then $\pi_1 = \pi_2 = \pi$ (say) and

$$P_1 - P_2 \dot{\sim} N\left(0\ ,\ \pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

   The value of π needs to be estimated from the observed valus $Y_1 = 175$ and $Y_2 = 143$, random samples of size n_1 and n_2. If H_0 is true, then P_1 and P_2 are both estimates of π, and we combine them to get a "pooled" estimate of π. The pooled estimate is called P, and is computed as a weighted average of P_1 and P_2:

$$P = (n_1P_1 + n_2P_2)/(n_1 + n_2) = (Y_1 + Y_2)/(n_1 + n_2).$$

   We can say, approximately, that

$$P_1 - P_2 \dot{\sim} N\left(0, P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

   so that the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \dot{\sim} N(0,1)$$

For our example, $P_1 = 175/318 = 0.550$, $P_2 = 143/307 = 0.466$, $n_1 = 318$, $n_2 = 307$. The pooled estimate of P is given by

$$P = (318 \times 0.550 + 307 \times 0.466)/(318 + 307) = (175 + 143)/625 = 0.509.$$

We compute the test statistic:

$$z = \frac{0.550 - 0.466}{\sqrt{0.509 \times 0.491 \times \left(\frac{1}{318} + \frac{1}{307}\right)}} = 2.10.$$

6. We reject H_0 because the test-statistic $z = 2.10$ satisfies $z > 1.96$, and conclude that the difference between the proportions is, at the 5% level, significant.

Example 13B: Miss Jones, the senior typist, made errors on 15 out of 125 pages of typing, whilst Miss Smith made errors on 44 pages out of 255 pages of typing. Is Miss Jones' error rate significantly lower than Miss Smith's?

1. $H_0 : \pi_1 = \pi_2$.
2. $H_1 : \pi_1 < \pi_2$.
3. The test statistic is

$$Z = (P_1 - P_2) \Big/ \sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

We have $P_1 = 15/125 = 0.120$, $P_2 = 44/255 = 0.173$, $n_1 = 125$, $n_2 = 255$. We compute the pooled estimate P:

$$P = (125 \times 0.120 + 255 \times 0.173)/(125 + 255) = (15 + 44)/380 = 0.155.$$

Thus

$$\begin{aligned} z &= (0.120 - 0.173) \Big/ \sqrt{0.155 \times 0.845 \left(\frac{1}{125} + \frac{1}{255}\right)} \\ &= -1.34. \end{aligned}$$

The test statistic is significant only at the 10% level. We conclude that the difference between the error rates is nearly significant ($z = -1.34$, $P < 0.10$).

Example 14C: In a random sample of 350 students, 47 were overweight, while in a random sample of 176 businessmen, 36 were overweight. Do these data support the hypothesis that a larger proportion of businessmen are overweight than students?

You can also test the hypothesis of Example 14C using the test of association in a 2×2 contingency table, which we learnt in Chapter 10. Check this statement. The two tests can be shown to be mathematically equivalent; both make the same assumptions, and will always lead to the same decision about accepting or rejecting the null hypothesis.

Random Sampling methods...

We have stated in many examples that we have "a sample" or "a random sample". It is now time to consider how random samples are obtained. An important branch of statistics is called **sampling theory**, and we will introduce a few of the concepts.

Two general types of sampling exist — **probability** and **non-probability**. Probability sampling includes simple random sampling, stratified, cluster, area, double (two-phase), multi-stage and sequential random sampling. Non-probability sampling includes judgement, systematic and quota sampling and is characterized by the fact that no check can be kept on sampling errors for estimates. Various combinations of these methods may, or sometimes must, be used to fit individual circumstances. Randomness should be sought wherever possible, but problems of non-response and organization sometimes mitigate against this ideal.

A standard method of drawing a **simple random sample** is the use of **random number** tables. Table 6 is an example of such a table. Tables of up to one million random digits have been prepared for this purpose, and there are standard computer package programs that will easily generate them.

Assume we wish to draw a simple random sample of 320 from a population of 8724 dwelling units, each of which has been allocated a number from 1 to 8724. We start by artibrarily choosing a starting point somewhere in the table, say the 6th row and the lst column in table 6. This starting point gives us the numbers 1164 4842 2873. We use these four digit numbers to select the sample. As long as we work through the table systematically (here along the rows) we will obtain random digits.

Thus we would get the numbers 1164, 4842, 2873, 6089, 9329, 7601, 5677, 7791, 5219, 7374 etc. If a number occurs which is greater than 8724 we ignore and delete it and select another until we have selected 320 distinct numbers between 1 and 8724. Thus we would delete 9329. If the population consists of data stored on a computerized accounting system or municipal records file, the procedure is simple. Units can also be chosen on an area or spatial basis, where each random number defines the coordinates of a block or quadrat. The important thing to remember is that each unit being potentially sampled should have an equal chance of being chosen. Consequently, samples of families chosen on an area basis should take into account varying population densities to ensure equal probabilities of selection for all families. All the statistical methods we have devised in this course are applicable to simple random samples. We mention briefly some other sampling strategies.

Stratification is the separation of an entire population into several groups (called strata), on the basis of some known information. Strata are constructed so that the elements within any stratum are very similar to one another, but also tend to be at least somewhat different from elements in the other strata. Stratifying the population enables us to eliminate the possibility of obtaining particular kinds of unrepresentative samples. By randomly sampling within each stratum we obtain a composite random sample that is suitably balanced across the strata.

The population and hence the sample may be stratified with regard to one or more classification variables, the number depending on practical limitations (e.g. gender, education level, type of dwelling being known for each person in a region of interest, before the sampling begins).

In cases where the proportions in the sample from each subpopulation are the same as the proportions in the population we say we have **proportional stratified random sampling**. When the proportions do not correspond, some form of weighting must usually be applied to each subsample in order to draw conclusions about the whole

population.

Often the stratification allows us to obtain very much better estimates and much narrower confidence intervals, than corresponding amounts spent on a simple random sample. Alternatively, we may save on sampling costs because we can achieve the same reliability margins as a simple random sample, but with a smaller total count of sample units.

Cluster random sampling focuses upon subgroups of a population that are conveniently identified and sampled. It may be easier to sample Cape Town addresses than to sample the Cape Town population. Clustering reduces the cost of taking a random sample by ensuring that the units with sampled clusters are geographically close to each other, and travel costs to obtain data are less than when individual persons have to be located and interviewed.

Whereas strata are sets of units constructed on the basis of additional knowledge about similarities and contrasts between units available ahead of the sampling, clusters are sets of units constructed purely upon a basis of convenient access.

If the householder population of a country are to be sampled with regard to housing questions, one might first choose a random sample from the list of towns and cities (as Stage 1) from the set of towns and cities. Stage 2 might involve random selection from the list of suburbs in each town chosen, Stage 3 a random sample from a list of blocks within each suburb selected, and at Stage 4 a random selection from a list of houses from each block.

An overall margin of error statement can be made by combining the variations or sampling errors at each of the four stages. This strategy is also known as **multi-stage** cluster random sampling. One can easily see that this method will be more convenient than a random sample from the list of all householders in the country. The convenience comes with some cost: the confidence intervals for cluster random sampling are wider than for simple random sampling.

Sometimes two or more random samples are taken from the population at distinct points in time. The first random sample is often used to give an indication of the sample size required for the second and main random sample, and to aid in its stratification. The questions of importance appear in the second sample. This strategy is called **double** or **two-phase** sampling. **Sequential** random sampling also involves a series of random samples, usually to keep a continuous check on some feature which may change with time, or to provide further information on a particular aspect. Depending on circumstances and the information required, each sample may or may not include the same units.

Systematic random sampling involves selecting a random starting point and then every k-th unit in a list or in a moving system thereafter, e.g., the twenty-third person leaving a train station, and then every hundredth person. This strategy will often give good results, but is subject to immeasurable and perhaps large errors if the interval between units sampled coincides with a cycle inherent in the population sampled. An obvious example of a difficulty is sampling every 14th (or 21st) unit of a population of daily petrol sales figures, when a regular pattern occurs over a seven day sales week. Other cycles may, unfortunately, not be as obvious.

In all forms of probability sampling, the problem of **non-response** is an important issue. Suppose we choose a random sample of voters using the electoral register and random numbers. The interviewers may have to make several calls to find people at home; they may have moved, be on holiday or in hospital; they may refuse to be interviewed. If we have a sample of 1000 and 900 respond, can we simply regard this as a random sample of 900? It would be dangerous to do so, because we would then be assuming that the non-responders hold similar views to the responders. The very fact that they

are non-responders makes them different to the others. The difference probably extends to their opinions on the subject of the questionnaire.

If we did proceed to assume that respondents and non-respondents were similar enough to ignore the 100, and treat the 900 as a random sample, we have an ethical obligation to report that assumption as a fundamental basis of our analysis. This strategy can be made explicit, but we will be inlikley to ever know whether or not it was valid.

In short, despite our best efforts to use random selection to increase the chance of near-representative samples, non-response within a well-chosen random sample may cause important parts of the information to be hidden from our analysis.

Non-random Sampling methods...

Judgement sampling selects units according to the views of the sampler. However experienced and objective an expert may be in choosing what he feels is a representative or balanced sample, the possibility of unintentional bias in the selections can never be ruled out. Thus the findings of his study will only have credibility with those people who believe the sample was balanced, and may have no credibility beyond that group.

Quota sampling attempts to remove the problem of non-response. The interviewer is given a description of the types of people to be interviewed and the number of each type required. The selection of actual individuals is then left to the interviewer. This subjective selection means that, like systematic and judgement sampling, randomness is not present, and there is no real check on the size of the error in sampling. The British Market Research Society, asked to explain the poor performance of almost all opinion polls in predicting results of the 1970 British general election, cited the use of quota sampling as a major reason. Although unintentional interviewer bias had in the past often tended to "even out", the only way in which this bias could consistently be avoided was through some form of random or probability sampling.

Designing sample surveys...

(a) Before embarking on a major survey it is best to arrange a **pilot** survey. This small-scale study will test the effectiveness and shortcomings of the questionnaire and give first-hand information on problems facing the interviewer. It is well worth the extra effort and time involved, as it enables one to revise ambiguous questions, insert reply categories which might have been overlooked, and even change to a more efficient survey procedure. If a major error is only discovered during or after the main survey, economic reasons usually prevent one from starting afresh, and the entire exercise can be largely a waste of time and money.

(b) **Questionnaires** should be carefully drawn up, taking into account the information that is really desired, the range of possible answers, and the types of people likely to be interviewed. Wherever possible, people concerned with the survey subject,the study population or study area should be consulted. Questions should be uncomplicated and kept to a minimum.
Data will often be entered directly into the computer from a questionnaire — the layout should be checked in advance for its suitability at the data capture stage. The questionnaire should also be discussed beforehand with the person responsible for summarizing and analysing the information, particularly if a computer is involved. The question of whether or not to process the information by computer should, of course, be considered at a very early stage. This decision will depend

largely on the size of the sample, the number of questions in the questionnaire, the complexity of the information required and the relative costs.

(c) **Confidence levels** are usually selected to be 95% or 99%, although in circumstances with high sampling costs, 90% may be the maximum achievable in practice.

(d) A **reliability margin** of about 5% is usually quite satisfactory, but anything larger than 10% is not of much practical use. A smaller margin demands a **much** greater sample. There is nowhere near a linear relationship between size of population and size of sample necessary for a fixed reliability margin $L\%$. There is no basis for the often held view that "5% (or 10%) of the population will always give a satisfactory sample." It should be borne in mind that the reliability margin is an absolute value and that a figure of $14\% \pm 3\%$ is relatively more reliable than one of $45\% \pm 5\%$, because the margin is smaller.

(e) We have so far restricted ourselves to a question with either a "yes" or "no" answer. The same theory holds for questions with **any number of categories** — if P_i is the percentage of replies to category i, we use the same theory with P_i instead of P.

(f) If any conclusions are to be drawn specifically for a **subgroup of the population**, e.g. all bachelors, then the reliability margin of any of these conclusions is dependent on the **numbers of the specific subgroup** in the sample and in the population. Although 400 may give an accurate balanced view of a population of 40 000, it is simply naïve and completely incorrect to expect 10 to represent an accurate and balanced view of a sub-population of 1000.

(g) These sample size figures all relate to a **randomly drawn** sample, without obvious **bias**. Even in the balanced case, different levels of **response** (acceptance and satisfactory completion of the questionnaire) from different types in a heterogeneous population may cause bias. Here suitable stratification prior to random selection might help in maintaining the correct balance. Questionnaires can be postal, by personal interview or postal with a personal follow-up. The postal method saves much time and costs, but suffers much more from problems of non-response.

(h) Finally, the above points all attempt to deal with and minimise possible inaccuracies caused by drawing a sample which is not statistically representative of the population. These selection errors may be minimal compared with those errors resulting from **poor interviewing, bad questions, incorrect transcription** and **faulty processing** and **flawed interpretation**. A moderate-sized random sample, well controlled and carefully processed, may well give better results than a complete census, the sheer size of which can cause very rushed interviewing and processing and a greater proportion of errors.

The classic sampling fiasco is that of the "Literary Digest" poll, which sampled over 2 million opinions on the winner of the 1936 American Presidential Election. The L.D. predicted only 40.9% of the votes for Roosevelt and a landslide win for Landon. Actual returns gave Roosevelt an overwhelming win with 60.7% of the vote. Sample size 2 million — final error 20%! This huge error was due to plain stupidity, in picking the sample from telephone directories and car registration files. The L.D. survey specialists may have effectively sampled the opinions of the upper-class and middle-class people, but left out the "ill-clad, ill-fed and ill-housed" lower-income classes who voted overwhelmingly in favour of Roosevelt's "New Deal" policies.

Example 15C: For each of the following situations describe an appropriate sampling technique.

(a) The Electricity Supply Commission is interested in evaluating the effect of an energy conservation advertising campaign. It wishes to estimate the proportion of households which are acting to reduce their electricity consumption.
(b) The City Council wishes to know the proportion of residents who favour the construction of a new city by-pass road.
(c) A chain store with outlets in nine regions wishes to estimate the number of bad debtors nationwide.
(d) The university library wants to estimate the proportion of books that have not been used for a year. A book is defined to have been used if it has been date-stamped within the past year.
(e) A manager of a commercial forest plantation wishes to determine the proportion of trees in a plantation that have been infested with an insect pest.

Solutions to examples...

3C $(0.518, 0.640)$ or $(51.8\%, 64.0\%)$ $L = 6.1\%$.

5C Materialist Party: $(53.4\%\ ,\ 58.2\%)$ $L = 2.4\%$.
Ecological Party: $(35.6\%\ ,\ 40.4\%)$ $L = 2.4\%$.
Undecided: $(5.0\%\ ,\ 7.4\%)$ $L = 1.2\%$.

8C 1590 (Use value of P closest to 0.5, i.e. the 0.22 of Brand B.)

11C $z = -2.282 < -1.64$, reject H_0.

14C $z = -2.086$, $P < 0.05$, significant.

Exercises on confidence intervals and sample sizes...

*11.1 In a sample of 500 garages it was found that 170 sold tyres at prices below those recommended by the manufacurer.

(a) Estimate the percentage of all garages selling tyres below the recommended price.
(b) Calculate the 95% confidence limits for this estimate.
(c) What is the reliability of the estimate?
(d) What size sample would have to be taken in order to estimate the percentage to within 2% at a 95% confidence level? Use the value obtained in (a) as a rough estimate of the true proportion.

*11.2 What size sample is needed in each of the following situations? Sampling will be done without replacement.

	Population size N	Confidence level	Reliability $L\%$	Rough estimate of π
(a)	infinite	95%	5%	50%
(b)	infinite	95%	1%	50%
(c)	infinite	95%	5%	25%
(d)	infinite	99%	5%	25%
(e)	infinite	95%	5%	75%
(f)	infinite	95%	1%	10%
(g)	3000	95%	3%	40%

(h)	10 000	95%	3%	40%
(i)	10 000	95%	1%	40%
(j)	10 000	99%	3%	20%
(k)	10 000	99%	2%	12%
(l)	40 000	95%	1%	8%
(m)	infinite	95%	2%	unknown
(n)	7000	95%	0.5%	97%

11.3 A simple random sample from a community of 6000 is to be asked a question to determine the proportion in favour of some proposal.

(a) What size sample is needed if the result is required to have a reliability of 3% at a 95% confidence level, and it is known in advance that the percentage in favour will exceed 70%?

(b) If 840 out of 1000 members of the community were in favour, find a 99% confidence interval for the community viewpoint on the proposal. What is the reliability?

11.4 The proportion of people owning citizen band radios in a town of 7500 people is to be estimated by means of a simple random sample, without replacement.

(a) What size sample is needed if it is required that the sample result be within 4% of the true result at a 99% confidence level?

(b) If there were 235 citizen band radio owners in a sample of 850 people, find a 95% confidence interval for the proportion in the town as a whole. What is the reliability of the estimate?

*11.5 A survey is to be made of student participation in sport within the three faculties of a university. The faculties have 1000, 4000 and 5000 students. The proportion of students participating in sport is known to be well under 30%.

(a) What size sample is needed within each faculty if a reliability of 3% is required at the 95% confidence level?

(b) What size sample is required if it is only necessary to estimate the overall proportion within the university with the same reliability and confidence level?

(c) Explain the difference between the overall sample sizes obtained in (a) and (b).

11.6 You read in a newspaper report that 53% of businessmen think that the economic climate is improving. Upon checking you learn that the sample size used in the research was 100. Find the 95% confidence interval and comment.

*11.7 A questionnaire has four questions, and the rough estimate of the proportions of interest are 10%, 70%, 80% and 5%. What size sample is required to achieve a reliability of 2% at the 90% confidence level in all four questions? The population may be assumed to be infinite.

11.8 In a random sample of 120 pages typed by Miss Brown there were 23 pages with typing errors.

(a) Find a 90% confidence interval for the proportion of Miss Brown's pages that have errors.

(b) How large a sample would we need if we wanted our confidence interval to have overall length 4%?

Exercises on hypothesis testing...

11.9 A manufacturer guarantees that only 10% of vehicles have brake defects in the first 10 000 km. Feeling that this is an underestimate, a consumer organization undertook an investigation which showed that 18 out of 110 vehicles examined had brake defects within 10 000 km. At the 1% significance level, do these data disprove the manufacturer's claim?

*11.10 A certain type of aircraft develops minor trouble in 4% of flights. Another type of aircraft develops similar trouble in 19 out of 150 flights. Investigate the performance of the two types of aircraft and comment on any significant difference.

11.11 The national average for the ownership of motorbikes by teenagers is 15%. In an affluent suburb which has 2000 teenagers, 45 out of a sample of 250 teenagers owned motorbikes. Test if this proportion is significantly more than the national average.

*11.12 After corrosion tests, 42 of 536 metal components treated with Primer A and 91 of 759 components treated with Primer B showed signs of rusting. Test the hypothesis that Primer A is superior to Primer B as a rust inhibitor at the 1% significance level.

11.13 In a sample of 540 wives of professional and salaried workers, 42% had visited their doctor at least once during the preceding 3 months. During the same period, of a sample of 270 wives of labourers and unskilled workers, 36% had visited a doctor. By the use of an appropriate statistical test, consider the validity of the assertion that middle-class wives are more likely to visit their doctors than the wives of working-class husbands.

11.14 In a sample of 569 wives of professional and salaried workers 45% attended weekly the local welfare centre with their infants. For a sample of 245 wives of agricultural workers, the corresponding proportion was 35%. Test the hypothesis that there is no difference between the two groups in respect of their attendance at such centres.

*11.15 Two groups, A and B, each consist of 100 people who have a disease. A serum is given to Group A but not to Group B (which is called the control group); otherwise, the two groups are treated identically. It is found that in Groups A and B, 75 and 65 people, respectively, recover from the disease. Test the hypothesis that the serum helps to cure the disease.

11.16 A sample poll of 300 voters from district A and 200 voters from district B showed that 168 and 96 respectively were in favour of a given candidate. At a level of significance of 5%, test the hypothesis that

(a) there is a difference between the districts
(b) the candidate is preferred in district A.

11.17 Two separate groups of sailors were randomly selected. One group of 350 sailors was given seasickness pills of Brand A and another group of 220 sailors was given pills of Brand B. The number of sailors in each group that became seasick during a very heavy storm were 57 and 28 respectively. Can one conclude, at a 5% significance level, that there is no real difference in the effectiveness of the pills?

*11.18 There are 3000 students in the Arts Faculty and 2500 in the Science Faculty of a university. In a sample of 350 arts students there were 186 non-smokers, while of 400 science students, 273 were non-smokers. Is there a difference between the proportions of non-smokers in the two faculties? (Develop a test that adjusts the variances for the finite population sizes.)

11.19 In a study to estimate the proportion of residents in a certain city and its suburbs who were opposed to the construction of a nuclear power plant, it was found that 48 out of 100 urban residents were opposed to the construction of the power plant, while 91 out of 125 suburban residents were opposed. Test whether the level of opposition to the nuclear power plant varies significantly between urban and suburban residents.

11.20 In an election, one ballot box at a polling station was found to have a broken seal, and there was concern that votes for a particular party, called ABC, had been removed and destroyed. None of the other boxes had broken seals. Of the remaining votes cast at that polling station, 32% were in favour of party ABC. The ballot box with the broken seal contained 543 votes, of which 134 were for party ABC. Does this information provide support for the allegation that the ballot box had been tampered with? What assumptions are required to conduct the test?

Exercises on sample surveys ...

*11.21 What is the aim of choosing a random sample? Mention two modifications that can be made to a simple random sampling plan, and outline the benefits of these modifications.

*11.22 A statistician wishes to assess the opinion that residents of a suburb have about their local bus service. Describe how he might go about obtaining a representative sample of their opinions.

*11.23 What objections would you raise if you were told to use, as a random sample of the households in the Cape Town area, the first three addresses on the top of each page of the Cape Peninsula telephone directory?

Solutions to exercises...

11.1 (a) 0.34 or 34% (b) (.382 , 0.298) or (38.2% , 29.8%)
(c) $L = 4.2\%$ (d) 2156.

11.2

(a)	385	(b)	9604	(c)	289	(d)	500
(e)	289	(f)	3458	(g)	764	(h)	930
(i)	4798	(j)	1059	(k)	1495	(ℓ)	2640
(m)	2401	(n)	2728				

11.3 (a) 780 (b) (0.813 , 0.867) or (81.3% , 86.7%) $L = 2.7\%$.

11.4 (a) approximately 911 (b) (0.248 , 0.305) $L = 2.83\%$.

11.5 (a) 473 out of 1000, 732 out of 4000 and 760 out of 5000, a total of 1965.
(b) 823.

11.6 (43.2% , 62.8%). The confidence interval is so long that it is of little use.

11.7 $n = 1412$. (Use the estimate closest to 0.50, i.e. 0.70).

11.8 (a) (13.3% , 25.1%) (b) $L = 2\%$, 1042.

11.9 $z = 2.23 < 2.33$, cannot reject H_0.

11.10 $z = 5.47$, $P < 0.0001$, very highly significant.

11.11 $z = 1.42$, $P < 0.10$, significant.

11.12 $z = -2.43 < -2.33$, reject H_0.

11.13 $z = 1.643$, $P = 0.05$, significant.

11.14 $z = 2.65$, $P < 0.01$, significant.

11.15 $z = 1.54$, $P < 0.10$, significant

11.16 (a) $z = 1.76 < 1.96$, cannot reject H_0.
(b) $z = 1.76 > 1.64$, reject H_0.

11.17 $z = 1.16 < 1.96$, cannot reject H_0.

11.18 The test statistic is

$$z = (P_1 - P_2) \Big/ \left[P(1-P) \left(\frac{(N_1 - n_1)}{n_1(N_1 - 1)} + \frac{(N_2 - n_2)}{n_2(N_2 - 1)} \right) \right]^{\frac{1}{2}}$$

$z = -4.56$, $P < 0.0001$, very highly significant.

11.19 There is a significant difference between the urban and suburban residents ($z = 3.804$, $P < 0.0005$).

11.20 There is evidence of removal of votes ($z = -3.66$, $P < 0.0005$). (One-sided alternative used, as suggested by the question.) The chief assumption is votes were randomly allocated to the boxes. If the boxes were not used simultaneously but sequentially, the proportion voting for party ABC is assumed to have remained constant at the polling station across all the two sets of boxes (with broken and unbroken seals).

Chapter 12

REGRESSION AND CORRELATION

KEYWORDS: Regression, scatter plot, dependent and independent variables, method of least squares, the normal equations, regression coefficients, correlation coefficient, residual standard deviation, confidence intervals for predictions, nonlinear regression, exponential growth.

Advertising and sales, wages and productivity, class record and final mark...

We often encounter problems in which we would like to describe the relationship between two or more variables. A person in marketing will want to determine the number of sales associated with advertising expenditure, an economist will want to know how wages effects the productivity of an industry, and you, as a student of statistics, should be able to define a more precise problem, and ask: "Given my class record, find a 95% confidence interval for my final mark."

In this chapter we will make a start at answering two questions:

1. **Is there really a relationship between the variables? (the correlation problem).** If the answer to this question is yes, we go on to ask:
2. **How do we predict values for one variable, given particular values for the other variable(s)? (the regression problem).**

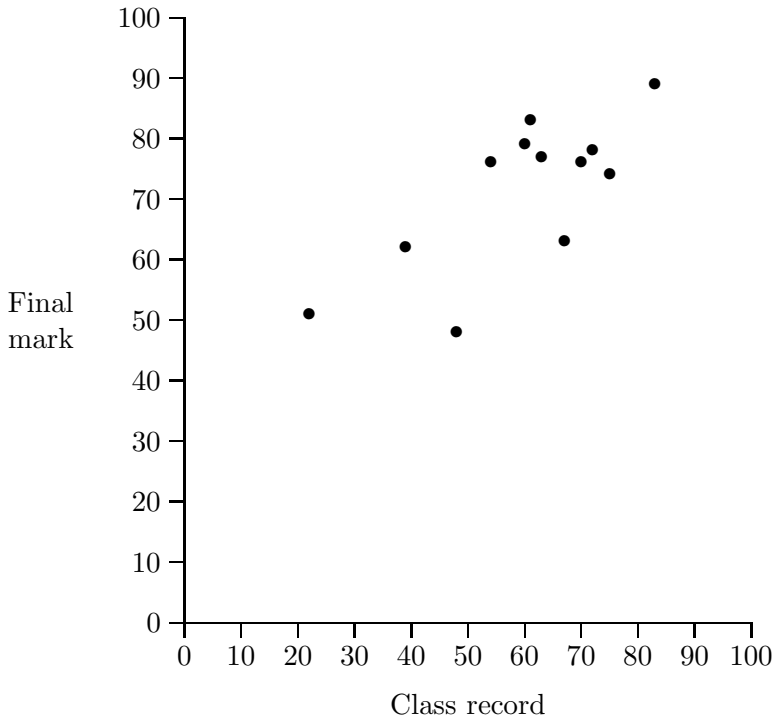
We consider the second question first.

Regression...

Example 1A: Suppose, as suggested above, we would like to predict a student's mark in the final examination from his class record. We gather the following data, for 12 students. (In practice we would use far more data; we are now just illustrating the method.)

Class record	Final mark
61	83
39	62
70	76
63	77
83	89
75	74
48	48
72	78
54	76
22	51
67	63
60	79

We depict the data graphically in what is known as a **scatter plot** , or **scattergram**; or **scatter diagram**:



A haphazard scattering of points in the scatter plot would show that no relationship exists. Here we have a distinct trend — as x increases, we see that y tends to increase too.

We are looking for an equation which describes the relationship between mid-year mark and final mark — so that for a given mark at the mid-year, x, we can **predict** the final mark y. The equation finding technique is called **regression analysis**. We call the variable to be predicted, y, the **dependent variable**, and x is called the **explanatory variable**.

The variable x is often called the **independent variable**, but this is a **very** poor name, because statisticians use the concept of **independence** in an entirely different context. Here, x and y are **not** (statistically) independent. If they were, it would be stupid to try to find a relationship between them!

Correlation...

To justify a regression analysis we need first to determine whether in fact there is a **significant** relationship between the two variables. This can be done by **correlation analysis**.

If all the data points lie close to the regression line, then the line is a very good fit and quite accurate predictions may be expected. But if the data is widely scattered, then the fit is poor and the predictions are inaccurate.

The goodness of fit depends on the **degree of association** or **correlation** between the two variables. This is measured by the **correlation coefficient**. We use r for the correlation coefficient and define it by the formula

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

For computational purposes, the most useful formula for the correlation coefficient is expressed in terms of SS_x, SS_{xy} and SS_y:

$$r = \frac{SS_{xy}}{\sqrt{SS_x\, SS_y}}.$$

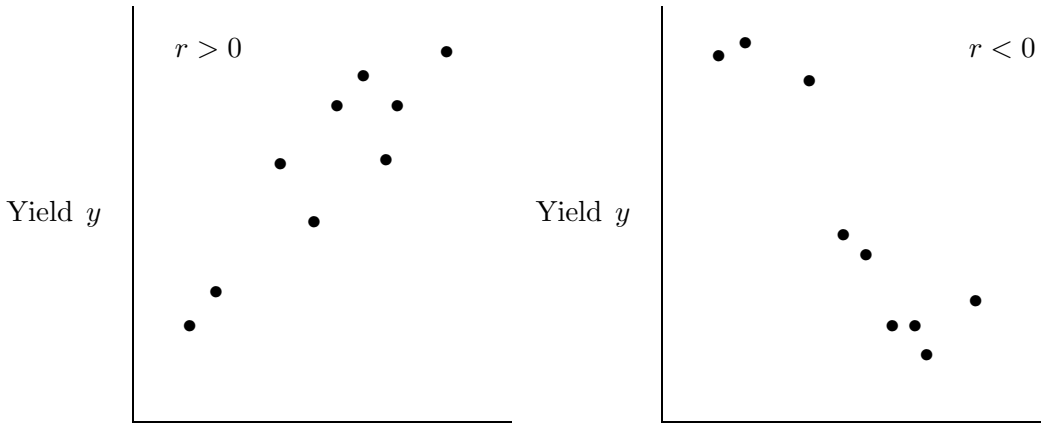
SUMS OF SQUARES

$$SS_x = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$
$$SS_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n}$$
$$SS_y = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2)}{n}$$

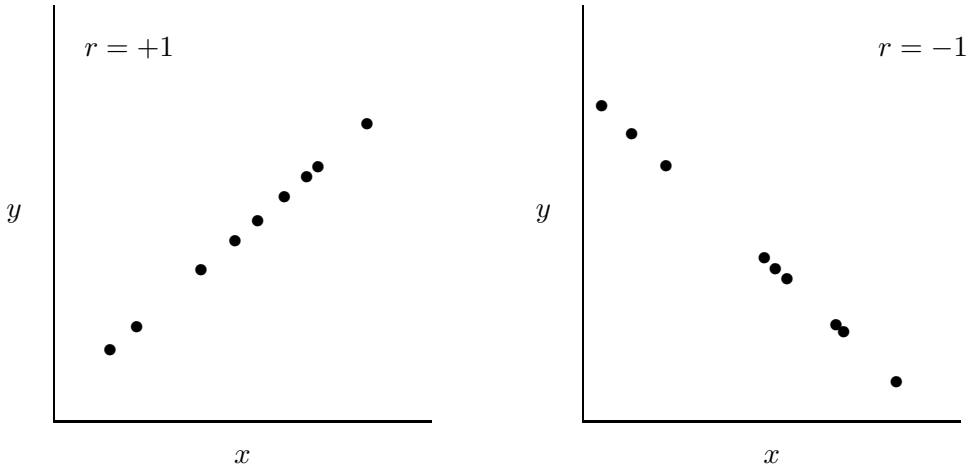
The letters "SS" stand for Sum of Squares, and the subscript(s) indicate(s) the variable(s) in the sum.

The correlation coefficient r **always** lies between -1 and $+1$. If r is positive, then the regression line has positive slope and as x increases so does y. If r is negative then the regression line has negative slope and as x increases, y decreases.

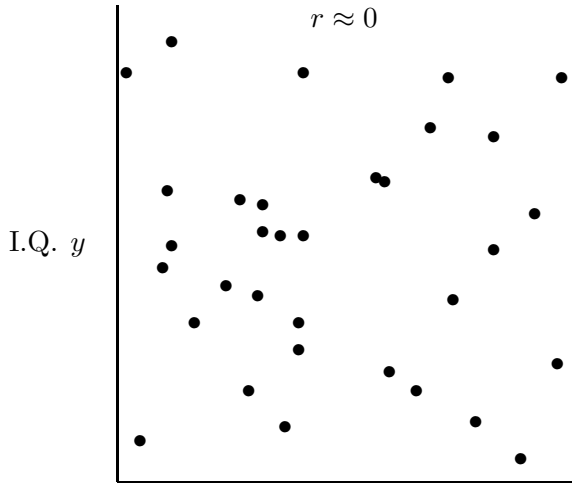
In the **scatter plot** on the left, both the correlation coefficient r and the slope coefficient of the regression line will be positive. In contrast, both r and slope will be negative in the plot on the right:



We now look at the extreme values of r. It can readily be shown that r cannot be larger than plus one or smaller than minus one. These values can be interpreted as representing "perfect correlation". If $r = +1$ then the observed data points must lie exactly on a straight line with positive slope, and if $r = -1$, the points lie on a line with negative slope:



Half way between $+1$, perfect positive correlation, and -1, perfect negative correlation, is 0. How do we get zero correlation? Zero correlation arises if there is no relationship between the variables x and y, as in the example below:



Thus, if the data cannot be used to predict y from x, r will be close to zero. If it can be used to make predictions r will be close to $+1$ or -1. How close to $+1$ or to -1 must r be to show statistically significant correlation? We have tables which tell us this.

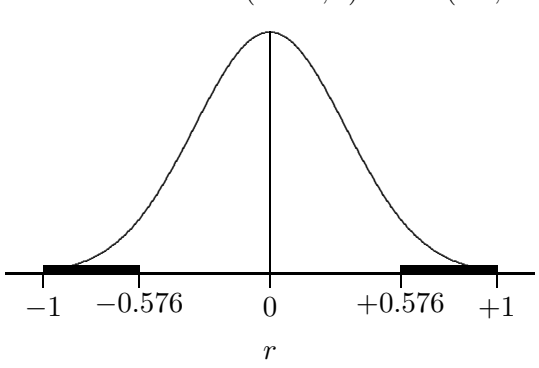
We express the above more formally by stating that r, which is the **sample** correlation coefficient estimates ρ, the **population** correlation coefficient. (ρ, "rho", is the small Greek "r" — we are conforming to our convention of using the Greek letter to signify the population parameter and the Roman letter for the estimate of the parameter.) What we need is a test of the null hypothesis that the population correlation coefficient is zero (i.e. no relationship) against the alternative that there is correlation; i.e. we need to test the null hypothesis $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$. Mathematical statisticians have derived the probability density function for r when the null hypothesis is true, and tabulated the appropriate critical values. Table 5 is such a table. The probability density function for r depends on the size of the sample, so that, as for the t-distribution, degrees of freedom is an issue.

The tables tell us, for various degrees of freedom, how large (or how small) r has to be in order to reject H_0. The alternative hypothesis above was two-sided, thus there must be **negative** as well as **positive** values of r that will lead to the rejection of the null hypothesis. Because the sampling distribution of r is symmetric, our tables only give us the "upper" percentage points. Unlike the earlier tests of hypotheses that we developed, there is no "extra" calculation to be done to compute the test statistic. In this case, r itself is the test statistic.

When the sample correlation coefficient r is computed from n pairs of points, the degrees of freedom for r are $n-2$. In terms of our "degrees of freedom rule", we lose two degrees of freedom, because we needed to calculate $\bar{x}$ and $\bar{y}$ before we could calculate r. We estimated two parameters, the mean of x and the mean of y, so we lose two degrees of freedom.

Example 2A: Suppose that we have a sample of 12 pairs of observations, i.e. 12 x-values and 12 y-values and that we calculate the sample correlation coefficient r to be 0.451. Is this significant at the 5% level, using a two-sided alternative?

1. $H_0 : \rho = 0$
2. $H_1 : \rho \neq 0$
3. Significance level : 5%
4. Because we have a two sided alternative we consult the 0.025 column of the table. We have 12 pairs of observations, thus there are 12-2 = 10 degrees of freedom. The critical value from Table 5 is 0.576. We would reject H_0 if the sample correlation coefficient lay either in the interval $(0.576, 1)$ or in $(-1, -0.576)$:



5. Our calculated value of r is 0.451.

6. Because the sample correlation coefficient r does not fall into the rejection region we are unable to say that there is significant correlation. Thus we would not be justified in performing a regression analysis.

Example 3C: In the following situations, decide whether the sample correlation coefficients r represent significant correlation. Use the modified hypothesis testing procedure, and give "observed" P-values.

(a) $n = 12,\ r = 0.66$, two-sided alternative hypothesis
(b) $n = 36,\ r = -0.25$, two-sided alternative hypothesis
(c) $n = 52,\ r = 0.52$, one-sided alternative hypothesis $H_1 : \rho > 0$
(d) $n = 87,\ r = 0.23$, one-sided alternative hypothesis $H_1 : \rho > 0$
(e) $n = 45,\ r = -0.19$, one-sided alternative hypothesis $H_1 : \rho < 0$.

Example 1A, continued: For the data given in Example 1A, test if there is a significant correlation between class record and final mark. Do the test at the 1% significant level.

We can justify using a one-sided alternative hypothesis — if this world is at all fair, if there is **any** correlation between class record and final mark it must be positive! Thus we have:

1. $H_0 : \rho = 0$

2. $H_1 : \rho > 0$

3. The sample correlation coefficient is

$$r = \frac{SS_{xy}}{\sqrt{SS_x\, SS_y}} = \frac{1764}{\sqrt{3119 \times 1748.67}} = 0.755$$

4. The sample size was $n = 12$, so we have 10 degrees of freedom. The correlation is significant at the 5% level ($r > r_{10}^{(0.05)} = 0.4973$), the 1% level ($r > r_{10}^{(0.01)} = 0.6581$), the 0.5% level ($r > r_{10}^{(0.005)} = 0.7079$), but not at the 0.1% level ($r > r_{10}^{(0.001)} = 0.7950$).

5. Our conclusion is that there is a strong positive relationship between the class record and the final mark ($r_{10} = 0.755,\ P < 0.005$).

Example 4B: A personnel manager wishes to investigate the relationship between income and education. He conducts a survey in which a random sample of 20 individuals born in the same year disclose their monthly income and their number of years of formal education. The data is presented in the first two columns of the table below.

Is there significant correlation between monthly income and years of education? Do the test at the 1% significance level.

Person	Years of formal education x	Annual income (1000's of rands) y	x^2	y^2	xy
1	12	4	144	16	48
2	10	5	100	25	50
3	15	8	225	64	120
4	12	10	144	100	120
5	16	9	256	81	144
6	15	7	225	49	105
7	12	5	144	25	60
8	16	10	256	100	160
9	14	7	196	49	98
10	14	6	196	36	84
11	16	8	256	64	128
12	12	6	144	36	72
13	15	9	225	81	135
14	10	4	100	16	40
15	18	7	324	49	126
16	11	8	121	64	88
17	17	9	289	81	153
18	15	11	225	121	165
19	12	4	144	16	48
20	13	5	169	25	65
$\sum$	275	142	3883	1098	2009

We complete the table by computing the terms x^2, y^2 and xy, and adding the columns.
Next, $\bar{x} = 275/20 = 13.75$ and $\bar{y} = 142/20 = 7.10$.
We now calculate SS_x, SS_{xy} and SS_y:

$$SS_x = \sum x^2 - (\sum x)^2/n = 3883 - 275^2/20 = 101.75$$
$$SS_{xy} = \sum xy - (\sum x)(\sum y)/n = 2009 - 275 \times 142/20 = 56.50$$
$$SS_y = \sum y^2 - (\sum y)^2/n = 1098 - 142^2/20 = 89.80.$$

1. $H_0 : \rho = 0$
2. $H_1 : \rho \neq 0$
3. Significance level : 1%
4. The sample size was $n = 20$, so the appropriate degrees of freedom is $20 - 2 = 18$. We will reject H_0 if the sample correlation coefficient is greater than $r_{18}^{(0.005)} = 0.5614$, or less than -0.5614 — remember that the alternative hypothesis here is two-sided.

5. We calculate r to be

$$r = SS_{xy}/\sqrt{SS_x SS_y} = 56.5/\sqrt{101.75 \times 89.8} = 0.591$$

6. The sample correlation coefficient lies in the rejection region. We have established a significant correlation between years of education and monthly income. We may use the regression line to make predictions.

Caution : cause and effect relationships...

A significant correlation coefficient does **not** in itself prove a **cause-effect relationship**. It only measures how the variables vary in relation to each other, e.g. the high correlation between smoking and lung cancer does not in itself **prove** that smoking causes cancer. The existence of a correlation between dental plaque and tooth decay does not prove that plaque causes tooth decay. Historically, over the past few decades, the price of gold has trended upwards through time — but time does not cause the price of gold to increase.

There might well be a cause-effect relationship between these variables — the statistician points out that a relationship exists, the research worker has to explain the mechanism. Thus an economist might note a high correlation between the price of milk and the price of petrol, but would not say that an increase in the price of petrol actually causes an increase in the price of milk. The relationship between petrol and milk prices is explained in terms of a third variable, the rate of inflation.

It is mainly in laboratory situations, where all other possible explanatory variables can be controlled by the researcher, that the strength of cause-effect relations between variables are most easily measured. Japanese quality engineers, under the leadership and inspiration of Genichi Taguchi, have been extremely successful in improving the quality of products by isolating the variables that "explain" defectiveness. This has been done, largely, by holding all variables in a manufacturing process constant, except for one experimental variable. If a correlation is found between the variable that is being experimented with and product quality, then this variable is a likely cause of defectiveness, and needs to be monitored carefully. On the other hand, if no correlation is found between a variable and product quality, then this variable is unlikely to be critical to the process, and the cheapest possible setting can be used.

Caution : linear relationships...

The correlation coefficient, r, is designed to measure the degree of **straight line (linear) trend**. If the data lies along some curved line, the correlation coefficient may be small, even though there is a very strong relationship; e.g. when the relationship is quadratic and the points lie close to a parabola, as in Figure 12.1.

The real relationship between x and y is part of a parabola, not a straight line. It is always a good idea to plot a scatter plot, and to determine whether the assumption of **linearity** is reasonable. We can deal with certain non-linear situations quite easily by making transformations. We return to this later.

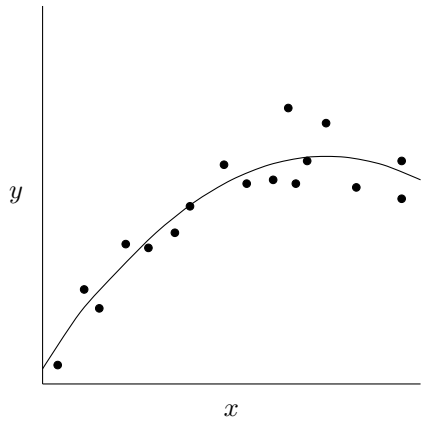


Figure 12.1:

Linear regression ...

We confine ourselves to the fitting of equations which are straight lines — thus we will consider **linear regression**. This is not as restrictive as it appears — many non-linear equations can be transformed into straight lines by simple mathematical techniques; and many relationships can be approximated by straight lines in the range in which we are interested.

The general formula for the straight line is

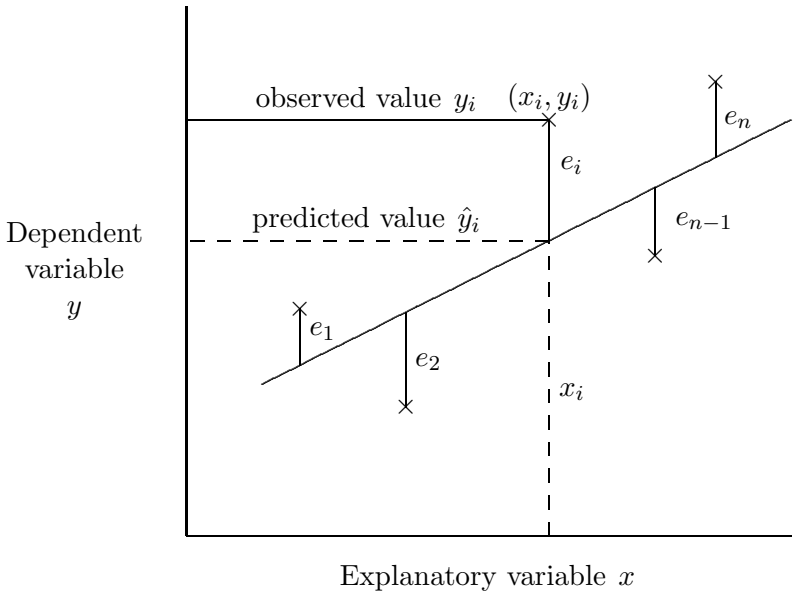
$$y = a + bx$$

The a value gives the y-intercept, and the b value gives the slope. When a and b are given numerical values the line is uniquely specified. The problem in linear regression is to find values for the **regression coefficients** a and b in such a way that we obtain the "best" fitting line that passes through the observations as closely as possible.

We must decide the criteria this "best" line should satisfy. Mathematically, the simplest condition is to stipulate that the sum of the squares of the vertical differences between the observed values and the fitted line should be a minimum. This is called the **method of least squares**, and is the criterion used almost universally in regression analysis.

Pictorially, we must choose the straight line in such a way that the sum of the squares of the e_i on the graph below is a minimum, i.e. we wish to minimize

$$\theta = \sum_{i=1}^{n} e_i^2$$



In general, we have n pairs of observations (x_i, y_i). Usually these points do not lie on a straight line. Note that e_i is the vertical difference between the observed value of y_i and the associated value on the straight line. Statisticians use the notation $\hat{y}_i$ (and say "y hat") for the point that lies on the straight line. Because it lies on the line, $\hat{y}_i = a + b\,x_i$. The difference is expressed as

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i).$$

Thus

$$\theta = \sum e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

We want to find values for a and b which minimize θ. The mathematical procedure for doing this is to differentiate θ firstly with respect to a and secondly to differentiate θ with respect to b. We then set each of the derivatives equal to zero. This gives us two equations in two "unknowns", a and b, and we ought to be able to "solve" them. Fortunately, the two equations turn out to be a pair of linear equations for a and b; this is a type of problem we learnt to do in our first years at high school! Technically, the derivatives are **partial derivatives**, and we use the standard mathematical notation for partial derivatives, $\frac{\partial\theta}{\partial a}$, instead of the more familiar notation $\frac{d\theta}{da}$.

The partial derivatives are:

$$\frac{\partial\theta}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial\theta}{\partial b} = -2 \sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0$$

Setting these partial derivatives equal to zero gives us the so-called **normal equations**:

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2$$

We can calculate $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} x_i y_i$ from our data, and n is the sample size. This gives us the numerical coefficients for the normal equations. The only unknowns are a and b.

By manipulating the normal equations algebraically, we can solve them for a and b to obtain

$$b = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{(\sum x)^2}{n}}$$
$$a = \bar{y} - b\bar{x} = \frac{\sum y - b \sum x}{n},$$

abbreviating $\sum_{i=1}^{n} x_i$ to $\sum x$, etc. It is convenient to use the quantities SS_x, SS_{xy} and SS_y as defined previously. In this notation, the regression coefficients a and b, can be written as

$$b = SS_{xy}/SS_x$$
$$a = \bar{y} - b\bar{x} = \frac{\sum y - b \sum x}{n},$$

and the straight line for predicting the values of the dependent variable y from the explanatory variable x is

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

These formulae are the most useful for finding the least squares linear regression equation $y = a + bx$.

As mentioned previously, the slope of the regression line and the correlation coefficient will always have the same sign (in the case of simple linear regression). In other words, in any one example, r, b and SS_{xy} must have the same sign:

$$r = \frac{SS_{xy}}{\sqrt{SS_x\, SS_y}} = b\sqrt{\frac{SS_x}{SS_y}}.$$

A COMPUTATIONAL SCHEME ...

We set out the procedure for calculating the regression coefficents in full. It is very arithmetic intensive, and is best done using a computer. But it is important to be able to appreciate what the computer is doing for you!

Example 1A, continued: The manual procedure for calculating the regression coefficients is a four-point plan.

1. We set out our data as in the following table, and sum the columns:

	x	y	x^2	y^2	xy
	61	83	$61^2 = 3721$	$83^2 = 6889$	$61 \times 83 = 5063$
	39	83	1521	3844	2418
	70	76	4900	5776	5320
	63	77	3969	5929	4851
	83	89	6889	7921	7387
	75	74	5625	5476	5550
	48	48	2304	2304	2304
	72	78	5184	6084	5616
	54	76	2916	5776	4104
	22	51	484	2601	1122
	67	63	4489	3969	4221
	60	79	3600	6241	4740
$\sum$	714	856	45 602	62 810	52 696

Thus $\sum x = 714$, $\sum y = 856$, $\sum x^2 = 45\,602$, $\sum y^2 = 62\,810$, and $\sum xy = 52\,696$.

2. Calculate the sample means:

$$\bar{x} = \sum x/n = 714/12 = 59.5 \quad \text{and} \quad \bar{y} = \sum y/n = 856/12 = 71.33$$

3. Calculate SS_x, SS_{xy} and SS_y:

$$SS_x = \sum x^2 - \frac{(\sum x)^2}{n} = 45\ 602 - \frac{714^2}{12} = 3119$$
$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 52\ 696 - \frac{714 \times 856}{12} = 1764$$
$$SS_y = \sum y^2 - \frac{(\sum y)^2}{n} = 62\ 810 - \frac{856^2}{12} = 1748.67$$

4. Calculate the regression coefficients a and b:

Thus

$$b = \frac{SS_{xy}}{SS_x} = \frac{1764}{3119} = 0.566$$

and

$$a = \bar{y} - b\bar{x} = 71.33 - 0.566 \times 59.5 = 37.65$$

Therefore the regression equation for making year-end predictions (y) from mid-year mark (x) is

$$\hat{y} = 37.65 + 0.566\,x$$

The hat notation is a device to remind you that this is an equation which is to be used for making predictions of the dependent variable y. This notation is almost universally used by statisticians. So if you obtain $x = 50\%$ at mid-year, you can predict a mark of

$$\hat{y} = 37.65 + 0.566 \times 50 = 66.0\%$$

for yourself at the end of the year.

We will defer the problem of placing a confidence interval on this prediction until later. In the meantime, $\hat{y}$ is a point estimate of the predicted value of the dependent variable. The quantity SS_y, which we have calculated above, will be used in forming confidence intervals (and also in correlation analysis).

Example 4B continued: Find the regression line for predicting monthly income (y) from years of formal education (x).

Thus, the regression coefficients are

$$b = SS_{xy}/SS_x = 56.50/101.75 = 0.555$$

and

$$a = \bar{y} - b\bar{x} = 7.10 - 0.555 \times 13.75 = -0.535$$

The required regression line for predicting monthly income from years of education is thus

$$\hat{y} = -0.535 + 0.555x.$$

We can use this equation to predict that the (average) monthly income of people with 12 years of education is

$$\hat{y} = -0.535 + 0.555 \times 12 = 6.125,$$

or R6125.

Example 5C: Suppose that for some reason the personnel manager wants to predict years of education from monthly income.

(a) Calculate this regression line. (Hint: interchange the roles of x and y.)
(b) Is this regression line the same as that obtained by making x the subject of the formula in the regression line $\hat{y} = -0.535+0.555x$ obtained in example 4B? Explain your results.

Some sums of squares...

The algebra of regression is rich in relationships between the various quantities. We only cover those relationships that are essential to what follows!

Consider the quantity

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2.$$

This measures the **sum of squares** of the differences between the observed values y_i and the predicted values $\hat{y}_i$. This quantity will be small if the observed values y_i fall close to the regression line $\hat{y} = a + bx$, and will be large if they do not. The term $y_i - \hat{y}_i$ is therefore called the **residual** or **error** for the ith observation. Thus we define

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$$

to be the **sum of squares due to error**.

We now do some algebra. Substitute $a = \bar{y} - b\bar{x}$ into $\sum(y_i - a - bx_i)^2$ to obtain:

$$\begin{aligned}
\text{SSE} &= \sum(y_i - a - bx_i)^2 \\
&= \sum(y_i - \bar{y} + b\bar{x} - bx_i)^2 \\
&= \sum((y_i - \bar{y}) - b(x_i - \bar{x}))^2 \\
&= \sum(y_i - \bar{y})^2 - 2b\sum(y_i - \bar{y})(x_i - \bar{x}) + b^2\sum(x_i - \bar{x})^2 \\
&= SS_y - 2b\,SS_{xy} + b^2\,SS_x
\end{aligned}$$

But $b = SS_{xy}/SS_x$, so, in a few more easy steps, we have

$$\text{SSE} = SS_y - b\,SS_{xy}.$$

The first term on the right-hand side measures the total variability of y — in fact, $SS_y/(n-1)$ is the sample variance of y. We call this term the **total sums of squares** and denote it SST, so that

$$\text{SST} = SS_y.$$

The second term on the right-hand side measures how much the total variability is reduced by the regression line. Thus the term $b\,SS_{xy}$ is known as the **sums of squares due to regression**, and is denoted SSR. So we can write

$$\text{SSE} = \text{SST} - \text{SSR} \quad \text{or} \quad \text{SST} = \text{SSR} + \text{SSE}.$$

This result is an important one, because it shows that we can decompose the total sum of squares, SST, into the part that is "explained" by the regression, SSR, and the remainder that is unexplained, SSE, the **error sum of squares** (or residual sum of squares).

Finally, we find two alternative expressions for $\text{SSR} = b\,SS_{xy}$ to be useful. First the short one, useful for calculations. Because $b = SS_{xy}/SS_x$, another expression for SSR is

$$\text{SSR} = \frac{SS_{xy}^2}{SS_x}$$

Secondly,

$$\begin{aligned}
\text{SSR} = b\,SS_{xy} &= b\sum(y_i - \bar{y})(x_i - \bar{x}) \\
&= b\sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(x_i - \bar{x}) \\
&= b\sum(y_i - \hat{y}_i)(x_i - \bar{x}) + \sum(\hat{y}_i - \bar{y})(bx_i - b\bar{x})
\end{aligned}$$

We now consider the first and second terms separately. We will show that the first term is equal to zero. We note that $\hat{y}_i = a + bx_i = \bar{y} - b\bar{x} + bx_i$ and substitute this in place of $\hat{y}_i$ in the first term:

$$\begin{aligned}
b\sum(y_i - \bar{y} + b\bar{x} - bx_i)(x_i - \bar{x}) &= b\left[\sum((y_i - \bar{y}) - b(x_i - \bar{x}))(x_i - \bar{x})\right] \\
&= b\left[(\sum(y_i - \bar{y})(x_i - \bar{x}) - b\sum(x_i - \bar{x})^2\right] \\
&= bSS_{xy} - b^2SS_x \\
&= (\frac{SS_{xy}}{SS_x})SS_{xy} - (\frac{SS_{xy}}{SS_x})^2SS_x \\
&= 0,
\end{aligned}$$

which is a lot of stress to prove that something is nothing! For the second term we note that $b\,x_i = \hat{y}_i - a$ and $b\,\bar{x} = \bar{y} - a$, so $(bx_i - b\bar{x}) = (\hat{y}_i - \bar{y})$ and we have that

$$\text{SSR} = \sum(\hat{y}_i - \bar{y})^2.$$

Thus, the partitioning of the total sum of squares

$$\text{SST} = \text{SSR} + \text{SSE}$$

can be written as:

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

Prediction intervals...

In the same way as the standard deviation played a key role in confidence intervals for means, an equivalent quantity called the **residual standard deviation** is used in calculating confidence intervals for our predictions — we call these **prediction intervals**. The residual standard deviation, also denoted s, measures the amount of variation left in the observed y-values after allowing for the effect of the explanatory variable x. The residual standard deviation is defined to be

$$\begin{aligned} s &= \sqrt{\frac{SSE}{n-2}} \\ &= \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} \\ &= \sqrt{\frac{\sum(y_i - a - bx_i)^2}{n-2}}. \end{aligned}$$

In the previous section, we showed that

$$\text{SSE} = SS_y - b\,SS_{xy}.$$

So a better formula for computational purposes is

$$s = \sqrt{\frac{SS_y - b\,SS_{xy}}{n-2}}.$$

Because the residual standard deviation is estimated, the t-distribution is used to form the prediction intervals. The degrees of freedom for the residual standard deviation are $n-2$ (we lose two degrees of freedom because two parameters are estimated by $\bar{x}$ and $\bar{y}$ before we can calculate the residual standard deviation). The prediction interval for a predicted value of y for a given value of x is

$$\left(\text{predicted value} \;-\; t_{n-2} \times \;s\; \times \sqrt{1 + \frac{1}{n} \;+\; \frac{(x-\bar{x})^2}{SS_x}}\;, \right.$$

$$\left. \text{predicted value} + t_{n-2} \times \;s\; \times \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}} \right).$$

The term $(x-\bar{x})^2$ in the prediction interval has the effect of widening the prediction interval when x is far from $\bar{x}$. Thus our predictions are most reliable when they are made for x-values close to $\bar{x}$, the average value of the explanatory variable. Predictions are less reliable when $(x-\bar{x})^2$ is large. In particular, it is unwise to **extrapolate**, that is, to make predictions that go beyond the range of the x-values in the sample, although in practice we are often required to do this.

Example 1A, continued: Find a 95% prediction interval for the predicted value of the final mark given a class record of 50%.

We first need to compute the residual standard deviation:

$$s = \sqrt{\frac{SS_y - b\,SS_{xy}}{n-2}} = \sqrt{\frac{174.67 - 0.566 \times 1764}{12-2}} = 8.66$$

Earlier we saw that the point estimate of the predicted value for y was 66% when $x = 50\%$.

The required value from the t-tables, $t_{10}^{(0.025)}$, is 2.228. Note also $\bar{x} = 59.5, SS_x = 3119$. Thus the 95% prediction interval is given by

$$(66.0 - 2.228 \times 8.66 \times \sqrt{1 + \frac{1}{12} + \frac{(50-59.5)^2}{3119}}\ ;\ 66.0 + 20.35) = (45.65, 86.35).$$

The coefficient of determination...

The earlier examples demonstrated how the sample correlation coefficient, r, can be used to test whether a significant linear relationship between x and y exists. A further useful statistic which is frequently reported in practice is the **coefficient of determination**. It is defined as

$$\text{coefficient of determination} = \frac{\text{SSR}}{\text{SST}}.$$

The coefficient of determination has a useful practical interpretation. It measures the proportion of the total variability of y which is "explained" by the regression line.

There is a simpler formula for the coefficient of determination. Above, we showed that $\text{SSR} = b\,SS_{xy}$ and $\text{SST} = SS_y$. Substituting, we have

$$\text{coefficient of determination} = \frac{b\,SS_{xy}}{SS_y} = \frac{S_{xy}^2}{SS_x\,SS_y} = r^2.$$

So we have the astonishing result that the coefficient of determination is simply the square of the correlation coefficient! For this reason the coefficient of determination in simple linear regression is denoted r^2.

Example 1A, continued: Compute and interpret the coefficient of determination, r^2, for the regression of the final examination mark on the class record.

The sample correlation coefficient was 0.755. So the coefficient of determination is

$$r^2 = 0.755^2 = 0.5700.$$

We interpret this by stating that 57% of the variation in marks in the final examination can be explained by variation in the marks of the class test.

Example 4B, continued: Compute and interpret the r^2 for the relationship between years of formal education and monthly income.

The coefficient of determination is

$$r^2 = 0.5911^2 = 0.3494.$$

Hence approximately 35% of the variation in monthly incomes can be explained by the variation in the years of formal education.

Summary of computational scheme for simple linear regression...

The key quantities in simple linear regression are summarized below.

x	y	x^2	y^2	xy
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$\overline{\sum x}$	$\overline{\sum y}$	$\overline{\sum y^2}$	$\overline{\sum y^2}$	$\overline{\sum xy}$

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum x \\
\bar{y} &= \frac{1}{n}\sum y \\
SS_x &= \sum x^2 - (\sum x)^2/n \\
SS_{xy} &= \sum xy - (\sum y)(\sum x)/n \\
SS_y &= \sum y^2 - (\sum y)^2/n \\
b &= SS_{xy}/SS_x \\
a &= \bar{y} - b\bar{x} \\
r &= SS_{xy}/\sqrt{SS_x\ SS_y} \\
s &= \sqrt{(SS_y - b\,SS_{xy})/(n-2)}
\end{aligned}$$

$$\text{Predicted values}: \ \hat{y} = a + bx$$

95% prediction intervals:

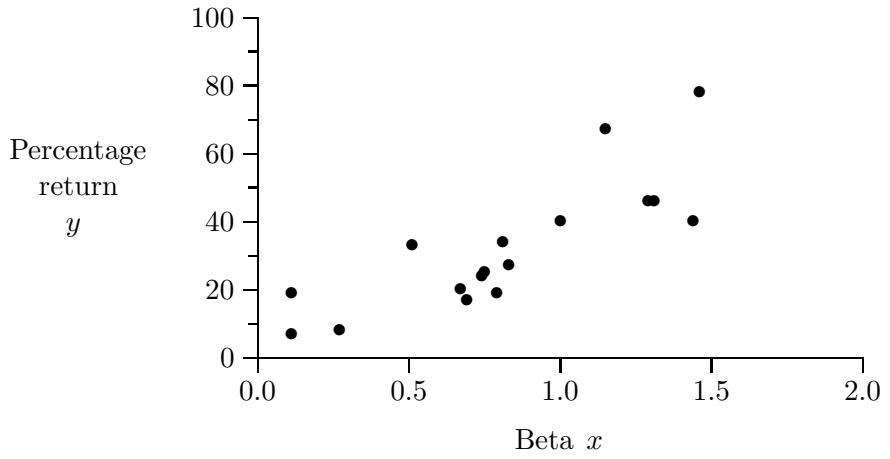
$$\text{predicted value } \pm t_{n-2}^{(0.025)} \times s \ \times \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$

Example 6B: In an investigation to determine whether the rewards on financial investments are related to the risks taken, data on a sample of 17 sector indices on the Johannesburg Stock Exchange was collected (Source: *Financial Risk Service*, D.J. Bradfield & D. Bowie). The data gives a "beta" value for each sector (x), a quantity widely used by financial analysts as a proxy for risk, and the reward (y) or return for each sector, calculated as the percentage price change over a 12 month period.

(a) Draw a scatter plot and discuss the visual appearance of the relationship.
(b) Calculate the correlation coefficient, and decide if the relationship between risk and return is significant?
(c) Find the regression line. What is the coefficient of determination of the regression.
(d) Find 95% prediction intervals for the predicted returns for investments with risks given by betas of 0.5, 0.8 and 1.5.

Sector	Beta x	Return (%) y	x^2	y^2	xy
1. Diamonds	1.31	46	1.716	2116	60.26
2. West Wits	1.15	67	1.323	4489	77.05
3. Metals & Mining	1.29	46	1664	2116	59.34
4. Platinum	1.44	40	2.074	1600	57.60
5. Mining Houses	1.46	78	2.132	6084	113.88
6. Evander	1.00	40	1.000	1600	40.00
7. Motor	0.74	24	0.548	576	17.76
8. Investment Trusts	0.67	20	0.449	400	13.40
9. Banks	0.69	17	0.476	289	11.73
10. Insurance	0.79	19	0.624	361	15.01
11. Property	0.27	8	0.073	64	2.16
12. Paper	0.75	25	0.563	625	18.75
13. Industrial Holdings	0.83	27	0.689	729	22.41
14. Beverages & Hotels	0.81	34	0.656	1156	27.54
15. Electronics	0.11	7	0.012	49	0.77
16. Food	0.11	19	0.012	361	2.09
17. Printing	0.51	33	0.260	1089	16.83
Totals	13.93	550	14.271	23704	556.58

(a) The scatter plot shows that the percentage return increases with increasing values of beta. Furthermore, the plot makes it clear that it is appropriate, over the observed range of values for beta, to fit a straight line to this data.



Having computed the basic sums at the foot of the table, we compute $\bar{x} = 0.8194$ and $\bar{y} = 32.35$. Then

$$SS_x = 14.271 - (13.93)^2/17 = 2.857$$
$$SS_{xy} = 556.58 - 13.93 \times 550/17 = 105.904$$
$$SS_y = 23704 - (550)^2/17 = 5909.882$$

(b) Using the modified hypothesis testing procedure:

1. $H_0 : \rho = 0$
2. $H_1 : \rho \neq 0$
3. $r = SS_{xy}/\sqrt{SS_x\, SS_y} = 105.904/(2.857 \times 5909.882)^{\frac{1}{2}} = 0.815$
4. Using the correlation coefficient table with $17 - 2 = 15$ degrees of freedom, we see that 0.815 is significant at the 0.1% level, because $0.815 > r_{15}^{(0.0005)} = 0.7247$.
5. We have established a significant relationship between risk and return ($r = 0.815,\ P < 0.001$).

(c) The coefficients for the regression line are given by:

$$b = \frac{SS_{xy}}{SS_x} = \frac{105.994}{2.857} = 37.068$$
$$a = \bar{y} - b\bar{x} = 32.35 - 37.068 \times 0.8194 = 1.976$$

The regression line is thus

$$\hat{y} = 1.976 + 37.068x.$$

The coefficient of determination is equal to $r^2 = 0.815^2 = 0.664$. Thus the regression line accounts for about two-thirds of the variability in return.

(d) Predicted return for an investment which had a beta of 0.5 is

$$\hat{y} = 1.976 + 37.068 \times 0.5 = 20.51\%$$

Similarly, the predicted return when $x = 0.8$ and $x = 1.5$ are 31.63% and 57.58%, respectively.

To find the prediction intervals we need to calculate the residual standard deviation:

$$s = \sqrt{(SS_y - bSS_{xy})/(n-2)} = \sqrt{(5909.882 - 37.068 \times 105.904)/15}$$
$$= 11.501.$$

We need $t_{15}^{(0.025)}$. From the t-tables, (Table 2) this is 2.131. We note that $\bar{x} = 0.8194$.
The 95% prediction interval for $x = 0.5$ is thus

$$(20.510 - 2.131 \times 11.501 \times \sqrt{1 + \frac{1}{17} + \frac{(0.5 - 0.8194)^2}{2.857}} \quad , \quad 20.510 + 25.641)$$

or $(-5.13\%, 46.15\%)$.
For $x = 0.8$ and $x = 1.5$, the only terms in the prediction interval formula that change are the first term (the predicted value) and the final term under the square root sign. When $x = 0.8$, the 95% prediction interval is

$$(31.630 - 2.131 \times 11.501 \times \sqrt{1 + \frac{1}{17} + \frac{(0.8 - 0.8194)^2}{2.857}} \quad , \quad 31.630 + 25.221)$$

or $(6.41\%, 56.85\%)$.
for $x = 1.5$, we have

$$(57.578 - 2.131 \times 11.501 \times \sqrt{1 + \frac{1}{17} + \frac{(1.5 - 0.8194)^2}{2.857}} \quad , \quad 57.578 + 27.081)$$

or $(30.50\%, 84.65\%)$.
Of the three values for which we have computed prediction intervals, $x = 0.8$ lies closest to the mean of the observed range of x-values, and therefore the associated prediction interval for $x = 0.8$ is the shortest of the three.

Example 7C: To check on the strength of certain large steel castings, a small test piece is produced at the same time as each casting, and its strength is taken as a measure of the strength of the large casting. To examine whether this procedure is satisfactory, i.e., the test piece is giving a reliable indication of the strength of the castings, 11 castings were chosen at random, and both they, and their associated test pieces were broken. The following were the breaking stresses:

Test piece (x) :	45	67	61	77	71	51	45	58	48	62	36
Casting (y) :	39	86	97	102	74	53	62	69	80	53	48

(a) Calculate the correlation coefficient, and test for significance.
(b) Calculate the regression line for predicting y from x.
(c) Compute and interpret the coefficient of determination.
(d) Find 90% prediction limits for the strength of a casting when $x = 60$.

Nonlinear curve fitting...

Not all relationships between variables are linear. An obvious example is the relationship between time (t) and population (p), a relationship which is frequently said to be **exponential**, i.e. of the form $p = ae^{bt}$, where a and b are "regression coefficients". Sometimes there are theoretical reasons for using a particular mathematical function to describe the relationship between variables, and sometimes the relationship that fits variables reasonably well has to be found by trial and error. The more complex the mathematical function, the more data points are required to determine the "regression coefficients" reliably. The simplest mathematical function useful in regression analysis is the straight line, giving rise to the linear regression we have been considering. Because of the simplicity of linear regression, we often make use of it as an approximation (over a short range) to a more complex mathematical function.

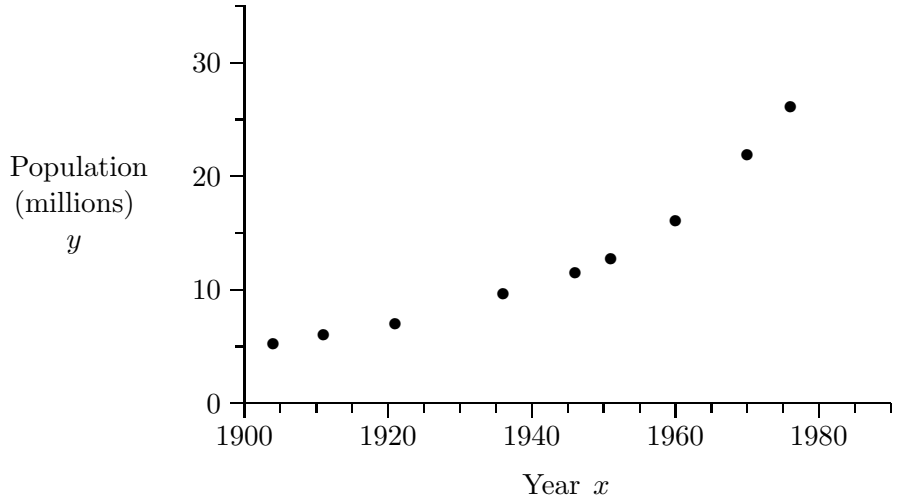
Sometimes, however, we can **transform** a nonlinear relationship into a linear one, and then apply the methods we have already learnt. It is these rather special situations which we consider next.

Exponential growth...

Example 8A: Fit an exponential growth curve of the form $y = ae^{bx}$ to the population of South Africa, 1904-1976:

Year (x) :	1904	1911	1921	1936	1946	1951	1960	1970	1976
Population : (millions)(y)	5.2	6.0	6.9	9.6	11.4	12.7	16.0	21.8	26.1

From a scatter plot it is clear that linear regression is inappropriate:



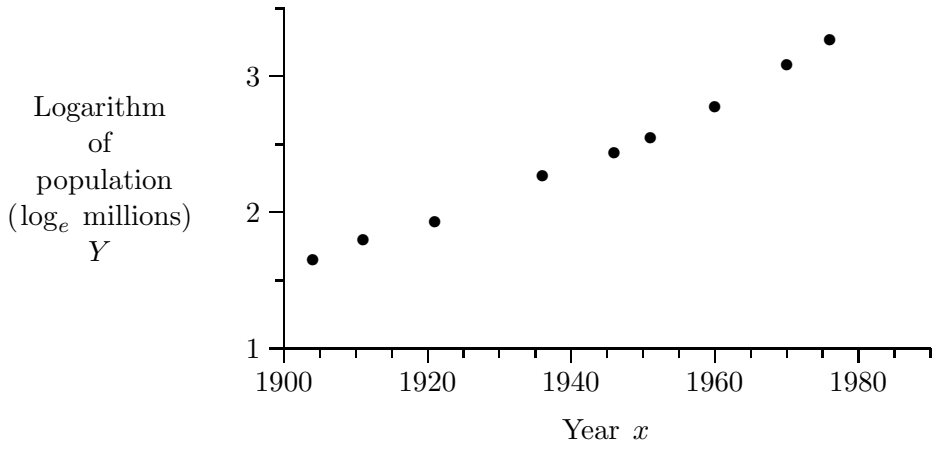
However, there is a straightforward way to transform this into a linear relation. Take natural logarithms on both sides of the equation $y = ae^{bx}$. This yields (remembering $\log_e e = 1$)

$$\log y = \log a + bx.$$

We now put $Y = \log y$ and $A = \log a$ we rewrite it as

$$Y = A + bx.$$

This is the straight line with which we are familiar! The scatter plot of the logarithm of population against year is plotted below; a straight line through the points now seems to be a satisfactory model.



By using our computational scheme for linear regression on the pairs of data values x and $Y(= \log y)$, we compute the regression coefficients A and b. Having done this, we can now transform back to the exponential growth curve we wanted. Because $Y = \log y$, we can write

$$\begin{aligned} y = & e^Y = e^{A+bx} \\ & = e^A e^{bx} = ae^{bx}, \end{aligned}$$

where $a = e^A$. The table below demonstrates the computational procedure for fitting exponential growth. (It is convenient here to take x to be years since 1900.)

x	y	$Y = \log y$	x^2	Y^2	xY
4	5.2	1.65	16	2.723	6.60
11	6.0	1.79	121	3.204	19.69
21	6.9	1.93	441	3.725	40.53
36	9.6	2.26	1296	5.108	81.36
46	11.4	2.43	2116	5.905	111.78
51	12.7	2.54	2601	6.452	129.54
60	16.0	2.77	3600	7.673	166.20
70	21.8	3.08	4900	9.486	215.60
76	26.1	3.26	5776	10.628	247.76
$\sum$ 375		21.72	20 867	54.904	1019.06

From these basic sums we first compute $\bar{x} = 41.67$ and $\bar{Y} = 2.41$ and then $SS_x = 5242$ and $SS_{xY} = 114.06$. Thus

$$\begin{aligned} b &= SS_{xY}/SS_x = 0.0218 \\ A &= \bar{Y} - b\bar{x} = 1.50. \end{aligned}$$

The regression line for predicting Y from x is

$$Y = 1.50 + 0.0218\,x$$

Transforming back to the exponential growth curve yields

$$\begin{aligned} y &= e^{1.50} \times e^{0.0218\,x} \\ &= 4.48\,e^{0.0218\,x}. \end{aligned}$$

In full knowledge that it is dangerous to use regression analysis to extrapolate beyond the range of x-values contained in our data, we risk an estimate of the population of South Africa in the year 2000, i.e. when $x = 100$:

$$y = 4.48\, e^{0.0218\times 100} = 39.6 \text{ million.}$$

(In 1980 the population demographers were in fact predicting a population of 39.5 million for the year 2000.)

Other relationships that can be transformed to be linear...

Example 9A: Transform the following functions into linear relationships.

(a) $y = ab^x$
(b) $y = ax^b$
(c) $ay = b^{-x}$

In each case we take natural logarithms:

(a) $\log y = \log a + x\, \log b$.
Putting $Y = \log y,\ A = \log a$ and $B = \log b$, this becomes

$$Y = A + xB,$$

which is linear.

(b) $\log y = \log a + b\ \log x$.
Put $Y = \log y,\ A = \log a$ and $X = \log x$ to obtain

$$Y = A + bX.$$

(c) $\log a + \log y = -x\ \log b$.
Put $A = \log a,\quad Y = \log y \quad$ and $\quad B = \log b$ to obtain

$$A + Y = -xB \quad \text{or} \quad Y = -A - xB.$$

Example 10B: To investigate the relationship between the curing time of concrete and the tensile strength the following results were obtained:

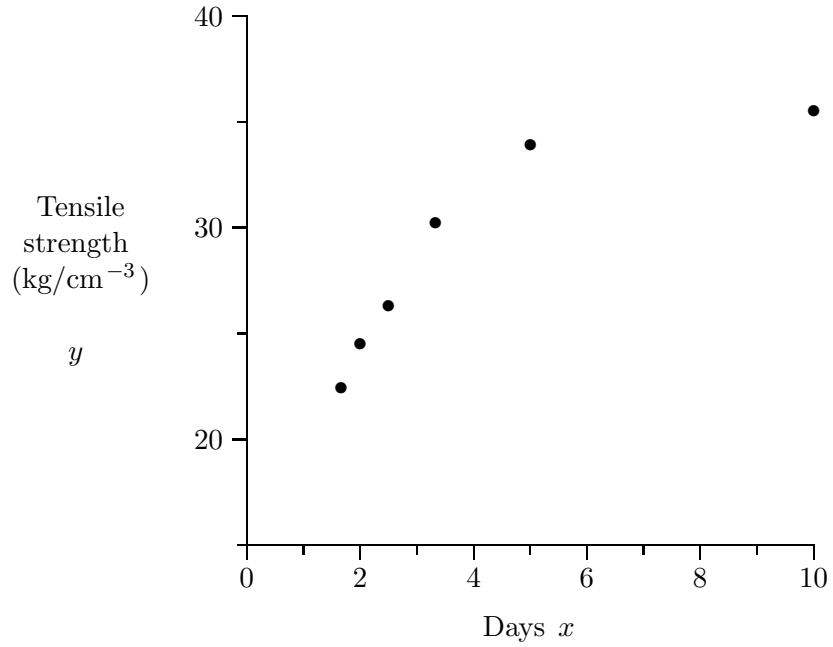
Curing time (days) (x):	$1\frac{2}{3}$	2	$2\frac{1}{2}$	$3\frac{1}{3}$	5	10
Tensile strength (kg/cm^2) (y):	22.4	24.5	26.3	30.2	33.9	35.5

(a) Draw a scatter plot.
(b) Assuming that the theoretical relationship between tensile strength (y) and curing time (x) is given by

$$y = ae^{\frac{b}{x}},$$

find the regression coefficients a and b.
(c) Predict the tensile strength after curing time of three days.

(a)



(b) Taking logarithms we obtain

$$\log y = \log a + b \times \frac{1}{x}$$

Put $Y = \log Y$, $A = \log A$ and $X = \frac{1}{x}$ to obtain the linear equation $Y = A + bX$. The computational scheme is:

x	y	$X = 1/x$	$Y = \log y$	X^2	Y^2	XY
$1\frac{2}{3}$	22.4	0.6	3.11	0.36	9.67	1.87
2	24.5	0.5	3.20	0.25	10.24	1.60
$2\frac{1}{2}$	26.3	0.4	3.27	0.16	10.69	1.31
$3\frac{1}{3}$	30.2	0.3	3.41	0.09	11.63	1.02
5	33.9	0.2	3.52	0.04	12.39	0.70
10	35.5	0.1	3.57	0.01	12.74	0.36
$\sum$		2.1	20.08	0.91	67.36	6.86

From these sums we compute the quantities

$$\bar{X} = 0.35 \qquad \bar{Y} = 3.347 \qquad SS_X = 0.175 \qquad SS_{XY} = -0.168$$

and then the slope coefficient and the y-intercept

$$b = SS_{XY}/SS_X = -0.960 \quad \text{and} \quad A = \bar{Y} - b\bar{X} = 3.683.$$

Thus the equation for predicting Y, the logarithm of the tensile strength y, from X, the reciprocal of curing time x, is

$$Y = 3.683 - 0.960\ X.$$

To express this in terms of the original variables x and y, we write

$$\begin{aligned} y = e^Y &= e^{A+bX} = e^A\, e^{bX} \\ &= e^{3.683} e^{-0.960X} \\ &= 39.766\, e^{-0.960/x}. \end{aligned}$$

(c) After 3 days curing, i.e. $x = 3$,

$$y = 39.766\, e^{-0.960/3} = 28.9 \text{ kg/cm}^2.$$

Example 11C: A company that manufactures gas cylinders is interested in assessing the relationships between pressure and volume of a gas. The table below gives experimental values of the pressure P of a given mass of gas corresponding to various values of the volume V. According to thermodynamic principles, the relationship should be of the form $PV^\gamma = C$, where γ and C are constants.

(a) Find the values of γ and C that best fit the data.
(b) Estimate P when $V = 100.0$.

Volume V	54.3	61.8	72.4	88.7	118.6	194.0
Pressure P	61.2	49.5	37.6	28.4	19.2	10.1

Multiple regression...

In many practical situations, it is more realistic to believe that more than one explanatory variable is related to the dependent variable. Thus the quality of the grape harvest depends not only on the amount of rain that falls during spring, but also probably on the hours of sunshine during summer, the amount of irrigation during summer, whether the irrigation was by sprinklers, furrows or a drip system, the amounts and types of fertilizers applied, the amounts and types of pesticides used, and even whether or not the farmer used scarecrows to frighten the birds away! Regression models that include more than one explanatory variable are called **multiple regression models**. Multiple regression should be seen as a straightforward extension of simple linear regression.

The general form of the multiple regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

where y is the dependent variable, x_1, x_2, ..., x_k are the explanatory variables and β_0, β_1, β_2, ..., β_k are the true regression coefficients, the population regression parameters.

The term e at the end of the regression model is usually called the **error**, but this is a bad name. It does not mean "mistake", but it is intended to absorb the variability in the dependent variable y which is not accounted for by the explanatory variables which we have measured and have included in the regression model.

The regression coefficients β_0, β_1, β_2, ..., β_k are unknown parameters (note the use of Greek letters once again for parameters) and need to be estimated. The data from which we are to estimate the regression coefficients consists of n sets of $k + 1$ numbers of the form: the observed values of the dependent variable y, and the associated values of the k explanatory variables x_i. For example, we would measure the quality of the grape harvest, together with all the values of the explanatory variables that led to that quality.

The least squares estimation approach...

As with simple linear regression, the regression coefficients need to be estimated. We are searching for a model of the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k,$$

where b_i is the sample estimate of the regression parameter β_i.

The method for estimating the regression coefficients in multiple regression is identical to that used for simple regression. As before, we use the method of least squares and minimize the sum of squares of the difference between the observed y-values and the "predicted" $\hat{y}$-values. Because there are now $k+1$ regression coefficients to estimate (and not only two as was the case with simple regression), the algebra gets messy, and the arithmetic gets extremely tedious. So we resort to getting computers to do the calculations on our behalf.

The primary reason for the computational difficulty in multiple regression as compared with simple regression, is that, instead of having only two partial derivatives and two equations to solve, we have to find $k+1$ partial derivatives and the regression coefficients are obtained as the solutions to a set of $k+1$ linear equations. But the key point is that the underlying philosophy remains the same: we minimize the sum of squared residuals.

There are no simple explicit formulae for the regression coefficients b_i for multiple regression. They are most conveniently expressed in terms of matrix algebra. But they are readily computed by a multitude of statistical software packages. We assume that you have access to a computer that will do the calculations, and we will take the approach of helping you to interpret the results.

Example 12A: To demonstrate how a regression equation can be estimated for two or more explanatory variables, we consider again Example 4B where the personnel manager was concerned with the analysis and prediction of monthly incomes.

Recall that in Example 4B the relationship between monthly income (y) and years of education (which we will now denote x_1) was estimated using the simple regression model:

$$\hat{y} = -0.535 + 0.555 x_1.$$

Because $r = 0.5911$ was significant, we had established that a relationship existed between the two variables. But the coefficient of determination $r^2 = 0.3494$, so that approximately 35% of the variability in incomes could be explained by this single variable, years of formal education.

But what about the remaining 65%? The personnel manager is curious to establish whether any other variable can be found that will help to reduce this unexplained variability and which will improve the goodness of fit of the model.

After some consideration, the personnel manager considers that "years of relevant experience" is another relevant explanatory variable that could also impact on their incomes. He gathers the extra data, and calls it variable x_2.

Person	Monthly income (R1000's) y	Years of formal education x_1	Years of experience x_2
1	4	12	2
2	5	10	6
3	8	15	16
4	10	12	23
5	9	16	12
6	7	15	11
7	5	12	1
8	10	16	18
9	7	14	12
10	6	14	5
11	8	16	17
12	6	12	4
13	9	15	20
14	4	10	6
15	7	18	7
16	8	11	13
17	9	17	12
18	11	15	3
19	4	12	2
20	5	13	6

Now the regression equation that includes both the effect of years of education (x_1) and the years of relevant experience (x_2) is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

The computer generated solution is

$$\hat{y} = 0.132 + 0.379 x_1 + 0.180 x_2$$

where x_1 is years of formal education and x_2 is years of relevant experience.

Notice that the coefficients of x_1 and the intercept have changed from the solution when only x_1 was in the model. Why? Because some of the variation previously explained by x_1 is now explained by x_2. (There is a mathematical result that says that the earlier regression coefficients will remain unchanged only if there is no correlation between x_1 and x_2.)

In multiple regression we can interpret the the magnitude of the regression coefficient b_i as the change induced in y by a change of 1 unit in x_i, holding all the other variables constant. In the above example, $b_1 = 0.379$ can be interpreted as measuring the change in income (y) corresponding to a one-year increase in years of formal education (while holding years of experience constant). And $b_2 = 0.180$ is the change in y induced by a one-year increase in experience (while holding years of formal education constant). Because y is measured in 1000s of rands per month, the regression model is telling us that each year of education is worth R379 per month, and that each year of experience is worth R180 per month!

One question that the personnel manager would ask is: "To what extent has the inclusion of the additional variable contributed towards explaining the variation of incomes?" For simple regression, the coefficient of determination, r^2, answered this question. We now have a multiple regression equivalent, called the **multiple coefficient**

of determination, denoted R^2, which measures the proportion of variation in the dependent variable y which is explained jointly by all the explanatory variables in the regression model. The computation of R^2 is not as straightforward as in simple regression, but is invariably part of the computer output. In our example, the computer printout that gave us the model

$$\hat{y} = 0.132 + 0.379x_1 + 0.180x_2$$

also told us that $R^2 = 0.607$. This means that 60.7% of the variability of incomes is explained jointly by x_1 (years of formal education) and x_2 (years of relevant experience). This represents a substantial improvement in the explanation of y (monthly incomes) from 35% to 61%. However, 39% of the variation in y remains unexplained, and is absorbed into the "error" term discussed earlier. This leads us into taking a closer look at the sources of variation in the dependent variable y.

Understanding the sources of variation in y ...

In simple regression, we partitioned the variation in the dependent variable y into two components, the variation explained by the regression and the unexplained variation, the variation due to "errors". The last paragraph of the previous section suggested that we can do the same in multiple regression.

The same partitioning of the total sum of squares that we developed for simple regression is applicable in the multiple regression situation (although the arithmetic involves a lot more number crunching!):

$$\text{SST} = \text{SSR} + \text{SSE}$$

or

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2.$$

The multiple coefficient of determination is defined in the same way as in the simple case. It is the ratio of the sum of squares explained by the regression and the total sum of squares,

$$R^2 = \frac{\text{SSR}}{\text{SST}}.$$

But in multiple regression, we do need to have a different notation for the multiple coefficient of determination (R^2 in place of r^2) because we no longer have a straightforward squaring of a sample correlation coefficient.

The partitioning of the total sum of squares also helps to provide insight into the structure of the multiple regression model. In the section that follows, we will see that the partitioning enables us to decide whether the equation generated by the multiple regression is "significant". To do this in simple regression, we merely calculated the correlation coefficient r and checked in Table 5 whether or not this value of r was significant. The approach is now quite different.

Testing for significant relationships ...

In multiple regression, the appropriate null and alternative hypotheses for testing whether or not there is a significant relationship between y and the x_i's are:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1 : one or more of the coefficients are non-zero.

If we reject the null hypothesis we can conclude that there is a significant relationship between the dependent variable y and at least one of the explanatory variables x_i.

The test statistic for this hypothesis is couched in terms of the ratio between the variance due to regression and that due to error. It is therefore not surprising that the **F-distribution** provides the critical values for the test statistic (recall chapter 9). The test statistic is calculated as one of the outputs in most regression software packages, and is usually presented as part of an **analysis of variance** or **ANOVA** table.

The ANOVA table summarizes the sources of variation, and usually has the following structure:

ANOVA table

Source	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean square (MS)	F
Regression	SSR	k	$\text{MSR} = \text{SSR}/k$	$F = \text{MSR}/\text{MSE}$
Error	SSE	$n-k-1$	$\text{MSE} = \text{SSE}/(n-k-1)$	
Total	SST	$n-1$		

The test statistic is the variance explained by the regression divided by the variance due to error. The distribution of the test statistic for the above null hypothesis is

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F_{k,n-k-1},$$

where n is the number of observations and k is the number of dependent variables. We reject the null hypothesis if the observed F-value is larger then the critical value in the F-tables.

Example 12A, continued: Test the significance of the regression model with dependent variable "monthly income" and explanatory variables "years of formal education" and "years of relevant experience". Perform the test at the 5% significance level.

The hypotheses for testing the significance of the regression can thus be formulated as:

1. $H_0 : \beta_1 = \beta_2 = 0$
2. H_1 : one or more of the coefficients are non-zero.
3. Significance level: 5%
4. Rejection region. Because $n = 20$ and $k = 2$, the test statistic has the $F_{k,n-k-1} = F_{2,17}$-distribution. We reject H_0 if the observed F-value is greater than the upper 5% point of $F_{2,17}$, i.e. if it is greater than

$$F_{2,17}^{0.05} = 3.59.$$

5. Test statistic. The ANOVA table obtained from the computer printout looks like this:

ANOVA table				
Source	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Square (MS)	F
Regression	54.5391	2	27.2696	13.15
Error	35.2609	17	2.0742	
Total	89.8000	19		

The observed value of the test statistic is given by $F = 27.2696/2.0742 = 13.15$.

6. Conclusion. Because $F = 13.35 > 3.59$, we reject H_0 at the 5% significance level and conclude that a significant relationship exists between monthly income, the dependent variable, and the explanatory variables, years of formal education and years of relevant experience. Thus the regression model can be used for predicting monthly income. As in the simple regression case, it is advisable not to extrapolate, and the values of the explanatory variables for which predictions are made ought to be within their ranges in the observed data.

   For example, the predicted monthly income for a person with $x_1 = 13$ years of education and $x_2 = 8$ years of experience is obtained by making the appropriate substitutions in the regression model:

$$\begin{aligned} \hat{y} &= 0.132 + 0.379x_1 + 0.180x_2 \\ &= 0.132 + 0.379 \times 13 + 0.180 \times 8 \\ &= 6.5. \end{aligned}$$

   The predicted monthly income for this person is R6 500.

A test for the significance of individual variables...

Frequently, we need to establish whether a single regression coefficient in a multiple regression model is significant in the regression. In the jargon of hypothesis testing, the appropriate null and alternative hypotheses are:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0.$

The test statistic for this null hypothesis is

$$\frac{b_i}{s_{b_i}} \sim t_{n-k-1},$$

where b_i is the estimated regression coefficient, and s_{b_i} is the standard deviation of the estimate of b_i. Both these values are calculated by the computer software and can be found in the printout.

Example 12A, continued: Test whether each of the explanatory variables x_1 and x_2 in the regression model for monthly income are significant. Do the tests at the 5% level.

To test if x_1, years of formal education, is significant the procedure is as follows:

1. $H_0 : \beta_1 = 0$
2. $H_1 : \beta_1 \neq 0$
3. Significance level: 5%
4. Rejection region. The appropriate degrees of freedom are $n-k-1 = 20-2-1 = 17$. The test is two-sided, because we want to reject the null hypothesis either if β_1 is significantly positive or significantly negative. So the critical value of the test statistic is $t_{17}^{(0.025)} = 2.110$, and we will reject H_0 if $|t| > 2.110$.
5. Test statistic. The values $b_1 = 0.379$ and $s_{b_1} = 0.152$ are computed by the regression programme. Hence $t = 0.379/0.152 = 2.49$
6. Conclusion. Because the observed t-value of 2.49 lies in the rejection region, we reject $H_0 : \beta_1 = 0$ at the 5% level, and conclude that the variable x_1, number of years of formal education is significant in the regression model.

A similar procedure is used to decide if x_2, years of relevant experience is significant.

1. $H_0 : \beta_2 = 0$
2. $H_1 : \beta_2 \neq 0$
3. Significance level: 5%
4. Rejection region. The critical value of the test statistic is once again $t_{17}^{(0.025)} = 2.110$, and we will reject H_0 if $|t| > 2.110$.
5. Test statistic. The values $b_2 = 0.180$ and $s_{b_2} = 0.054$ are computed by the regression programme, and $t = 0.180/0.054 = 3.33$.
6. Conclusion. Because the observed t-value lies in the rejection region, we reject $H_0 : \beta_2 = 0$ at the 5% level, and conclude that the variable x_1, number of years of relevant experience is also significant in the regression model.

Example 13B: A market research company is interested in investigating the relationship between monthly sales income and advertising expenditure on radio and television for a specific area. The following data is gathered.

Sample number	Monthly sales (R1000s)	Radio advertising (R1000s)	Television advertising (R1000s)
1	105	0.5	6.0
2	99	1.0	3.0
3	104	0.5	5.0
4	101	1.5	3.5
5	106	2.3	4.0
6	103	1.3	4.5
7	103	3.2	3.5
8	104	1.5	4.0
9	100	0.8	3.0
10	98	0.6	2.8

(a) Find the regression equation relating monthly sales revenue to radio and television advertising expenditure.
(b) Estimate and interpret R^2.
(c) Test the regression equation for significance at the 5% significance level.
(d) Test, for each variable individually, whether radio and television advertising expenditure are significant in the model at the 5% level.
(e) Comment on the effectiveness of radio vs television advertising for this industry.

You will need the following information, extracted from a computer printout, to answer the questions.

Table of estimated coefficients		
Variable	Estimated coefficient	Estimated standard deviation
Radio	1.6105	0.4687
Television	2.3414	0.4041
Intercept	90.9725	

ANOVA table				
Source	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Square (MS)	F
Regression	54.194	2	27.097	19.15
Error	9.906	7	1.415	
Total	64.100	9		

(a) From the table of estimated coefficients, the regression model is

$$\hat{y} = 90.97 + 1.61x_1 + 2.34x_2$$

where y is monthly sales revenue, x_1 is the radio advertising expenditure and x_2 is the television advertising expenditure.

(b) The coefficient of determination $R^2 = \text{SSR/SST} = 54.19/64.10 = 0.845$. Hence 84.5% of the variation in sales volume can be explained by variation in radio and television advertising expenditures.

(c) To test the regression for significance:

1. $H_0 : \beta_1 = \beta_2 = 0$
2. H_1 : At least one of the β_i are non-zero, where $i = 1, 2$.
3. Significance level: 5%.
4. Rejection region. Reject H_0 if observed $F > F_{2,7}^{(0.05)} = 4.74$
5. Test statistic. $F = \text{MSR/MSE} = 27.097/1.415 = 19.15$.
6. Because the observed F exceeds 4.74 we reject H_0 at the 5% level and conclude that a significant relationship exists between sales revenue and advertising expenditure.

(d) To test the individual coefficients for significance. Firstly, for advertising expenditure on radio.

1. $H_0 : \beta_1 = 0$
2. $H_1 : \beta_1 \neq 0$
3. Significance level: 5%.
4. Reject H_0 if observed $|t| > t_7^{(0.025)} = 2.365$.
5. Test statistic: $t = b_1/s_{b_1} = 1.611/0.469 = 3.43$.
6. Reject H_0, and conclude that sales revenue is related to expenditure on radio advertising.

Secondly, for advertising expenditure on television.

1. $H_0 : \beta_2 = 0$
2. $H_1 : \beta_2 \neq 0$
3. Significance level: 5%.
4. Again, reject H_0 if observed $|t| > t_7^{(0.025)} = 2.365$.
5. Test statistic: $t = b_2/s_{b_1} = 2.341/0.404 = 5.80$.
6. Reject H_0, and conclude that sales revenue is related to expenditure on television advertising.

(e) Recall that the regression coefficients can be interpreted as the magnitude of the impact that a unit change in x_i has on y (holding the other variables constant). In this example, both explanatory variables are measured in the same units. Because $b_1 < b_2$, we have a suggestion that expenditure on television advertising is more effective than expenditure on radio advertising. But, beware, we have not tested this difference statistically — we need to test $H_0 : \beta_1 = \beta_2$ against the alternative $H_1 : \beta_1 < \beta_2$.

Example 14C: A chain of department stores wants to examine the effectiveness of its part-time employees. Data on sales as well as number of hours worked during one weekend and months of experience was collected for 10 randomly selected part-time sales persons.

Person	Number of sales	Number of hours worked	Months of experience
1	4	5	1
2	2	4	2
3	15	12	6
4	9	10	6
5	11	9	8
6	8	8	10
7	14	13	12
8	17	14	15
9	16	12	14
10	2	4	3

(a) Write down the regression equation relating number of sales (y) to number of hours worked (x_1) and months of experience (x_2).
(b) Compute and interpret R^2.
(c) Test the regression equation for significance at the 5% significance level.
(d) Test, for x_1 and x_2 individually, that they are significant in the model at the 5% level.
(e) Do you think that the experience of part-time employees makes a difference to their number of sales? Explain.

You will require the following information extracted from the relevant computer printout.

Table of estimated coefficients		
Variable	Estimated coefficient	Estimated standard deviation
Hours worked	1.3790	0.2270
Experience	0.0998	0.1716
Intercept	-3.5176	

ANOVA table			
Source	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Square (MS)
Regression	282.708	2	
Error	12.892	7	
Total	295.599	9	

Person	Monthly income (R1000's) y	Years of formal education x_1	Years of experience x_2	Gender x_3
1	4	12	2	1
2	5	10	6	0
3	8	15	16	1
4	10	12	23	0
5	9	16	12	1
6	7	15	11	0
7	5	12	1	1
8	10	16	18	1
9	7	14	12	0
10	6	14	5	0
11	8	16	17	1
12	6	12	4	1
13	9	15	20	0
14	4	10	6	0
15	7	18	7	1
16	8	11	13	1
17	9	17	12	0
18	11	15	3	1
19	4	12	2	0
20	5	13	6	0

The use of dummy variables for qualitative variables ...

In the discussion thus far on regression, all the explanatory variables have been quantitative. Occasions frequently arise, however, when we want to include a categorical or qualitative variable as an explanatory variable in a multiple regression model. We demonstrate in this section how categorical variables can be included in the model.

Example 12A, continued: Assume now that the personnel manager who was trying to find explanatory variables to predict monthly income now believes that, in addition to years of formal education and years of relevant experience, gender may have a bearing on an individual's monthly income. Test whether the gender variable is significant in the model (at the 5% level).

The personnel manager first gathers the gender information on each person in the sample. The next step is to convert the qualitative variable "gender" into a **dummy variable**. A dummy variable consists only of 0's and 1's. In this example, we will code gender as a dummy variable x_3. We put $x_3 = 0$ if the gender is female and $x_3 = 1$ if the gender is male. Results of this coding are shown in the final column of the table. The variable x_3 can be thought of as a "switch" which is turned "on" or "off" to indicate the two genders.

We now proceed as if the dummy variable x_3 was just an ordinary explanatory variable in the regression model, and estimate the regression coefficients using the standard least squares method to obtain the estimated model:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3.$$

Now if gender is significant in the model, b_3 will be significantly different from zero. The interpretation of b_3 is the estimated difference in monthly incomes between males

and females (holding the other explanatory variables constant). In the way that we have coded x_3, then a positive value for b_3 would indicate that males are estimated to earn more than females, and vice versa.

The computer printout contains the table of estimated regression coefficients.

Table of estimated coefficients

Variable	Estimated coefficient	Estimated standard deviation	Computed t-value
Years of education	0.321	0.156	2.06
Years of experience	0.192	0.054	3.56
Gender	0.838	0.664	1.26
Intercept	0.381		

The multiple regression model is therefore

$$\hat{y} = 0.381 + 0.321x_1 + 0.192x_2 + 0.838x_3,$$

At face value, the regression model suggests that males earn R838 more than females.

The ANOVA table is as follows:

ANOVA table

Source	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Square (MS)	F
Regression	57.7390	3	19.246	9.60
Error	32.0610	16	2.004	
Total	89.8000	19		

The observed F-value of 9.60 needs to be compared with the five per cent point of $F_{3,16}$.

From the F-tables, this is 3.24, so the multiple regression model is significant.

The multiple coefficient of determination is computed as $R^2 = \text{SSR}/\text{SST} = 57.739/89.8 = 0.643$. As a result of the inclusion of the additional variable, gender, the multiple correlation coefficient has increased from 60.7% (when we had two explanatory variables) to 64.3%. This seems a relatively small increase, especially when compared with the increase in the coefficient of determination in going from one explanatory variable (34.9%) to two (60.7%). This leads us to ask whether the addition of the third explanatory variable x_3 was worthwhile. We can do this using the methods of the previous section — we test if the coefficient associated with gender is significantly different from zero.

Following our standard layout for doing a test of a statistical hypothesis (for the last time!), the appropriate null and alternative hypotheses are:

1. $H_0 : \beta_3 = 0$

2. $H_0 : \beta_3 \neq 0$
3. Significance level: 5%
4. The degrees of freedom are $n - k - 1 = 20 - 3 - 1 = 16$. We reject H_0 if $|t| > t_{16}^{0.025} = 2.120$.
5. The test statistic is
$$t = \frac{b_3}{s_{b_3}} = \frac{0.838}{0.664} = 1.26,$$
which lies in the acceptance region.
6. We conclude that we cannot reject H_0, and that we have found no difference in monthly income between the sexes. We concludet that β_3 is not significantly different from zero.

In this example, the qualitative variable had only two categories, male and female. If there are more than two categories, it is necessary to include further dummy variables. The number of dummy variables needed is always one less than the number of categories. For example, if we decided that we also wanted to include "marital status" as an explanatory variable with the three categories "single", "married" and "divorced", we could use two dummy variables x_4 and x_5 to code each individual as follows:

$x_4 = 0$	$x_5 = 0$	for married
$x_4 = 1$	$x_5 = 0$	for single
$x_4 = 0$	$x_5 = 1$	for divorced

Example 15C: In example 6B, we investigated the relationship between rewards and risks on the Johannesburg Stock Exchange (JSE). It is well known that the JSE can be divided into three major categories of shares: Mining (M), Finance (F) and Industrial (I). The shares of example 6B fall into the following categories:

(a) Design a system of dummy variables which accommodates the three share categories in the regression model. Show how each sector is coded.
(b) Use a multiple regression computer package to compute the regression analysis with return (y) as the dependent variable using, as explanatory variables, the beta (x_1) and the share category effects. Interpret the regression coefficients.
(c) Test whether the share category effects are significant.

Sector	Return (%) y	Beta x_1	Industry
1. Diamonds	46	1.31	M
2. West Wits	67	1.15	M
3. Metals & Mining	46	1.29	M
4. Platinum	40	1.44	M
5. Mining Houses	78	1.46	F
6. Evander	40	1.00	M
7. Motor	24	0.74	I
8. Investment Trusts	20	0.67	F
9. Banks	17	0.69	F
10. Insurance	19	0.79	F
11. Property	8	0.27	F
12. Paper	25	0.75	I
13. Industrial Holdings	27	0.83	I
14. Beverages & Hotels	34	0.81	I
15. Electronics	7	0.11	I
16. Food	19	0.11	I
17. Printing	33	0.51	I

Solutions to examples...

3C (a) $P < 0.05$ ($P < 0.02$ is also correct) (b) $P < 0.20$, but this would not be considered "significant." (c) $P < 0.0005$ (d) $P < 0.025$ (e) $P < 0.20$, but not significant!

5C (a) $x = 9.284 + 0.629y$
(b) Making x the subject of the formula yields $x = 0.964 + 1.802y$.
Note that the "method of least squares" chooses the coefficients a and b to minimize **vertical distances**, i.e. parallel to the $y-$axis. Thus it is not "symmetric" in its treatment of x and y. Interchanging the roles of x and y gives rise to a new arithmetical problem, and hence a different solution.

7C (b) $r = 0.704,\ 0.01 < P < 0.02$, a significant relationship exists.
(c) $r^2 = 0.4958$. Almost 50% of the variation in the breaking stress of the casting can be explained by the breaking stress of the test piece.
(d) $y = 4.50 + 1.15\ x$ (e) $(44, 103)$.

11C (a) $\gamma = 1.404$ $C = 15978.51$ $log(C) = 9.679$, thus $PV^{1.404} = 15978.51$ (b) 24.86

14C (a) $\hat{y} = -3.518 + 1.379x_1 + 0.100x_2$.

(b) $R^2 = 0.956$. The percentage of variation in sales explained by the two explanatory variables is 95.6%.

(c) $F = 76.75 > 4.74$, hence a significant relationship exists.

(d) For x_1: $t = 6.074 > 2.365$, significant in the model.
For x_2: $t = 0.581 < 2.365$, not significant in the model.

(e) The result suggests that experience is not an important factor for the performance of part-time salespersons.

15C (a) The final two columns show the two dummy variables needed to code the share categories.

Sector	Return (%)	Beta	Dummy	
	y	x_1	x_2	x_3
1. Diamonds	46	1.31	0	0
2. West Wits	67	1.15	0	0
3. Metals & Mining	46	1.29	0	0
4. Platinum	40	1.44	0	0
5. Mining Houses	78	1.46	1	0
6. Evander	40	1.00	0	0
7. Motor	24	0.74	0	1
8. Investment Trusts	20	0.67	1	0
9. Banks	17	0.69	1	0
10. Insurance	19	0.79	1	0
11. Property	8	0.27	1	0
12. Paper	25	0.75	0	1
13. Industrial Holdings	27	0.83	0	1
14. Beverages & Hotels	34	0.81	0	1
15. Electronics	7	0.11	0	1
16. Food	19	0.11	0	1
17. Printing	33	0.51	0	1

(b) $\hat{y} = -0.578 + 39.078x_1 - 1.346x_2 + 3.172x_3$

(c) In each case, compare with $t_{13}^{0.025} = 2.160$, and reject if $|t| > 2.160$. For x_1, $t = 3.887$, significant. For x_2, $t = -0.150$, not significant. For x_3, $t = 0.320$, not significant. Hence the share categories do not have an influence on return. In other words, if the beta (risk) is held constant, no additional reward can be expected from being in a different category. This finding is consistent with the idea in finance that you reduce risk by diversifying. Therefore you cannot expect any additional rewards for only being invested in one area.

EXERCISES...

For each example it is helpful to plot a scatter plot.

12.1 The marks of 10 students in two class tests are given below.

(a) Calculate the correlation coefficient and test it for significance at the 5% level.
(b) Find the regression line for predicting y from x.
(c) What mark in the second test do you predict for a student who got 70% for the first test?

1st class test (x)	65	78	52	82	92	89	73	98	56	76
2nd class test (y)	39	43	21	64	57	47	27	75	34	52

*12.2 The table below shows the mass (y) of potassium bromide that will dissolve in $100\,\text{m}\ell$ of water at various temperatures (x).

Temperature x°C	0	20	40	60	80
Mass y (g)	54	65	75	85	96

(a) Find the regression line for predicting the y from x.
(b) Find the correlation coefficient, and test it for significance.
(c) Find 95% prediction limits for predicting y when $x = 50°C$.

12.3 Fit a linear regression to the trading revenue of licensed hotels in South Africa, 1970–1979. Test for significant correlation. Forecast the trading revenue for 1982, and find a 90% prediction interval for your forecast.

Year (x)	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
Trading revenue (millions of rands) (y)	146	170	186	209	245	293	310	337	360	420

12.4 Investigate the relationship between the depth of a lake (y) and the distance from the lake shore (x). Use the following data, which gives depths at 5 metre intervals along a line at right angles to the shore.

Distance from shore (m) (x)	5	10	15	20	25	30	35	40	45
Depth (m) (y)	2	4	6	9	13	22	37	57	94

Fit a nonlinear regression of the form $y = ae^{bx}$.

12.5 Fit a growth curve of the form $y = ab^x$ to the per capita consumption of electricity in South Africa, 1950-1978. Use the following data:

Year(x)	1950	1955	1960	1965	1969	1973	1975	1977	1978
per capita consumption (kWh/yr)(y)	733	960	1125	1483	1810	2295	2533	3005	3272

*12.6 Fit a linear regression to the number of telephones in use in South Africa, 1968–1979. Test for significant correlation. Forecast the number of telephones for 1984, and find a 90% prediction interval for this forecast.

Year(x)	1968	1969	1970	1971	1972	1973
Number of telephones (millions) (y)	1.24	1.31	1.51	1.59	1.66	1.75
Year(x)	1974	1975	1976	1977	1978	1979
Number of telephones (millions) (y)	1.86	1.98	2.11	2.24	2.36	2.59

*12.7 The number of defective items produced per unit of time, y, by a certain machine is thought to vary directly with the speed of the machine, x, measured in 1000s of revolutions per minute. Observations for 12 time periods yielded results which are summarized below:

$$\sum x = 60 \qquad \sum x^2 = 504 \qquad \sum xy = 1400$$
$$\sum y = 200 \qquad \sum y^2 = 4400$$

(a) Fit a linear regression line to the data.
(b) Calculate the correlation coefficient, and test it for significance at the 5% level.
(c) Calculate the residual standard deviation.
(d) Find a 95% prediction interval for the number of defective items when the machine is running at firstly 2000, and secondly 6000 revolutions per minute.

*12.8 20 students were given a mathematical problem to solve and a short poem to learn. Student i received a mark x_i for his progress with the problem and a mark y_i for his ability to memorize the poem.

The data is given, in summary form, below:

$$\sum x_i = 996 \qquad \sum y_i = 1\ 101$$
$$\sum x_i^2 = 60\ 972 \qquad \sum y_i^2 = 73\ 681$$
$$\sum x_i y_i = 64\ 996$$

(a) Compute the correlation coefficient, and test it for significance at the 1% level of significance.
(b) Compute the regression line for predicting the ability to memorize poetry from the ability to solve mathematical problems.
(c) Compute the regression line for predicting problem solving ability from ability to memorize poetry.
(d) Why is the line obtained in (c) not the same as the line found by making x the subject of the formula in the line obtained in (b)?

12.9 The following data is obtained to reveal the relationship between the annual production of pig iron (x) and annual numbers of pigs slaughtered (y)

Year	A	B	C	D	E	F	G	H	I	J
x : Production of pig iron (millions of metric tons)	8	7	5	6	9	8	11	9	10	12
y : Number of pigs slaughtered (10 000s)	16	12	8	10	20	14	18	17	21	20

The following information is also provided:

$\sum x = 85 \quad \sum x^2 = 765 \quad \sum y = 156 \quad \sum y^2 = 2614 \quad \sum xy = 1405$

(a) Construct a scatter plot.

(b) Compute the least squares regression line, making production of pig iron the independent variable.
(c) Determine the correlation coefficient, and test for significance at the 5% significance level.
(d) Interpret your results, considering whether your findings make sense in the light of the nature of the two variables.

12.10 Daily turnover (y) and price (x) of a product were measured at each of 10 retail outlets. The data was combined and summarized and the least squares regression line was found to be $y = 8.7754 - 0.211\ x$. The original data was lost and only the following totals remain:

$$\sum x = 55 \qquad \sum x^2 = 385 \qquad \sum y^2 = 853.27$$

(a) Find the correlation coefficient r between x and y and test for significance at a 1% significance level.
(b) Find the residual standard deviation.

12.11 An agricultural company is testing the theory that, for a limited range of values, plant growth (y) and fertilizer (x) are related by an equation of the form $y = ax^b$. Plant growth and fertilizer values were recorded for a sample of 10 plants and after some calculation the following results were obtained:

$$n = 10 \qquad \sum x = 54 \qquad \sum y = 1427 \qquad \sum xy = 10\ 677$$
$$\sum(\log x) = 6.5 \quad \sum(\log y) = 20.0 \quad \sum(\log x\ \log y) = 14.2$$
$$\sum(\log x)^2 = 5.2 \quad \sum(\log y)^2 = 41.7 \quad \sum x^2 = 366$$
$$\sum y^2 = 301\ 903$$

(a) Use these results to fit the equation to this data.
(b) Use your equation to predict plant growth when $x = 10$.

*12.12 A famous economist has a theory that the quantity of a commodity produced is related to the price of the commodity by the function

$$Q = a\,b^P$$

where Q is the quantity produced, and P is the price.

The economist studies the market for 8 months, and notes the following results:

P	Q
1.35	10.35
1.45	11.25
1.50	11.95
1.55	12.55
1.70	14.10
1.90	17.25
2.05	19.85
2.10	20.30

(a) Calculate the values of the constants a and b that best fit the data. You will need some of the following:

$$\sum P = 13.60 \qquad \sum Q = 117.60 \qquad \sum P^2 = 23.68$$
$$\sum Q^2 = 1836.48 \quad \sum PQ = 207.73 \quad \sum \log Q = 9.24$$
$$\sum (\log Q)^2 = 10.77 \quad \sum P \log Q = 15.94$$

(The logarithms have been taken to base 10)

(b) Estimate Q when $P = 2$.

12.13 Given a set of data points $(x_i\,,\, y_i)$, and that the slopes of the regression lines for predicting y from x and x from y are b_1 and b_2 respectively, show that the correlation coefficient is given by

$$r = \sqrt{b_1 b_2}$$

AN EXERCISE ON TRANSFORMATIONS TO LINEARITY...

12.14 Find transformations that linearize the following.

(a) The logistic growth curve

$$y = \frac{1}{1 + ae^{bx}}.$$

(Hint: consider $1 - y$, then take logarithms. Note also that $0 < y < 1$, and that y is interpreted as the proportion of the asymptote [the final value] grown at time x. The curve is shaped like an "S" squashed to the right!)

(b) The Beverton-Holt stock-recruit relationship used in fisheries models

$$y = ax/(b + x)$$

where y is the recruitment value, and x the size of the parent stock.

(c) The Ricker stock-recruit model

$$y = axe^{b^x}$$

where x and y have the same meanings as in (b).

(d) The Gompertz growth curve

$$y = (e^a)^{b^x}$$

where $b > 0$, and y is the proportion of the asymptote.

(e) The von Bertalanffy growth curve

$$y = (1 - aw^{bx})^{-3}$$

where $a > 0$, and y is the proportion of the asymptote. (Hint: consider $(1 - y^{\frac{1}{3}})^{-1}$.)

The growth curves (d) and (e) are similar to the logistic growth curve (a). They differ in that growth in the initial period is faster, then slower in the middle, and the approach to the asymptote is more gradual.

EXERCISES ON MULTIPLE REGRESSION ...

12.15 Shown below is a partial computer output from a regression analysis of the form

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3.$$

Table of estimated coefficients		
Variable	Estimated coefficient	Estimated standard deviation
x_1	0.044	0.011
x_2	1.271	0.418
x_3	0.619	0.212
Intercept	0.646	

ANOVA table			
Source	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Square (MS)
Regression	22.012	3	C
Error	A	B	0.331
Total	24.000	9	

(a) Find values for A, B, and C in the ANOVA table.
(b) Compute and interpret R^2, the multiple coefficient of determination.
(c) Test the regression equation for significance at the 5% level.
(d) Test each of the explanatory variables for significance (5% level).

12.16 A large corporation conducted an investigation of the job satisfaction (y) of its employees, as a function of their length of service (x_1) and their salary (x_2). Job satisfaction (y) was rated on a scale from 0 to 10, x_1 was measured in years and x_2 was measured in R1000s per month. The following regression model was estimated from a sample of employees.

$$\hat{y} = 4.412 - 0.615x_1 + 0.353x_2.$$

Part of the computer output was as follows:

Table of estimated coefficients		
Variable	Estimated coefficient	Estimated standard deviation
Length of service (x_1)		0.187
Salary (x_2)		0.061
Intercept		

ANOVA table			
Source	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Square (MS)
Regression	7.646	C	D
Error	B	7	F
Total	A	E	

$R^2 = 0.8364$

(a) Interpret the implications of the signs of b_1 and b_2 for job satisfaction.
(b) Complete the entries A – F in the ANOVA table. What was the sample size?
(c) Test the overall regression for significance at the 1% level.
(d) Compute the appropriate t-ratios, and test the explanatory variables individually for significance.

12.17 In an analysis to predict sales of for a chain of stores, the planning department gathers data for a sample of stores. They take account of the number of competitors in a 2 km radius (x_1), the number of parking bays within a 100 m radius (x_2), and whether or not there is an automatic cash-dispensing machine on the premises (x_3), where $x_3 = 1$ if a cash-dispensing machine is present, and $x_3 = 0$ if it is not. They estimate the multiple regression model

$$\hat{y} = 11.1 - 3.2x_1 + 0.4x_2 + 8.5x_3,$$

where y is daily sales in R1000s.

(a) What does the model suggest is the effect of having an automatic cash-dispensing machine on the premises?
(b) Estimate sales for a store which will be opened in an area with
  (i) 3 competitors within 2 km, 56 parking bays within 100 m and a cash-dispensing machine;
  (ii) 1 competitor, 35 parking bays and no cash-dispensing machine.

Solutions to exercises...

12.1 (a) $r = 0.84 > 0.6319$, reject H_0.

(b) $y = -24.58 + 0.93\ x$.

(c) When $x = 70 \quad y = 40.25$.

12.2 (a) $y = 54.2 + 0.52\ x$

(b) $r = 0.9998, \quad P < 0.001$, very highly significant.

(c) $s = 0.365$, prediction interval: $(78.9, 81.5)$.

12.3 $r = 0.9922, \quad P < 0.001$, very highly significant. $y = 133.91 + 29.71\ x$, taking x as years since 1970. For 1982, $x = 12$, $y = 490.42$, and the 90% prediction interval is (461.03, 519.81).

12.4 $y = 1.395\ e^{0.0930\ x}$

12.5 $y = 705.71 \times 1.054^x$, taking x as years since 1950.

$r = 0.997, P < 0.001$, significant.

12.6 $y = 1.214 + 0.116\ x$, taking x as years since 1968. In 1984, $x = 16$, $\hat{y} = 3.06$ million telephones. The 90% prediction interval is $(2.94, 3.19)$. $r = 0.994$, $P < 0.001$, significant.

12.7 (a) $y = 6.86 + 1.96\ x$

(b) $r = 0.857$, $P < 0.001$, significant.

(c) $s = 5.31$.

(d) 10.78 ± 12.57, 18.63 ± 12.35.

12.8 (a) $r = 0.8339$, $P < 0.001$, significant.

(b) $y = 10.329 + 0.894\ x$.

(c) $x = 6.971 + 0.778\ y$.

(d) The method of least squares is not symmetric in its treatment of the dependent and independent variables.

12.9 (b) $y = -0.200 + 1.859\ x$.

(c) $r = 0.902 > 0.6319$, reject H_0.

(d) There is not a cause and effect relationship between the variables. The relationship between the variables is caused by a third variable, possibly either "population" or "standard of living".

12.10 (a) $r = -0.1159 > -0.7646$, cannot reject H_0.

(b) 5.81.

12.11 (a) $y = 3.32\ x^{1.231}$.

(b) When $x = 10, \quad y = 56.5$.

12.12 (a) $\log_{10} Q = 0.4507 + 0.4143\, P$

or $Q = 10^{0.4507\ 0.4143P} = 2.823 \times 2.596^P$

(b) $Q = 19.02$.

12.15 (a) $A = 1.988,\ B = 6,\ C = 7.337$

(b) $R^2 = 0.9172$, or 91.7% of the variation accounted for by the explanatory variables.

(c) $F = 22.17 > F_{3,6}^{(0.05)} = 4.76$, hence a significant relationshop exists.

(d) In each case, explanatory variable is significant if $|t| > 2.447$. For x_1, $t = 4.00$, for x_2, $t = 3.04$, and for x_3, $t = 2.92$. Hence all three explanatory variables are significant in the regression model.

12.16 (a) The longer the service the less the job satisfaction. Larger salaries are associated with greater job satisfaction.

(b) $A = 9.142,\ B = 1.496,\ C = 2,\ D = 3.823,\ E = 9,\ F = 0.214$, and $n = 10$.

(c) $F = 17.86 > F_{2,7}^{(0.01)} = 9.55$, hence a significant relationship exists.

(d) In each case the explanatory variable is significant if $|t| > 2.365$. For x_1, $t = -3.289$, for x_2, $t = 5.787$. Hence both explanatory variables are significant in the regression model.

12.17 (a) R8500 per day increase in sales (b) (i) R32 400 (ii) R21 900

SUMMARY OF THE PROBABILITY DISTRIBUTIONS

ALL IN ONE PLACE!...

In this section, we summarize the probability distributions introduced in this book. The probability mass functions and probability density functions are given, as well as their means and variances.

BINOMIAL DISTRIBUTION — Discrete

We have n independent trials, each trial has two outcomes, success of failure, and Pr[success] $= p$ for all trials. The random variable X is the number of successes in n trials; $n \geq 1$ must be an integer, and $0 \leq p \leq 1$. Then X has the binomial distribution, i.e. $X \sim B(n, p)$, with probability mass function

$$\begin{aligned} p(x) &= \binom{n}{x} p^x (1-p)^{n-x} && x = 0, 1, \ldots, n \\ &= 0 && \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = np \qquad \mathrm{Var}[X] = np(1-p)$$

POISSON DISTRIBUTION — Discrete

Events occur at random in time, with an average rate of λ events per time period (or space). The random variable X is a count of the number of events occurring during a fixed interval of time (or space). The time period (or amount of space) referred to in the rate must be the same as the time period (or space) in which events are counted. Then X has the Poisson distribution with parameter $\lambda > 0$, i.e. $X \sim P(\lambda)$, and has probability mass function

$$\begin{aligned} p(x) &= \frac{e^{-\lambda}\lambda^x}{x!} && x = 0, 1, 2, \ldots \\ &= 0 && \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = \lambda \qquad \mathrm{Var}[X] = \lambda$$

EXPONENTIAL DISTRIBUTION Continuous

As for the Poisson distribution, events occur at random with average rate λ per unit of time (or space). Let the continuous random variable X be the interval between two events. X has the exponential distribution with parameter λ, i.e. $X \sim E(\lambda)$, with probability density function

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} \quad x \geq 0 \\ &= 0 \qquad\quad \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = \frac{1}{\lambda} \qquad \mathrm{Var}[X] = \frac{1}{\lambda^2}$$

NORMAL DISTRIBUTION Continuous

The normal distribution with parameters $\mu(-\infty < \mu < \infty)$ and $\sigma^2(\sigma > 0)$ is a continuous random variable X, denoted $X \sim N(\mu, \sigma^2)$, with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

$$\mathrm{E}[X] = \mu \qquad \mathrm{Var}[X] = \sigma^2$$

The standard normal distribution is a continuous random variable Z, and has $\mu = 0, \sigma = 1$. The probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}z^2} \qquad -\infty < x < \infty$$

$$\mathrm{E}[X] = 0 \qquad \mathrm{Var}[X] = 1$$

NEGATIVE BINOMIAL DISTRIBUTION Discrete

As for the binomial distribution, we have independent trials, each with two outcomes, success or failure; Pr[success] $= p$ for each trial. Fix the number of successes r, and let the random variable X be the number of failures obtained **before** the rth success. Then X has the negative binomial distribution with parameters r and p, i.e. $X \sim NB(r,\ p)$, with probability mass function

$$\begin{aligned} p(x) &= \binom{x+r-1}{x}\ p^r\ q^x \quad x = 0, 1, 2, \ldots \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = \frac{r(1-p)}{p} \qquad \mathrm{Var}[X] = \frac{r(1-p)}{p^2}$$

GEOMETRIC DISTRIBUTION Discrete

Under the same conditions as for the negative binomial distribution, let the random variable X be the number of trials **before** the first success. Then X has the geometric distribution with parameter p, $X \sim G(p)$, and has probability mass function

$$\begin{aligned} p(x) &= pq^x \quad x = 0,\ 1,\ 2,\ \ldots \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = \frac{(1-p)}{p} \qquad \mathrm{Var}[X] = \frac{(1-p)}{p^2}$$

HYPERGEOMETRIC DISTRIBUTION Discrete

Given N objects, of which M are are of type 1 and $N-M$ of type 2, a sample of size n $(n \leq N)$ is drawn. Let the random variable X be the number of objects of type 1 in the sample. Then X has the hypergeometric distribution, $X \sim H(N,\ M,\ n)$, with probability mass function

$$\begin{aligned} p(x) &= \binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n} \quad x = 0,\ 1,\ \ldots, n \\ &= 0 \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = \frac{nM}{N} \qquad \mathrm{Var}[X] = n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$$

UNIFORM DISTRIBUTION Continuous

If the continuous random variable X is equally likely to take on any value in the interval (a, b), then X has the uniform distribution, $X \sim U(a, b)$, with probability density function

$$\begin{aligned} f(x) &= \frac{1}{b-a} \quad a \leq x \leq b \\ &= 0 \qquad\quad \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = \frac{a+b}{2} \qquad \mathrm{Var}[X] = \frac{(b-a)^2}{12}$$

THE t, F and χ^2 DISTRIBUTIONS Continuous

For completeness sake, we state the probability density functions of these three distributions. The t-distribution with parameters n, the degrees of freedom, is a continous random variable with probability density function

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} \qquad -\infty < x < \infty$$

$$\mathrm{E}[X] = 0 \qquad \mathrm{Var}[X] = \frac{n}{n-2}$$

The F-distribution with parameters n_1 and n_2, the degrees of freedom for the numerator and denominator respectively, is a continuous random variable with probability density function

$$\begin{aligned} f(x) &= \frac{\Gamma[(n_1+n_2)/2]}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{(n_1/2)} \frac{x^{(n_1/2)-1}}{(1+\frac{n_1}{n_2}x)^{(n_1+n_2)/2}} & 0 < x < \infty \\ &= 0 & \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = \frac{n_2}{n_2-2} \qquad \mathrm{Var}[X] = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$$

The χ^2-distribution with parameter n, the degrees of freedom, is a continuous random variable with probability density function

$$\begin{aligned} f(x) &= \frac{1}{2^{n/2}\Gamma(n/2)} \; x^{n/2-1}e^{-x/2} & 0 < x < \infty \\ &= 0 & \text{otherwise} \end{aligned}$$

$$\mathrm{E}[X] = n \qquad \mathrm{Var}[X] = 2n$$

$\Gamma(n)$ is defined to be $(n-1)!$ if n is an integer. If $n + \frac{1}{2}$ is an integer, then

$$\Gamma(n+\frac{1}{2}) = \frac{1 \times 3 \times 5 \times \cdots \times (2n-1)}{2^n}\sqrt{\pi}$$

so that

$$\Gamma(4) = 3! = 6 \qquad \text{and} \qquad \Gamma(2.5) = \frac{1 \times 3}{2^2}\sqrt{\pi} = 1.329$$

TABLES

TABLE 1. STANDARD NORMAL DISTRIBUTION: Areas under the standard normal curve between 0 and z, i.e. $\Pr[0 < Z < z]$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998	0.49998	0.49998	0.49998

TABLE 2. t-DISTRIBUTION: One sided critical values, i.e. the value of t_{df}^{P} such that $P = \Pr[t_{df} > t_{df}^{P}]$, where df is the degrees of freedom

	Probability Level P										
df	0.4	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	0.727	1.376	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	4.773	5.894	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.689
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.660
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.256	0.530	0.853	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.255	0.530	0.853	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.255	0.530	0.853	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	0.255	0.529	0.852	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37	0.255	0.529	0.851	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	0.255	0.529	0.851	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39	0.255	0.529	0.851	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
45	0.255	0.528	0.850	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520
50	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
70	0.254	0.527	0.847	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
80	0.254	0.526	0.846	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
100	0.254	0.526	0.845	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
110	0.254	0.526	0.845	1.289	1.659	1.982	2.361	2.621	2.865	3.166	3.381
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
140	0.254	0.526	0.844	1.288	1.656	1.977	2.353	2.611	2.852	3.149	3.361
160	0.254	0.525	0.844	1.287	1.654	1.975	2.350	2.607	2.847	3.142	3.352
180	0.254	0.525	0.844	1.286	1.653	1.973	2.347	2.603	2.842	3.136	3.345
200	0.254	0.525	0.843	1.286	1.653	1.972	2.345	2.601	2.838	3.131	3.340
z	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

TABLE 3. CHI-SQUARED DISTRIBUTION: One sided critical values, i.e. the value of $\chi^{2(P)}_{df}$ such that $P = \Pr[\chi^2_{df} > \chi^{2(P)}_{df}]$, where df is the degrees of freedom, for $P > 0.5$

	Probability Level P									
df	0.9995	0.999	0.9975	0.995	0.99	0.975	0.95	0.9	0.8	0.6
1	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.016	0.064	0.275
2	0.001	0.002	0.005	0.010	0.020	0.051	0.103	0.211	0.446	1.022
3	0.015	0.024	0.045	0.072	0.115	0.216	0.352	0.584	1.005	1.869
4	0.064	0.091	0.145	0.207	0.297	0.484	0.711	1.064	1.649	2.753
5	0.158	0.210	0.307	0.412	0.554	0.831	1.145	1.610	2.343	3.656
6	0.299	0.381	0.527	0.676	0.872	1.237	1.635	2.204	3.070	4.570
7	0.485	0.599	0.794	0.989	1.239	1.690	2.167	2.833	3.822	5.493
8	0.710	0.857	1.104	1.344	1.647	2.180	2.733	3.490	4.594	6.423
9	0.972	1.152	1.450	1.735	2.088	2.700	3.325	4.168	5.380	7.357
10	1.265	1.479	1.827	2.156	2.558	3.247	3.940	4.865	6.179	8.295
11	1.587	1.834	2.232	2.603	3.053	3.816	4.575	5.578	6.989	9.237
12	1.935	2.214	2.661	3.074	3.571	4.404	5.226	6.304	7.807	10.182
13	2.305	2.617	3.112	3.565	4.107	5.009	5.892	7.041	8.634	11.129
14	2.697	3.041	3.582	4.075	4.660	5.629	6.571	7.790	9.467	12.078
15	3.107	3.483	4.070	4.601	5.229	6.262	7.261	8.547	10.307	13.030
16	3.536	3.942	4.573	5.142	5.812	6.908	7.962	9.312	11.152	13.983
17	3.980	4.416	5.092	5.697	6.408	7.564	8.672	10.085	12.002	14.937
18	4.439	4.905	5.623	6.265	7.015	8.231	9.390	10.865	12.857	15.893
19	4.913	5.407	6.167	6.844	7.633	8.907	10.117	11.651	13.716	16.850
20	5.398	5.921	6.723	7.434	8.260	9.591	10.851	12.443	14.578	17.809
21	5.895	6.447	7.289	8.034	8.897	10.283	11.591	13.240	15.445	18.768
22	6.404	6.983	7.865	8.643	9.542	10.982	12.338	14.041	16.314	19.729
23	6.924	7.529	8.450	9.260	10.196	11.689	13.091	14.848	17.187	20.690
24	7.453	8.085	9.044	9.886	10.856	12.401	13.848	15.659	18.062	21.652
25	7.991	8.649	9.646	10.520	11.524	13.120	14.611	16.473	18.940	22.616
26	8.537	9.222	10.256	11.160	12.198	13.844	15.379	17.292	19.820	23.579
27	9.093	9.803	10.873	11.808	12.878	14.573	16.151	18.114	20.703	24.544
28	9.656	10.391	11.497	12.461	13.565	15.308	16.928	18.939	21.588	25.509
29	10.227	10.986	12.128	13.121	14.256	16.047	17.708	19.768	22.475	26.475
30	10.804	11.588	12.765	13.787	14.953	16.791	18.493	20.599	23.364	27.442
31	11.388	12.196	13.407	14.458	15.655	17.539	19.281	21.434	24.255	28.409
32	11.980	12.810	14.055	15.134	16.362	18.291	20.072	22.271	25.148	29.376
33	12.576	13.431	14.709	15.815	17.073	19.047	20.867	23.110	26.042	30.344
34	13.180	14.057	15.368	16.501	17.789	19.806	21.664	23.952	26.938	31.313
35	13.788	14.688	16.032	17.192	18.509	20.569	22.465	24.797	27.836	32.282
36	14.401	15.324	16.700	17.887	19.233	21.336	23.269	25.643	28.735	33.252
37	15.021	15.965	17.373	18.586	19.960	22.106	24.075	26.492	29.635	34.222
38	15.644	16.611	18.050	19.289	20.691	22.878	24.884	27.343	30.537	35.192
39	16.272	17.261	18.732	19.996	21.426	23.654	25.695	28.196	31.441	36.163
40	16.906	17.917	19.417	20.707	22.164	24.433	26.509	29.051	32.345	37.134
45	20.136	21.251	22.899	24.311	25.901	28.366	30.612	33.350	36.884	41.995
50	23.461	24.674	26.464	27.991	29.707	32.357	34.764	37.689	41.449	46.864
60	30.339	31.738	33.791	35.534	37.485	40.482	43.188	46.459	50.641	56.620
70	37.467	39.036	41.332	43.275	45.442	48.758	51.739	55.329	59.898	66.396
80	44.792	46.520	49.043	51.172	53.540	57.153	60.391	64.278	69.207	76.188
90	52.277	54.156	56.892	59.196	61.754	65.647	69.126	73.291	78.558	85.993
100	59.895	61.918	64.857	67.328	70.065	74.222	77.929	82.358	87.945	95.808
110	67.631	69.790	72.922	75.550	78.458	82.867	86.792	91.471	97.362	105.632
120	75.465	77.756	81.073	83.852	86.923	91.573	95.705	100.624	106.806	115.465
140	91.389	93.925	97.591	100.655	104.034	109.137	113.659	119.029	125.758	135.149
160	107.599	110.359	114.350	117.679	121.346	126.870	131.756	137.546	144.783	154.856
180	124.032	127.011	131.305	134.884	138.821	144.741	149.969	156.153	163.868	174.580
200	140.659	143.842	148.426	152.241	156.432	162.728	168.279	174.835	183.003	194.319

TABLE 3, continued. CHI-SQUARED DISTRIBUTION: One sided critical values, i.e. the value of $\chi^{2(P)}_{df}$ such that $P = \Pr[\chi^2_{df} > \chi^{2(P)}_{df}]$, where df is the degrees of freedom, for $P < 0.5$

	Probability Level P									
df	0.4	0.2	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.708	1.642	2.706	3.841	5.024	6.635	7.879	9.140	10.827	12.115
2	1.833	3.219	4.605	5.991	7.378	9.210	10.597	11.983	13.815	15.201
3	2.946	4.642	6.251	7.815	9.348	11.345	12.838	14.320	16.266	17.731
4	4.045	5.989	7.779	9.488	11.143	13.277	14.860	16.424	18.466	19.998
5	5.132	7.289	9.236	11.070	12.832	15.086	16.750	18.385	20.515	22.106
6	6.211	8.558	10.645	12.592	14.449	16.812	18.548	20.249	22.457	24.102
7	7.283	9.803	12.017	14.067	16.013	18.475	20.278	22.040	24.321	26.018
8	8.351	11.030	13.362	15.507	17.535	20.090	21.955	23.774	26.124	27.867
9	9.414	12.242	14.684	16.919	19.023	21.666	23.589	25.463	27.877	29.667
10	10.473	13.442	15.987	18.307	20.483	23.209	25.188	27.112	29.588	31.419
11	11.530	14.631	17.275	19.675	21.920	24.725	26.757	28.729	31.264	33.138
12	12.584	15.812	18.549	21.026	23.337	26.217	28.300	30.318	32.909	34.821
13	13.636	16.985	19.812	22.362	24.736	27.688	29.819	31.883	34.527	36.477
14	14.685	18.151	21.064	23.685	26.119	29.141	31.319	33.426	36.124	38.109
15	15.733	19.311	22.307	24.996	27.488	30.578	32.801	34.949	37.698	39.717
16	16.780	20.465	23.542	26.296	28.845	32.000	34.267	36.456	39.252	41.308
17	17.824	21.615	24.769	27.587	30.191	33.409	35.718	37.946	40.791	42.881
18	18.868	22.760	25.989	28.869	31.526	34.805	37.156	39.422	42.312	44.434
19	19.910	23.900	27.204	30.144	32.852	36.191	38.582	40.885	43.819	45.974
20	20.951	25.038	28.412	31.410	34.170	37.566	39.997	42.336	45.314	47.498
21	21.992	26.171	29.615	32.671	35.479	38.932	41.401	43.775	46.796	49.010
22	23.031	27.301	30.813	33.924	36.781	40.289	42.796	45.204	48.268	50.510
23	24.069	28.429	32.007	35.172	38.076	41.638	44.181	46.623	49.728	51.999
24	25.106	29.553	33.196	36.415	39.364	42.980	45.558	48.034	51.179	53.478
25	26.143	30.675	34.382	37.652	40.646	44.314	46.928	49.435	52.619	54.948
26	27.179	31.795	35.563	38.885	41.923	45.642	48.290	50.829	54.051	56.407
27	28.214	32.912	36.741	40.113	43.195	46.963	49.645	52.215	55.475	57.856
28	29.249	34.027	37.916	41.337	44.461	48.278	50.994	53.594	56.892	59.299
29	30.283	35.139	39.087	42.557	45.722	49.588	52.335	54.966	58.301	60.734
30	31.316	36.250	40.256	43.773	46.979	50.892	53.672	56.332	59.702	62.160
31	32.349	37.359	41.422	44.985	48.232	52.191	55.002	57.692	61.098	63.581
32	33.381	38.466	42.585	46.194	49.480	53.486	56.328	59.046	62.487	64.993
33	34.413	39.572	43.745	47.400	50.725	54.775	57.648	60.395	63.869	66.401
34	35.444	40.676	44.903	48.602	51.966	56.061	58.964	61.738	65.247	67.804
35	36.475	41.778	46.059	49.802	53.203	57.342	60.275	63.076	66.619	69.197
36	37.505	42.879	47.212	50.998	54.437	58.619	61.581	64.410	67.985	70.588
37	38.535	43.978	48.363	52.192	55.668	59.893	62.883	65.738	69.348	71.971
38	39.564	45.076	49.513	53.384	56.895	61.162	64.181	67.063	70.704	73.350
39	40.593	46.173	50.660	54.572	58.120	62.428	65.475	68.383	72.055	74.724
40	41.622	47.269	51.805	55.758	59.342	63.691	66.766	69.699	73.403	76.096
45	46.761	52.729	57.505	61.656	65.410	69.957	73.166	76.223	80.078	82.873
50	51.892	58.164	63.167	67.505	71.420	76.154	79.490	82.664	86.660	89.560
60	62.135	68.972	74.397	79.082	83.298	88.379	91.952	95.344	99.608	102.697
70	72.358	79.715	85.527	90.531	95.023	100.425	104.215	107.808	112.317	115.577
80	82.566	90.405	96.578	101.879	106.629	112.329	116.321	120.102	124.839	128.264
90	92.761	101.054	107.565	113.145	118.136	124.116	128.299	132.255	137.208	140.780
100	102.946	111.667	118.498	124.342	129.561	135.807	140.170	144.292	149.449	153.164
110	113.121	122.250	129.385	135.480	140.916	147.414	151.948	156.230	161.582	165.436
120	123.289	132.806	140.233	146.567	152.211	158.950	163.648	168.081	173.618	177.601
140	143.604	153.854	161.827	168.613	174.648	181.841	186.847	191.565	197.450	201.680
160	163.898	174.828	183.311	190.516	196.915	204.530	209.824	214.808	221.020	225.477
180	184.173	195.743	204.704	212.304	219.044	227.056	232.620	237.855	244.372	249.049
200	204.434	216.609	226.021	233.994	241.058	249.445	255.264	260.735	267.539	272.422

TABLE 4.1. 5% critical values for the F-DISTRIBUTION, i.e. the value of $F^{(0.05)}_{\text{NUM,DEN}}$ where NUM and DEN are the numerator and denominator degrees of freedom respectively

	NUM (Numerator Degrees of Freedom)													
DEN	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	244.7	245.4
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04
31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15	2.11	2.08	2.05	2.03
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07	2.04	2.01
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13	2.09	2.06	2.03	2.00
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05	2.02	1.99
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04	2.01	1.99
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.07	2.03	2.00	1.98
37	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.14	2.10	2.06	2.02	2.00	1.97
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.99	1.96
39	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08	2.04	2.01	1.98	1.95
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	2.01	1.97	1.94	1.92
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.84
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88	1.84	1.82
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.90	1.86	1.83	1.80
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79
110	3.93	3.08	2.69	2.45	2.30	2.18	2.09	2.02	1.97	1.92	1.88	1.84	1.81	1.78
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.83	1.80	1.78
140	3.91	3.06	2.67	2.44	2.28	2.16	2.08	2.01	1.95	1.90	1.86	1.82	1.79	1.76
160	3.90	3.05	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.81	1.78	1.75
180	3.89	3.05	2.65	2.42	2.26	2.15	2.06	1.99	1.93	1.88	1.84	1.81	1.77	1.75
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80	1.77	1.74

TABLE 4.1, continued. 5% critical values for the F-DISTRIBUTION

	NUM (Numerator Degrees of Freedom)													
DEN	15	16	17	18	19	20	22	24	27	30	40	60	100	200
1	245.9	246.5	246.9	247.3	247.7	248.0	248.6	249.1	249.6	250.1	251.1	252.2	253.0	253.7
2	19.43	19.43	19.44	19.44	19.44	19.45	19.45	19.45	19.46	19.46	19.47	19.48	19.49	19.49
3	8.70	8.69	8.68	8.67	8.67	8.66	8.65	8.64	8.63	8.62	8.59	8.57	8.55	8.54
4	5.86	5.84	5.83	5.82	5.81	5.80	5.79	5.77	5.76	5.75	5.72	5.69	5.66	5.65
5	4.62	4.60	4.59	4.58	4.57	4.56	4.54	4.53	4.51	4.50	4.46	4.43	4.41	4.39
6	3.94	3.92	3.91	3.90	3.88	3.87	3.86	3.84	3.82	3.81	3.77	3.74	3.71	3.69
7	3.51	3.49	3.48	3.47	3.46	3.44	3.43	3.41	3.39	3.38	3.34	3.30	3.27	3.25
8	3.22	3.20	3.19	3.17	3.16	3.15	3.13	3.12	3.10	3.08	3.04	3.01	2.97	2.95
9	3.01	2.99	2.97	2.96	2.95	2.94	2.92	2.90	2.88	2.86	2.83	2.79	2.76	2.73
10	2.85	2.83	2.81	2.80	2.79	2.77	2.75	2.74	2.72	2.70	2.66	2.62	2.59	2.56
11	2.72	2.70	2.69	2.67	2.66	2.65	2.63	2.61	2.59	2.57	2.53	2.49	2.46	2.43
12	2.62	2.60	2.58	2.57	2.56	2.54	2.52	2.51	2.48	2.47	2.43	2.38	2.35	2.32
13	2.53	2.51	2.50	2.48	2.47	2.46	2.44	2.42	2.40	2.38	2.34	2.30	2.26	2.23
14	2.46	2.44	2.43	2.41	2.40	2.39	2.37	2.35	2.33	2.31	2.27	2.22	2.19	2.16
15	2.40	2.38	2.37	2.35	2.34	2.33	2.31	2.29	2.27	2.25	2.20	2.16	2.12	2.10
16	2.35	2.33	2.32	2.30	2.29	2.28	2.25	2.24	2.21	2.19	2.15	2.11	2.07	2.04
17	2.31	2.29	2.27	2.26	2.24	2.23	2.21	2.19	2.17	2.15	2.10	2.06	2.02	1.99
18	2.27	2.25	2.23	2.22	2.20	2.19	2.17	2.15	2.13	2.11	2.06	2.02	1.98	1.95
19	2.23	2.21	2.20	2.18	2.17	2.16	2.13	2.11	2.09	2.07	2.03	1.98	1.94	1.91
20	2.20	2.18	2.17	2.15	2.14	2.12	2.10	2.08	2.06	2.04	1.99	1.95	1.91	1.88
21	2.18	2.16	2.14	2.12	2.11	2.10	2.07	2.05	2.03	2.01	1.96	1.92	1.88	1.84
22	2.15	2.13	2.11	2.10	2.08	2.07	2.05	2.03	2.00	1.98	1.94	1.89	1.85	1.82
23	2.13	2.11	2.09	2.08	2.06	2.05	2.02	2.01	1.98	1.96	1.91	1.86	1.82	1.79
24	2.11	2.09	2.07	2.05	2.04	2.03	2.00	1.98	1.96	1.94	1.89	1.84	1.80	1.77
25	2.09	2.07	2.05	2.04	2.02	2.01	1.98	1.96	1.94	1.92	1.87	1.82	1.78	1.75
26	2.07	2.05	2.03	2.02	2.00	1.99	1.97	1.95	1.92	1.90	1.85	1.80	1.76	1.73
27	2.06	2.04	2.02	2.00	1.99	1.97	1.95	1.93	1.90	1.88	1.84	1.79	1.74	1.71
28	2.04	2.02	2.00	1.99	1.97	1.96	1.93	1.91	1.89	1.87	1.82	1.77	1.73	1.69
29	2.03	2.01	1.99	1.97	1.96	1.94	1.92	1.90	1.88	1.85	1.81	1.75	1.71	1.67
30	2.01	1.99	1.98	1.96	1.95	1.93	1.91	1.89	1.86	1.84	1.79	1.74	1.70	1.66
31	2.00	1.98	1.96	1.95	1.93	1.92	1.90	1.88	1.85	1.83	1.78	1.73	1.68	1.65
32	1.99	1.97	1.95	1.94	1.92	1.91	1.88	1.86	1.84	1.82	1.77	1.71	1.67	1.63
33	1.98	1.96	1.94	1.93	1.91	1.90	1.87	1.85	1.83	1.81	1.76	1.70	1.66	1.62
34	1.97	1.95	1.93	1.92	1.90	1.89	1.86	1.84	1.82	1.80	1.75	1.69	1.65	1.61
35	1.96	1.94	1.92	1.91	1.89	1.88	1.85	1.83	1.81	1.79	1.74	1.68	1.63	1.60
36	1.95	1.93	1.92	1.90	1.88	1.87	1.85	1.82	1.80	1.78	1.73	1.67	1.62	1.59
37	1.95	1.93	1.91	1.89	1.88	1.86	1.84	1.82	1.79	1.77	1.72	1.66	1.62	1.58
38	1.94	1.92	1.90	1.88	1.87	1.85	1.83	1.81	1.78	1.76	1.71	1.65	1.61	1.57
39	1.93	1.91	1.89	1.88	1.86	1.85	1.82	1.80	1.77	1.75	1.70	1.65	1.60	1.56
40	1.92	1.90	1.89	1.87	1.85	1.84	1.81	1.79	1.77	1.74	1.69	1.64	1.59	1.55
45	1.89	1.87	1.86	1.84	1.82	1.81	1.78	1.76	1.73	1.71	1.66	1.60	1.55	1.51
50	1.87	1.85	1.83	1.81	1.80	1.78	1.76	1.74	1.71	1.69	1.63	1.58	1.52	1.48
60	1.84	1.82	1.80	1.78	1.76	1.75	1.72	1.70	1.67	1.65	1.59	1.53	1.48	1.44
70	1.81	1.79	1.77	1.75	1.74	1.72	1.70	1.67	1.65	1.62	1.57	1.50	1.45	1.40
80	1.79	1.77	1.75	1.73	1.72	1.70	1.68	1.65	1.63	1.60	1.54	1.48	1.43	1.38
90	1.78	1.76	1.74	1.72	1.70	1.69	1.66	1.64	1.61	1.59	1.53	1.46	1.41	1.36
100	1.77	1.75	1.73	1.71	1.69	1.68	1.65	1.63	1.60	1.57	1.52	1.45	1.39	1.34
110	1.76	1.74	1.72	1.70	1.68	1.67	1.64	1.62	1.59	1.56	1.50	1.44	1.38	1.33
120	1.75	1.73	1.71	1.69	1.67	1.66	1.63	1.61	1.58	1.55	1.50	1.43	1.37	1.32
140	1.74	1.72	1.70	1.68	1.66	1.65	1.62	1.60	1.57	1.54	1.48	1.41	1.35	1.30
160	1.73	1.71	1.69	1.67	1.65	1.64	1.61	1.59	1.56	1.53	1.47	1.40	1.34	1.28
180	1.72	1.70	1.68	1.66	1.64	1.63	1.60	1.58	1.55	1.52	1.46	1.39	1.33	1.27
200	1.72	1.69	1.67	1.66	1.64	1.62	1.60	1.57	1.54	1.52	1.46	1.39	1.32	1.26

TABLE 4.2. 2.5% critical values for the F-DISTRIBUTION, i.e. the value of $F^{(0.025)}_{\text{NUM,DEN}}$ where NUM and DEN are the numerator and denominator degrees of freedom respectively

	NUM (Numerator Degrees of Freedom)													
DEN	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	648	799	864	900	922	937	948	957	963	969	973	977	980	983
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.41	39.42	39.43
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.30	14.28
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.72	8.68
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.49	6.46
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.16	4.13
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.83	3.80
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.39	3.36
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.24	3.21
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.08
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	3.01	2.98
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.92	2.89
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.79	2.75
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.73	2.70
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.68	2.65
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.60	2.56
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.56	2.53
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.53	2.50
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.50	2.47
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.48	2.44
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54	2.49	2.45	2.42
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.51	2.47	2.43	2.39
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.41	2.37
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.48	2.43	2.39	2.36
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34
31	5.55	4.16	3.57	3.23	3.01	2.85	2.73	2.64	2.56	2.50	2.44	2.40	2.36	2.32
32	5.53	4.15	3.56	3.22	3.00	2.84	2.71	2.62	2.54	2.48	2.43	2.38	2.34	2.31
33	5.51	4.13	3.54	3.20	2.98	2.82	2.70	2.61	2.53	2.47	2.41	2.37	2.33	2.29
34	5.50	4.12	3.53	3.19	2.97	2.81	2.69	2.59	2.52	2.45	2.40	2.35	2.31	2.28
35	5.48	4.11	3.52	3.18	2.96	2.80	2.68	2.58	2.50	2.44	2.39	2.34	2.30	2.27
36	5.47	4.09	3.50	3.17	2.94	2.78	2.66	2.57	2.49	2.43	2.37	2.33	2.29	2.25
37	5.46	4.08	3.49	3.16	2.93	2.77	2.65	2.56	2.48	2.42	2.36	2.32	2.28	2.24
38	5.45	4.07	3.48	3.15	2.92	2.76	2.64	2.55	2.47	2.41	2.35	2.31	2.27	2.23
39	5.43	4.06	3.47	3.14	2.91	2.75	2.63	2.54	2.46	2.40	2.34	2.30	2.26	2.22
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.25	2.21
45	5.38	4.01	3.42	3.09	2.86	2.70	2.58	2.49	2.41	2.35	2.29	2.25	2.21	2.17
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.26	2.22	2.18	2.14
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.13	2.09
70	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24	2.18	2.14	2.10	2.06
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.16	2.11	2.07	2.03
90	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19	2.14	2.09	2.05	2.02
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.12	2.08	2.04	2.00
110	5.16	3.82	3.24	2.90	2.68	2.53	2.40	2.31	2.23	2.17	2.11	2.07	2.02	1.99
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.10	2.05	2.01	1.98
140	5.13	3.79	3.21	2.88	2.66	2.50	2.38	2.28	2.21	2.14	2.09	2.04	2.00	1.96
160	5.12	3.78	3.20	2.87	2.65	2.49	2.37	2.27	2.19	2.13	2.07	2.03	1.99	1.95
180	5.11	3.77	3.19	2.86	2.64	2.48	2.36	2.26	2.19	2.12	2.07	2.02	1.98	1.94
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	2.06	2.01	1.97	1.93

TABLE 4.2, continued. 2.5% critical values for the F-DISTRIBUTION

	NUM (Numerator Degrees of Freedom)													
DEN	15	16	17	18	19	20	22	24	27	30	40	60	100	200
1	985	987	989	990	992	993	995	997	1000	1001	1006	1010	1013	1016
2	39.43	39.44	39.44	39.44	39.45	39.45	39.45	39.46	39.46	39.46	39.47	39.48	39.49	39.49
3	14.25	14.23	14.21	14.20	14.18	14.17	14.14	14.12	14.10	14.08	14.04	13.99	13.96	13.93
4	8.66	8.63	8.61	8.59	8.58	8.56	8.53	8.51	8.48	8.46	8.41	8.36	8.32	8.29
5	6.43	6.40	6.38	6.36	6.34	6.33	6.30	6.28	6.25	6.23	6.18	6.12	6.08	6.05
6	5.27	5.24	5.22	5.20	5.18	5.17	5.14	5.12	5.09	5.07	5.01	4.96	4.92	4.88
7	4.57	4.54	4.52	4.50	4.48	4.47	4.44	4.41	4.39	4.36	4.31	4.25	4.21	4.18
8	4.10	4.08	4.05	4.03	4.02	4.00	3.97	3.95	3.92	3.89	3.84	3.78	3.74	3.70
9	3.77	3.74	3.72	3.70	3.68	3.67	3.64	3.61	3.58	3.56	3.51	3.45	3.40	3.37
10	3.52	3.50	3.47	3.45	3.44	3.42	3.39	3.37	3.34	3.31	3.26	3.20	3.15	3.12
11	3.33	3.30	3.28	3.26	3.24	3.23	3.20	3.17	3.14	3.12	3.06	3.00	2.96	2.92
12	3.18	3.15	3.13	3.11	3.09	3.07	3.04	3.02	2.99	2.96	2.91	2.85	2.80	2.76
13	3.05	3.03	3.00	2.98	2.96	2.95	2.92	2.89	2.86	2.84	2.78	2.72	2.67	2.63
14	2.95	2.92	2.90	2.88	2.86	2.84	2.81	2.79	2.76	2.73	2.67	2.61	2.56	2.53
15	2.86	2.84	2.81	2.79	2.77	2.76	2.73	2.70	2.67	2.64	2.59	2.52	2.47	2.44
16	2.79	2.76	2.74	2.72	2.70	2.68	2.65	2.63	2.59	2.57	2.51	2.45	2.40	2.36
17	2.72	2.70	2.67	2.65	2.63	2.62	2.59	2.56	2.53	2.50	2.44	2.38	2.33	2.29
18	2.67	2.64	2.62	2.60	2.58	2.56	2.53	2.50	2.47	2.44	2.38	2.32	2.27	2.23
19	2.62	2.59	2.57	2.55	2.53	2.51	2.48	2.45	2.42	2.39	2.33	2.27	2.22	2.18
20	2.57	2.55	2.52	2.50	2.48	2.46	2.43	2.41	2.38	2.35	2.29	2.22	2.17	2.13
21	2.53	2.51	2.48	2.46	2.44	2.42	2.39	2.37	2.33	2.31	2.25	2.18	2.13	2.09
22	2.50	2.47	2.45	2.43	2.41	2.39	2.36	2.33	2.30	2.27	2.21	2.14	2.09	2.05
23	2.47	2.44	2.42	2.39	2.37	2.36	2.33	2.30	2.27	2.24	2.18	2.11	2.06	2.01
24	2.44	2.41	2.39	2.36	2.35	2.33	2.30	2.27	2.24	2.21	2.15	2.08	2.02	1.98
25	2.41	2.38	2.36	2.34	2.32	2.30	2.27	2.24	2.21	2.18	2.12	2.05	2.00	1.95
26	2.39	2.36	2.34	2.31	2.29	2.28	2.24	2.22	2.18	2.16	2.09	2.03	1.97	1.92
27	2.36	2.34	2.31	2.29	2.27	2.25	2.22	2.19	2.16	2.13	2.07	2.00	1.94	1.90
28	2.34	2.32	2.29	2.27	2.25	2.23	2.20	2.17	2.14	2.11	2.05	1.98	1.92	1.88
29	2.32	2.30	2.27	2.25	2.23	2.21	2.18	2.15	2.12	2.09	2.03	1.96	1.90	1.86
30	2.31	2.28	2.26	2.23	2.21	2.20	2.16	2.14	2.10	2.07	2.01	1.94	1.88	1.84
31	2.29	2.26	2.24	2.22	2.20	2.18	2.15	2.12	2.08	2.06	1.99	1.92	1.86	1.82
32	2.28	2.25	2.22	2.20	2.18	2.16	2.13	2.10	2.07	2.04	1.98	1.91	1.85	1.80
33	2.26	2.23	2.21	2.19	2.17	2.15	2.12	2.09	2.05	2.03	1.96	1.89	1.83	1.78
34	2.25	2.22	2.20	2.17	2.15	2.13	2.10	2.07	2.04	2.01	1.95	1.88	1.82	1.77
35	2.23	2.21	2.18	2.16	2.14	2.12	2.09	2.06	2.03	2.00	1.93	1.86	1.80	1.75
36	2.22	2.20	2.17	2.15	2.13	2.11	2.08	2.05	2.01	1.99	1.92	1.85	1.79	1.74
37	2.21	2.18	2.16	2.14	2.12	2.10	2.07	2.04	2.00	1.97	1.91	1.84	1.77	1.73
38	2.20	2.17	2.15	2.13	2.11	2.09	2.05	2.03	1.99	1.96	1.90	1.82	1.76	1.71
39	2.19	2.16	2.14	2.12	2.10	2.08	2.04	2.02	1.98	1.95	1.89	1.81	1.75	1.70
40	2.18	2.15	2.13	2.11	2.09	2.07	2.03	2.01	1.97	1.94	1.88	1.80	1.74	1.69
45	2.14	2.11	2.09	2.07	2.04	2.03	1.99	1.96	1.93	1.90	1.83	1.76	1.69	1.64
50	2.11	2.08	2.06	2.03	2.01	1.99	1.96	1.93	1.90	1.87	1.80	1.72	1.66	1.60
60	2.06	2.03	2.01	1.98	1.96	1.94	1.91	1.88	1.85	1.82	1.74	1.67	1.60	1.54
70	2.03	2.00	1.97	1.95	1.93	1.91	1.88	1.85	1.81	1.78	1.71	1.63	1.56	1.50
80	2.00	1.97	1.95	1.92	1.90	1.88	1.85	1.82	1.78	1.75	1.68	1.60	1.53	1.47
90	1.98	1.95	1.93	1.91	1.88	1.86	1.83	1.80	1.76	1.73	1.66	1.58	1.50	1.44
100	1.97	1.94	1.91	1.89	1.87	1.85	1.81	1.78	1.75	1.71	1.64	1.56	1.48	1.42
110	1.96	1.93	1.90	1.88	1.86	1.84	1.80	1.77	1.73	1.70	1.63	1.54	1.47	1.40
120	1.94	1.92	1.89	1.87	1.84	1.82	1.79	1.76	1.72	1.69	1.61	1.53	1.45	1.39
140	1.93	1.90	1.87	1.85	1.83	1.81	1.77	1.74	1.70	1.67	1.60	1.51	1.43	1.36
160	1.92	1.89	1.86	1.84	1.82	1.80	1.76	1.73	1.69	1.66	1.58	1.50	1.42	1.35
180	1.91	1.88	1.85	1.83	1.81	1.79	1.75	1.72	1.68	1.65	1.57	1.48	1.40	1.33
200	1.90	1.87	1.84	1.82	1.80	1.78	1.74	1.71	1.67	1.64	1.56	1.47	1.39	1.32

TABLE 4.3. 1% critical values for the F-DISTRIBUTION, i.e. the value of $F^{(0.01)}_{\text{NUM,DEN}}$ where NUM and DEN are the numerator and denominator degrees of freedom respectively

	NUM (Numerator Degrees of Freedom)													
DEN	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6126	6143
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.41	99.42	99.42	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.98	26.92
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74
31	7.53	5.36	4.48	3.99	3.67	3.45	3.28	3.15	3.04	2.96	2.88	2.82	2.77	2.72
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.86	2.80	2.74	2.70
33	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91	2.84	2.78	2.72	2.68
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.82	2.76	2.70	2.66
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.80	2.74	2.69	2.64
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86	2.79	2.72	2.67	2.62
37	7.37	5.23	4.36	3.87	3.56	3.33	3.17	3.04	2.93	2.84	2.77	2.71	2.65	2.61
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.75	2.69	2.64	2.59
39	7.33	5.19	4.33	3.84	3.53	3.30	3.14	3.01	2.90	2.81	2.74	2.68	2.62	2.58
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.61	2.56
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.67	2.61	2.55	2.51
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.63	2.56	2.51	2.46
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.44	2.39
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.51	2.45	2.40	2.35
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.48	2.42	2.36	2.31
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.45	2.39	2.33	2.29
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.43	2.37	2.31	2.27
110	6.87	4.80	3.96	3.49	3.19	2.97	2.81	2.68	2.57	2.49	2.41	2.35	2.30	2.25
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.28	2.23
140	6.82	4.76	3.92	3.46	3.15	2.93	2.77	2.64	2.54	2.45	2.38	2.31	2.26	2.21
160	6.80	4.74	3.91	3.44	3.13	2.92	2.75	2.62	2.52	2.43	2.36	2.30	2.24	2.20
180	6.78	4.73	3.89	3.43	3.12	2.90	2.74	2.61	2.51	2.42	2.35	2.28	2.23	2.18
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27	2.22	2.17

TABLE 4.3, continued. 1% critical values for the F-DISTRIBUTION

	NUM (Numerator Degrees of Freedom)													
DEN	15	16	17	18	19	20	22	24	27	30	40	60	100	200
1	6157	6170	6181	6191	6201	6209	6223	6234	6249	6260	6286	6313	6334	6350
2	99.43	99.44	99.44	99.44	99.45	99.45	99.46	99.46	99.46	99.47	99.48	99.48	99.49	99.49
3	26.87	26.83	26.79	26.75	26.72	26.69	26.64	26.60	26.55	26.50	26.41	26.32	26.24	26.18
4	14.20	14.15	14.11	14.08	14.05	14.02	13.97	13.93	13.88	13.84	13.75	13.65	13.58	13.52
5	9.72	9.68	9.64	9.61	9.58	9.55	9.51	9.47	9.42	9.38	9.29	9.20	9.13	9.08
6	7.56	7.52	7.48	7.45	7.42	7.40	7.35	7.31	7.27	7.23	7.14	7.06	6.99	6.93
7	6.31	6.28	6.24	6.21	6.18	6.16	6.11	6.07	6.03	5.99	5.91	5.82	5.75	5.70
8	5.52	5.48	5.44	5.41	5.38	5.36	5.32	5.28	5.23	5.20	5.12	5.03	4.96	4.91
9	4.96	4.92	4.89	4.86	4.83	4.81	4.77	4.73	4.68	4.65	4.57	4.48	4.41	4.36
10	4.56	4.52	4.49	4.46	4.43	4.41	4.36	4.33	4.28	4.25	4.17	4.08	4.01	3.96
11	4.25	4.21	4.18	4.15	4.12	4.10	4.06	4.02	3.98	3.94	3.86	3.78	3.71	3.66
12	4.01	3.97	3.94	3.91	3.88	3.86	3.82	3.78	3.74	3.70	3.62	3.54	3.47	3.41
13	3.82	3.78	3.75	3.72	3.69	3.66	3.62	3.59	3.54	3.51	3.43	3.34	3.27	3.22
14	3.66	3.62	3.59	3.56	3.53	3.51	3.46	3.43	3.38	3.35	3.27	3.18	3.11	3.06
15	3.52	3.49	3.45	3.42	3.40	3.37	3.33	3.29	3.25	3.21	3.13	3.05	2.98	2.92
16	3.41	3.37	3.34	3.31	3.28	3.26	3.22	3.18	3.14	3.10	3.02	2.93	2.86	2.81
17	3.31	3.27	3.24	3.21	3.19	3.16	3.12	3.08	3.04	3.00	2.92	2.83	2.76	2.71
18	3.23	3.19	3.16	3.13	3.10	3.08	3.03	3.00	2.95	2.92	2.84	2.75	2.68	2.62
19	3.15	3.12	3.08	3.05	3.03	3.00	2.96	2.92	2.88	2.84	2.76	2.67	2.60	2.55
20	3.09	3.05	3.02	2.99	2.96	2.94	2.90	2.86	2.81	2.78	2.69	2.61	2.54	2.48
21	3.03	2.99	2.96	2.93	2.90	2.88	2.84	2.80	2.76	2.72	2.64	2.55	2.48	2.42
22	2.98	2.94	2.91	2.88	2.85	2.83	2.78	2.75	2.70	2.67	2.58	2.50	2.42	2.36
23	2.93	2.89	2.86	2.83	2.80	2.78	2.74	2.70	2.66	2.62	2.54	2.45	2.37	2.32
24	2.89	2.85	2.82	2.79	2.76	2.74	2.70	2.66	2.61	2.58	2.49	2.40	2.33	2.27
25	2.85	2.81	2.78	2.75	2.72	2.70	2.66	2.62	2.58	2.54	2.45	2.36	2.29	2.23
26	2.81	2.78	2.75	2.72	2.69	2.66	2.62	2.58	2.54	2.50	2.42	2.33	2.25	2.19
27	2.78	2.75	2.71	2.68	2.66	2.63	2.59	2.55	2.51	2.47	2.38	2.29	2.22	2.16
28	2.75	2.72	2.68	2.65	2.63	2.60	2.56	2.52	2.48	2.44	2.35	2.26	2.19	2.13
29	2.73	2.69	2.66	2.63	2.60	2.57	2.53	2.49	2.45	2.41	2.33	2.23	2.16	2.10
30	2.70	2.66	2.63	2.60	2.57	2.55	2.51	2.47	2.42	2.39	2.30	2.21	2.13	2.07
31	2.68	2.64	2.61	2.58	2.55	2.52	2.48	2.45	2.40	2.36	2.27	2.18	2.11	2.04
32	2.65	2.62	2.58	2.55	2.53	2.50	2.46	2.42	2.38	2.34	2.25	2.16	2.08	2.02
33	2.63	2.60	2.56	2.53	2.51	2.48	2.44	2.40	2.36	2.32	2.23	2.14	2.06	2.00
34	2.61	2.58	2.54	2.51	2.49	2.46	2.42	2.38	2.34	2.30	2.21	2.12	2.04	1.98
35	2.60	2.56	2.53	2.50	2.47	2.44	2.40	2.36	2.32	2.28	2.19	2.10	2.02	1.96
36	2.58	2.54	2.51	2.48	2.45	2.43	2.38	2.35	2.30	2.26	2.18	2.08	2.00	1.94
37	2.56	2.53	2.49	2.46	2.44	2.41	2.37	2.33	2.28	2.25	2.16	2.06	1.98	1.92
38	2.55	2.51	2.48	2.45	2.42	2.40	2.35	2.32	2.27	2.23	2.14	2.05	1.97	1.90
39	2.54	2.50	2.46	2.43	2.41	2.38	2.34	2.30	2.26	2.22	2.13	2.03	1.95	1.89
40	2.52	2.48	2.45	2.42	2.39	2.37	2.33	2.29	2.24	2.20	2.11	2.02	1.94	1.87
45	2.46	2.43	2.39	2.36	2.34	2.31	2.27	2.23	2.18	2.14	2.05	1.96	1.88	1.81
50	2.42	2.38	2.35	2.32	2.29	2.27	2.22	2.18	2.14	2.10	2.01	1.91	1.82	1.76
60	2.35	2.31	2.28	2.25	2.22	2.20	2.15	2.12	2.07	2.03	1.94	1.84	1.75	1.68
70	2.31	2.27	2.23	2.20	2.18	2.15	2.11	2.07	2.02	1.98	1.89	1.78	1.70	1.62
80	2.27	2.23	2.20	2.17	2.14	2.12	2.07	2.03	1.98	1.94	1.85	1.75	1.65	1.58
90	2.24	2.21	2.17	2.14	2.11	2.09	2.04	2.00	1.96	1.92	1.82	1.72	1.62	1.55
100	2.22	2.19	2.15	2.12	2.09	2.07	2.02	1.98	1.93	1.89	1.80	1.69	1.60	1.52
110	2.21	2.17	2.13	2.10	2.07	2.05	2.00	1.96	1.92	1.88	1.78	1.67	1.58	1.50
120	2.19	2.15	2.12	2.09	2.06	2.03	1.99	1.95	1.90	1.86	1.76	1.66	1.56	1.48
140	2.17	2.13	2.10	2.07	2.04	2.01	1.97	1.93	1.88	1.84	1.74	1.63	1.53	1.45
160	2.15	2.11	2.08	2.05	2.02	1.99	1.95	1.91	1.86	1.82	1.72	1.61	1.51	1.42
180	2.14	2.10	2.07	2.04	2.01	1.98	1.94	1.90	1.85	1.81	1.71	1.60	1.49	1.41
200	2.13	2.09	2.06	2.03	2.00	1.97	1.93	1.89	1.84	1.79	1.69	1.58	1.48	1.39

TABLE 4.4. 0.5% critical values for the F-DISTRIBUTION, i.e. the value of $F^{(0.005)}_{\text{NUM,DEN}}$ where NUM and DEN are the numerator and denominator degrees of freedom respectively

	NUM (Numerator Degrees of Freedom)													
DEN	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	16212	19997	21614	22501	23056	23440	23715	23924	24091	24222	24334	24427	24505	24572
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4
3	55.55	49.80	47.47	46.20	45.39	44.84	44.43	44.13	43.88	43.68	43.52	43.39	43.27	43.17
4	31.33	26.28	24.26	23.15	22.46	21.98	21.62	21.35	21.14	20.97	20.82	20.70	20.60	20.51
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.49	13.38	13.29	13.21
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.13	10.03	9.95	9.88
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.27	8.18	8.10	8.03
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.10	7.01	6.94	6.87
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.31	6.23	6.15	6.09
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.75	5.66	5.59	5.53
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.32	5.24	5.16	5.10
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.99	4.91	4.84	4.77
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.72	4.64	4.57	4.51
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.51	4.43	4.36	4.30
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.33	4.25	4.18	4.12
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.18	4.10	4.03	3.97
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	4.05	3.97	3.90	3.84
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.94	3.86	3.79	3.73
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.84	3.76	3.70	3.64
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.76	3.68	3.61	3.55
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.68	3.60	3.54	3.48
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.61	3.54	3.47	3.41
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.55	3.47	3.41	3.35
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.50	3.42	3.35	3.30
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.45	3.37	3.30	3.25
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.40	3.33	3.26	3.20
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.36	3.28	3.22	3.16
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.32	3.25	3.18	3.12
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.29	3.21	3.15	3.09
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.25	3.18	3.11	3.06
31	9.13	6.32	5.20	4.59	4.20	3.92	3.71	3.55	3.42	3.31	3.22	3.15	3.08	3.03
32	9.09	6.28	5.17	4.56	4.17	3.89	3.68	3.52	3.39	3.29	3.20	3.12	3.06	3.00
33	9.05	6.25	5.14	4.53	4.14	3.86	3.66	3.49	3.37	3.26	3.17	3.09	3.03	2.97
34	9.01	6.22	5.11	4.50	4.11	3.84	3.63	3.47	3.34	3.24	3.15	3.07	3.01	2.95
35	8.98	6.19	5.09	4.48	4.09	3.81	3.61	3.45	3.32	3.21	3.12	3.05	2.98	2.93
36	8.94	6.16	5.06	4.46	4.06	3.79	3.58	3.42	3.30	3.19	3.10	3.03	2.96	2.90
37	8.91	6.13	5.04	4.43	4.04	3.77	3.56	3.40	3.28	3.17	3.08	3.01	2.94	2.88
38	8.88	6.11	5.02	4.41	4.02	3.75	3.54	3.39	3.26	3.15	3.06	2.99	2.92	2.87
39	8.85	6.09	5.00	4.39	4.00	3.73	3.53	3.37	3.24	3.13	3.05	2.97	2.90	2.85
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	3.03	2.95	2.89	2.83
45	8.71	5.97	4.89	4.29	3.91	3.64	3.43	3.28	3.15	3.04	2.96	2.88	2.82	2.76
50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.90	2.82	2.76	2.70
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.82	2.74	2.68	2.62
70	8.40	5.72	4.66	4.08	3.70	3.43	3.23	3.08	2.95	2.85	2.76	2.68	2.62	2.56
80	8.33	5.67	4.61	4.03	3.65	3.39	3.19	3.03	2.91	2.80	2.72	2.64	2.58	2.52
90	8.28	5.62	4.57	3.99	3.62	3.35	3.15	3.00	2.87	2.77	2.68	2.61	2.54	2.49
100	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	2.66	2.58	2.52	2.46
110	8.21	5.56	4.52	3.94	3.57	3.30	3.11	2.95	2.83	2.72	2.64	2.56	2.50	2.44
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.62	2.54	2.48	2.42
140	8.13	5.50	4.47	3.89	3.52	3.26	3.06	2.91	2.78	2.68	2.59	2.52	2.45	2.40
160	8.10	5.48	4.44	3.87	3.50	3.24	3.04	2.88	2.76	2.66	2.57	2.50	2.43	2.38
180	8.08	5.46	4.42	3.85	3.48	3.22	3.02	2.87	2.74	2.64	2.56	2.48	2.42	2.36
200	8.06	5.44	4.41	3.84	3.47	3.21	3.01	2.86	2.73	2.63	2.54	2.47	2.40	2.35

TABLE 4.4, continued. 0.5% critical values for the F-DISTRIBUTION (continued)

	NUM (Numerator Degrees of Freedom)													
DEN	15	16	17	18	19	20	22	24	27	30	40	60	100	200
1	24632	24684	24728	24766	24803	24837	24892	24937	24997	25041	25146	25254	25339	25399
2	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5
3	43.08	43.01	42.94	42.88	42.83	42.78	42.69	42.62	42.54	42.47	42.31	42.15	42.02	41.92
4	20.44	20.37	20.31	20.26	20.21	20.17	20.09	20.03	19.95	19.89	19.75	19.61	19.50	19.41
5	13.15	13.09	13.03	12.98	12.94	12.90	12.84	12.78	12.71	12.66	12.53	12.40	12.30	12.22
6	9.81	9.76	9.71	9.66	9.62	9.59	9.53	9.47	9.41	9.36	9.24	9.12	9.03	8.95
7	7.97	7.91	7.87	7.83	7.79	7.75	7.69	7.64	7.58	7.53	7.42	7.31	7.22	7.15
8	6.81	6.76	6.72	6.68	6.64	6.61	6.55	6.50	6.44	6.40	6.29	6.18	6.09	6.02
9	6.03	5.98	5.94	5.90	5.86	5.83	5.78	5.73	5.67	5.62	5.52	5.41	5.32	5.26
10	5.47	5.42	5.38	5.34	5.31	5.27	5.22	5.17	5.12	5.07	4.97	4.86	4.77	4.71
11	5.05	5.00	4.96	4.92	4.89	4.86	4.80	4.76	4.70	4.65	4.55	4.45	4.36	4.29
12	4.72	4.67	4.63	4.59	4.56	4.53	4.48	4.43	4.38	4.33	4.23	4.12	4.04	3.97
13	4.46	4.41	4.37	4.33	4.30	4.27	4.22	4.17	4.12	4.07	3.97	3.87	3.78	3.71
14	4.25	4.20	4.16	4.12	4.09	4.06	4.01	3.96	3.91	3.86	3.76	3.66	3.57	3.50
15	4.07	4.02	3.98	3.95	3.91	3.88	3.83	3.79	3.73	3.69	3.59	3.48	3.39	3.33
16	3.92	3.87	3.83	3.80	3.76	3.73	3.68	3.64	3.58	3.54	3.44	3.33	3.25	3.18
17	3.79	3.75	3.71	3.67	3.64	3.61	3.56	3.51	3.46	3.41	3.31	3.21	3.12	3.05
18	3.68	3.64	3.60	3.56	3.53	3.50	3.45	3.40	3.35	3.30	3.20	3.10	3.01	2.94
19	3.59	3.54	3.50	3.46	3.43	3.40	3.35	3.31	3.25	3.21	3.11	3.00	2.91	2.85
20	3.50	3.46	3.42	3.38	3.35	3.32	3.27	3.22	3.17	3.12	3.02	2.92	2.83	2.76
21	3.43	3.38	3.34	3.31	3.27	3.24	3.19	3.15	3.09	3.05	2.95	2.84	2.75	2.68
22	3.36	3.31	3.27	3.24	3.21	3.18	3.12	3.08	3.03	2.98	2.88	2.77	2.69	2.62
23	3.30	3.25	3.21	3.18	3.15	3.12	3.06	3.02	2.97	2.92	2.82	2.71	2.62	2.56
24	3.25	3.20	3.16	3.12	3.09	3.06	3.01	2.97	2.91	2.87	2.77	2.66	2.57	2.50
25	3.20	3.15	3.11	3.08	3.04	3.01	2.96	2.92	2.86	2.82	2.72	2.61	2.52	2.45
26	3.15	3.11	3.07	3.03	3.00	2.97	2.92	2.87	2.82	2.77	2.67	2.56	2.47	2.40
27	3.11	3.07	3.03	2.99	2.96	2.93	2.88	2.83	2.78	2.73	2.63	2.52	2.43	2.36
28	3.07	3.03	2.99	2.95	2.92	2.89	2.84	2.79	2.74	2.69	2.59	2.48	2.39	2.32
29	3.04	2.99	2.95	2.92	2.88	2.86	2.80	2.76	2.70	2.66	2.56	2.45	2.36	2.29
30	3.01	2.96	2.92	2.89	2.85	2.82	2.77	2.73	2.67	2.63	2.52	2.42	2.32	2.25
31	2.98	2.93	2.89	2.86	2.82	2.79	2.74	2.70	2.64	2.60	2.49	2.38	2.29	2.22
32	2.95	2.90	2.86	2.83	2.80	2.77	2.71	2.67	2.61	2.57	2.47	2.36	2.26	2.19
33	2.92	2.88	2.84	2.80	2.77	2.74	2.69	2.64	2.59	2.54	2.44	2.33	2.24	2.16
34	2.90	2.85	2.81	2.78	2.75	2.72	2.66	2.62	2.56	2.52	2.42	2.30	2.21	2.14
35	2.88	2.83	2.79	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.39	2.28	2.19	2.11
36	2.85	2.81	2.77	2.73	2.70	2.67	2.62	2.58	2.52	2.48	2.37	2.26	2.17	2.09
37	2.83	2.79	2.75	2.71	2.68	2.65	2.60	2.56	2.50	2.46	2.35	2.24	2.14	2.07
38	2.82	2.77	2.73	2.70	2.66	2.63	2.58	2.54	2.48	2.44	2.33	2.22	2.12	2.05
39	2.80	2.75	2.71	2.68	2.64	2.62	2.56	2.52	2.46	2.42	2.31	2.20	2.11	2.03
40	2.78	2.74	2.70	2.66	2.63	2.60	2.55	2.50	2.45	2.40	2.30	2.18	2.09	2.01
45	2.71	2.66	2.62	2.59	2.56	2.53	2.47	2.43	2.37	2.33	2.22	2.11	2.01	1.93
50	2.65	2.61	2.57	2.53	2.50	2.47	2.42	2.37	2.32	2.27	2.16	2.05	1.95	1.87
60	2.57	2.53	2.49	2.45	2.42	2.39	2.33	2.29	2.23	2.19	2.08	1.96	1.86	1.78
70	2.51	2.47	2.43	2.39	2.36	2.33	2.28	2.23	2.17	2.13	2.02	1.90	1.80	1.71
80	2.47	2.43	2.39	2.35	2.32	2.29	2.23	2.19	2.13	2.08	1.97	1.85	1.75	1.66
90	2.44	2.39	2.35	2.32	2.28	2.25	2.20	2.15	2.10	2.05	1.94	1.82	1.71	1.62
100	2.41	2.37	2.33	2.29	2.26	2.23	2.17	2.13	2.07	2.02	1.91	1.79	1.68	1.59
110	2.39	2.35	2.31	2.27	2.24	2.21	2.15	2.11	2.05	2.00	1.89	1.77	1.66	1.56
120	2.37	2.33	2.29	2.25	2.22	2.19	2.13	2.09	2.03	1.98	1.87	1.75	1.64	1.54
140	2.35	2.30	2.26	2.22	2.19	2.16	2.11	2.06	2.00	1.96	1.84	1.72	1.60	1.51
160	2.33	2.28	2.24	2.20	2.17	2.14	2.09	2.04	1.98	1.93	1.82	1.69	1.58	1.48
180	2.31	2.26	2.22	2.19	2.15	2.12	2.07	2.02	1.97	1.92	1.80	1.68	1.56	1.46
200	2.30	2.25	2.21	2.18	2.14	2.11	2.06	2.01	1.95	1.91	1.79	1.66	1.54	1.44

TABLE 5. CORRELATION COEFFICIENT. Critical values of the correlation coefficients for one-sided tests of the null hypothesis $H_0 : \rho = 0$ (where degrees of freedom = sample size - 2)

Deg. of	Probability Level (P)										
Freedom	0.4	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.3090	0.5878	0.8090	0.9511	0.9877	0.9969	0.9995	0.9999	1.0000	1.0000	1.0000
2	0.2000	0.4000	0.6000	0.8000	0.9000	0.9500	0.9800	0.9900	0.9950	0.9980	0.9990
3	0.1577	0.3197	0.4919	0.6870	0.8054	0.8783	0.9343	0.9587	0.9740	0.9859	0.9911
4	0.1341	0.2735	0.4257	0.6084	0.7293	0.8114	0.8822	0.9172	0.9417	0.9633	0.9741
5	0.1186	0.2427	0.3803	0.5509	0.6694	0.7545	0.8329	0.8745	0.9056	0.9350	0.9509
6	0.1075	0.2204	0.3468	0.5067	0.6215	0.7067	0.7887	0.8343	0.8697	0.9049	0.9249
7	0.0990	0.2032	0.3208	0.4716	0.5822	0.6664	0.7498	0.7977	0.8359	0.8751	0.8983
8	0.0922	0.1895	0.2998	0.4428	0.5494	0.6319	0.7155	0.7646	0.8046	0.8467	0.8721
9	0.0867	0.1783	0.2825	0.4187	0.5214	0.6021	0.6851	0.7348	0.7759	0.8199	0.8470
10	0.0820	0.1688	0.2678	0.3981	0.4973	0.5760	0.6581	0.7079	0.7496	0.7950	0.8233
11	0.0780	0.1607	0.2552	0.3802	0.4762	0.5529	0.6339	0.6835	0.7255	0.7717	0.8010
12	0.0746	0.1536	0.2443	0.3646	0.4575	0.5324	0.6120	0.6614	0.7034	0.7501	0.7800
13	0.0715	0.1474	0.2346	0.3507	0.4409	0.5140	0.5923	0.6411	0.6831	0.7301	0.7604
14	0.0688	0.1419	0.2260	0.3383	0.4259	0.4973	0.5742	0.6226	0.6643	0.7114	0.7419
15	0.0664	0.1370	0.2183	0.3271	0.4124	0.4821	0.5577	0.6055	0.6470	0.6940	0.7247
16	0.0643	0.1326	0.2113	0.3170	0.4000	0.4683	0.5425	0.5897	0.6308	0.6777	0.7084
17	0.0623	0.1285	0.2049	0.3077	0.3887	0.4555	0.5285	0.5751	0.6158	0.6624	0.6932
18	0.0605	0.1248	0.1991	0.2992	0.3783	0.4438	0.5155	0.5614	0.6018	0.6481	0.6788
19	0.0588	0.1214	0.1938	0.2914	0.3687	0.4329	0.5034	0.5487	0.5886	0.6346	0.6652
20	0.0573	0.1183	0.1888	0.2841	0.3598	0.4227	0.4921	0.5368	0.5763	0.6219	0.6524
21	0.0559	0.1154	0.1843	0.2774	0.3515	0.4132	0.4815	0.5256	0.5647	0.6099	0.6402
22	0.0546	0.1127	0.1800	0.2711	0.3438	0.4044	0.4716	0.5151	0.5537	0.5986	0.6287
23	0.0534	0.1102	0.1760	0.2653	0.3365	0.3961	0.4622	0.5052	0.5434	0.5879	0.6178
24	0.0522	0.1078	0.1723	0.2598	0.3297	0.3882	0.4534	0.4958	0.5336	0.5776	0.6074
25	0.0511	0.1056	0.1688	0.2546	0.3233	0.3809	0.4451	0.4869	0.5243	0.5679	0.5974
26	0.0501	0.1036	0.1655	0.2497	0.3172	0.3739	0.4372	0.4785	0.5154	0.5587	0.5880
27	0.0492	0.1016	0.1624	0.2451	0.3115	0.3673	0.4297	0.4705	0.5070	0.5499	0.5789
28	0.0483	0.0997	0.1594	0.2407	0.3061	0.3610	0.4226	0.4629	0.4990	0.5415	0.5703
29	0.0474	0.0980	0.1567	0.2366	0.3009	0.3550	0.4158	0.4556	0.4914	0.5334	0.5621
30	0.0466	0.0963	0.1540	0.2327	0.2960	0.3494	0.4093	0.4487	0.4840	0.5257	0.5541
31	0.0458	0.0947	0.1515	0.2289	0.2913	0.3440	0.4032	0.4421	0.4770	0.5184	0.5465
32	0.0451	0.0932	0.1491	0.2254	0.2869	0.3388	0.3972	0.4357	0.4703	0.5113	0.5392
33	0.0444	0.0918	0.1468	0.2220	0.2826	0.3338	0.3916	0.4296	0.4639	0.5045	0.5322
34	0.0437	0.0904	0.1446	0.2187	0.2785	0.3291	0.3862	0.4238	0.4577	0.4979	0.5254
35	0.0431	0.0891	0.1425	0.2156	0.2746	0.3246	0.3810	0.4182	0.4518	0.4916	0.5189
36	0.0425	0.0878	0.1405	0.2126	0.2709	0.3202	0.3760	0.4128	0.4461	0.4856	0.5126
37	0.0419	0.0866	0.1386	0.2097	0.2673	0.3160	0.3712	0.4076	0.4406	0.4797	0.5066
38	0.0414	0.0855	0.1368	0.2070	0.2638	0.3120	0.3665	0.4026	0.4353	0.4741	0.5007
39	0.0408	0.0844	0.1350	0.2043	0.2605	0.3081	0.3621	0.3978	0.4301	0.4686	0.4950
40	0.0403	0.0833	0.1333	0.2018	0.2573	0.3044	0.3578	0.3932	0.4252	0.4634	0.4896
45	0.0380	0.0785	0.1257	0.1903	0.2429	0.2876	0.3384	0.3721	0.4028	0.4394	0.4647
50	0.0360	0.0744	0.1192	0.1806	0.2306	0.2732	0.3218	0.3542	0.3836	0.4188	0.4432
60	0.0328	0.0679	0.1088	0.1650	0.2108	0.2500	0.2948	0.3248	0.3522	0.3850	0.4079
70	0.0304	0.0628	0.1007	0.1528	0.1954	0.2319	0.2737	0.3017	0.3274	0.3583	0.3798
80	0.0284	0.0588	0.0942	0.1430	0.1829	0.2172	0.2565	0.2830	0.3072	0.3364	0.3568
90	0.0268	0.0554	0.0888	0.1348	0.1726	0.2050	0.2422	0.2673	0.2903	0.3181	0.3375
100	0.0254	0.0525	0.0842	0.1279	0.1638	0.1946	0.2301	0.2540	0.2759	0.3025	0.3211
110	0.0242	0.0501	0.0803	0.1220	0.1562	0.1857	0.2196	0.2425	0.2635	0.2890	0.3068
120	0.0232	0.0479	0.0769	0.1168	0.1496	0.1779	0.2104	0.2324	0.2526	0.2771	0.2943
140	0.0214	0.0444	0.0712	0.1082	0.1386	0.1648	0.1951	0.2155	0.2343	0.2572	0.2733
160	0.0201	0.0415	0.0666	0.1012	0.1297	0.1543	0.1826	0.2019	0.2195	0.2411	0.2562
180	0.0189	0.0391	0.0628	0.0954	0.1223	0.1455	0.1723	0.1905	0.2072	0.2276	0.2419
200	0.0179	0.0371	0.0595	0.0905	0.1161	0.1381	0.1636	0.1809	0.1968	0.2162	0.2298

TABLE 6. RANDOM NUMBERS

1842	7248	4572	2884	7994	8904	5441	3710	5437	9180	9723	6911	8996	2226	7527
8411	4445	2115	9302	8052	8852	4543	9079	6915	6536	8237	4318	8966	4303	0906
3329	8483	3260	1151	4112	3867	2605	9180	5773	7800	5221	9812	7016	1899	6825
8342	5339	2103	0546	5312	8447	8218	1429	3901	1889	4345	3038	5025	9947	2428
9577	8769	3404	6018	1481	6641	6127	1974	3478	7315	3645	2797	5889	4978	5694
9260	1773	2373	3943	9952	8827	5374	7016	6559	4518	3037	4093	2778	4245	6863
6854	2385	3769	1505	0099	6685	9859	4800	6095	2221	2675	8894	7105	9735	6481
8061	8246	8313	5338	7002	2006	3462	8117	2844	3971	7014	4159	6087	3157	1573
2275	4043	4137	7436	0080	8413	7718	1638	9699	1970	7233	6456	5343	4214	0817
2453	4814	9120	0116	3097	3390	5672	6287	0050	4622	3817	7440	4984	0997	9950
7995	3188	3782	1284	6682	2307	9949	0605	4959	7336	9878	9360	7677	8020	3299
3329	2605	6814	2584	6088	7177	9473	9354	6024	5855	6495	9472	2205	1426	3400
3252	8779	1573	7410	5579	5135	7530	3195	0501	5631	3160	7705	5810	9037	7870
3308	4331	6087	8428	9502	8536	3902	3728	7258	3659	1172	4274	9787	3963	5425
3601	4656	9303	0513	2630	1838	3909	1890	5102	9500	5080	4738	1819	5066	3187
0098	5981	7971	1254	6413	5339	8307	7689	3546	4765	3620	6393	7912	8686	5065
8778	1995	4396	2959	3856	2016	6395	7847	2998	9491	3545	9076	5068	8692	2283
9149	1478	8770	3335	2938	7809	9490	9559	5777	1619	7598	8395	9259	3674	8593
8457	0337	8879	9435	0645	9860	4881	6051	2708	8592	6646	2767	3833	2632	2945
8495	2970	2976	1370	0204	7878	0410	1245	8268	4209	2157	8786	2832	4930	1139
9095	7574	0253	0485	5710	0443	1373	5404	2240	5112	2216	3844	7258	8395	8609
4716	0519	8983	1821	8257	5865	1486	3161	2184	6693	1157	2886	9580	0847	0018
0617	4677	1549	0465	3564	5332	3100	8316	1760	8025	5405	9183	9612	1306	7833
5138	9398	6245	3911	6015	1197	1071	1173	1272	4543	0762	7544	6000	7456	1904
8031	7280	1808	4177	1206	0794	3982	0841	6275	6927	5398	4793	8390	7872	8592
4685	8378	4340	5990	2445	9772	2856	4678	0456	2882	9745	9610	3496	2231	0204
1665	7644	3124	4258	3538	3387	8440	9950	5852	7082	7773	6085	7727	0785	9639
4963	7079	1734	4744	7366	1140	0655	7800	6761	6722	1563	9289	6811	7166	5880
9853	3148	9325	5830	0608	2755	1234	1817	4631	9784	7150	2934	6797	7577	4499
2125	6406	4992	6884	2229	4184	8952	2055	4474	3272	7140	9858	1481	0486	0561
7216	9245	6514	7537	0072	0474	6964	1369	6495	3049	1766	8207	6281	1673	3482
1250	5258	3548	8209	3016	0327	6596	6928	6625	8323	0292	5613	8384	7724	5840
0273	1903	0740	9576	4333	0474	5134	7357	6126	4217	0093	4184	7233	5458	2687
5217	1289	5780	9525	0266	4078	2553	0840	4898	2647	0413	9460	5841	5663	9496
5941	9609	5224	0675	3021	9678	6256	0463	0280	7587	8163	0920	0262	2477	4040
3318	6272	9021	1377	3149	9188	3245	2329	7923	2314	7092	1620	4299	3635	4807
5328	1193	0257	7574	9949	5199	1733	7618	9301	7056	9198	7181	4557	0355	2796
9697	1200	6706	8268	9867	9767	7285	1439	4446	0862	1676	0673	8778	2847	8831
6125	1531	1358	6706	1945	9360	4487	0223	8121	2470	9005	5668	6895	6050	1809
1167	4222	4936	4307	6683	4544	2686	3813	2458	8805	7690	8647	2056	4648	8243
8131	1481	6890	4665	1319	3973	7018	5907	3183	7687	8607	4695	3019	1039	4847
3646	5913	7150	1635	9083	9251	7835	6734	9149	4086	3521	5614	5243	6171	6925
2834	7394	8833	6211	6265	0290	0614	7567	5574	0261	2231	1637	8183	5952	3186
1275	8353	7434	1597	0759	2526	6989	5181	9568	3035	8052	3531	2905	0290	6655
3584	0939	1925	9936	6171	6291	9337	7515	2669	8732	4515	7913	9694	5179	3324
6847	2855	5258	8267	5406	2241	5103	6257	5163	5027	8543	5279	3970	6601	6435
1165	9721	3268	5588	4202	0905	3912	9584	0299	0328	4663	4729	4336	0791	6693
3550	6718	9068	2148	7720	7061	4316	8889	8286	4229	2406	7974	2445	7841	0896
8461	1677	3950	9131	6396	6897	4147	6032	7152	2010	0841	2457	9989	7943	8435
4416	5971	9962	7448	9187	8090	9518	0475	6221	3120	4345	6585	6205	8054	7288
1064	7573	6604	2980	5491	0816	7375	3370	2999	0813	6664	2380	7989	9032	7767
7777	7686	4441	1537	4447	6389	3353	1626	2837	6568	9476	1532	6040	5244	1436
2534	3655	7238	3066	3409	8208	6580	1837	6552	2221	7779	0149	0171	7566	6909
5210	9652	2835	3636	2114	3373	6244	4405	7593	2323	3062	0774	1321	2709	3984
8936	1425	9412	8837	6370	9426	3140	1934	7892	9781	0873	4402	6353	8856	5787