Preface

Some supplementary material, including comments on the Activities will be found at the url

`http://stats.lse.ac.uk/knott`

If you do not have access to the internet, please contact the LSE External Study Office, Houghton Street, London WC2 2AE to request a paper copy.

In spite of careful proof-reading, there are bound to be errors remaining in the text. Most of them will be trivial but annoying, but some may be more gross or more dangerous. I would be grateful for information on all errors from any reader to act as a source of corrections.

MK

5 May 2002 London

Contents

Chapter 1

Introduction

How to use this subject guide; time management; calculators; examinations and examination technique; textbooks and tables; studying statistics.

The Subject Guide and how to use it

This guide does not attempt to offer a complete treatment. There are very many well-written textbooks that cover this subject, and it would be foolish to compete with them. You will need to buy at least one textbook and consult several others from time to time. The choice of the main textbook is your personal choice, though some students will have a teacher to guide them. There are many good textbooks besides those recommended in this guide and you should be prepared to look in bookshops and libraries for texts that help you. A critical part of a good statistics text is the collection of problems for students, and you may want to look at several different texts just to see a lot of problems on some tricky topic. The guide is there mainly to describe the syllabus in some detail, and to show what level of understanding is expected, and should not be used as a main source of help but as a preliminary to more detailed work with the textbooks.

The subject guide is divided into 14 chapters which should be worked in the order given. There is little point in rushing past material half understood to reach the later chapters, the presentation being somewhat sequential, and not a series of self-contained topics. You should be familiar with the earlier chapters, and have some understanding of them before moving to the later ones.

The following procedure is recommended:

1. Read the introductory comments.
2. Read the appropriate section of your text.
3. Study the notes and examples.
4. Go carefully through the learning outcomes.
5. Attempt some of the problems from your text.

6. Refer back to this subject guide, or to the text, or to supplementary texts to improve understanding to the point where it becomes possible to work confidently through the problems.

The last two steps are the most important. It is easy to think that one has understood the text after reading it, but **working problems is the crucial test of understanding**. Problem solving should take most of your study time.

Each chapter of the guide has suggestions for reading from several texts. Usually, a student will only need to read the material in the main text, but it may be helpful from time to time to look at another one.

Time management

About one-third of your self-study time should be spent reading textbooks and the rest should be spent doing problems. An internal student would expect maybe 15 hours of formal teaching and another 50 hours of private study to be enough to cover the subject. Of the 50 hours of private study about 33 hours should be spent on trying problems (which may well require more reading) and about 17 hours on initial study of the textbook and subject guide.

Calculators and computers

You will need to provide yourself with a good calculator that has built-in routines for means, standard deviations and regression. It may be best for examinations if it is not programmable, because such machines are not allowed for use in examinations. The models change all the time, so it is hard to recommend one, but something as good as the Casio Scientific Calculator fx-570s is fine for the built-in routines. More expensive graphical calculators with the capacity to carry out symbolic algebra, and to plot data are in the shops. They are not necessary for this subject.

Those students aiming to carry out serious statistical analysis (beyond the level of this subject) will probably use some Statistics package such as MINITAB, or SPSS. It is not necessary for this subject to have such software available, but those who may have it could sometimes use it in this subject with profit.

Examination paper

The examination is by a two-hour unseen question-paper. No books may be taken into the examination, but the use of calculators is permitted, and statistical tables and a formula sheet are provided. A sample examination paper is provided in Appendix A on page 167.

The examination paper has a variety of questions, some quite short, some longer. All questions must be answered correctly for full marks. You may use your calculator whenever you feel it appropriate, always remembering that the examiner can give marks only for what appears on the examination script.

There is not much that can be helpfully said about examination technique specific to this paper. As always it is important to manage time carefully and not to stick on one question - move on and forget the question that went wrong. If English language is a problem it may be easier to give examples than to attempt an abstract description.

Textbooks and tables

Essential reading and statistical tables

Newbold, P. *Statistics for Business and Economics*. (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0].

Statistical tables

Lindley, D.V. and W.F. Scott *New Cambridge Statistical Tables*. (Cambridge: Cambridge University Press, 1995) second edition) [ISBN 0-521-48485-5].

While Newbold can act as an essential text, there are many that are as good. One looks at the range of textbooks that cover this subject with admiration for their excellence. You are encouraged to look at those given below, and at any you find. It may be necessary to look at several texts for any topic, and you may find the approach of one text suits you better than that of another. Some of the larger books now come with a disk or CD-ROM of additional material. One example of a computer-based approach with lively demonstrations is:

Doane, D.P., K. Mathieson and R.L. Tracy, *Visual Statistics 2.0*. (Irwin McGraw-Hill, 2000) [ISBN 0-07-240094-3].

There is even more computer-based teaching material available fairly freely over the web. For a well-produced on-line textbook one could try CAST by D.Stirling from

`http://cast.massey.ac.nz/CASTprog/index.html`

or HyperStatistics Online by D.M. Lane at

`http://davidmlane.com/hyperstat`

which has links to other teaching resources, as well as a web-based course.

The statistical tables are those distributed for use in the examination. It is essential that you get familiar with these tables rather than those at the end of a text.

Further reading

These are all excellent textbooks. You may not need to read them if you have Newbold's book, but from time to time it may be useful to look at one of them for help on some topic.

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6].

Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971].

Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7].

Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4].

Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7].

Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188].

Studying Statistics

Why study statistics?

By successfully completing this subject you will be familiar with the key ideas of statistics that are accessible to a student with moderate mathematical competence. You will understand the ideas of randomness and variability, and the way in which these link to probability theory to allow the use of a systematic and logical collection of statistical techniques of great practical importance in many applied areas. [1] This subject will give a grounding in probability theory and some grasp of the most common statistical methods.

[1] *The examples in this guide will concentrate on the Social Sciences, but the methods are important for physical sciences too.*

The material in this subject is necessary as a preparation for some subjects you may study later on in your degree. These subjects will not always require the detail that is discussed in this guide, but they will need an understanding of the ideas, and these can only be absorbed by seeing how they emerge in detailed technique.

How to study statistics

For statistics you need some familiarity with abstract mathematical concepts and yet enough common sense to see how to use those ideas in real-life applications. The concepts needed for probability and for statistical inference are hard to absorb by just reading them in a book. You need to read, then think a little, then try some problems, and then read and think some more. This procedure should be repeated until the problems are easy to do; **you should not spend a long time reading and forget about the problems.**

Chapter 2

Elementary probability theory. This is needed for all reasoning about random variation.

Probability

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 3

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], chapter 2, but they approach probability through random variables, rather than directly
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], chapter 4
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], chapter 5
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], sections 4.1 and 4.2
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], chapter 3
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 3, but not 3-7.

Introduction

You may find the approach a bit too abstract to start with, but it will be very satisfying to understand these basic ideas and to be able to apply them to problems which may look quite challenging . Probability is very important for statistics because it is the rules of probability that allow one to reason about uncertainty, and at the basis of statistics lies the idea of uncertainty or randomness. Independence and conditional probability are profound ideas, but they must be fully understood in order to think clearly about any statistical investigation.

Random experiment, sample space, event

Random experiment

The outcome of a *random experiment* is a result that is one of a known set of *outcomes*.

Sample space

The *sample space* Ω is the set of elements that are all the different possible outcomes of a random experiment. We write

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_r, \dots\}.$$

Note that all the *elementary outcomes* ω_i are different by definition, and that the order they are written in the braces does not matter. The sample space may be finite or infinite.

Event

An *event* A is a subset of the sample space Ω. We write $A \subset \Omega$. Note that if ω_i is an elementary outcome, then

$$\omega_i \in \Omega$$

but that the event $A = \{\omega_i\}$ with only outcome ω_i is a subset of Ω, that is

$$\{\omega_i\} \subset \Omega.$$

There are two special events that are often needed for coherence of the theory. The *null event*, written $\emptyset = \{\}$, has no outcome. It can never occur. The sample space Ω can also be thought of as an event, the *universal event*. It always occurs.

Example 2.1. Suppose that we toss a coin twice. We can use the representations HH, HT, TH, TT for the elementary outcomes of this experiment, where HT, for instance, means heads on the first toss and tails on the second toss. Then the sample space is $\Omega = \{HH, HT, TH, TT\}$. The event 'Heads on the first toss' is the subset $A = \{HH, HT\}$.
■

Example 2.2. Suppose that you arrive randomly at the station. There is a train once an hour. The random experiment is to observe the number of (rounded up) minutes that you wait before the train leaves. The elementary outcomes here are the integers 1 to 60, and the sample space is $\Omega = \{1, 2, \dots, 60\}$. The event 'You wait less than 10 minutes' is the subset $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
■

Example 2.3. A random experiment weighs a box of chocolates that is somewhere between 400g and 500g. The elementary outcomes are all the real numbers between 400 and 500, that is the open interval $(400, 500)$, so $\Omega = (400, 500)$. The event 'The box is at least 450g' is the subset $A = [450, 500)$, which is a half-open interval containing all real numbers greater than or equal to 450, yet less than 500. [1]

■

[1] *The use of different brackets and braces is important here to distinguish lists of outcomes from open or half-open intervals.*

Example 2.4. A population of interest has four members: Mabel, Belmont, Gertrude and Elsie. A random experiment selects a sample of size two from this population without replacement. The sample space is

$$\Omega = \{\{\text{Mabel}, \text{Belmont}\}, \{\text{Mabel}, \text{Gertrude}\}, \{\text{Mabel}, \text{Elsie}\}, \{\text{Belmont}, \text{Gertrude}\}, \{\text{Belmont}, \text{Elsie}\}, \{\text{Gertrude}, \text{Elsie}\}\}.$$

The event "The sample includes Gertrude" is the subset

$$\{\{\text{Mabel}, \text{Gertrude}\}, \{\text{Belmont}, \text{Gertrude}\}, \{\text{Gertrude}, \text{Elsie}\}\}.$$

■

Example 2.4 shows that elementary outcomes can be themselves sets. For instance $\omega_1 = \{\text{Mabel}, \text{Belmont}\}$.

Complement, union, intersection

Complement

If A is an event, then A^c, the *complement* of A, is the event that is the subset of all elementary outcomes in Ω but not in A.

Union

If A and B are events, then the event 'at least one of A, B', which is written $A \cup B$, occurs if the result of the random experiment is an elementary outcome that is in A, in B, or in both A and B.

Intersection

If A and B are events, then the event 'both A and B', which is written $A \cap B$, occurs if the result of the random experiment is an elementary outcome both in A and in B.[2] Two events that have no outcomes in common are said to be *disjoint* or *mutually exclusive*. Two events A and B are disjoint if and only if $A \cap B = \emptyset$.[3]

Activity 2.1. Why is $\Omega = \{1, 1, 2\}$, not a sensible way to try to define a sample space?

■

[2] *These ideas are part of elementary set theory, but they are not always understood. Take care!*

[3] *This is a different idea from independence.*

Activity 2.2. Write out all the events for the sample space $\Omega = \{a, b, c\}$. (There are eight of them.)
■

Probability and its axioms

Probabilities are non-negative numbers defined for events. We write the probability of event A as $P(A)$. We think of $P(A)$ as the probability that event A occurs – that is that the random experiment has an outcome in A. There are two fundamental properties (or axioms) for probabilities from which all other properties of probabilities may be deduced. They are:

1. $P(\Omega) = 1$

2. If $A_1, A_2, A_3, \ldots$ are disjoint events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots ...) = P(A_1) + P(A_2) + P(A_3) + \ldots .$$

You should be able to prove all the other well-known properties of probabilities from these two axioms. The second axiom says that probabilities are additive for unions of disjoint events. The left-hand side is probability that at least one of the events $A_1, A_2, \ldots$ occurs. The first axiom scales probabilities to lie in the range $[0, 1]$. [4]

[4] *Only a numbskull produces work which shows a probability outside the range* $[0, 1]$*, but we are all numbskulls from time to time.*

The simplest rules for probabilities that follow from the axioms are

$$P(\emptyset) = 0,$$

$$P(A^c) = 1 - P(A).$$

If A, B are *disjoint* events, then

$$P(A \cup B) = P(A) + P(B).$$

Also, if events A, B are such that $A \subset B$ (that is if A occurs then B occurs), then

$$P(A) \leq P(B).$$

The best known result for two general events A, B is

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

Activity 2.3. Prove, using only the axioms of probability given, that $P(\emptyset) = 0$. (Hint: $\emptyset \cup \emptyset = \emptyset$ and $\emptyset \cap \emptyset = \emptyset$.)
■

Conditional probability, independence

The most distinctive and subtle part of probability theory and practice stems from the ideas of conditioning and independence. Conditional probability is hard to work with because it can give results that, although correct, may yet be counter-intuitive.

Conditional probability

Suppose that we have a sample space Ω, probabilities defined for all events, and two particular events A, B. How should we change the assessment of $P(A)$ if we are told that event B (with $P(B) > 0$) has occurred?

If B has occurred, we know that the result of the random experiment is not in B^c. The sample space is reduced from Ω to B by the knowledge that B has occurred. The *relative* probability of events that are subsets of B is not, however, changed by knowing that B has occurred.

We write $P(A|B)$ for the probability of A given B has occurred. It is called the *conditional probability* of A given B. We can easily find a formula for $P(A|B)$ just using the fact that the relative probabilities of subsets of B are unchanged. Obviously, $P(A|B) = P(A \cap B|B)$, and $P(B|B) = 1$. Since both the events $A \cap B$ and B are subsets of B, their relative probabilities are unchanged if found conditional on the occurrence of event B, so

$$P(A|B) = P(A \cap B|B) = \frac{P(A \cap B|B)}{P(B|B)} = \frac{P(A \cap B)}{P(B)}.$$

So we have the result

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

We can't define the conditional probability $P(A|B)$ unless $P(B) > 0$. Notice also that $P(A|B)$ is not usually the same as $P(B|A)$. For instance, the probability that you have very short hair given that you have just had your hair cut is not necessarily the same as the probability that you have just had your hair cut if you have short hair. The assumption in legal applications that these conditional probabilities are equal is called the 'prosecutor's fallacy'. It goes along the fallacious lines: *There is a one in a million chance that the DNA at the scene of the crime does not belong to the accused. So the odds are a million to one that the accused was at the crime scene.*

Activity 2.4. If all elementary outcomes are equally likely, and $\Omega = \{a, b, c, d\}$, $A = \{a, b, c\}$, $B = \{c, d\}$, find $P(A|B)$ and $P(B|A)$.
■

Activity 2.5. Identify the events and then the conditional probabilities wrongly equated in the prosecutors fallacy above.
■

Independent events

If two events A, B with $P(B) > 0$ are such that $P(A|B) = P(A)$, then A and B are said to be *independent* events. The condition is equivalent to

$$P(A \cap B) = P(A)P(B) \tag{2.1}$$

which can also be used to define independence between A, B and covers the case $P(B) = 0$. In this form the definition of independence is symmetric in A, B and covers the cases when $P(A) = 0$, $P(B) = 0$ Equation (2.1) is the preferred way to define independence for two events.

The idea here is that B has no effect on A, so knowing that B occurs does not change the assessment of $P(A)$.

Activity 2.6. Suppose that we toss a coin twice. The sample space is $\Omega = \{HH, HT, TH, TT\}$, where the elementary outcomes are defined in the obvious way - for instance HT is heads on the first toss and tails on the second toss. Show that if all four elementary outcomes are equally likely, then the events 'Heads on the first toss' and 'Heads on the second toss' are independent.
■

Activity 2.7. Show that if A and B are disjoint events, and are also independent, then $P(A) = 0$ or $P(B) = 0$.
■

Independence for several events

A collection of events $\{A_1, A_2, \ldots, A_k\}$ is said to be *independent* if the product rule

$$P[A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_k] = P(A_1)P(A_2)P(A_3)\ldots P(A_k)$$

holds **and** if every subset of the k events is independent. [5]

[5] *This is a rather difficult recursive definition, but it does not simplify.*

Example 2.5. Write down the condition for three events A, B and C to be independent.

Answer

Applying the product rule above for $k = 3$, we must have

$$P[A \cap B \cap C] = P(A)P(B)P(C),$$

Then since all subsets of two events from A, B and C must be independent, we must have

$$P[A \cap B] = P[A]P[B], P[A \cap C] = P[A]P[C], P[B \cap C] = P[B]P[C].$$

One must check that all **four** conditions hold to verify independence of A, B and C.
■

Example 2.6. It can be cold in London. Four impoverished teachers dress to feel warm. Teacher A has a hat and a scarf and gloves, B has only a hat, C only a scarf and D has only gloves. One teacher out of the four is selected at random. Show that though each **pair** of events $H =$ 'the teacher selected has a hat', $S =$ 'the teacher selected has a scarf', and $G =$ 'the teacher selected has gloves' are independent, all **three** of these events are not independent.[6]

[6] *Think carefully about the difference between pairwise independence and full independence.*

Solution

Two teachers have a hat, two teachers have a scarf, and two teachers have gloves, so

$$P(H) = 2/4 = 1/2, P(S) = 2/4 = 1/2, P(G) = 2/4 = 1/2.$$

Only one teacher has both hat and scarf, so

$$P(H \cap S) = 1/4$$

and similarly

$$P(H \cap G) = 1/4, P(S \cap G) = 1/4.$$

From these results

$$P(H \cap S) = P(H)P(S), P(H \cap G) = P(H)P(G), P(S \cap G) = P(S)P(G),$$

and the events are pairwise independent. But one teacher has hat, scarf and gloves, so

$$P(H \cap S \cap G) = 1/4 \neq P(H)P(S)P(G),$$

so the three events are not independent. If the selected teacher has a hat and a scarf, then we **know** that teacher has gloves. There is no independence for all three events together.

■

Bayes' theorem

Suppose that the events B_i, $i = 1, \ldots, n$ partition Ω, that is they are pairwise disjoint events such that $\cup_{i=1}^{n} B_i = \Omega$. Then **Bayes' Theorem** says that, for an event A,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}.$$

Activity 2.8. Prove this result from first principles.

■

Sometimes calculations for Bayes' theorem are arranged on a *tree diagram*. There is an example in Figure 2.1 of Example 2.7. This technique may clarify what is necessary for such calculations in complicated applications.

Example 2.7. A statistics teacher knows from past experience that a student who does the homework consistently has a probability 0.95 of passing the exam, whereas a student who does not do the homework has a probability 0.30 of passing.

1. If 25% of students do their homework consistently, what percentage can expect to pass?

2. If a student chosen at random from the group gets a pass, what is the probability that the student has done the homework consistently?

Answer

Here the random experiment is to choose a student at random, and to record whether that student passes (G) or fails (B), and whether that student has done the homework consistently (H) or has not (L). The sample space is $\Omega = \{GH, GL, BH, BL\}$. We use the events Pass $= \{GH, GL\}$, and Fail $= \{BH, BL\}$. We consider the sample space partitioned by Homework $= \{GH, BH\}$, and No Homework $= \{GL, BL\}$.

The first part of the example asks for the denominator of Bayes' Theorem:

$$\begin{aligned} P(\text{Pass}) &= P(\text{Pass}|\text{Homework})P(\text{Homework}) \\ &\quad + P(\text{Pass}|\text{No Homework})P(\text{No Homework}) \\ &= 0.95 \times 0.25 + 0.30(1 - 0.25) \\ &= 0.2375 + 0.225 \\ &= 0.4625 = 46.25\%. \end{aligned}$$

Now applying Bayes' Theorem

$$\begin{aligned} P(\text{Homework}|\text{Pass}) &= P(\text{Homework} \cap \text{Pass})/P(\text{Pass}) \\ &= P(\text{Pass}|\text{Homework})P(\text{Homework})/P(\text{Pass}) \\ &= 0.95 \times 0.25/0.4625 \\ &= 0.5135. \end{aligned}$$

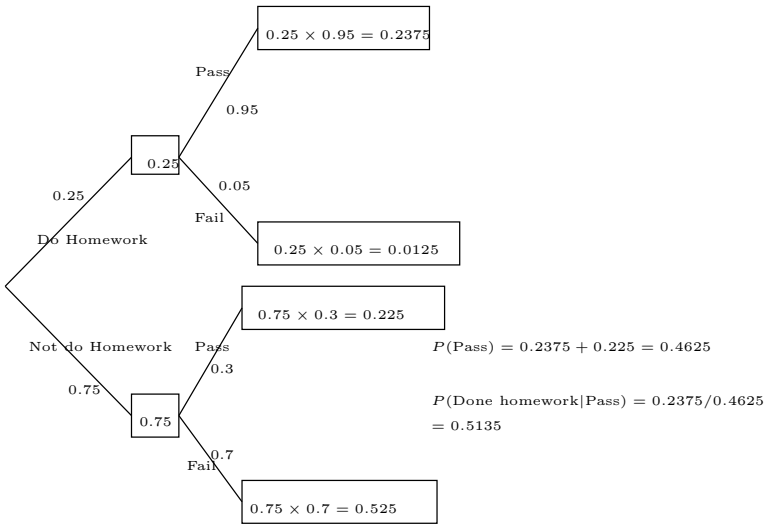
Or we could arrange the calculations in a tree diagram as in Figure 2.1

■

It can help to understand Bayes' Theorem and the tree diagrams if one forgets about probability for a moment and thinks about classifying a population.

Example 2.8. In Example 2.7 one could think of 10000 students. These may be classified into 2375 who do their homework and pass, 125 who do their homework and fail, 2250 who do not do their homework and pass, and 5250 who do not do their homework and fail. These numbers are found by simple calculations. For example, 25% of the 10000 do their homework, so there are 2500 of those. Of that 2500, 95% pass, so 2375 do their homework and pass. Then $2375 + 2250 = 4625$ pass, and of

Figure 2.1: Tree diagram for Example 2.7



those passes, 2375 have done their homework. These figures are in tabular form in Table 2.1.
■

Table 2.1: Results for Example 2.8

	Do Homework	Do not do homework
Pass	2375	2250
Fail	125	5250

Activity 2.9. Plagiarism is a serious problem for assessors of course-work. One check on plagiarism is to compare the course-work with a standard text. If the course-work has plagiarised that text, then there will be a 95% chance of finding exactly two phrases that are the same in both course-work and text, and a 5% chance of finding three or more. If the work is not plagiarised, then these probabilities are both 50%.

Suppose that 5% of course-work is plagiarised. An assessor chooses some course-work at random. What is the probability that it has been plagiarised if it has exactly two phrases in the text?[7] And if there are three phrases? Did you manage to get a

[7] *Try making a guess before doing the calculation*

roughly correct guess of these results before calculating?
■

Permutations and combinations

If the elementary outcomes in the sample space are equally likely outcomes for the random experiment, then a calculation of probability for an event reduces to counting the number of elementary outcomes for which that event occurs and dividing by the total number of elementary outcomes.

In some applications the equally likely elementary outcomes are either *permutations* or *combinations* of some collection of tokens (for instance, letters in the English alphabet). Permutations are ordered arrangements, whereas combinations are unordered arrangements.

The set of all permutations of two letters from the first four letters of the English alphabet is

$$\{ab, ac, ad, bc, bd, cd, dc, db, da, cb, ca, ba\}.$$

There are four ways to choose the first letter and three ways to choose the second letter giving 12 permutations. Generalising gives the formula for the number of permutations $(m)_r$ of r objects from m as

$$(m)_r = m(m-1)\dots(m-r+1) = \frac{m!}{(m-r)!},$$

where $m! = m(m-1)(m-2)\dots(3)(2)(1)$ (and is called m *factorial*). [8]

[8] *Note that $m!$ is the number of permutations of all m objects out of m. By convention we take $0! = 1$.*

The set of all combinations of two letters from the first four letters of the English alphabet is

$$\{ab, ac, ad, bc, bd, cd\}.$$

Each of these combinations is arranged to give two of the previous permutations. Generalising this argument gives the formula for the number of combinations $\binom{m}{r}$ of r objects from m as

$$\binom{m}{r} = \frac{m!}{r!(m-r)!}. \tag{2.2}$$

Sampling without replacement

An important use of these countings is to find the probability that a sample of size n taken at random without replacement from a population with R red balls and $N-R$ green balls has r red balls. It is

$$\frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}. \tag{2.3}$$

To obtain this result we note that the (unordered) samples are equally likely combinations of n out of N balls, so that there are $\binom{N}{n}$ equally likely samples. A sample with r red balls has $n-r$ green balls. Each of the $\binom{R}{r}$ ways of choosing the unordered red balls is combined with any one of the $\binom{N-R}{n-r}$ ways of choosing the unordered green balls to give

$$\binom{R}{r}\binom{N-R}{n-r}$$

different samples with r red balls. The ratio of the number of such samples to the total number of samples gives the result.

Intuitively, one should get the same probability in the following fashion: suppose that we begin by separating the N balls into n for the sample and $N-n$ for the others. Colour red a randomly chosen R out of the N balls, colouring the others green. What is the probability that r balls in the sample are coloured red? Using the same argument as before, the probability is

$$\frac{\binom{n}{r}\binom{N-n}{R-r}}{\binom{N}{R}}. \tag{2.4}$$

You can check directly that (2.3) and (2.4) are the same by using the expression (2.2) for binomial coefficients in terms of factorials.

Activity 2.10. A box contains three red balls and two green balls. Two balls are taken from it without replacement. What is the probability that 0 balls taken are red? And 1 ball? And 2 balls? Show that the probability that the first ball taken is red is the same as the probability that the second ball taken is red.
■

Miscellaneous examples

Example 2.9. A, B and C throw a die in that order until a six appears. The person who throws the first six wins. What are their respective chances of winning?

Answer

We must assume that the game finishes with probability one (it would be proved in a more advanced subject). If A, B and C all throw and fail to get a six, then their respective chances of winning are as at the start of the game. We can call each completed set of three throws a round. Let us denote the probabilities of winning by

$P(A)$, $P(B)$ and $P(C)$. Then

$$\begin{aligned}
P(A) &= P(A \text{ wins on the first throw}) \\
&\quad + P(A \text{ wins in some round after first}) \\
&= 1/6 + P(A, B \text{ and } C \text{ fail on 1st throw and A wins after 1st round}) \\
&= 1/6 + P(A, B, C \text{ fail in 1st round}) \\
&\quad \times P(A \text{ wins after 1st round} | A, B, C \text{ fail in 1st round}) \\
&= 1/6 + P(\text{No six in first 3 throws}) P(A) \\
&= 1/6 + (5/6)^3 P(A) \\
&= 1/6 + (125/216) P(A).
\end{aligned}$$

So $(1 - 125/216)P(A) = 1/6$, and $P(A) = 216/(91 \times 6) = 36/91$.

Similarly,

$$\begin{aligned}
P(B) &= P(B \text{ wins in first round}) \\
&\quad + P(B \text{ wins after first round}) \\
&= P(A \text{ fails with first throw and } B \text{ throws a six on first throw}) \\
&\quad + P(\text{ All fail in 1st round and } B \text{ wins after 1st round}) \\
&= P(A \text{ fails with 1st throw}) Pr(B \text{ throws a six with 1st throw}) \\
&\quad + P(\text{All fail in 1st round}) P(B \text{ wins after 1st} | \text{All fail in 1st}) \\
&= (5/6)(1/6) + (5/6)^3 P(B).
\end{aligned}$$

So, $(1 - 125/216)P(B) = 5/36$, and $P(B) = 5(216)/(91 \times 36) = 30/91$.

In the same way, $P(C) = (5/6)(5/6)(1/6)(216/91) = 25/91$.

Notice that $P(A) + P(B) + P(C) = 1$. You may, on reflection, think that this rather long solution could be shortened, by considering the relative winning chances of A, B, C.

■

Example 2.10. In men's singles tennis, matches are played on the best of five sets principle. Thus the first player to win three sets wins the match, and a match may consist of three, four or five sets. Assuming that two players are perfectly evenly matched, and that sets are independent events, calculate the probabilities that a match lasts three sets, four sets and five sets.

Answer

Suppose that the two players are A and B. We calculate the probability that A wins a three, four or five set match, and then, since the players are evenly matched,

double these probabilities for the final answer.

$$P(A \text{ wins in 3 sets}) = P(A \text{ wins first set} \cap A \text{ wins second set} \cap A \text{ wins third set}).$$

Since the sets are independent,

$$\begin{aligned} P(A \text{ wins in 3 sets}) &= P(A \text{ wins first set})P(A \text{ wins second set})P(A\text{wins third set}) \\ &= (1/2)(1/2)(1/2) = 1/8. \end{aligned}$$

So the total probability that the game lasts three sets is

$$2(1/8) = 1/4.$$

If A wins in four sets, the possible winning patterns are

$$BAAA,\ ABAA,\ AABA.$$

Each of these patterns has probability $(1/2)^4$ by using the same argument as in the case of 3 sets. So the probability that A wins in four sets is $3(1/16) = 3/16$. The total probability of a match lasting four sets is $2 \times 3/16 = 3/8$.

The probability of a five-set match should be $1 - 3/8 - 1/4 = 3/8$, but let us check this directly. The winning patterns for A in a five-set match are

$$BBAAA,\ BABAA,\ BAABA,\ ABBAA,\ ABABA,\ AABBA.$$

Each of these has probability $(1/2)^5$ because of the independence of the sets. So the probability that A wins in five sets is $6(1/32) = 3/16$. The total probability of a five-set match is thus $3/8$ as before.

■

Learning outcomes

After working through this chapter you should be able to:

1. discuss the fundamental ideas of random experiments, sample spaces and events
2. list the axioms of probability and be able to deduce all the common rules for probabilities from them
3. explain conditional probability, and the idea of independent events
4. prove Bayes' Theorem and to use it to find conditional probabilities
5. list the formulae for the number of combinations and permutations of r tokens out of m, and be able to routinely apply such results in problems.

Sample examination questions

1. Alphonse, Bertille and Clarence play a simple game with a fair die. Alphonse tosses the die and observes the number on the uppermost face, which the other two do not see. Bertille tries to guess that number. If she is right, she wins. If she is wrong, then Clarence tries to guess the number and so win the game. If neither Bertille nor Clarence guess correctly, then Alphonse wins the game.

   (a) What is the chance that Clarence wins the game?

   (b) What is the conditional probability that Clarence wins the game given that Alphonse does not win?

   (Elements of Statistics 2001 Zone A)

2. In a large sociology class, 10% of all students get an A grade in their end of year examination. 60% of all students had missed no classes. The examiners checked and found that 10% of all students who had got an A grade had missed no classes.

   (a) What is the probability that a student with an A grade had missed no classes?

   (b) What is the probability that a student who missed at least one class did not get an A grade?

   (c) Are the events 'got an A grade' and 'missed no classes' mutually exclusive? Explain.

   (Elements of Statistics 1998 Zone A)

3. State and prove Bayes' theorem. (Elements of Statistics 1998 Zone A)

4. In six independent tosses of a fair coin, what is the probability that there are at least three successive heads somewhere in the sequence?
   (Elements of Statistics 1998 Zone A)

5. If A and B are events such that $P(A|B^c) = 2P(A|B)$ and $P(B^c) = 2P(B)$, show that $P(B^c|A) = 0.8$. (The event B^c is the complement of event B.)
   (Elements of Statistics 1998 Zone B)

6. Show that if $P(A \mid B) = P(A \mid B^c)$ then A and B are independent.
   (Elements of Statistics 1997 Zone A)

7. A box of 50 coloured light bulbs consists of 5 blue bulbs, 20 pink bulbs, 20 yellow bulbs and 10 white bulbs. If 4 bulbs are selected at random without replacement, what is the probability that:

(a) all 4 bulbs are white

(b) all 4 bulbs are the same colour

(c) all four bulbs are different colours

(d) no bulb is yellow?

(Elements of Statistics 1997 Zone B)

8. Annabel is twice as likely to go shopping as Barbara. Carmel is three times as likely to go shopping when Annabel meets her than she is when Barbara meets her. How much more likely is Annabel to shop than Barbara when each has been met by Carmel?
(Elements of Statistics 1997 Zone B)

Chapter 3

Univariate distributions

This chapter covers the distributions of most importance for direct application: normal, exponential, uniform, binomial, Poisson.

Essential reading

Newbold, P. *Statistics for Business and Economics*. (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapters 4 (but not 4.4) and 5.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], chapter 2, 3, 4 and 5 but the approach is rather different in this book

Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], chapters 5 and 6.

Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], chapters 6 and 7.

Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4]chapter 4, but fewer distributions covered.

Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7]chapters 4 and 5, but fewer distributions covered.

Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188]chapter 4, but fewer distributions covered.

Introduction

This chapter introduces the two most useful standard distributions for counts - the binomial distribution and the Poisson distribution. These are so often used that everyone should be familiar with them. We also look at the most important[1]

distributions for measured data - the normal distribution, the exponential distribution and the uniform distribution. *Univariate* means one variable. These distributions are for only one quantity; if two quantities are used we need a bivariate distribution, see Chapter 4.

For all these distributions one needs to know the mean and the variance, and how

[1] *There are other distributions, of great importance for inference such as the t distribution and the F distribution. These will appear when they become necessary.*

to find simple probabilities.

Random variables

A random variable is a function that acts on the elementary outcomes in the sample space Ω to produce real numbers. We use upper-case letters like X, Y to represent random variables.

Example 3.1. Suppose that we use Ω from Example 2.1 on page 8 for tossing two coins. One useful random variable X is defined by

$$X(HH) = 2,$$
$$X(HT) = 1,$$
$$X(TH) = 1,$$
$$X(TT) = 0.$$

This random variable is 'The number of heads'. It takes values $0, 1, 2$.
■

Example 3.2. Suppose that we use Ω from Example 2.3 on page 9, which is the open interval of all real numbers from 400 to 500. A simple random variable X is defined for $\omega \in (400, 500)$ by

$$X(\omega) = \omega/1000.$$

This random variable is 'The weight in kg'. It takes values from $(.4, .5)$.
■

We can see from the examples that the random variable is a function, and it takes values that are real numbers. We use notation like x for a value of the random variable X.[2]

[2] *Although in some parts of statistics it is usual to blur the distinction between a random variable and its values, in this subject we shall carefully keep these two ideas completely distinct.*

The 'randomness' of a random variable comes from the random experiment, which gives rise to one of the outcomes ω in Ω. In Example 3.1 on page 24 we know that for independent tosses of a fair coin

$$P[X = 2] = 0.25,$$

because $X = 2$ if and only if the event $\{HH\}$ occurs, and this event has probability 0.25. In general we define the probabilities for random variables from the probabilities already defined for events - a probability calculated for a random variable comes in principle from the definition of probability for the events arising from Ω.

Of course, once we start to think about the random variables, they take on a life of their own, and we don't always bother to think about Ω. One useful concept is $F_X(x)$, the *cumulative distribution function (cdf)* of the random variable X. It is defined by

$$F_X(x) = P(X \leq x)$$

for all values $-\infty < x < \infty$. This is the *left-hand tail probability* for the distribution of X.

Activity 3.1. Show that for a discrete random variable X taking integer values x, $P(X = x) = F_X(x) - F_X(x-1)$.
■

Binomial random variable

Suppose that a random experiment records the result of n trials, each of which is either a success or a failure. The sample space contains all the different ordered sequences ω of length n of successes and failures. For instance, with 0 for a failure and 1 for a success, and with $n = 4$ one possible ω is 0110. This corresponds to a failure followed by two successes, followed by a failure. Suppose that the n trials are independent, and that the probability of success in each trial is π. Define a random variable X (generalising Example 3.1) by

$$X(\omega) = \text{ number of successes in } \omega.$$

For instance, with $\omega = 0110$ we have $X(\omega) = 2$. Then X is a *binomial random variable*. We usually describe it as counting the number of successes in n independent trials each with probability of success π. For $x = 0, 1, \ldots, n$ there are $\binom{n}{x}$ elementary outcomes ω with x successes in Ω, and each of these has probability $\pi^x(1-\pi)^{n-x}$. We can therefore write

$$P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \qquad x = 0, 1, \ldots, n. \tag{3.1}$$

The description of the probabilities in (3.1) is called the *Binomial Distribution* for n independent trials[3] with probability of success π. We use Binomial(n, π) for short. For a particular choice of n, π these probabilities may be found with a calculator or looked up in tables. If there is only one trial, instead of talking about a binomial distribution, we sometimes say we have a *Bernoulli trial*. A Bernoulli trial gives either 0 success or 1 success. If we add together n independent Bernoulli trials, we get a random variable with a binomial distribution.

[3] *One must take care in applications that the trials really are independent, and that they all have the same probability of success.*

Activity 3.2. The greengrocer has a very large (effectively infinite) pile of oranges on his stall. The pile of fruit is a mixture of 50% old fruit with 50% new fruit; one can't tell which are old and which are new. However, 20% of old oranges are mouldy inside, but only 10% of new oranges are mouldy. Suppose that you choose 5 oranges at random. Is it true that the number of mouldy oranges in your sample has a binomial distribution with $n = 5$ and $\pi = 0.15$?
■

Example 3.3. Tube trains on the Northern Line have a probability 0.1 of failure between Golders Green and King's Cross. Supposing that the failures are all independent, what is the probability that out of 10 journeys between Golders Green and King's Cross more than 8 do not have a breakdown.

The probability of no breakdown on one journey is $\pi = 0.9$, and the number of journeys without breakdown, X, has a Binomial$(10, 0.9)$ distribution. We want $P(X > 8)$ which is, from a hand calculator,

$$\begin{aligned} P(X = 9) + P(X = 10) &= \binom{10}{9}(0.9)^9(0.1) + \binom{10}{10}(0.9)^{10}(0.1)^0 \\ &= 0.38742 + 0.34868 \\ &= 0.73610. \end{aligned}$$

From the *New Cambridge Statistical Tables*, it is just as easy. One must work with the number of failures $Y = 10 - X$, where a failure has probability 0.1. Since $P(X > 8) = P(Y \leq 1)$ we can look in Table 1 under $n = 10$, $r = 1$ and $p = 0.1$. The table value is 0.7631, which agrees with our previous result.

■

Example 3.4. Suppose that the normal rate of infection of a certain disease in cattle is 25%. To test a new serum which may prevent infection three experiments are carried out. The test for infection is not always valid for some particular cattle, so the experimental results are incomplete - we can't always tell whether a cow is infected or not.

1. 10 animals are injected: all 10 remain free from infection
2. 17 animals are injected: more than 15 remain free from infection: there are two doubtful cases
3. 23 animals are infected: more than 20 remain free from infection: there are three doubtful cases.

Which experiment provides the strongest evidence in favour of the serum?

Answer

These experiments involve tests on different cattle, which one might expect to behave independently of one another. The probability of infection without injection with the serum might also reasonably be assumed to be the same for all cattle. So the distribution that we need here is the binomial distribution. If the serum has no effect, then the probability of infection for each of the cattle is 0.25.

One way to assess the evidence of the three experiments is to calculate the probability of the result of the experiment if the serum had no effect at all. If it has an effect, then one would expect larger numbers of cattle to remain free from infection, so the experimental results as given do provide some clue as to whether the serum has an effect, in spite of their incompleteness.

1. With 10 trials, the probability of 0 infected if the serum has no effect is

$$(0.75)^{10} = 0.0563.$$

2. With 17 trials, the probability of 16 or 17 remaining uninfected if the serum has no effect is, using the binomial probabilities

$$\binom{17}{0}(0.75)^{17} + \binom{17}{1}(0.25)(0.75)^{16}$$

$$= (0.75)^{17} + 17(0.25)(0.75)^{16}$$

$$= 0.0075 + 0.0426 = 0.0501.$$

3. With 23 trials, the probability of 21, 22 or 23 remaining uninfected if the serum has no effect is, again using the binomial probabilities

$$\binom{23}{0}(0.75)^{23} + \binom{23}{1}(0.25)(0.75)^{22} + \binom{23}{2}(0.25)^2(0.75)^{21}$$

$$= (0.75)^{23} + 23(0.25)(0.75)^{22} + 23(22)(0.25)^2(0.75)^{21}/2$$

$$= 0.0013 + 0.0103 + 0.0376 = 0.0492.$$

The most surprising looking event in these three experiments is that of experiment 3, and so one might believe that this experiment offered the most support for the serum. A certain degree of caution is needed with this argument. It would also be possible to use the binomial tables for the calculations, but a hand calculator is enough here.
■

Poisson random variable

A Poisson random variable X is useful for applications where the observations are counts. If the mean of the distribution is μ, then the probability function is

$$P(X = x) = e^{-\mu}\mu^x/x! \tag{3.2}$$

for $x = 0, 1, 2, \ldots$.

The distribution can be used for counts of random events in a given time period, assuming that the numbers of events in disjoint time intervals are independent, and that the probability of an event occurring in a time interval of length t is proportional to μt.

The Poisson distribution with mean $n\pi$ can be used to approximate a binomial distribution from n trials and probability of success π when n is large and π is small. [4]

[4] *For instance, we might use this approximation for binomial distributions not covered by the Tables.*

Example 3.5. In a large industrial plant there is a serious accident on average every two days. What is the chance that there will be exactly two serious accidents in a given week? The chance of two or more? If I go to work there for a four-week period what is the probability that no serious accident occurs while I am there?

Here we have counts of random events over time, which is a typical application for the Poisson distribution. We are assuming that accidents are equally likely to occur at any time and are independent. The mean for the Poisson distribution is 0.5 per day.

The probability of exactly two accidents in a given week is found by using the Poisson tables for $\mu = 5 \times 0.5 = 2.5$ (5 working days a week assumed). From Table 2 of the *New Cambridge Statistical Tables* with $\mu = 2.50$ and $r = 2$, $r = 1$ we get

$$P(X \leq 2) = 0.5438.$$

$$P(X \leq 1) = 0.2873.$$

So the probability of exactly two accidents in a week is $0.5438 - 0.2873 = 0.2565$.

For the probability of two or more accidents,

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - 0.2873 \\ &= 0.7127 \end{aligned}$$

If you go to the works, and do not change the probabilities simply by being there (you might bring bad luck, or be superbly efficient) then over 4 weeks there are 20 working days, and the probability of no accident comes from a Poisson random variable with mean 10. The probability is therefore

$$e^{-10}10^0/0! = e^{-10} = 0.00004540.$$

You are very likely to be there when there is an accident.

■

Activity 3.3. The chance that a lottery ticket has a winning number is .0000001. If 10,000,000 people buy tickets that are independently numbered, what is the probability of 0 winner? Of 1 winner? (Hint, use a Poisson distribution with mean 1. The answers are 0.37, 0.37.)

■

Uniform random variable

We turn now to random variables with values that are measured, and so can lie anywhere in some interval (a, b). (The very simplest is a uniform random variable on the interval $(0, 1)$.) The sample space is the interval (a, b), and the random experiment is

to choose one of the real numbers ω in that interval in such a way that there is no bias in favour of any particular number. The uniform random variable X just gives back ω, that is $X(\omega) = \omega$. Roughly speaking, X is a random number in the interval (a, b).

Random variables with values that are measured are usually called *continuous* to distinguish them from those which have counted values such as the Poisson random variable. Random variables with counted values are called *discrete*.

Probabilities like $P(X = x)$ are no use to describe continuous random variables. For a uniform random variable on the interval $(0, 1)$ it is fairly intuitive that $P(X = 0.5) = 0$ – there is a zero probability of observing the value 0.5 even though this is a possible value for X.[5] Similarly, for any other particular value x, we have $P(X = x) = 0$. However, there is a non-zero probability for an interval of non-zero length. We have $P(X \leq 0.5) = 0.5$.

[5] *Note that $X = 0.5$ is not impossible, but it does have probability 0.*

Activity 3.4. Show that for a uniform random variable on $(0, 1)$, if the probability $P(X = x)$ is the same for all x between 0 and 1, then it must be equal to 0 for all those x (otherwise the probability of the sample space is not equal to 1, but infinite). ■

Similarly if X is a uniform random variable on the interval (a, b), and $a < c < d < b$ then

$$P(c < X \leq d) = \frac{d - c}{b - a}.$$

Notice that if we take

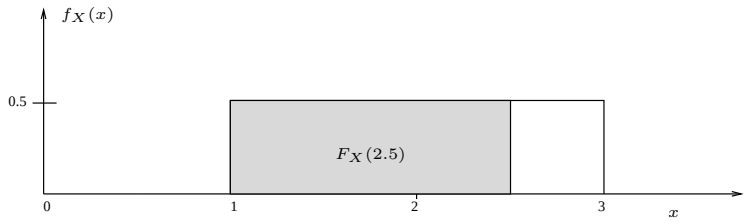
$$f_X(x) = \frac{1}{b - a},$$

then

$$P(c < X \leq d) = F_X(d) - F_X(c) = \int_c^d f_X(x)dx = \int_c^d \frac{1}{b - a}dx = \frac{d - c}{b - a}$$

so we can think of $f_X(x)$ as a *probability density function* for X. This is pictured in Figure 3.1 below.

Figure 3.1: Density function for an uniform random variable on the range 1 to 3

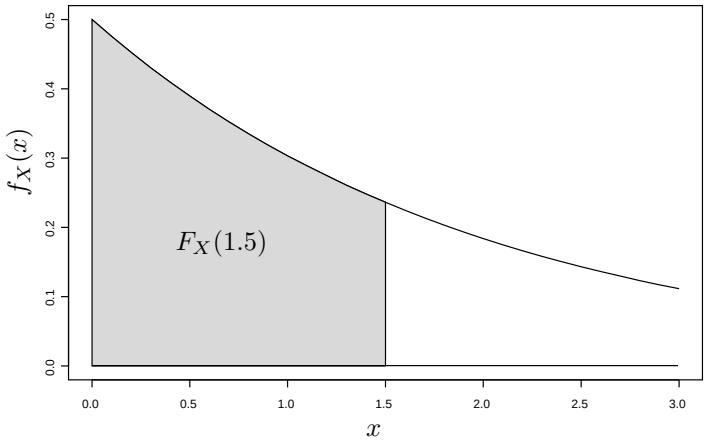


Activity 3.5. Suppose that X is uniformly distributed on $(0, 1)$.
What is $P(X > 0.2)$, $P(X \geq 0.2)$, $P(X^2 > 0.04)$. (Hint: All these probabilities are the same.)
■

Exponential random variable

An exponential random variable with mean $1/\lambda$ is a continuous random variable taking non-negative values. See Figure 3.2 on page 30 for a picture.

Figure 3.2: Density function for an exponential random variable with mean 2



The exponential random variable cdf for $x \geq 0$ is

$$F_X(x) = 1 - e^{-\lambda x}.$$

If $0 < c < d$, then

$$P(c < X \leq d) = \int_c^d \lambda e^{-\lambda x} dx$$

and so the exponential random variable has probability density function

$$f_X(x) = \lambda e^{-\lambda x} \tag{3.3}$$

for $x \geq 0$. [6].

[6] *Notice that for $\lambda > 1$ this density function is greater than 1 at $x = 0$.*

Example 3.6. Suppose that the service time for a customer at a fast-food outlet has an exponential distribution with mean 3 minutes. What is the probability that a customer waits more than 4 minutes?

The probability of waiting fewer than 4 minutes is

$$F_X(4) = 1 - e^{-4/3}$$

and so the probability of waiting more than 4 minutes is

$$e^{-4/3} = 0.2636.$$

■

Normal random variable

A normal random variable with mean μ and variance σ^2 is often said to have an $N(\mu, \sigma^2)$ distribution. The *standard normal random variable*, Z, has an $N(0, 1)$ distribution. The random variable Z has a probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

This is usually written as

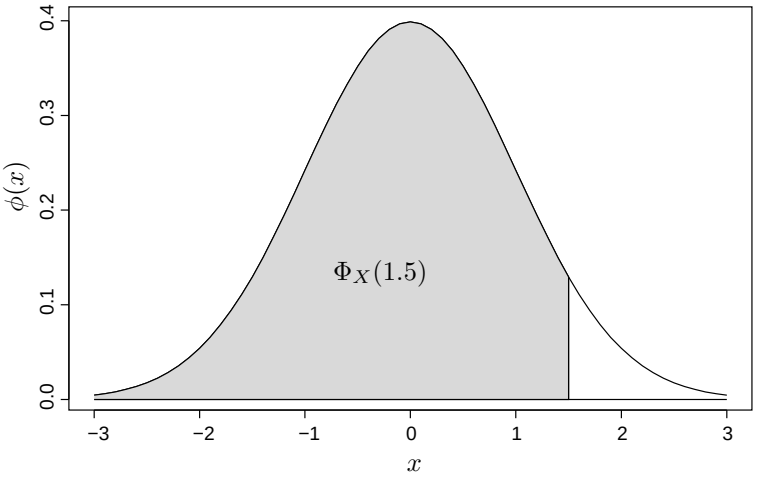
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{3.4}$$

The cdf for an $N(0, 1)$ random variable is usually written

$$\Phi(x) = P(X \leq x) = \int_{-\infty}^{x} \phi(x)dx.$$

It is tabulated in Table 4 of the *New Cambridge Statistical Tables* , and can be used to find probabilities for any normal distribution using the fact that if X has an $N(\mu, \sigma^2)$ distribution, then $\frac{X-\mu}{\sigma}$ has an $N(0, 1)$ distribution. The density function is in Figure 3.3 below.

Figure 3.3: Density function for a standard normal random variable



Example 3.7. Suppose that a population of men's heights is normally distributed with a mean of 68 inches,[7] and standard deviation of 3 inches. Find the proportion of men who are:

[7] *An inch is about 2.54 cm.*

1. Under 66 inches

2. Over 72 inches

3. Between 66 and 72 inches.

Answer

The two cut-off values are 66 and 72. Converted to standard units these are

$$(66 - 68)/3 = -2/3 \text{ and } (72 - 68)/3 = 4/3.$$

The right-hand tail probability for $z = 4/3 = 1.33$ is from the Table 4 of the *New Cambridge Statistical Tables* with $x = 1.33$ given by $1 - \Phi(1.33) = 1 - 0.9082 = 0.092$. The left-hand tail probability for $z = -2/3 = -0.66$ is the same as $1 -$ the left-hand tail probability for $z = 0.66$, and this from Table 4 with $x = 0.66$ is $1 - \Phi(0.66) = 1 - 0.7454 = 0.255$.

So the answers are

1. 25.5%

2. 9.2%

3. $100 - 25.5 - 9.2 = 65.3\%$.

■

Example 3.8. Two statisticians disagree about the distribution of IQ scores for a population under study. Both agree that the distribution is normal, and that the standard deviation is 15, but A says that 5% of the population have IQ scores greater than 134.6735, whereas B says that 10% of the population have IQ scores greater than 109.224. What is the difference between the mean IQ score as assessed by A and that as assessed by B?

Answer

The standardised z value giving 5% in the upper tail is from Table 5 with $P = 5$ given[8] by $x(P) = 1.6449$, and for 10% it is given for $P = 10$ by $x(P) = 1.2816$. So, converting to the scale for IQ scores, the values are

[8] *P is already a **percentage**.*

$$1.6449 \times 15 = 24.6735$$

$$1.2816 \times 15 = 19.224.$$

Write the means according to A, B as μ_A, μ_B respectively. Then

$$\mu_A + 24.6735 = 134.6735,$$

so

$$\mu_A = 110,$$

whereas

$$\mu_B + 19.224 = 109.224$$

so $\mu_B = 90$.

The difference $\mu_A - \mu_B = 110 - 90 = 20$.

■

Example 3.9. If 95% of a normal population lies between 151 and 249, what are the mean and the standard deviation of the population?

If the mean is μ and the standard deviation σ, then working with standardised values,

$$1 - \Phi((249 - \mu)/\sigma) - [1 - \Phi((151 - \mu)/\sigma)] = 0.95.$$

Here there is only one equation for two unknowns, so one must make more assumptions to arrive at a unique answer. Consider two possible assumptions:

a) If one assumes an equal amount of probability in each tail, then the interval $(151, 249)$ is symmetric about μ, and so $\mu = (151 + 249)/2 = 200$, and then $(249 - 200)/\sigma = 1.9600$, so $\sigma = 25$. The value $x(P) = 1.9600$ is from Table 5 of the *New Cambridge Statistical Tables* with $P = 2.5$.

b) If one assumes that the left-hand tail probability is 1%, and the right-hand tail probability is 4%, then from Table 5 with $P = 4.0$, $x(P) = 1.7507$ and with $P = 1.0$, $x(P) = 2.3263$, so

$$(249 - \mu)/\sigma = 1.7507$$

and

$$(151 - \mu)/\sigma = -2.3263.$$

Solving the equations for μ and σ,

$$249 - 151 = (1.7507 + 2.3263)\sigma,$$

so $\sigma = 24.0372$.

Substituting back in the first equation gives

$$249 - (24.0372 \times 1.7507) = \mu,$$

so $\mu = 206.9179$.

The fairly large difference in the assumptions has led to much the same answer as before. This is because the tail of the normal distribution falls off so sharply on either side.

■

Activity 3.6. If X is a random variable with a standard normal distribution, what is $P(X^2 > 3.84)$?

■

Expected value of a random variable

The *expected value*, $E(X)$, of a random variable X, also called the *mean* of X, is defined for a discrete random variable as

$$E(X) = \sum_x xP(X = x) \tag{3.5}$$

where the summation is over all values of X. For a continuous random variable it is defined as

$$E(X) = \int xf_X(x)dx, \tag{3.6}$$

where we integrate over the values x for which the density function $f_X(x)$ is greater than 0. $E(X)$ is also called the *mean value* of X, or the *mean of the distribution* of X. It is a sort of central value for the distribution of X. A standard notation for is

$$E(X) = \mu_X.$$

When there is no ambiguity, we sometimes write EX instead of $E(X)$.

Example 3.10. Suppose that X take values -1, 0, 1 with probabilities 0.5, 0.3, 0.2 respectively.

The expected value of X is

$$E(X) = -1 \times 0.5 + 0 \times 0.3 + 1 \times 0.2 = -0.5 + 0.2 = -0.3.$$

[9] ■

[9] *Although X is always an integer, its mean is not an integer.*

Example 3.11. If the density function of X is $2e^{-2x}$ for $x > 0$, then the mean of X is

$$E(X) = \int_0^\infty x 2e^{-2x} dx = [-xe^{-2x} - e^{-2x}/2]_0^\infty = 0.5.$$

This confirms a special case of the mean for the exponential distribution on page 30.
■

Activity 3.7. Using (3.3) generalise Example 3.11 to show that the mean of the exponential distribution is $1/\lambda$.
■

Activity 3.8. Find the mean of the Bernoulli trial distribution, where $X = 0$ with probability $1 - \pi$ and $X = 1$ with probability π.
■

The mean of X can be thought of as a centre of gravity for the distribution for X. Thinking of a discrete random variable X, consider a uniform beam with a measurement scale marked on it. For all values x of X, arrange weights equal to $P(X = x)$ at scale point x on the beam. Then $E(X)$ is the point at which the weighted beam will balance on a pivot. It is the centre of gravity of the weights $P(X = x)$ at positions x.

Activity 3.9. What is the expected value of X if the only possible value of X is 0? (We have $P(X = 0) = 1$, so X is effectively a constant, even though it is called a random variable.)
■

Expected value of a function of a random variable

We can define the expected value for a function, say $h(X)$, of a random variable X. The definition (3.5) for discrete random variables becomes

$$E(h(X)) = \sum_x h(x) P(X = x), \tag{3.7}$$

whereas (3.6) for continuous random variables becomes

$$E(h(X)) = \int h(x) f_X(x) dx. \tag{3.8}$$

Example 3.12. Applying the result to the function $h(X) = X - E(X)$ we see easily that $E(X - E(X)) = 0$.
■

Example 3.13. Continuing Example 3.10, with $h(X) = X^2$

$$E(X^2) = (-1)^2 \times 0.5 + (0)^2 \times 0.3 + (1)^2 \times 0.2 = 0.5 + 0.3 = 0.8.$$

■

From (3.7) and (3.8), if a and b are constants then

$$E(a + bX) = a + bE(X). \tag{3.9}$$

This is an important rule to remember for routine evaluation of expected values. Another rule of great importance (but not so easy to prove) is that, if X, Y are two random variables,

$$E[X + Y] = E(X) + E(Y). \tag{3.10}$$

[10] *There is no difference in meaning between $E[X + Y]$ and $E(X + Y)$, but one form may be easier to read than the other*

Activity 3.10. Show, using equation (3.10) that $E[X - E(X)] = 0$.
■

Variance and standard deviation

The *variance*, $\mathrm{var}\, X$ of a random variable X is the average squared deviation of X from its mean. We write

$$\mathrm{var}\, X = E[(X - E(X))^2].$$

The variance is measured in squared units of measurement for X. It measures the spread of values of X around its mean, so that large values of the variance are from a widely spread distribution.

There is a standard notation for the variance of X:

$$\mathrm{var}\, X = \sigma_X^2.$$

To avoid ambiguity we sometimes write $\mathrm{var}(X)$ or $\mathrm{var}[X]$ instead of $\mathrm{var}\, X$.

A more interpretable measure of spread is the *standard deviation* of X which is the non-negative square root of the variance of X. The standard deviation is measured in the same units as X. The standard deviation of X is sometimes written σ_X.

It is often convenient to calculate a variance by using the result

$$\mathrm{var}\, X = E(X^2) - [E(X)]^2$$

This result is easy to prove:

$$\begin{aligned}\text{var}\, X &= E[(X - E(X))^2] \\ &= E[X^2 - 2XE(X) + (E(X))^2]\end{aligned}$$

so using (3.10) and (3.9)

$$\begin{aligned}\text{var}\, X &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - [E(X)]^2\end{aligned}$$

[11]

[11] *It is a common mistake not to distinguish between $E(X^2)$ and $[E(X)]^2$.*

Example 3.14. Continuing Examples 3.10 and 3.13, on pages 35, 36

$$\text{var}\, X = E(X^2) - [E(X)]^2 = 0.8 - (0.2)^2 = 0.8 - 0.04 = 0.76.$$

■

Activity 3.11. Continuing Activity 3.8 on page 35, show that the variance of the Bernoulli trial distribution is $\pi(1 - \pi)$.
■

Activity 3.12. Show that if $\text{var}\, X = 0$ then $P(X = \mu_X) = 1$. (We say in this case that X is **almost surely** equal to its mean.)
■

Activity 3.13. Find the variance of the exponential distribution with mean $1/\lambda$.
■

Learning outcomes

After working through this chapter you should be able to:

1. give the formal definition of a random variable, and distinguish between a random variable and the values it takes
2. explain the difference between continuous and discrete random variables.
3. discuss the basic distributions such as uniform, exponential normal, Poisson, binomial and calculate probabilities of events for such random variables
4. find the mean and the variance of simple random variables whether continuous or discrete
5. show how to prove and use simple properties of expected values and variances.

Sample examination questions

1. A company which manufactures drink dispensing machines sets the fill level at 198cc. The standard deviation is 4cc. Assuming that the fill levels have a normal distribution,

   (a) What proportion of drinks will have less than 195cc?

   (b) What is the probability that a random sample of 50 drinks has a mean value greater than 199cc?

   (c) The company claims that an average drink is 200cc. What percentage of the sample means is 200cc or more if samples of size 36 are taken?

   (d) explain briefly why you would or would not buy this dispensing machine.

   (Elements of Statistics 1998 Zone A)

2. Suppose that X has a Poisson distribution with mean λ.

   (a) Find by summation the mean of X.

   (b) Find also the variance of X.

   (Elements of Statistics 2001 Zone A)

3. The distribution of random variable X has density function

   $$f_X(x) = 1/3$$

   where $-1 < x < 2$.

   (a) Find by integration the mean of X.

   (b) Find also the variance of X.

   What is the $P[X > 1|X > 0]$? (Elements of Statistics 2001 Zone B)

4. If W is a Poisson random variable with mean 2, what is $P(W > 3|W > 1)$? (Elements of Statistics 2001 Zone B)

5. X is a random variable with $P(X = 0) = 0.1$, $P(X = 1) = 0.3$, $P(X = 2) = 0.4$. X can also take the value of 3, but no other values. What is $E[X^2]$? (Elements of Statistics 2000 Zone B)

6. If $x_1 = 3, x_2 = 2, x_3 = 4, x_4 = 2, x_5 = 5$, and all are equally likely values for X, what is $E[X(X - 1)]$? (Elements of Statistics 2000 Zone A)

Chapter 4

Bivariate distributions

How to think about two random variables together. Marginal and conditional distributions. Covariance and correlation.

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], section 4.4.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], a little material in 11.1.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods.* (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], not much in this book.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 5.

Introduction

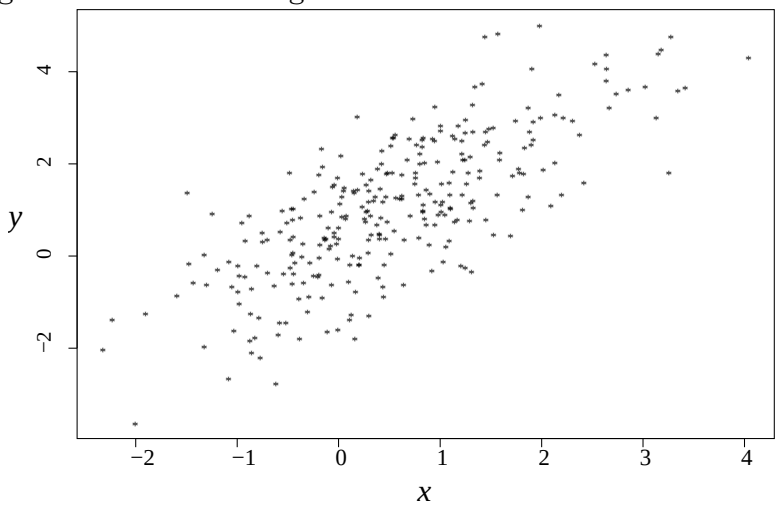
Almost all applications of statistical methods deal with several measurements on the same, or connected items. To think statistically about several measurements on a randomly selected item one must understand some of the concepts for joint distributions for random variables. This chapter looks at distributions for two random variables. *bivariate means **two** variables*. [1] It is fairly easy to extend most of the ideas to more than two. The idea of covariance, and the results on linear combinations of two random variables are fundamental for all further work.

[1] *Two random variables are more exciting to think about than one. New ideas are needed. One must get used to thinking about scatter plots and graphs.*

Two random variables

It is easy to set up the framework for this chapter. We have two random variables X and Y, which we think about together. Each outcome ω in the sample space gives simultaneously a pair of values $(X(\omega), Y(\omega)) = (x, y)$ for the two random variables.

Figure 4.1: A scatter diagram of values for two random variables



One can think of the pair of values (x, y) as plotted on a scatter diagram. Repetition of the random experiment leads to a cloud of points in that diagram, as each outcome gives a new pair of values (x, y). Figure 4.1 shows a typical scatter diagram for values of independent observations of two continuous random variables. The continuous nature of the variables forces each pair of values to be distinct from every other. Looking at Figure 4.1, two natural and fundamental questions arise:

- What do the values of X look like if we forget about Y (and vice-versa)?
- What do the values of Y look like just for those pairs of values for which X is, say, 1.2?

Activity 4.1. How would the scatter diagram change if the random variables were discrete rather than continuous?
■

If we look just at the values of X forgetting about Y, then we are looking at the univariate distribution of X, just as in Chapter 3. Because this univariate distribution is thought of as coming from the joint distribution of X and Y it is called a *marginal distribution*[2] for X.

If we hold X at, say, 1.2 and think about the values for Y, we are looking at a *conditional distribution* of Y given $X = 1.2$.[3]

Bivariate distributions are usually described by the joint probability function if they are discrete,

$$p(x, y) = P[(X, Y) = (x, y)]$$

[2] *There is much quaint antique terminology in use in statistics.*

[3] *Conditioning on $X = 1.2$ is an idealised concept for continuous X.*

or by a joint density function $f_{X,Y}(x, y)$ which is a non-negative and such that

$$P(a < X \leq b \cap c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)dxdy.$$

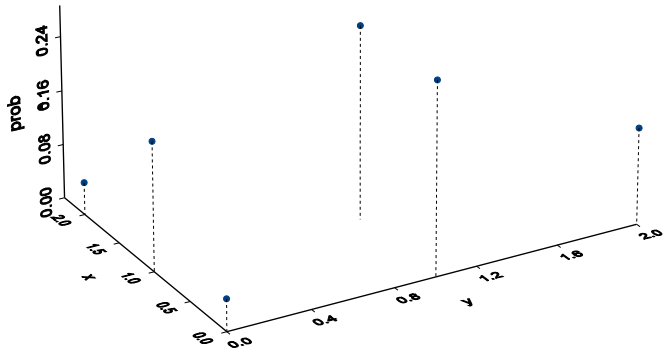
Example 4.1. Let us first look at a simple example of a discrete bivariate distribution. Suppose that a box contains 3 red balls, 2 blue balls and 2 green balls. We select two balls at random and without replacement. We will look at the bivariate distribution of X the number of red balls chosen, and Y the number of blue balls chosen.

Only nine different pairs of values for (X, Y), namely $(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (2, 0)$ can have probability > 0 . All the different pairs can be arranged as the cells in a 3×3 table, each cell being filled with the probability of that pair of values occurring. The result is shown at Table 4.1, and graphically in Figure 4.2.

Table 4.1: Probabilities for joint distribution of (X, Y) in Example 4.1

		Y		
		0	1	2
	0	1/21	4/21	1/21
X	1	6/21	6/21	0
	2	3/21	0	0

Figure 4.2: Probabilities for Example 4.1



Adding across the rows of Table 4.1 gives the marginal distribution of X. For instance,

$$P(X = 0) = P[(X, Y) = (0, 0)] + P[(X, Y) = (0, 1)] + P[(X, Y) = (0, 2)]$$

$$= 1/21 + 4/21 + 1/21 = 6/21 = 2/7.$$

One can write the calculation of the marginal probabilities as a formula

$$P[X = x] = \sum_y P[X = x, Y = y],$$

and similarly

$$P[Y = y] = \sum_x P[X = x, Y = y].$$

[4] *One can similarly find a conditional distribution for Y when $X = 0$ and when $X = 2$.*

The conditional distribution of Y given $X = 0$ is found[4] from the probabilities in the row for $X = 0$ of Table 4.1. Each probability is divided by the row total $P(X = 0) = 6/21 = 2/7$ so that the row sum is 1. So:

$$P(Y = 0|X = 0) = P(X = 0, Y = 0)/P(X = 0) = (1/21)/(6/21) = 1/6,$$

$$P(Y = 1|X = 0) = P(X = 0, Y = 1)/P(X = 0) = (4/21)/(6/21) = 4/6 = 2/3,$$

$$P(Y = 2|X = 0) = P(X = 0, Y = 2)/P(X = 0) = (1/21)/(6/21) = 1/6.$$

■

Activity 4.2. Write down the marginal distribution of Y, and the conditional distributions of X given Y.

■

Activity 4.3. Nothing in principle stops us from thinking about the joint distribution of (Y, Y), though this is a fairly pointless thing to do except for consistency of approach. Suppose that Y is as in Example 4.1. Write down the table of probabilities for the joint distribution (Y, Y).

■

Example 4.2. Show that the marginal distributions of a bivariate distribution are not enough to fix the bivariate distribution itself.

Answer

Here we must show that there are two distinct bivariate distributions with the same marginal distributions. It is easiest to think of the simplest case where X and Y each take only two values, say 0, 1.

Suppose the marginal distributions for X and Y are the same, with $p(0) = p(1) = 1/2$. Then one possible bivariate distribution with these margins is the one for which there is independence[5] between X and Y. This has $p(x, y) = p(x)p(y)$ for all x, y. Writing it in full:

[5] *See the next section.*

$$p(0, 0) = p(1, 0) = p(0, 1) = p(1, 1) = 1/2 \times 1/2 = 1/4.$$

The table of probabilities for this choice of independence is shown in Table 4.2.

Table 4.2: Probabilities for a independent X and Y in Example 4.2

		Y	
		0	1
	0	1/4	1/4
X	1	1/4	1/4

Table 4.3: Probabilities for non-independent X and Y in Example 4.2

		Y	
		0	1
	0	0.2	0.3
X	1	0.3	0.2

Trying some other value for $p(0,0)$, like 0.2, gives Table 4.3.

The construction of these probabilities is done by making sure the row and column totals are equal to 0.5, and so we now have a second distribution with the same marginal distributions as the first.

This example is very simple, but one can almost always construct many bivariate distributions with the same marginal distributions even for continuous random variables.

■

Independence

We say that discrete random variables X and Y are *independent*[6] if for **all** pairs of values (x, y)

$$P[X = x, Y = y] = P[X = x]P[Y = y]. \tag{4.1}$$

It follows from (4.1) that for all x, y

$$P(X \leq x \cap Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y). \tag{4.2}$$

Equation (4.2) can act as a definition for the independence of X, Y whether those variables are continuous or discrete, but we usually prefer in the case of continuous independent variables to characterise independence through the property that one can find a joint density function $f_{X,Y}(x, y)$ for (X, Y) such that for all x, y

$$= f_X(x)f_Y(y)$$

Activity 4.4. Why isn't (4.1) any good for continuous variables?

■

[6] *Notice that this is a precise technical definition for a concept used much more imprecisely in ordinary speech.*

Example 4.3. Referring back to Example 4.1 on page 41 we can see immediately that X and Y are not independent because, for instance,

$$P[X = 0, Y = 1] = 4/21 \neq 2/7 \times 10/21 = P[X = 0]P[Y = 1].$$

On the other hand if we look at the different joint distribution for (X, Y) in Table 4.4, we can check that the independence property does hold.

Table 4.4: Probabilities for joint distribution of (X, Y) for independence

		Y		
		0	1	2
	0	$20/147$	$20/147$	$2/147$
X	1	$40/147$	$40/147$	$4/147$
	2	$10/147$	$10/147$	$1/147$

■

Slightly rearranging (4.1) gives

$$P[X = x, Y = y]/P[Y = y] = P[X = x].$$

The left hand side is $P[X = x|Y = y]$ giving an alternative definition of independence through

$$P[X = x|Y = y] = P[X = x]$$

which must hold for all values (x, y) such that $P[Y = y] \neq 0$. One may interchange the roles of X and Y in this second definition of independence, which seems at first sight to be asymmetric in the two random variables.

Activity 4.5. Show that if $P(\{X \leq x\} \cap \{Y \leq y\}) = (1 - e^{-x})(1 - e^{-y})$ for all $x, y > 0$, then X and Y are independent random variables, each with an exponential distribution. (Hint: use (4.2).)
■

Expected values

There is not much new about expected values for the case of two random variables distributed jointly. The expected value of a function $h(X, Y)$ of two random variables is defined for discrete random variables by

$$E[h(X, Y)] = \sum_x \sum_y h(x, y)P[X = x, Y = y],$$

and for continuous random variables by

$$E[h(X,Y)] = \int_x \int_y h(x,y) f(x,y) dx dy$$

where $f(x,y)$ is the joint density function of (X,Y) (which integrates to 1 over all values (x,y)).

Example 4.4. Of course, if $h(X,Y)$ is really a function of X alone, say $h(X,Y) = h(X)$, then we get back the definition of the previous chapter (3.7) because

$$\begin{aligned} E[h(X,Y)] &= \sum_x \sum_y h(x,y) P[X=x, Y=y] \\ &= \sum_x \sum_y h(x) P[X=x, Y=y] \\ &= \sum_x h(x) \sum_y P[X=x, Y=y] \\ &= \sum_x h(x) P[X=x] \\ &= E[h(X)]. \end{aligned}$$

■

Properties of expected values

From Example 4.4 and (3.9) it follows that the expected value of a constant c is c.

If X and Y are **independent**, the expected value $E[h(X)g(Y)]$ of the product of a function $h(X)$ and a function $g(Y)$ is the product $E[h(X)]E[g(Y)]$ of the expected values.[7] This follows for discrete variables because

$$\begin{aligned} E[h(X)g(Y)] &= \sum_x \sum_y h(x)g(y) P[X=x, Y=y] \\ &= \sum_x \sum_y h(x)g(y) P[X=x] P[Y=y] \\ &= \sum_x h(x) P[X=x] \sum_y g(y) P[Y=y] \\ &= E[h(X)] E[g(Y)]. \end{aligned}$$

The proof looks very similar for continuous distributions, but uses the joint density function and integrals instead of sums.

[7] *This result is not true without the assumption of independence, or some other simplifying assumption.*

Covariance

The *covariance*, written σ_{XY} or $\mathrm{cov}(X, Y)$, of two jointly distributed random variables (X, Y) is defined by

$$\sigma_{XY} = \mathrm{cov}(X, Y) = E[(X - E(X))(Y - E(Y))].$$

That covariance is positive when X and Y have a direct association and negative when they are inversely related. Obviously, $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$. If X and Y are independent, then $\mathrm{cov}(X, Y) = 0$, for

$$\begin{aligned}\mathrm{cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[X - E(X)]E[Y - E(Y)] \\ &= 0 \times 0 = 0.\end{aligned}$$

If the covariance of X, Y is zero, then we say that X and Y are *uncorrelated*.[8]

[8] *Note that this is not the same as saying that they are independent.*

Activity 4.6. There are other ways to write the covariance. Show that

$$\mathrm{cov}(X, Y) = E[XY] - E[X]E[Y],$$

and

$$\mathrm{cov}(X, Y) = E[(X - E(X))Y] = E[X(Y - E(Y))].$$

■

Covariances are useful to find the variances for sums of random variables. A general result is that for constant a, b

$$\mathrm{var}(aX + bY) = a^2 \,\mathrm{var}(X) + 2ab \,\mathrm{cov}(X, Y) + b^2 \,\mathrm{var}(Y).$$

This is easy to see, because the right-hand side is

$$a^2 E[(X - E(X))^2] + 2abE[(X - E(X))(Y - E(Y))] + b^2 E[(Y - E(Y))^2]$$

which is

$$E[a^2(X - E(X))^2 + 2ab(X - E(X))(Y - E(Y)) + b^2(Y - E(Y))^2].$$

This simplifies to

$$\begin{aligned}E[(a(X - E(X)) + b(Y - E(Y)))^2] &= E[(aX + bY - E(aX + bY))^2] \\ &= \mathrm{var}(aX + bY).\end{aligned}$$

There are important special cases: $a = 1, b = 1$ gives

$$\mathrm{var}(X + Y) = \mathrm{var}\, X + 2\,\mathrm{cov}(X, Y) + \mathrm{var}\, Y,$$

while $a = 1, b = -1$ gives

$$\text{var}(X - Y) = \text{var}\, X - 2\,\text{cov}(X, Y) + \text{var}\, Y.$$

Notice that if the covariance of X and Y is positive, then

$$\text{var}(X + Y) > \text{var}\, X + \text{var}\, Y$$

and if the covariance of X and Y is negative, then

$$\text{var}(X + Y) < \text{var}\, X + \text{var}\, Y.$$

If the covariance of X and Y is zero, then

$$\text{var}(X + Y) = \text{var}(X - Y) = \text{var}\, X + \text{var}\, Y.$$

In particular if X and Y are **independent** then the variance of the sum of X and Y is the sum of their variances.

Notice that we could **define** covariance through

$$\text{cov}(X, Y) = [\text{var}(X + Y) - \text{var}(X - Y)]/4.$$

This alternative definition has the advantage of making it clear from the known properties of variances that if (X, Y) is independent of (W, V), which implies that $X + Y$ is independent of $W + V$, and that $X - Y$ is independent of $W - V$, then

$$\text{cov}(X + W, Y + V) = \text{cov}(X, Y) + \text{cov}(W, V). \tag{4.3}$$

Notice also that

$$\text{cov}(X, X) = \text{var}(X).$$

Although the covariance between X and Y gives some idea of their dependence, usually one prefers to use the *correlation coefficient* ρ to measure dependence. This is defined as

$$\rho = \text{cov}(X, Y)/\sqrt{\text{var}\, X\, \text{var}\, Y}.$$

This is unaffected by either the location or the scale of measurement of X, Y and always lies between -1 and 1.

Activity 4.7. Suppose that $\text{var}(X) = \text{var}(Y) = 1$, and that X and Y have correlation coefficient ρ. Show that it follows from $\text{var}(X - \rho Y) \geq 0$ that $\rho^2 \leq 1$.
■

Example 4.5. The distribution of a random variable X is as below:

x	-1	0	1
$P(X = x)$	a	b	a

Show that X and X^2 are uncorrelated.

Answer

This is an example of two random variables X and $Y = X^2$ that are uncorrelated, but obviously dependent. The bivariate distribution of X, Y in this case is *singular* because of the complete functional dependence between them.

$$\begin{aligned} E(X) &= -1 \times a + 0 \times b + 1 \times a = 0, \\ E(X^2) &= +1 \times a + 0 \times b + 1 \times a = 2a, \\ E(X^3) &= -1 \times a + 0 \times b + 1 \times a = 0. \end{aligned}$$

We must show that the covariance is zero.

$$\begin{aligned} \text{Covariance}(X, Y) &= E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2) \\ &= 0 - 0 \times 2a \\ &= 0. \end{aligned}$$

There are many possible choices for a, b that give a valid probability distribution, for instance $a = 0.25, b = 0.5$.

■

Example 4.6. A fair coin is thrown n times, each throw being independent of the ones before. Put R = the number of heads, and S = the number of tails. Find the covariance of R and S. What is the correlation of R and S ?

Answer

One can go about this in a straightforward way. If X_i is the number of heads, and Y_i is the number of tails, on the ith throw then the distribution of X_i and Y_i is given by

Y_i	0	1
X_i		
0	0	0.5
1	0.5	0

$$\begin{aligned} E(X_i) &= 0 \times 0.5 + 1 \times 0.5 = 0.5 = E(Y_i) \\ E(X_i^2) &= 0 \times 0.5 + 1 \times 0.5 = 0.5 \\ &= E(Y_i^2) \end{aligned}$$

$$\operatorname{var} X_i = 0.5 - (0.5)^2 = 0.25 = \operatorname{var} Y_i$$

$$E(X_i Y_i) = 0 \times 0.5 + 0 \times 0.5 = 0$$

$$\text{cov}(X_i, Y_i) = E(X_i Y_i) - E(X_i)E(Y_i) = 0 - 0.25 = -0.25.$$

Now, since $R = \sum X_i$ and $S = \sum Y_i$, and we can add covariances of independent X_is and Y_is, just like means and variances (see (4.3)), then

$$\text{cov}(R, S) = -0.25n.$$

Since $R + S = n$ is a fixed quantity, there is a complete linear dependence between R and S. We have $R = n - S$. So the correlation between R and S should be -1. This can be checked directly, since

$$\text{var}\, R = 0.25n = \text{var}\, S$$

(add the variances of X_is or Y_is). The correlation between R and S works out as $-0.25n/0.25n = -1$.

■

Example 4.7. Suppose that X and Y have a bivariate distribution. Find the covariance of the new random variables $W = aX + bY$, $V = cX + dY$ where a, b, c and d are constants.

Answer

Covariance of W and V is

$$\begin{aligned} E(WV) - E(W)E(V) &= E(acX^2 + bdY^2 + \{ad + bc\}XY) \\ &\quad + (acE(X)^2 + bdE(Y)^2 + \{ad + bc\}E(X)E(Y)) \end{aligned}$$

$$\begin{aligned} &= ac(E(X^2) - E(X)^2) + bd(E(Y^2) - E(Y)^2) \\ &\quad + \{ad + bc\}(E(XY) - E(X)E(Y)) \end{aligned}$$

$$= ac\sigma_X^2 + bd\sigma_Y^2 + \{ad + bc\}\sigma_{XY}.$$

■

Example 4.8. Following on from the last example, show that if the variances of X and Y are the same, then $W = X + Y$ and $V = X - Y$ are uncorrelated.

Answer

Here we have $a = b = c = 1$, $d = -1$. Substituting in the formula found for the last example,

$$\sigma_{WV} = \sigma_X^2 - \sigma_Y^2 = 0.$$

There is no assumption here that X and Y are independent. It is not true that W and V are independent without further restriction on X, Y.

■

Learning outcomes

After working through this chapter you should be able to:

1. show what a scatterplot is
2. put the probabilities for a discrete bivariate distribution in a table
3. define marginal and conditional distributions, and find them for a discrete bivariate distribution
4. know how to define and check for the independence of two random variables
5. define and work out expected values for functions of two random variables and know how to prove simple properties for the expected values
6. give the definition of covariance and correlation for two random variables and calculate covariance and correlation for a discrete bivariate distribution
7. show how to prove and apply the formulae for the covariance of the sum of two random variables.

Sample examination questions

1. X and Y are random variables with Normal distributions with mean 0, variance 1 and correlation coefficient 0.5. What is $P(X+Y>2)$? Assume that $X+Y$ has a normal distribution. (Elements of Statistics 2001 Zone B)
2. X and Y are independent random variables with Normal distributions with mean 0 and variance 1. For some choice of $c>0$ $P(X+cY>4.2732)=0.15$. What is c?
3. Prove that
$$\operatorname{var}(X+Y)=\operatorname{var}X+\operatorname{var}Y+2\operatorname{cov}(X,Y).$$
(Elements of Statistics 1998 Home)
4. The distribution of (X,Y) is specified in the following table:

X	Y	Probability
1	6	1/3
2	5	1/3
3	4	1/3

Find the correlation coefficient of X, Y.

(Elements of Statistics 1998 Home)
5. Why is a correlation coefficient used to measure linear association rather than covariance?

6. Find the correlation coefficient of X and X^2 where X is a binomial random variable from 3 trials with probability of success 0.5. (Elements of Statistics 1997 Overseas)

Chapter 5

The distribution of statistics over repeated random samples.

Sampling distributions

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 6.

Further reading

Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], chapter 7.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], parts of chapter 8, but the presentation is organised differently in this book.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], 5.1 and 5.2.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 6.

Introduction

We now work on the topic that makes statistics a distinct subject. The idea of sampling and of a sampling distribution for a statistic such as the mean must be understood by all users of statistics. Students should do a few Monte Carlo sampling experiments themselves to get a good intuitive feel for sampling distributions, because the concept is a bit slippery. For those fortunate enough to have access to a computer, there is software available to help. See for instance D. P. Doane, K. Mathieson and R. L. Tracy, *Visual Statistics 2.0* (Irwin McGraw-Hill, 2000) [ISBN 072400943]

Mean and variance of a sample mean

We first recall some things you will have seen already in Statistics 1.

The sampling distribution of the sample mean over random samples taken either

with or without replacement has a mean equal to the population mean μ. That is

$$E\bar{X} = EX = \mu. \tag{5.1}$$

The variance of the sampling distribution of the sample mean in random samples of size n taken **without** replacement from a population of size N is given by

$$\text{var}\,\bar{X} = \frac{\sigma^2}{n}\frac{N-n}{N-1}, \tag{5.2}$$

where σ^2 is the population variance. The variance of the sampling distribution of the sample mean in random samples of size n taken **with** replacement from a population of size N is given by

$$\text{var}\,\bar{X} = \frac{\sigma^2}{n}. \tag{5.3}$$

Notice that for samples of size greater than 1, the second variance is greater than the first. The most important consequence of (5.2) and (5.3) is that for sample size n greater than 1, the variance of the sample mean is less than the variance of a single observation. This implies that we can get a better idea of the population mean from a sample mean than from a single observation.[1]

[1] *Non-statisticians often use sample means but rarely know why they do so: here is one reason.*

Often, (5.3) is given in the form

$$\text{standard deviation}\ \bar{X} = \frac{\sigma}{\sqrt{n}},$$

and the special name *standard error* is given to this standard deviation of a sample mean.[2]

[2] *Be careful to distinguish between standard error and standard deviation. The latter is more generally used.*

Activity 5.1. Continuing Activities 3.8 and 3.11, use (5.1) and (5.3) to show that the mean and variance of the binomial distribution in (3.1) (which is the same as the sum of n independent Bernoulli trials) has mean $n\pi$ and variance $n\pi(1-\pi)$.
■

Activity 5.2. From the results in 5.1, by fixing $\lambda = n\pi$, and allowing n to increase, and π to get small, discover that the mean and variance of a Poisson distribution are both λ.
■

Sampling from a normal population

The normal distribution is often used as a sort of idealised version of the distribution of a random variable of interest X over some (very large, effectively infinite) population. Of course, a continuous random variable must of necessity be an idealised representation for anything in our naturally discrete world.

The normal distribution has the advantage that it has very simple properties. For instance if X has an $N(\mu_x, \sigma_x^2)$ distribution and an independent random variable Y has an $N(\mu_y, \sigma_y^2)$ distribution then $X + Y$ has an

$$N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2) \tag{5.4}$$

distribution.[3]

[3] *This is hard to prove, but it is a very important result, which you should remember.*

One of the most important properties of the normal distribution, which follows from this result, is that the sampling distribution of the sample mean $\bar{X}$ of a random sample of size n from a normal population $N(\mu, \sigma^2)$ is **exactly** $N(\mu, \sigma^2/n)$. This is a precise mathematical result, which goes farther than the results for $E\bar{X}$ and $\text{var}\,\bar{X}$ obtained earlier, though it is consistent with them. One can find exact probabilities for the sampling distribution of $\bar{X}$ in this normal case.

Example 5.1. Suppose that the heights of students are normally distributed with a mean of 68.5 inches and a standard deviation of 2.7 inches. If 200 random samples of size 25 are drawn from this population and their means recorded to the nearest tenth of an inch, determine:

1. The expected mean and standard deviation of the sampling distribution of the mean.
2. The expected number of recorded sample means that fall between 67.9 and 69.2 inclusive.
3. The expected number of recorded sample means falling below 67.0.

Answer

1. The sampling distribution of the mean of 25 observations has the same mean as the population, which is 68.5 inches, and standard deviation from (5.3) of $2.7/5 = 0.54$.
2. The samples are random, so one can't be sure just how many will have means recorded between 67.9 and 69.2 inches. One can work out the probability that a recorded mean will so lie, from the sampling distribution of the sample mean which is normal with mean 68.5 and standard deviation 0.54. That probability is the probability that the sample mean lies between 67.85 and 69.25, allowing for the crude recording of the means. The corresponding z values are

$$(67.85 - 68.5)/0.54 = -1.20 \text{ and } (69.25 - 68.5)/0.54 = 1.39.$$

Table 4 of the *New Cambridge Statistical Tables* has the left-hand tail probabilities for the positive z values. For $z = 1.39$, the left-hand tail probability is

0.9177. For $z = 1.20$ the left-hand tail probability is 0.8849, so the left-hand tail probability for $z = -1.20$ is $1 - 0.8849 = 0.1151$. The probability between the two z values is

$$0.9177 - 0.1151 = 0.8026.$$

Since there are 200 samples drawn, you can now think of each as a single trial. The recorded mean lies between 67.9 and 69.2 with probability 0.8026 at each trial. We are dealing with a binomial distribution with $n = 200$ trials and probability of success $\pi = 0.8026$. The expected number of successes is

$$n\pi = 200 \times 0.8026 = 160.52.$$

3. Similarly, the probability that a recorded mean lies below 67.0 is the probability that the sample mean lies below 66.95. The z value is $(66.95 - 68.5)/0.54 = -2.87$. We want the left-hand tail probability for $z = -2.87$, which is the right-hand tail probability for $z = 2.87$. From Table 4 of the *New Cambridge Statistical Tables* this is 0.00205. So the expected number of sample means out of 200 recorded below 67.0 is $200 \times 0.00205 = 0.41$.

■

The Central Limit Theorem

A useful result from the mathematical theory for distributions says roughly that for a large enough sample size n, the sampling distribution of the sample mean $\bar{X}$ from a random sample of size n without replacement from a population of values for X is close to the normal distribution $N(\mu, \sigma^2/n)$, where the population values of X have mean μ and variance σ^2.

This is an approximate version of the precise result for samples from $N(\mu, \sigma^2)$, but holding much more generally, for the population is not restricted to be normal.[4] To be more precise, the theorem says that, for each fixed value x,

[4] *We do, however assume it has a mean and a variance.*

$$P\left[\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right] \to \Phi(x)$$

as $n \to \infty$.

We could apply the Central Limit Theorem to get approximately the same results for Example 5.1 even if the population of values of X were not normal.

Example 5.2. We can use the central limit theorem to understand (5.4). Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables each with mean μ_x and variance σ_x^2, and that $Y_1, Y_2, \ldots, Y_n$ are independent random variables each with mean μ_y and variance σ_y^2. Then applying the central limit theorem we see that as n gets large, the distribution of $\bar{X}$ gets close to $N(\mu_x, \sigma_x^2/\sqrt{n})$, and that the distribution of

$\bar{Y}$ gets close to $N(\mu_y, \sigma_y^2/\sqrt{n})$. On the other hand, we can see that $\bar{X} + \bar{Y}$ is the mean of $X_i + Y_i$ for $i = 1, \ldots, n$. Since $X_i + Y_i$ has mean $\mu_x + \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$, the Central Limit Theorem tells us that as n gets large $\bar{X} + \bar{Y}$ has a $N(\mu_x + \mu_y, (\sigma_x^2 + \sigma_y^2)/\sqrt{n})$ distribution.
■

Application to the binomial distribution

A binomial random variable Y coming from n independent trials, each with probability of success π, may be thought of as

$$Y = \sum_{i=1}^{n} X_i,$$

where X_i is a random variable taking the value 1 if the ith trial is a success and 0 if the ith trial is a failure. Since the trials are independent, so are the random variables X_i. The proportion of successes is $\bar{X}$. The collection $\{X_1, X_2, \ldots, X_n\}$ is a random sample with replacement from the population that has the value 1 with probability π and value 0 with probability $1 - \pi$. This population has[5] mean π and variance $\pi(1 - \pi)$, so the mean of the proportion of successes in n trials is π, and the variance is $\pi(1 - \pi)/n$.

[5] *You should be able to verify the value for the mean and the variance.*

The Central Limit theorem says that for large n, the proportion of successes has approximately a normal distribution with mean π and variance $\pi(1-\pi)/n$. In practice to use this result one makes a *continuity correction* that improves the approximation. This adjusts the integer values of the number of successes by 0.5, to allow for the fact that the normal approximation will put the probability for r successes around both sides of r, say round about from $r - 0.5$ to $r + 0.5$.

Example 5.3. Compare the normal distribution approximation to the exact values for the right-hand tail probabilities for the binomial distribution with 100 trials and probability of success 0.1.

Answer

r	$P(R \geq r)$	$z = \frac{r-0.5-10}{3}$	$P(Z > z)$
1	0.999973	−3.1667	0.999229
2	0.999678	−2.8333	0.997697
3	0.998055	−2.5000	0.993790
4	0.992164	−2.1667	0.984870
5	0.976289	−1.8333	0.966624
6	0.942423	−1.5000	0.933193
7	0.882844	−1.1667	0.878327
8	0.793949	−0.8333	0.797672
9	0.679126	−0.5000	0.691462
10	0.548710	−0.1667	0.566184
11	0.416844	0.1667	0.433816
12	0.296967	0.5000	0.308538
13	0.198179	0.8333	0.202328
14	0.123877	1.1667	0.121673
15	0.072573	1.5000	0.066807
16	0.039891	1.8333	0.033376
17	0.020599	2.1667	0.015130
18	0.010007	2.5000	0.006210
19	0.004581	2.8333	0.002303
20	0.001979	3.1667	0.000771
21	0.000808	3.5000	0.000233
22	0.000312	3.8333	0.000063
23	0.000114	4.1667	0.000015
24	0.000040	4.5000	0.000003
25	0.000013	4.8333	0.000001
26	0.000004	5.1667	0.000000
27	0.000001	5.5000	0.000000
28	0.000000	5.8333	0.000000

The second column gives the exact binomial probabilities that the number of successes is greater than or equal to the numbers, r, in the first column. It will be found to fewer decimal places in the *New Cambridge Statistical Tables*. The third column gives the corresponding z values for the standard normal approximation to the binomial. The required mean is $100(0.1) = 10$, and the variance is $100(.1)(.9) = 9$. The z value for $r = 1$ is $(1 - 0.5 - 10)/3$, which is -3.1667. The continuity correction 0.5 means that effectively the value $r = 1$ is changed to $r = 0.5$, because we want the probability of number of successes greater than or equal to r. The fourth column gives

the right-hand normal probabilities corresponding to the z values in column three. You can find these from the *New Cambridge Statistical Tables* , but not so accurately.

Although the agreement between columns two and four is not too bad, you may think it is not as close as you would like for some applications. Much better approximations are available. The binomial distribution is asymmetric except for probability of success 0.5, and can't be very accurately approximated by the symmetric normal distribution using the approach above.

■

The approximation of the sampling distribution of a sample mean by a normal distribution can be surprisingly accurate even for small sample sizes like $n = 10$ and when the population sampled is far from normal. However, it can be a poor approximation for a very skew population, or one with tails that fall off very slowly.

Learning outcomes

After working through this chapter you should be able to:

1. show how random sampling a population gives rise to a sampling distribution for a sample statistic
2. prove and apply the results for the mean and variance of the sampling distribution of the sample mean from a random sample with replacement
3. state and understand the Central Limit Theorem and have an intuitive grasp of when the limit is likely to provide a good approximation to the distribution of the sample mean
4. use the Central Limit Theorem to approximate a binomial distribution, including a continuity correction
5. define and to be able to derive the mean and variance of the binomial distribution.

Sample examination questions

1. Write down the sample space of samples of size two without replacement from the population of three persons A, B and C.
2. Show that the variance of the mean of a random sample of size n taken from a large population is equal to the population variance divided by the sample size. (Elements of Statistics 2000 Zone A)
3. Show that the binomial distribution with n trials and probability of success π has mean $n\pi$ and variance $n\pi(1-\pi)$. (Elements of Statistics 1999 Zone A)

How to guess about the population when you have a sample.

Chapter 6

Point estimation

Essential reading

Newbold, P. *Statistics for Business and Economics*. (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 7.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], chapter 6.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 7.

Introduction

In this chapter we begin to look at one of the distinguishing topics of statistics - how to make inferences about a population from a random sample taken from that population. The concepts are subtle and need time to get accustomed to. At first sight it looks very obvious what to do. It is tempting to think, for instance, that to estimate a population mean one should 'obviously' use the sample mean, and to estimate a population variance one should use the sample variance. We must do better than this as statisticians, and not allow promising similarities of name to distract from thought and understanding.

Let us think about estimating the mean of a population from which we have a random sample taken with replacement. Suppose we use the 'obvious' estimate - the sample mean. How is that value connected with the population mean? When we look at the value of the sample mean it is one observation selected at random from the sampling distribution of the random variable $\bar{X}$. This is all we can say about the sample mean - it is an observation from a sampling distribution.

Sampling distributions

[1] *Here is the first subtle point.*

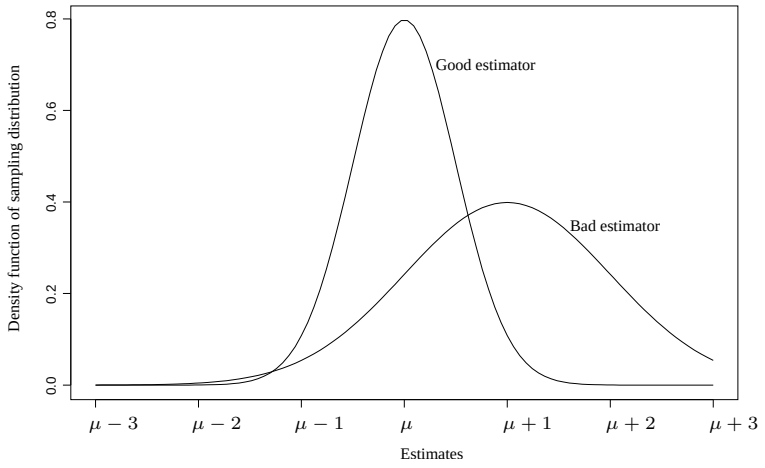
It follows that the only basis[1] for using the sample value $\bar{x}$ of $\bar{X}$ as an estimate of the population mean comes from properties of the sampling distribution of $\bar{X}$. We call the random variable $\bar{X}$ an *estimator* to distinguish it from its value for a particular sample which is the *estimate*. We must look for the properties of estimators rather than estimates and, to move to a more general setting, an estimator will provide for each sample an estimate of the population quantity of interest. We call population quantities of interest *parameters*, and label them with Greek letters such as θ.

Good estimators

[2] *Note that an estimate is just a number from a particular sample: it has no interesting properties.*

A good estimator (let's call it T) will be such that its sampling distribution is concentrated tightly around the population quantity of interest (let's call it θ).[2] A good estimator will, for most random samples, give estimates not too far from θ. Figure 6.1 shows the density functions of the sampling distributions of two estimators of a population mean μ.

Figure 6.1: A 'good' estimator and a 'bad' estimator



Bias, variance and mean squared error

To describe how closely the sampling distribution of an estimator T is concentrated around the parameter θ, we can use the mean and variance of the sampling distribution.

An estimator T is said to be *unbiased* for a parameter θ if

$$E[T] = \theta.$$

The sampling distribution of an unbiased[3] estimator T of the parameter θ has its mean at θ. On average, over all the possible random samples, it is centred at θ. The *bias* of an estimator T of parameter θ is defined as

$$\text{Bias}(T) = E[T] - \theta.$$

An *unbiased estimator* has bias zero.

[3] *Notice that there is no suggestion that every estimator should be unbiased.*

Example 6.1. The sample mean $\bar{X}$ is an unbiased estimator of the population mean μ because $E[\bar{X}] = \mu$. We could think of using $\bar{X} + 100$ as an estimator of μ, but this would have bias 100.
■

The variance of the sampling distribution of an estimator T gives some idea of how far from $E[T]$ the value of T is likely to be for a random sample. If the bias of T is small, then the variance will show how likely T is to be close to θ.

Example 6.2. The sample mean $\bar{X}$ from a random sample of size n has variance σ^2/n when the population variance is σ^2. As $\bar{X}$ is an unbiased estimator of the population mean μ, we see that as n increases we are more likely to find the value of $\bar{X}$ close to μ. Of course, $\bar{X} + 100$ has the same variance as $\bar{X}$, but this is an unsuitable estimator for μ because even for large n it becomes closely concentrated around $\mu + 100$.
■

A simple measure of closeness for an estimator T to the parameter θ is the *mean squared error* of T. This is defined by

$$\text{MSE}(T) = E[(T-\theta)^2] = \text{var}\, T + [\text{Bias}(T)]^2. \tag{6.1}$$

This measure combines variance and bias to give a composite measure that includes both. If an estimator is unbiased, its MSE is the same as its variance.

Activity 6.1. Prove the two expressions in 6.1 are equal.
■

Example 6.3. The sample mean $\bar{X}$ from a random sample of size n has variance σ^2/n when the population variance is σ^2. As $\bar{X}$ is an unbiased estimator of the population mean μ, $\text{MSE}(T) = \sigma^2/n$. The alternative estimator, $\bar{X} + 100$ has mean squared error $\sigma^2/n + 10000$. Since $\bar{X}$ has a smaller mean squared error than $\bar{X} + 100$ we prefer to use **that** to estimate μ.
■

Example 6.4. A precautionary example: if we **know for sure** that μ is between 1 and 2, and σ^2 is equal to 100 then the estimator T, which takes the value 1.5 for every sample, may have smaller MSE than the sample mean.
■

Activity 6.2. Work out why the above statement is true.
■

Here is another example to think about.

Example 6.5. A small business runs with four people. What is the best estimate of the average pay θ for the business that can be found from a sample of three people taken without replacement from the four workers?

Answer

This is a question with no clear answer. Unfortunately, using mean squared error will not always allow one to choose a best estimator. Let us compare the results for two different pay structures, and find for for these two structures an estimate better than the sample mean.

Suppose, for the moment, that each of the four workers is paid £40 per day. Then the sample mean $\bar{X}$ for the sample of size 3 will be equal to the population mean $\theta = £40$, so the sample mean has mean squared error zero. For this pay structure the sample mean can't be improved on, because it is always exactly right.

Now suppose that the boss is paid £130 and each of the three other workers is paid £10. The average pay for the business is still $\theta = £40$. There are four equally likely possible samples of size three taken without replacement. The samples give incomes $(130, 10, 10)$ three times and $(10, 10, 10)$ once. The sample means are 50 three times and 10 once. The mean squared error of the sample mean is

$$[3(50 - 40)^2 + (10 - 40)^2]/4 = 300.$$

On the other hand, one could use an estimate T that weights the smallest and largest observation half as much as the central observation. So each sample $(100, 10, 10)$ gives an estimate $(130/4 + 10/2 + 10/4) = 40$ whereas the sample $(10, 10, 10)$ gives an estimate $(10/4 + 10/2 + 10/4) = 10$. The estimator T has mean squared error

$$[3(40 - 40)^2 + (10 - 40)^2]/4 = 225.$$

This second estimator T is also exactly correct for the first pay structure considered, so if the only pay structures were the two considered, then T is a better estimator than the sample mean $\bar{X}$.

If other possible pay structures were taken into account, then it would become impossible to choose which of the two estimators was best - it would depend what the pay structure was like.
■

Activity 6.3. Find the mean squared error for T and $\bar{X}$ of Example 6.5 for a pay structure in which the boss is paid £100, and the three other workers £20.
■

Minimum variance unbiased estimators

If we confine attention to estimators that are unbiased for θ, we can sometimes find one estimator that has a smaller variance than all the others. This estimator is called the *minimum variance unbiased estimator* (MVUE) of θ. For instance, if we have a random sample from a population with a normal distribution the sample mean $\bar{X}$ is the MVUE of the population mean μ.

Learning outcomes

After working through this chapter you should be able to:

1. describe the performance of an estimator through its sampling distribution
2. use the concepts of bias and variance of an estimator
3. give the definition of mean squared error and calculate it for simple estimators
4. discuss the idea of a minimum variance unbiased estimator.

Sample examination questions

1. Write down the expected value of the square of the mean of a random sample in terms of the population mean and variance, and use this result to display an unbiased estimate of the square of the population mean based on the square of the sample mean and the sample variance. (Elements of Statistics 2001 Zone A)
2. Explain why we must consider both bias and variance when judging the performance of an estimator. (Elements of Statistics 2001 Zone B)
3. Give two reasons why, for a sample of size 10, one might not wish to use the sample range divided by 3.078 to estimate the population standard deviation, even though this estimate is unbiased for a random sample from a normal distribution. (Elements of Statistics 2001 Zone B)
4. Define the mean squared error of an estimator. (Elements of Statistics 2000 Zone A)
5. What is an unbiased estimator? (Elements of Statistics 2000 Zone B)

6. The mean of a random sample is an unbiased estimator of the population mean. Why do we prefer the mean from a sample of size 20 to the mean of a sample of size 10 when estimating the population mean. (Elements of Statistics 2000 Zone B)

How accurate is the estimate?

Chapter 7

Interval estimation

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 8.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], chapter 7.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods.* (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], parts of chapters 8, 9 and 10.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], chapter 9.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], 6.1 and parts of chapters 7 and 8.
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], some parts of chapters 6 and 9.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 8.

Introduction

The point estimates of Chapter 6 are very important, but they are usually supplemented in practice by some indication of their accuracy. We need to know whether the sample is likely to give us an estimate close to the population value. To tell us this we use *confidence intervals*.

We saw in Chapter 6 that one way to judge whether an estimator is good is to see how closely its sampling distribution lies around the population value. Suppose that our idea is now to give an interval estimator for a parameter θ of the form (T_l, T_u) where T_l and T_u give, for each sample, lower and upper bounds for θ. We shall be happy that (T_l, T_u) sets reasonable bounds for θ if there is a high probability that the joint distribution of T_l and T_u over repeated random samples is such that for say, 95%,

of the samples we have $T_l \leq \theta \leq T_u$. We say that the confidence interval (T_l, T_u) then **covers** θ with a probability of 95% over repeated samples. It is called a 95% confidence interval. Notice that it is the interval itself that is randomly changing over repeated samples, but that μ is a fixed unknown parameter value. It is clearer to say that the interval covers θ with probability 0.95 than to say that there is probability 0.95 that θ is in the interval, because we want to emphasise the variable nature of the interval and the fixed nature of θ.

Example 7.1. Suppose that we want to estimate the mean μ of a normal population with a variance known to be 1 from a random sample of size n. You can see that the interval $(\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n})$ is a 95% confidence interval for μ, because the probability that this interval covers μ is the same as the probability that $z = (\bar{X} - \mu)/(1/\sqrt{n})$ lies between -1.96 and 1.96.
■

Activity 7.1. Justify the last statement in Example 7.1.
■

Intervals for the mean of a normal population

Known variance

It is easy to generalise Example 7.1 to find a $100(1-\alpha)\%$ confidence interval for the mean μ of a normal population with known variance σ^2. We know that (see page 55)

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{7.1}$$

has a standard normal distribution. So if $z_{\alpha/2}$ is the upper $100\alpha/2\%$ point of the standard normal distribution, then

$$P\left[-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] = 100(1-\alpha)\%. \tag{7.2}$$

[1] *You should know from your mathematics course exactly how this works.*

Inverting this inequality leads immediately[1] to the coverage property wanted

$$P\left[\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right] = 100(1-\alpha)\%. \tag{7.3}$$

Notice that the interval is shorter as the sample size n increases, and longer for larger variances σ^2.

Activity 7.2. Why don't we always choose a very high confidence for the interval?
■

Unknown variance

Usually we don't know the variance, and must estimate it with the sample variance S^2 (here we use upper case S to show that it is thought of as a random variable). Just

using S in place of σ in (7.3) or equivalently (7.2) would not give the right coverage probabilities, because S is random, whereas σ is fixed. Fortunately there is an easy fix. Instead of (7.1) we can use the fact that

$$t_{(n-1)} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{7.4}$$

has a Student's t distribution with $n-1$ degrees of freedom, see (7.6) In just the same way as before we then get a $100(1-\alpha)\%$ confidence interval for μ from the coverage property

$$P[\bar{X} - t_{\alpha/2,(n-1)}S/\sqrt{n} \leq \mu \leq \bar{X} + t_{\alpha/2,(n-1)}S/\sqrt{n}] = 100(1-\alpha)\%. \tag{7.5}$$

A little more distribution theory

The χ^2 distribution

The sum of the squares of ν independent standard normal random variables has a χ^2 distribution with ν degrees of freedom, usually written $\chi^2_{(\nu)}$. This is a thoroughly investigated family of distributions which has been tabulated. The mean of the distribution is ν and the variance 2ν. The cumulative distribution function is tabulated in Table 7 of the *New Cambridge Statistical Tables* , and percentage points appear in Table 8. You should make yourself familiar with Table 8. A few of the density functions are shown in Figure 7.1.

It is a remarkable fact that if $Y = (n-1)S^2/\sigma^2$ (where S^2 is the sample variance from a random sample of size n from a normal distribution with mean μ and variance σ^2), then Y has a $\chi^2_{(n-1)}$ distribution. Furthermore Y is independent of the mean $\bar{X}$ of that random sample.

Student's t distribution

If Z has a standard normal distribution, and W is an independent random variable with a $\chi^2_{(\nu)}$ distribution then

$$\frac{Z}{\sqrt{W/\nu}} \tag{7.6}$$

has a Student's t distribution with ν degrees of freedom[2]. Student's t distribution has percentage points in Table 10 of the *New Cambridge Statistical Tables* . For $\nu \geq 30$ the Student's t distribution is almost indistinguishable from the standard normal distribution. Pictures of a few density functions appear in Figure 7.2.

[2] *The ν tells us which Student's t distribution to use; 'degrees of freedom' is a charmingly old-fashioned name for it.*

Figure 7.1: Density functions for χ^2 distributions

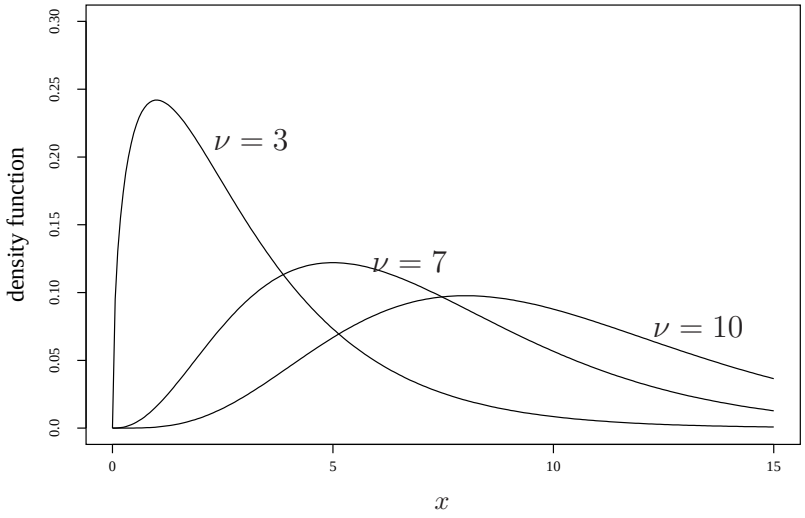
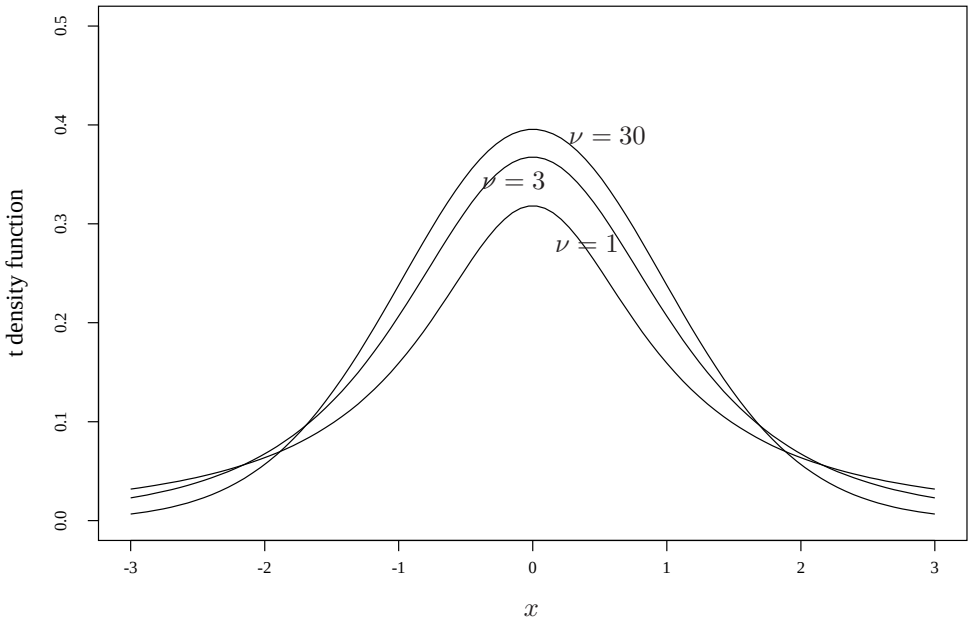


Figure 7.2: Density functions for Student's t distributions



The Student's t distribution for the expression in (7.4) follows from the properties given for the χ^2 distribution and (7.6).

Activity 7.3. Prove the last statement.
■

Because Student's t and the standard normal distribution are nearly the same for $\nu \geq 30$, one need not use the Student's t distribution for (7.4) for large sample sizes, but may replace it by the standard normal distribution just as one would if the variance were exactly known.

Example 7.2. Suppose that 9 bags of salt granules are selected from the supermarket shelf at random and weighed. The weights in grams are 812.0, 786.7, 794.1, 791.6, 811.1, 797.4, 797.8, 800.8 and 793.2. Give a 95% confidence interval for the mean of all the bags on the shelf. Assume the population is normal.

Answer

Here we have a random sample of size $n = 9$. The mean is 798.30. The sample variance[3] is $s^2 = 72.76$, which gives a sample standard deviation $s = 8.53$. From Table 10 of the *New Cambridge Statistical Tables*, the upper 2.5% point of the Student's t distribution with $n-1 = 9-1 = 8$ degrees of freedom is found with $P = 2.5$ and $\nu = 8$ as $t_{0.025,(8)} = 2.306$. The 95% confidence interval is therefore from (7.5)

[3] *You can find the mean and sample variance with a hand-calculator.*

$$(798.30 - 2.306 \times 8.53/\sqrt{9}, 798.30 + 2.306 \times 8.53/\sqrt{9})$$

which is

$$(798.30 - 6.56, 798.30 + 6.56) = (791.74, 804.86).$$

It is sometimes more useful to write this as 798.30 ± 6.56.

Note that even if we do not assume the population is normal (that assumption is never really true) the Central Limit Theorem might suggest that the confidence interval is nearly right. A larger confidence would increase the length of the interval, so we trade off increased certainty of coverage against a longer interval.
■

Example 7.3. Continuing Example 7.2 suppose we are now told that σ, the population standard deviation,[4] is known to be 10g. Give a confidence interval using this information.

[4] *The standard deviation is measured in the same units as the original sample.*

Answer

Now we can use (7.2) to set the intervals. From Table 5 of the *New Cambridge Statistical Tables* with $P = 2.5$ we get the upper 2.5% point of the standard normal distribution $z_{0.025} = 1.96$ so from (7.2) we get the 95% interval for the population mean as

$$(798.30 - 1.96 \times 10/\sqrt{9}, 798.30 + 1.96 \times 10/\sqrt{9})$$

which is

$$(798.30 - 6.53, 798.30 + 6.53) = (791.77, 804.83).$$

It is sometimes more useful to write this as 798.30 ± 6.53.
■

Intervals for mean differences

In many applications we are comparing two populations by looking at samples from each of them. In particular we may wish to set a confidence interval for the difference between the means of the two populations. There are two rather different common varieties of samples: paired samples and independent samples.

Paired samples

Here the two samples are of equal size, and the observations occur in pairs. Each pair is linked through it being a repeated observation on a common entity, which may be a person, a year, or anything else that may give rise to a repeated observation. We are really most interested in the differences between the repeated observations.

The procedure to follow in a case like this is to calculate the difference between each pair of measurements, and then to use the differences as a single sample to set a confidence interval for the population mean difference between the paired observations. The assumption made is that the pairwise differences are a random sample from some normal distribution.

Example 7.4. Ten soldiers visit the rifle range on two different weeks. The first week their scores are:

$$67 \quad 24 \quad 57 \quad 55 \quad 63 \quad 54 \quad 56 \quad 68 \quad 33 \quad 43$$

The second week they score

$$70 \quad 38 \quad 58 \quad 58 \quad 56 \quad 67 \quad 68 \quad 77 \quad 42 \quad 38$$

Give a 95% confidence interval for the improvement in scores from week one to week two.

Answer

This is a case of paired samples, for the scores are repeated observations for each soldier, and there is good reason to think that the soldiers will differ from each other in their shooting skill. So we work with the individual differences between the scores. We shall have to assume that the pairwise differences are a random sample from a normal distribution.[5] The differences are:

$$3 \quad 14 \quad 1 \quad 3 \quad -7 \quad 13 \quad 12 \quad 9 \quad 9 \quad -5$$

[5] *Why is this obviously not exactly true?*

Effectively we now have a single sample of size 10, and want a 95% confidence interval for the mean of the population from which these differences are drawn. For this we use a Student's t interval. The sample mean of the differences is 5.2, and[6] $s^2 = 54.84$. So $s = 7.41$, and the 95% t interval for the difference in the means is

[6] *Use a hand calculator.*

$$5.2 \pm 2.26(7.41)/\sqrt{10} = 5.2 \pm 5.3 = (-0.1, 10.5).$$

Table 10 of the *New Cambridge Statistical Tables* for the Student's t distribution with $P = 2.5$ and $\nu = 9$ provides the critical value 2.26. Notice that this interval includes 0 and even some negative values, so that one does not feel confident that there is any real improvement. The interval is much wider than is intuitive at first sight. It is an example of how the routine use of statistical methods can correct bad judgements about uncertainty.

■

Independent samples

These samples are taken quite independently from two populations. There is no link between observations in one sample and the other. The samples may be of different sizes. The simplest assumption to make is that both populations have normal distributions, though their means and variances may be different. We may denote the two population distributions by $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$. The corresponding samples (of size m and n) give rise to sample means $\bar{x}$ and $\bar{y}$ and sample variances s_x^2 and s_y^2.

With the simple assumptions that have been made, the sampling distribution of $\bar{X} - \bar{Y}$ is - see page 55 - a normal distribution with mean $\mu_x - \mu_y$ and variance $\sigma_x^2/m + \sigma_y^2/n$, so

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/m + \sigma_y^2/n}} \tag{7.7}$$

has a standard normal distribution. If σ_x^2 and σ_y^2 are known, then we can use the same method to find a confidence interval as for a single sample. Although the starting-point is a two-sample problem, once one arrives at (7.7) one is just using the same approach as for one sample.

Example 7.5. Two similar machines are making components of a particular length in mm. Give a 90% confidence interval for the difference in average length between the components produced by the two machines. Assume that the populations are normal with variance 9 for the first machine and 16 for the second machine.

Machine A:	23.7	23.0	22.2	24.0	21.2	23.1	27.1	24.0
Machine B:	23.4	15.3	30.9	18.8	25.3	25.2	32.1	

Answer

For machine A, $\bar{x}_A = 23.54$. For machine B, $\bar{x}_B = 24.43$. Since the variances are known, the confidence interval can come from the normal distribution. From Table 5 of the *New Cambridge Statistical Tables* with $P = 5.0$, the upper 5% point of the standard normal distribution is $z_{0.05} = 1.6449$. The standard deviation of the difference of the sample means is

$$\sqrt{\sigma_A^2/n_A + \sigma_B^2/n_B} = \sqrt{9/8 + 16/7} = 1.847.$$

The 90% confidence interval for the difference between machines A and B is

$$23.54 - 24.43 \pm 1.6449 \times 1.847 = -0.89 \pm 3.04 = (-3.93, 2.15).$$

There is no certainty of any difference between the population means because 0 lies in this interval.

■

If the variances are not known, a simple result is only possible if the population variances are assumed equal. This is a restrictive assumption that may not be true in practice. For large samples (say more than 30 observations in each) one may avoid this assumption by treating the sample variances as being known values for the population variances. If the samples are small then, assuming that both variances equal σ^2, a good estimator of σ^2 is

$$S^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}, \tag{7.8}$$

[7] *Two degrees of freedom are lost because we use two sample variances.*

where $(n_x + n_y - 2)S^2/\sigma^2$ has a χ^2 distribution with $n_x + n_y - 2$ degrees of freedom[7], and is independent of $\bar{X} - \bar{Y}$. From (7.6) and (7.7) we can now use a Student's t distribution to set a $100(1-\alpha)$% confidence interval for $\mu_x - \mu_y$ as

$$\left(\bar{X} - \bar{Y} - t_{\alpha,(n_x+n_y-2)} S\sqrt{1/n_x + 1/n_y}, \bar{X} - \bar{Y} + t_{\alpha,(n_x+n_y-2)} S\sqrt{1/n_x + 1/n_y}\right). \tag{7.9}$$

Example 7.6. Continuing Example 7.5, suppose that the variances are not known, but are assumed to be equal. Then a 90% interval from (7.9) may be found as follows:

[8] *From a hand calculator.*

The sample standard deviations[8] are $s_A = 1.73$ and $s_B = 6.03$, so the estimate of the common variance is from (7.8)

$$s^2 = \frac{7 \times 1.73^2 + 6 \times 6.03^2}{7 + 6} = 18.39$$

[9] *A common mistake is to confuse s and s^2 here.*

so[9] that $s = 4.29$. From Table 10 of the *New Cambridge Statistical Tables* with $P = 5\%$ and $\nu = 13$, $t_{0.05,(13)} = 1.771$. The 90% confidence interval for the difference between the means for the two machines is

$$23.54 - 24.43 \pm 1.771 \times 4.29\sqrt{1/8 + 1/7} = -0.89 \pm 3.93 = (-4.82, 3.04).$$

■

Confidence intervals for proportions

A common application of interval estimation is to measure the accuracy of estimates of population proportions. Opinion polls to discover the proportion of electors saying that they will vote for a particular party are one example of an investigation of a population proportion. The journalists reporting the opinion poll results rarely assume that their readers are sophisticated enough to understand an interval estimate.

Interval for a single proportion

We can refer the distribution theory for proportions back to the binomial distribution on page 25, and to the application of the central limit theorem to the binomial distribution on page 57. If the population proportion is π, and a random sample of size n has proportion of successes p, then for large n it is approximately true that

$$\frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$$

has a standard normal distribution. An even rougher approximation says that

$$\frac{p-\pi}{\sqrt{p(1-p)/n}}$$

has approximately a standard normal distribution. This result is used to give approximate confidence intervals in a similar way that the earlier result (7.1) was used to set intervals for a mean. The approximate $100(1-\alpha)\%$ confidence interval for π has the form

$$p \pm z_\alpha \sqrt{p(1-p)/n}. \tag{7.10}$$

Unfortunately, even for large sample sizes such as $n = 100$, the coverage probability of a 95% interval based on (7.10) can be as low as 86% if π is close to 0 or to 1. One suggestion to improve the coverage probabilities when aiming at a 95% confidence interval is to use 2.00 instead of $z_{0.05} = 1.96$ and to 'add 2 successes and 2 failures'. In other words, use

$$\tilde{p} \pm 2\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}. \tag{7.11}$$

where for r successes in n trials, $\tilde{n} = n + 4$ and $\tilde{p} = (r+2)/(n+4)$.

Example 7.7. A cigarette manufacturing firm finds that in a random sample of 200 smokers there are 42 who smoke Brand A. Give a 95% confidence interval for the population proportion of smokers who smoke brand A.

Answer

Using (7.10) the interval is

$$0.21 \pm 1.96\sqrt{0.0008295} = 0.21 \pm 0.056 = (0.154, 0.266).$$

Using (7.11) the interval is

$$0.216 \pm 2\sqrt{0.0008292} = 0.216 \pm 0.058 = (0.158, 0.273).$$

These intervals are not very different. Smaller n, and p closer to 0 or 1, would lead to greater divergence. ■

Activity 7.4. The results in (7.10) work only for large n, and π not too close to 0 or to 1. If n is very small then it becomes more difficult to obtain intervals with approximately a 95% confidence. Show that if $n = 1$, and we use the confidence interval $(0.1, 1)$ when there is a success, and $(0, 0.9)$ when there is a failure, we attain confidence of **at least** 90%, though the actual confidence percentage achieved varies with the true value π.
■

Differences between proportions

If we are comparing proportions on the basis of **independent** random samples, then another rough and ready approximation is used. Suppose that the two samples are of size n_1 and n_2, with sample proportions p_1 and p_2. Then, if the two population proportions are π_1 and π_2 the variance of $p_1 - p_2$ is $\pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2$ and one would expect that

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

would have approximately a standard normal distribution. This result can be used to give an approximate confidence interval for the difference between two proportions. The $100(1-\alpha)$% interval is

$$\pi_1 - \pi_2 \pm z_\alpha \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}. \qquad (7.12)$$

Example 7.8. A cigarette manufacturing firm claims that its brand A line of cigarettes outsells its brand B line by 8%. If it is found that 42 of 200 smokers buy brand A, and 18 of 150 buy B, compute a 95% confidence interval for the difference between the proportions of sales of the two brands. Is the manufacturer's claim plausible?

Answer

We assume that the smokers are independent random samples from a large population, so that the formula (7.12) can be used. The sample proportions are

$$p_1 = 42/200 = 0.21$$
$$p_2 = 18/150 = 0.12.$$

The variances for the sample proportions are estimated by

$$p_1(1-p_1)/200 = 0.21(0.79)/200 = 0.0008295$$
$$p_2(1-p_2)/150 = 0.12(0.88)/150 = 0.000704.$$

The variance of the difference between the sample proportions is therefore estimated by $0.0008295+0.000704 = 0.0015335$, and the standard deviation by $\sqrt{0.0015335} = 0.03916$. The 95% confidence interval for the difference in the population proportions is

$$0.21 - 0.12 \pm 1.96(0.03916) = 0.09 \pm 0.077.$$

The difference of 8% claimed by the manufacturer lies in this 95% interval, so one can't say it is unreasonable, but the interval for the difference is again very wide and includes values much different from 8%.
■

Learning outcomes

After working through this chapter you should be able to:

1. explain the coverage property of a confidence interval
2. find confidence intervals for means of normal populations, and for differences of means of two normal populations, both when variance(s) are known and when they are unknown
3. find confidence intervals for proportions and differences of proportions
4. describe the link between intervals and distribution theory, and discuss the assumptions made to justify the use of the various intervals.

Sample examination questions

1. Why do we work out a confidence interval for the difference between the means of two populations rather than comparing the separate intervals for each population mean? (Elements of Statistics 1998 Zone B)
2. A random sample of 10 observations from a normal distribution with mean μ and variance σ^2 gives a sample mean of 1.2. An independent random sample of size 20 from the same population has sample variance 3.6. Find a 90% confidence interval for μ. (Elements of Statistics 1998 Zone B)

Parts of the sample examination questions in Chapter 8 are also covered in this chapter, but it may be better to read Chapter 8 before trying them.

How to decide if a statement about a population is true.

Chapter 8

Hypothesis testing

Essential reading

Newbold, P. *Statistics for Business and Economics*. (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 9.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], chapter 8.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], parts of chapters 8, 9 and 10.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], chapters 10 and 11.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], parts of chapters 6, 7 and 8.
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], chapter 7 and parts of chapter 9.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 9.

Introduction

We often need to answer questions about a population such as 'Is the mean of the population greater than 2?', or 'Is there any difference between the performance of two operatives?'. In statistics we try to base our answer to these questions on information in samples. Since the questions are implicitly or explicitly about populations, we are here concerned with statistical inference.

Though the theory presented in this chapter has severe deficiencies, it is useful as an introduction to the area, and is still widely used in applications. Drawing attention to the two types of error that can be made when testing hypotheses makes an important contribution to life-skills.

The theory of tests of hypotheses is necessarily linked to that for confidence intervals, but for clarity in understanding it is best to try to distinguish clearly between the two areas.

Hypotheses

A statement, which may be true or false, about a parameter of a population is called a *hypothesis*.

Example 8.1. Suppose that the population under consideration is that of boys between 7 and 8 years old in the UK. One hypothesis of interest might be that the mean weight of the population is less than 50kg.

Another hypothesis might be that exactly[1] 50% of that population can manage to add $1/3$ to $1/4$ correctly.
■

[1] *Think carefully if an exact value like 50% makes any sense here.*

Activity 8.1. Why does it make no sense to use a hypothesis like $\bar{x} = 2$?
■

Null and alternative hypotheses

A peculiarity of the classical theory of testing is that we pick out one hypothesis as our baseline (this is the *Null Hypothesis* H_0) and set up another usually less precise but more interesting hypothesis as its competitor (this is the *Alternative Hypothesis* H_1). The Null Hypothesis is one that we really don't want to reject when it is true. The idea (which may not be altogether realistic) is that if the Alternative Hypothesis is rejected when it is true, that is less important. There is no overlap between the Null and Alternative Hypothesis. They cannot both be true.

Example 8.2. Suppose that we have a long-used and well-tested treatment for stomach ulcers. The average length of treatment to cure the condition using this treatment is known to be 6 months. Now **Kcauq Laboratories** has a brand-new treatment that it says is better. A suitable Null Hypothesis might be that the average time μ to a cure for the new treatment is 6 months
H_0: $\mu = 6$
whereas the Alternative Hypothesis could be that average time μ to a cure for the new treatment is less than 6 months
H_1: $\mu < 6$.
■

One-sided and two-sided alternative hypotheses

For a parameter θ, and a given value θ_0, if H_1 is of the form $\theta > \theta_0$ or of the form $\theta < \theta_0$, then it is said to be *one-sided*. If H_1 is of the form $\theta \neq \theta_0$ then it is said to be *two-sided*.

Example 8.3. The Alternative Hypothesis that the mean weight of boys between 7 and 8 is less that 50kg is a one-sided hypothesis. The Alternative Hypothesis that the Probability of 'Heads' is not equal to 0.5 is a two-sided hypothesis.
■

Test statistics and critical regions

To carry out a test we calculate from the data the value t of a *test statistic*. If the test statistics falls in the *critical region* we reject H_0 in favour of H_1. Otherwise we do not reject H_0 (which is retained as our working hypothesis). The critical region is often described using a *critical value* that is a percentage point from the distribution of the test statistic.

Example 8.4. Suppose that the Null Hypothesis is that the population mean μ is 5, and the Alternative Hypothesis is that μ is greater than 5. We could use as the test statistic $T = \bar{X}$ the sample mean from a sample of size n. One possible critical region is all values for $T = \bar{X} - 5$ greater than $2z_\alpha/\sqrt{n}$. The percentage point z_α is used to define the critical value $2z_\alpha/sqrtn$.

If $\bar{x} - 5$ from the sample is greater than $2z_\alpha/\sqrt{n}$ we would reject H_0 in favour of H_1; otherwise if $\bar{x} - 5 \leq 2z_\alpha/\sqrt{n}$ we would retain H_0 as a working hypothesis.
■

One- and two-tailed tests

A critical region that is the upper tail of a distribution, or that is the lower tail of a distribution, gives a *one-tailed test*. A critical region that contains both an upper tail of a distribution **and** its lower tail gives a *two-tailed test*. Often, a one-sided Alternative Hypothesis leads to a one-tailed test and a two-sided Alternative Hypothesis leads to a two-tailed test, but that pattern is not universal.

Example 8.5. Referring back to Example 8.4, the critical region $\bar{X} - 5 > 2z_\alpha/\sqrt{n}$ is an upper tail of the distribution of $\bar{X}$, so this is a one-tailed test.

Another test might have its critical region defined by $|\bar{X} - 5| > 2z_\alpha/\sqrt{n}$. This critical region contains both an upper tail of the distribution of $\bar{X} - 5$ (where $\bar{X} - 5 > 2z_\alpha/\sqrt{n}$) as well as a lower tail (where $\bar{X} - 5 < -2z_\alpha/\sqrt{n}$). It defines a two-tailed test.
■

Type I and type II errors

If the Null Hypothesis is rejected when it is true, then a *Type I* error has occurred. If the Null Hypothesis is not rejected when it is false, then a *Type II* error has occurred. It is very important, and it is a significant contribution to clear thought about testing hypotheses, to realise that there are **two** different kinds of error that can be made.

Level and power

The probability of making a Type I error with a test is called the *level*, or *significance level* of the test. It is controlled by the tester to some fixed amount, conventionally 1%, 5% or 10%. The level[2] of the test is usually labelled α by statisticians. The probability of making a Type II error is usually labelled β. Its complement $(1-\beta)$ is called the *power* of the test. The tester may obtain a high power by using a large enough sample size, but power is not directly controlled at the time of testing. Tests recommended in the textbooks are chosen so that they have the highest power for their chosen level.[3]

[2] *Be careful not to talk about a 95% level for a test.*

[3] *The tests suggested by the texts are not arbitrary.*

Activity 8.2. Of 100 clinical trials, 5 have shown that wonder-drug zap2 is better than the standard treatment (aspirin). Should we be excited by these results?

Of the 1000 clinical trials of 1000 different drugs this year 30 trials found drugs that seem better than the standard treatments with which they were compared. The television news reports only the results of those 30 'successful' trials. Should we believe the television news reports?

A child welfare officer says that she has a test that always reveals when a child has been abused, and she suggests it be put into general use. What is she saying about Type I and Type II errors for her test?

■

Testing hypotheses about population means

Suppose that we have a random sample from a normal population with mean μ and variance σ^2. It is a routine application to test hypotheses about the population mean μ. We may want to test such a hypothesis either when σ^2 is known, or when it is unknown.

Known variance

If the variance σ^2 is known, we can use the result on page 68 that (7.1) has a standard normal distribution. Suppose that the Null Hypothesis is
H_0: $\mu = \mu_0$.
Then when H_0 is true, from (7.1) it follows that

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

has a standard normal distribution. If the Alternative Hypothesis H_1 is that $\mu > \mu_0$, then for a $100\alpha\%$ level test use the critical region

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha, \tag{8.1}$$

which is a one-tailed test for a one-sided Alternative Hypothesis.

If the Alternative Hypothesis H_1 is that $\mu < \mu_0$, then for a $100\alpha\%$ level test use the critical region

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha, \tag{8.2}$$

which is again a one-tailed test for a one-sided Alternative Hypothesis.

If the Alternative Hypothesis H_1 is that $\mu \neq \mu_0$, then for a $100\alpha\%$ level test use the critical region

$$\frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2} \tag{8.3}$$

which is a two-tailed test for a two-sided Alternative Hypothesis.

Activity 8.3. You should be able to verify that for each of these critical regions the probability of rejecting H_0 when H_0 is true is $100\alpha\%$.
■

Notice carefully the difference between the one-sided and two-sided Alternative Hypotheses. The critical regions are carefully chosen to give high power. They will give as large a probability as possible of rejecting H_0 when it is false - that is when H_1 is true - for a test of level α.

Activity 8.4. Why don't we use a two-tailed test for a one-sided Alternative Hypothesis?
■

Example 8.6. Suppose that we have a random sample of size $n = 4$ of weights in g of eating plums from a large population having a normal distribution with known variance 1. The observations are $27.1, 28.1, 27.0, 28.0$ giving a sample mean $\bar{x} = 27.55$ and a sample standard deviation of $s = 0.580$.[4] Suppose we want to test
H_0: $\mu = 25$
against
H_1: $\mu > 25$.
There are not enough observations to give a test with a good power, but if we use a test at the 5% level, then the critical region is, from (8.1),

[4] *Use a hand-calculator.*

$$\frac{\bar{X} - 25}{1/\sqrt{4}} > z_{0.05}$$

which, using Table 5 of the *New Cambridge Statistical Tables* with $P = 5.0$ to get $z_{0.05} = 1.6449$, is

$$\frac{\bar{X} - 25}{1/2} = 5.1 > 1.6449.$$

Since the observed value of the test statistic falls in this critical region, we **reject** the null hypothesis that the population mean is 25.

If, instead the Null Hypothesis were to be
H_0: $\mu = 27$
and the Alternative Hypothesis were to be
H_1: $\mu \neq 27$
then from (8.3) the critical region for a 5% level test takes the form

$$\frac{|\bar{X} - 27|}{1/2} > z_{0.025}$$

and, again using Table 5 of the *New Cambridge Statistical Tables* with $P = 2.5$ to get $z_{0.025} = 1.9600$, this is

$$\frac{|\bar{X} - 27|}{1/2} > 1.96.$$

Since the observed value of the test statistic is $\frac{|27.55-27|}{1/2} = 1.1$, which does not fall into either tail of the critical region, we do not reject the Null Hypothesis that the population mean is 27g.
■

Example 8.7. A manufacturer has developed a new fishing line that he claims has a breaking strength of 7kg, with a standard deviation of 0.25kg. Assume that the standard deviation figure is correct. Suppose that we carry out a test, at the 5% level, of the null hypothesis that the mean is 7kg against the alternative that it is less than 7kg. Find the sample size that is necessary for the test to have 90% power if the true breaking strength is 6.95kg.

Answer

Assume that the population is normal with standard deviation 0.25kg. The test statistic is, from (8.2),

$$(\bar{X} - 7)/(0.25/\sqrt{n}).$$

The critical value for a 5% level one-sided test of
H_0: $\mu = 7$
against
H_1: $\mu < 7$
is taken from Table 5 of the *New Cambridge Statistical Tables* with $P = 5.0$ with a change of sign. It is -1.6449. The probability of rejecting the Null Hypothesis with the 5% level test is

$$P\left[\frac{\bar{X} - 7}{0.25/\sqrt{n}} < -1.6449\right]$$

which is equivalent to

$$P\left[\frac{\bar{X} - 6.95}{0.25/\sqrt{n}} < \frac{7 - 6.95}{0.25/\sqrt{n}} - 1.6449\right]$$

The statistic on the left of the inequality has a standard normal distribution when the breaking strength has mean 6.95 kg, so for a power of 90%, one must have the right-hand side of the above inequality equal to the upper 10% point of the standard normal distribution, which is 1.2816, from Table 5 of the *New Cambridge Statistical Tables* with $P = 5.0$.

So

$$\begin{aligned} 0.2\sqrt{n} - 1.6449 &= 1.2816 \\ \sqrt{n} &= 5(1.2816 + 1.6449) = 14.63 \\ n &= 14.63^2 = 214.11. \end{aligned}$$

To be sure of sizes at least as small as those required, we should use a sample of size 215. Notice the rather large sample size that is required. One of the important effects of interval estimation, or hypothesis testing theory, is that investigators are encouraged to use sample sizes large enough to come to rational decisions.
■

Unknown variance

When the variance is not known all the tests are based on the fundamental distributional result of (7.4), so that percentage points from Student's t distributions are used instead of percentage points from normal distributions, and the sample standard deviation is used instead of σ. In all other respects the tests remain the same as those for unknown variances.

Example 8.8. Continuing Example 8.6 let us now suppose that the population variance is not known. In that case σ is estimated by the sample standard deviation $s = 0.580$. Suppose again that we want to test
H_0: $\mu = 25$
against
H_1: $\mu > 25$.
If we use a test at the 5% level, then the critical region is from (7.4)

$$\frac{\bar{X} - 25}{S/\sqrt{4}} > t_{0.05,(3)} = 2.353.$$

(Using Table 10 of the *New Cambridge Statistical Tables* with $P = 5$ and $\nu = 3$ we get $t_{0.05,(3)} = 2.353$.) The value of the test statistic is

$$\frac{27.55 - 25}{0.580/2} = 8.79 > 2.353.$$

Since the observed value of the test statistic falls in the critical region, we **reject** the null hypothesis that the population mean is 25g.

If, instead the Null Hypothesis were to be
H_0: $\mu = 27$
and the Alternative Hypothesis were to be
H_1: $\mu \neq 27$
then from (7.4) the critical region for a 5% level test takes the form

$$\frac{|\bar{X} - 27|}{S/2} > t_{0.025,(3)} = 3.182.$$

(using again Table 10 of the *New Cambridge Statistical Tables* with $P = 2.5$ and $\nu = 3$ to get $t_{0.025,(3)} = 3.182$). The value of the test statistic is

$$\frac{27.55 - 27}{0.580/2} = 1.90 \not> 3.182.$$

Since the observed value of the test statistic does not fall into either tail of the critical region, we do not reject the Null Hypothesis that the population mean is 27g.

There is no particular reason why the results of this test should agree with those of the previous ones in Example 8.6.

■

Link to Confidence Intervals

Two-tailed tests are closely similar to confidence intervals. To test H_0: $\mu = 27$ against H_1: $\mu \neq 27$, at the 5% level, we could find a 95% confidence interval for μ, and reject H_0 if that interval did not include the value 27.

Activity 8.5. There is no obvious link between confidence intervals and one-tailed tests. What sort of confidence interval would one need to define to have such a link?

■

Two-sample tests

Once again, as in Chapter 7, we must distinguish between paired samples and random samples, and between those cases where the variances are known and those where the variances are unknown. That said, all one needs to do is to refer to the basic distributional results, for instance 7.7 and 7.8, and then use these as in the last two sections.

Example 8.9. The weights of a group of five-week-old chickens, reared on a high protein diet are 336, 421, 310, 446, 390 and 434 g; the weights of a second group of 5 chickens similarly reared except for their low protein diet are 224, 275, 393, 282 and 365 g. Is there evidence that the additional protein has increased the weight of the chickens?

Answer

We have to imagine that the chickens are random observations of large somewhat hypothetical populations of chickens in which we are interested, and that distributions of weights under the two different diets are normal distributions with possibly different means, but the same variance. The Null Hypothesis is that the two diets give the same mean population weight. The Alternative Hypothesis is that the first diet gives a greater population mean weight.

The sample means are 389.5 and 307.8, and the sample standard[5] deviations are 55.40 and 69.45. The estimate of the common variance is, from (7.8)

[5] *Use a hand-calculator.*

$$s^2 = \frac{5 \times 55.40^2 + 4 \times 69.45^2}{5+4} = 62.03^2$$

The critical region for a test at the 5% level is

$$\frac{\bar{X} - \bar{Y}}{S\sqrt{1/6 + 1/5}} > t_{0.05,(9)}$$

From Table 10 of the *New Cambridge Statistical Tables* with $P = 5$ and $\nu = 9$ we get $t_{0.05,(9)} = 1.833$.

The value of the test statistic is

$$\frac{389.5 - 307.8}{62.03\sqrt{1/6 + 1/5}} = 2.17 > 1.833.$$

The test statistic falls in the critical region, so with a test at the 5% level we reject the hypothesis that the mean weights are equal in favour of the alternative hypothesis that the mean weight for the first diet is greater.

■

Activity 8.6. Suppose that we have two independent samples from normal populations with known variances. We want to test the null hypothesis that the two populations have the same mean against the alternative that the means are different. One could use each sample by itself to write down a 95% confidence interval for the corresponding population mean. One could reject H_0 if those intervals did not overlap. What would be the significance level of this test?

■

p-values

Some statisticians prefer to calculate *p-values* when carrying out tests. The p-value is the smallest level of the test for which the Null Hypothesis would be rejected.

Example 8.10. Continuing Example 8.9, the value of the test statistic is 2.17, to be checked against a Student's t distribution with 9 degrees of freedom.

Looking in Table 9 of the *New Cambridge Statistical Tables*, which gives the left-hand tail probabilities of the t distribution with $\nu = 9$ degrees of freedom, we

find that the right-hand tail probability at 2.2 is $1 - 0.9723 = 2.77\%$. There is some evidence therefore of a difference in weight. This is a one-sided p-value because we are checking to see if the difference between population means is positive. Calculating the p-value exactly (not possible from the tables) gives the value 2.9%.
■

Using p-values seems to be closer to what most statistics users want to know. A very small p-value, say .0012, means that the Null hypothesis would be rejected at the 5% level, or at the 1% level, or indeed at the 0.5% level. So the p-value seems a less arbitrary way to carry out a test than than to fix on a particular level such as 5%. However, it is hard to situate the use of p-values coherently in the framework of classical testing theory, so one should perhaps be cautious in emphasising them.

Tests for binomial probabilities of success

One- and two-sample tests for population proportions π can be based on the approximate distributions used already for confidence intervals in (7.10) and (7.12). These approximate distributions give rise to the tests in just the same way as those for the normal populations that were used before in this chapter. For instance to test at the $100\alpha\%$ level with a random sample of size n the null hypothesis
H_0: $\pi = \pi_0$
against the Alternative Hypothesis
H_1: $\pi \neq \pi_0$
we reject H_0 if

$$\frac{|p - \pi_0|}{\sqrt{\pi_0(1-\pi_0)/n}} > z_{\alpha/2}.$$

For the two sample tests (which come from (7.12)), there is a slight twist that has to be remembered. The critical region is decided by the distribution of the test statistic under the Null Hypothesis, so if the Null Hypothesis is that two population proportions π_1 and π_2 are equal, the best guess when H_0 is true of the common proportion is

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2},$$

where the sample proportions are p_1 and p_2 and the sample sizes n_1 and n_2. So to test at the $100\alpha\%$ level
H_0: $\pi_1 = \pi_2$
against the Alternative Hypothesis
H_1: $\pi_1 \neq \pi_2$
we use the critical region

$$\frac{|p_1 - p_2|}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}} > z_{\alpha/2}.$$

Example 8.11. Suppose that of 200 randomly chosen households in the UK 15% have cable television, and of 150 randomly chosen households from France 6% have cable television. Is the proportion of UK households with cable television different from that in France?

Answer

Suppose that the population proportion of households with cable television in the UK is π_1, and in France π_2. The Null Hypothesis is
H_0: $\pi_1 = \pi_2$
and the Alternative Hypothesis is
H_1: $\pi_1 \neq \pi_2$
The estimate of the common proportion when H_0 is true is

$$\frac{200 \times 0.15 + 150 \times 0.06}{200 + 150} = 0.1114.$$

If we test at the 5% level of significance, then the test statistic is

$$\frac{0.15 - 0.06}{\sqrt{0.1114(1 - 0.1114)(1/200 + 1/150)}} = 2.65.$$

The critical value is $z_{0.025} = 1.96$, which is exceeded by the observed value of the test statistic. We therefore reject the hypothesis that the proportions of households with cable television are the same in the UK and France.
■

Learning outcomes

After working through this chapter you should be able to:

1. define and use the terminology of statistical testing
2. carry out statistical tests of all the types covered in this chapter
3. calculate the power of some of the simpler tests
4. explain the way in which the rejection regions of tests follow from the distributional results, taking into account the level and considerations of power.

Sample examination questions

1. The table below shows the annual salaries in dollars of randomly selected faculty in public educational institutions and private educational institutions.

Public	52127	57380	34122	8334	35730	22411	40196
Private	40807	26448	48970	52411	20223	39421	40102
Public	28528	10562	33666				
Private	46461	32557					

(a) Find a 90% confidence interval for the difference between population mean annual salaries in the public and private institutions.

(b) Test the Null Hypothesis that mean salary for the private institutions is 1000 dollars more than in the public institutions against the alternative that the mean for the private institutions is more than 1000 dollars greater.

(c) State carefully the assumptions you have made in arriving at the test and confidence interval.

(Elements of Statistics 1999 Zone B)

2. Primary school children with reading problems were randomly divided into a control group and a group that received special reading teaching. The results of a subsequent reading test for all the children are given below:

Control	42	43	55	26	62	37	33	41	19	54	20	
	85	46	10	17	60	53	42	37	42	55	28	48
Special	24	43	58	71	43	49	61	44	67	49		
Teaching	53	56	59	52	62	54	57	33	46	43	57	

(a) Find a 99% confidence interval for the difference in score between the controls and the specially taught group.

(b) Test at the 10% level the null hypothesis that there is no difference between the two groups.

(Elements of Statistics 1998 Zone B)

One-way and two-way tables of measurements.

Chapter 9

Analysis of variance

Essential reading

Newbold, P. *Statistics for Business and Economics*. (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 15.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], section 14.1.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], chapter 14.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7],chapter 12.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], chapters 12 and 13.
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], chapter 11.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 10.

Introduction

Analysis of variance (ANOVA) is a popular tool that has an applicability and power that we can only start to appreciate in this course. The idea of analysis of variance is to investigate how variation in structured data can be split into pieces associated with components of that structure. We look only at one-way and two-way classifications, providing tests and confidence intervals that are widely used in practice.

One-way analysis of variance

One-way analysis of variance looks to see how much of the variation in grouped data comes from differences between the groups, and how much is just random ob-

servational error. There can be any number of groups, that may be of different sizes (each group with at least two observations). A typical application of one-way analysis of variance would be to investigate whether three different types of growing conditions make any difference to the yield of an agricultural crop, and if so, how great those differences are. The observations would be the yields of many different experimental plots, grouped according to the growing condition that applied to them.

To describe analysis of variance accurately one needs a lot of notation; it is annoying, but worth spending time on. Suppose that we have n random observations classified into k different groups, so that there are n_i observations in group i for $i = 1, \ldots, k$. We shall assume that all the observations are independent of each other, and that the distribution from which those in group i are taken is $N(\mu_i, \sigma^2)$. [1] Notice that the population mean μ_i may be different for each group, but that the variance is the **same** for all observations in all groups. Since we assume that the observations are from normal distributions, they are not counts - so application of analysis of variance to tables of counted data (contingency tables) will not usually be possible. For counted data the variance is often proportional to the mean, and so not the same for all the counts. Outliers or wild observations also invalidate the assumption of normal distributions, so one needs to check there are none of those (for instance a wrong recording[2] of 8 instead of 80).

[1] *Subscripts are unattractive, but it's hard to do without them.*

[2] *Such errors easily occur during transcription to a spreadsheet.*

We have a collection of k independent samples, one from each of k normal populations. The common feature is the variance σ^2. One idea of one-way analysis of variance is to exploit the knowledge about σ^2 in all k samples to obtain better knowledge about the variability of estimates of μ_is than would be possible from one sample considered by itself.

Consider the n_i observations in group i. We can label these[3]

$$X_{i1}, X_{i2}, \ldots, X_{in_i}.$$

[3] *Two subscripts are even less attractive than one.*

The sample mean for group i, $\bar{X}_i$, is a good estimator of the population mean μ_i for group i. The sample variance for group i, S_i^2, is a good estimate for σ^2, the common variance for every group.

We put together the information about σ^2 in all the k groups to form a new estimator S^2 of σ^2

$$\begin{aligned} S^2 &= \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \cdots + (n_k-1)S_k^2}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)} \\ &= \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \cdots + (n_k-1)S_k^2}{n-k}. \end{aligned} \tag{9.1}$$

The degrees-of-freedom here are $n-k$ because we start with n observations giving n degrees-of-freedom and lose one for each of the group means used to calculate an S_i^2. It is usual in ANOVA to call the positive quantity $(n-k)S^2$ the *Within Groups Sum of Squares*.[4] Obviously we can write

[4] *Sometimes called the Error Sum of Squares.*

$$(n-k)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2.$$

The Within Groups Sum of Squares is an unbiased estimator of $(n-k)\sigma^2$. The *Mean Within Groups Sum of Squares* S^2 is an unbiased estimator of σ^2

A very important Null Hypothesis is that there are no differences between the means μ_i. That is

$$H_0: \qquad \mu_1 = \mu_2 = \cdots = \mu_k.$$

Suppose that we want to test H_0 against the very general Alternative Hypothesis

$$H_1: \qquad \text{Not all the } \mu_i \text{ are equal.}$$

If there are no differences between the population group means, then we could take all the observations from all the groups together as one sample of size n and get an estimator of σ^2 from the sample variance S_T^2 for that large sample, where

$$S_T^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2}{n-1}$$

and $\bar{X}$ is the mean of observations in all the groups. The positive quantity $(n-1)S_T^2$ is called the *Total Sum of Squares*. It has $(n-1)$ degrees-of-freedom. It is

$$(n-1)S_T^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2.$$

If all the population group means are equal, the Total Sum of Squares is an unbiased estimator of $(n-1)\sigma^2$.

The difference between the Total Sum of Squares and the Within Groups Sum of Squares is called the *Between Groups Sum of Squares*, $(k-1)S_B^2$. It may also be obtained by replacing each of the observations X_{ij} by its group mean $\bar{X}_i$ and calculating the total sum of squares for this new collection of n 'observations'. The Between Groups Sum of Squares may be written:

$$(k-1)S_B^2 = \sum_{i=1}^{k} n_i(\bar{X}_i-\bar{X})^2.$$

It has $(k-1)$ degrees-of-freedom; if all the population group means are equal it is an unbiased estimator of $(k-1)\sigma^2$. If the population group means are not all equal, then $(k-1)S_B^2$ is an unbiased estimator of

$$(k-1)\sigma^2 + \sum(\mu_i-\mu)^2$$

where $\mu = \sum_{i=1}^{k} n_i\mu_i/n$. So when the population group means are not all equal the Between Groups Sum of Squares is expected to be larger than $(k-1)\sigma^2$. The *Mean Between Groups Sum of Squares*, S_B^2 is an unbiased estimator of σ^2 if all the population group means are equal.

Sum of Squares Identity

It was stated above that the Total Sum of Squares was the sum of the Between Groups Sum of squares and the Within Groups Sum of Squares. It is possible to prove this result with elementary algebra.

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(X_{ij}-\bar{X}_i)+(\bar{X}_i-\bar{X})]^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i-\bar{X})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})^2. \qquad (9.2)\end{aligned}$$

The cross-product terms produced by the squaring vanish because $\sum_{j=1}^{n_i}(\bar{X}_i-\bar{X})$ is zero.

This sum of squares identity allows any one of the three Sums of Squares to be calculated from the other two.

F-test

Consider again the test that there are no differences between the means μ_i:

$$H_0: \quad \mu_1=\mu_2=\cdots=\mu_k$$

against the very general Alternative Hypothesis:

$$H_1: \quad \text{Not all the } \mu_i \text{ are equal.}$$

The suggested test at the $100\alpha\%$ level of significance is to reject H_0 if

$$F=\frac{S_B^2}{S^2}>F_{\alpha,k-1,n-k},$$

where the right-hand side is the upper $100\alpha\%$ point of the F distribution with $(k-1)$ degrees-of-freedom for the numerator and $(n-k)$ degrees-of-freedom for the denominator. (You can find a little more information about the F distribution on page 97.) This test is a one-tailed test, but the Alternative Hypothesis is **not** one-sided.

Example 9.1. Table 9.1 on page 95 shows the scores for crunchiness of four competing breakfast cereals. Each cereal was assessed once by several testers, but no tester was used twice because of possible order effects and fatigue.

Is there enough evidence to conclude that there are differences in crunchiness between the four cereals?

Table 9.1: Crunchiness scores

Cereals			
1	2	3	4
9.3	13.4	12.5	14.0
10.8	12.2	14.7	15.6
8.4	12.4	12.9	14.1
9.7	12.8	11.8	
9.5	12.2		
7.9			
9.5			

Answer

We will assume that the observations are independent random observations from normal populations with means $\mu_1, \mu_2, \mu_3, \mu_4$, and with the same variance σ^2. There are no obvious outliers or wild observations that make this assumption look wrong.

From the whole set of observations, with a calculator, the sample variance is $s_T^2 = 4.8843$, leading to the Total Sum of Squares $18s_T^2 = 87.9168$. Also, one rapidly obtains: $\bar{x}_1 = 9.3, \bar{x}_2 = 12.6, \bar{x}_3 = 12.975, \bar{x}_4 = 14.5667$ and $s_1^2 = 0.8767, s_2^2 = 0.2600, s_3^2 = 1.5292, s_4^2 = 0.8033$ and so from (9.1) the Mean Within Group Sum of Squares

$$s^2 = \frac{6 \times 0.8767 + 4 \times 0.2600 + 3 \times 1.5292 + 2 \times 0.8033}{6 + 4 + 3 + 2}$$
$$= 12.4944/15 = 0.8330.$$

The Within Groups Sum of Squares[5] is $15s^2 = 12.4944$. It can also be thought of as $18\times$ the sample variance obtained from the differences of all the observations from the appropriate group mean. Those differences can be written as in Table 9.2 below:

[5] *Sums of squares are always positive. Do not be content with a negative value.*

The difference between these two sums of squares gives the Between Groups Sum of Squares $(4-1)s_B^2 = 87.9168 - 12.4944 = 75.4224$ and the Mean Between Groups Sum of Squares $s_B^2 = 75.424/(4-1) = 25.1408$. One can also obtain the Between Groups Sum of Squares by finding $18\times$ the sample variance of replacement data where the original observations are replaced by the group means. These replacement data are shown in Table 9.3 below:

We will carry out a test of

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4$$

against the alternative

$$H_1: \text{Not all the population group means are equal}$$

Table 9.2: Differences from group means

Cereals			
1	2	3	4
0.0	0.8	-0.475	-0.5667
1.5	-0.4	1.725	1.0333
-0.9	-0.2	-0.075	-0.4667
0.4	0.2	-1.175	
0.2	-0.4		
-1.4			
0.2			

Table 9.3: Group Means for finding BGSS

Cereals			
1	2	3	4
9.3	12.6	12.975	14.5667
9.3	12.6	12.975	14.5667
9.3	12.6	12.975	14.5667
9.3	12.6	12.975	
9.3	12.6		
9.3			
9.3			

at the 5% level of significance. The value of the test statistic is $F = 25.1408/0.8330 = 30.1810$. We will reject H_0 if this value is greater than the upper 5% point $F_{0.05,3,15}$. From Table 12(b) (**not** 12(e)) of the *New Cambridge Statistical Tables* with $\nu_1 = 3$ and $\nu_2 = 15$ we find $F_{0.05,3,15} = 2.490$. So we reject the Null Hypothesis and find that there is good reason to believe that there are differences in crunchiness between the cereals.

These results are usually summarised in an *analysis of variance table*, and one should always present the results in that form. The table is shown below:

■

Activity 9.1. Think carefully about the last Tables 9.1 to 9.3. The sum of squares identity (9.2) says that the total sum of squares from all the data in the first table is obtained by adding together the total sums of squares of each of the last two tables. Also, the entries in the first table are the sums of the corresponding entries in the last two tables. We have broken down the original grouped observations into the last table, which reflects the group structure, and the second table, which has numbers which

Table 9.4: Analysis of variance table

Source	Degrees of Freedom	Sums of Squares	Mean Sums of Squares	F ratio
Between Groups	3	75.424	25.1408	30.1810
Within Groups	15	12.4944	0.8330	
Total	18	87.9168		

look randomly distributed around zero. How would these tables change if groups 3 and 4 were put together as one group?
■

The F-Distribution

It seems best to summarise the properties of the F-distribution for convenient reference. If U and V are independent random variables with χ^2 distributions, with ν_1 and ν_2 degrees-of-freedom respectively, then the ratio

$$F = \frac{U/\nu_1}{V/\nu_2}$$

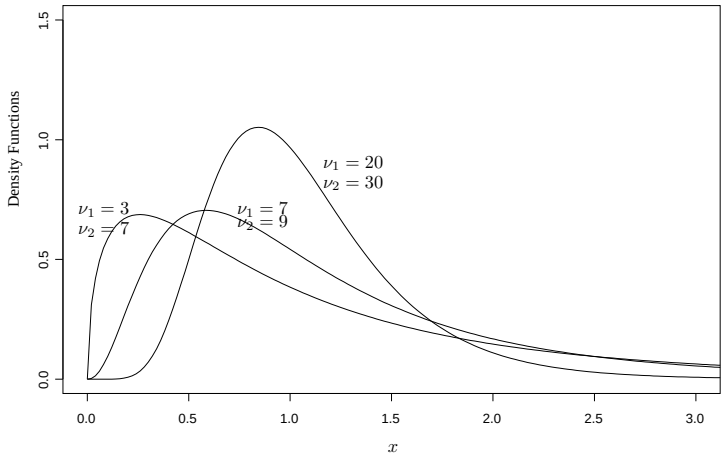
has an F-distribution with ν_1 degrees-of-freedom for the numerator and ν_2 degrees-of-freedom for the denominator. The distribution has mean $\nu_2/(\nu_2 - 2)$ for $\nu_2 > 2$. Pictures of the density functions of some F-distributions are shown in Figure 9.1.

One can see that the distribution looks more like a normal distribution for large values of the degrees-of-freedom.

The square of a random variable with a Student's t distribution with ν degrees-of-freedom has an F distribution with $\nu_1 = 1$ and $\nu_2 = \nu$, as can be seen from (7.6). The F-distribution comes into the one-way analysis of variance because, when the Null Hypothesis is true, the Between Group and Within Group sums of squares each divided by σ^2 have independent χ^2 distributions.

Activity 9.2. If you fix ν_1 and let ν_2 become large, the F distribution becomes close to what distribution?
■

Figure 9.1: Density functions for F distributions



Confidence intervals and tests for population group means

Single intervals

Confidence intervals and tests for a population group mean, or for differences of population group means, can be easily written down as in Chapter 7 using the result that

$$\frac{\bar{X}_i - \mu_i}{S/\sqrt{n_i}}$$

has a Student's t distribution with $(n-k)$ degrees-of-freedom, and for $i \neq j$

$$\frac{(\bar{X}_i - \bar{X}_j) - (\mu_i - \mu_j)}{S\sqrt{1/n_i + 1/n_j}}$$

has a Student's t distribution with $(n-k)$ degrees-of-freedom. The last result generalises those in Chapter 7 for two independent random samples.

Activity 9.3. Suppose that in a one-way ANOVA you reject the Null Hypothesis that all the population group means are equal with a test at the 5% level. Does it follow that a 95% confidence interval for the difference between at least one pair of population group means will not include 0?
■

Simultaneous intervals

We may want to think about setting intervals for all possible pairs of differences between any two μ_is. It would be unwise to use a 95% confidence for each of many separate intervals and yet look at them all together, since purely by chance 5 out of 100 such intervals might well not cover the population values. One could adjust by using a larger confidence level. If we use r simultaneous intervals then, for at least a confidence of $100(1-\alpha)$ % that they all simultaneously cover the parameter values it is enough to take each individual interval with $100(1-\alpha/r)$% confidence.

Activity 9.4. Check that using $100(1-0.05/r)$% gives r simultaneous intervals with at least 95% confidence level.
■

There are more specialised ways of obtaining valid sets of simultaneous confidence intervals in the analysis of variance. One method, due to Scheffé, is to use a $100(1-\alpha)$% set of simultaneous intervals for every contrast between the μ_is. These are simultaneous intervals for every linear combination $\sum_{i=1}^{k} d_i\mu_i$, where $\sum_{i=1}^{k} d_i = 0$. A typical interval in the set is

$$\sum_{i=1}^{k} d_i\bar{X}_i \pm S\sqrt{(k-1)F_{\alpha,k-1,n-k}\sum d_i^2/n_i}. \tag{9.3}$$

In the particular case of a difference in population group means $\mu_i - \mu_j$ we have all $d_i = 0$ except for $d_i = -d_j = 1$, and the confidence interval is

$$\bar{X}_i - \bar{X}_j \pm S\sqrt{(k-1)F_{\alpha,k-1,n-k}(1/n_i + 1/n_j)}.$$

Example 9.2. Continuing Example 9.1, a 95% confidence interval for $\mu_1 - \mu_2$ is

$$9.3 - 12.6 \pm 0.913 t_{0.025,(15)}\sqrt{1/7 + 1/5}$$

where from Table 10 of the *New Cambridge Statistical Tables* with $P = 2.5$ and $\nu = 15$ we get $t_{0.025,(15)} = 2.13$ and so the interval is

$$-3.3 \pm 1.14.$$

On the other hand if we want a 95% confidence simultaneous set of intervals of the Scheffé type, then the corresponding interval is

$$9.3 - 12.6 \pm 0.913\sqrt{3F_{0.05,3,15}(1/7 + 1/5)}.$$

From Table 12(b) of the *New Cambridge Statistical Tables* with $\nu_1 = 3$, $\nu_2 = 15$ we get $F_{0.05,3,15} = 3.287$ and so the interval is

$$-3.3 \pm 1.68.$$

This is longer than the previous one, as it must be to keep the confidence property for the whole set of confidence intervals for contrasts.

■

Activity 9.5. In Example 9.2 use (9.3) to give the interval in the simultaneous set for the contrast with $d_1 = -0.854953$, $d_2 = 0.203993$, $d_3 = 0.237256$, $d_4 = 0.413704$. Do you notice anything that connects your interval with the analysis of variance table on page 97?

■

Two-way analysis of variance

Suppose we have a two-way table of continuous random variables with r rows and c columns. The ith row and j'th column meet in the cell (i, j) that contains the random variable X_{ij}. The expected value of that random variable is defined as $EX_{ij} = \mu_{ij}$. We shall suppose that the observations are independent normal random variables with means μ_{ij}, all with the same variance σ^2. Two-way analysis of variance is based on the further assumption that the cell population means have an additive structure:

$$\mu_{ij} = \mu + \alpha_i + \beta_j. \tag{9.4}$$

Here μ is the *overall mean*, the *row effects* α_i add to zero, and the *column effects* β_j add to zero. This additive structure is one of the simplest ways to explain the numbers in a two-way array. It is not certain to be appropriate for all applications, for when there is additivity we may **without ambiguity** talk about differences between rows and about differences between columns. If we look at the differences between population cell means for two cells in the same row and in column j_1 and column j_2, when (9.4) holds, that difference is the same whichever row we look at. Similarly for the difference between two population cell means in the same column but in rows i_1 and i_2. Another way to think about the same property is to say that for any choices i_1, i_2, j_1, j_2 the *interaction*

$$\mu_{i_1 j_1} - \mu_{i_1 j_2} - \mu_{i_2 j_1} + \mu_{i_2 j_2}$$

is zero.

Activity 9.6. In the following table of population means, what is the difference between means in column 1 and column 2? What is the difference between means in row 2 and row 3? What is the interaction for the 2×2 table of cells in columns 1 and 2 and rows 2 and 3? Does this table show an additive structure for row and column effects?

1.0	2.0	10.0
2.1	3.1	4.3
3.2	4.2	5.4

■

The idea of two-way analysis of variance is to split the variation among the observations X_{ij} in the table into parts associated with the row effects, the column effects, and random error.

Just as with the one-way analysis we can think about several different ways of estimating σ^2. The jth column has mean $\bar{X}_{.j} = \sum_{i=1}^{r} X_{ij}/c$ which has expected value

$$\begin{aligned} E\sum_{i=1}^{r} X_{ij}/c &= \sum_{i=1}^{r} E[X_{ij}]/c \\ &= \sum_{i=1}^{r} [\mu + \alpha_i + \beta_j]/c \\ &= \mu + \beta_j \end{aligned}$$

since the sum of the α_is is 0.

If the column effects β_j are all equal to 0, then the column means all have normal distributions with mean μ and variance σ^2/r. It follows that when the β_j are all zero, the sample variance, S_C^2, calculated for the set of column means is an unbiased estimator of σ^2/r, so that rS_C^2, called the *Mean Between Columns Sum of Squares* is an unbiased estimator of σ^2. It is usual to call $r(c-1)S_C^2$ the *Between Columns Sum of Squares*. It has $(c-1)$ degrees-of-freedom.

Similarly, if the row effects α_i are all equal to 0, then S_R^2 (the sample variance of the set of row means $\bar{X}_{i.}$) is such that cS_R^2 the *Mean Between Rows Sum of Squares* is an unbiased estimator of σ^2, and the *Between Rows Sum of Squares* is $c(r-1)S_R^2$ with degrees-of-freedom $(r-1)$.

The *Total Sum of Squares* is defined in the same way as for the one-way analysis of variance as $(rc-1)S_T^2$, where S_T^2 is the sample variance of all the observations in the table. It has $(rc-1)$ degrees-of-freedom.

The difference between the Total Sum of Squares and the sum of the Between Columns Sum of Squares and the Between Rows Sum of Squares is $(r-1)(c-1)S^2$, the *Residual Sum of Squares*.[6] It has degrees of freedom $(r-1)(c-1)$. It is the analogue of the Within Groups Sum of Squares in the one-way analysis of variance. If one divides the Residual Sum of Squares by its degrees-of-freedom one gets S^2, the *Mean Residual Sum of Squares*.

[6] *Also called the Error Sum of Squares.*

Tests for row effects and column effects

One Null Hypothesis of interest is that all the row effects α_i, $i = 1, \ldots, r$ are zero. If that is true the row classification is redundant. We may write this Null Hypothesis as

$$H_0 : \quad \alpha_i = 0 \text{ for all } i,$$

and a suitable Alternative Hypothesis is

$$H_1: \quad \text{Some } \alpha_i \text{ is not zero.}$$

When H_0 is true,

$$F_R = \frac{cS_R^2}{S^2}$$

has an F-distribution with $(r-1)$ degrees-of-freedom for the numerator and $(r-1)(c-1)$ degrees-of-freedom for the denominator. The suggested $100\alpha\%$ level test is to reject H_0 if

$$F_R \geq F_{\alpha,(r-1),(r-1)(c-1)}.$$

This is a one-tailed test. Large values of F_R lead to the Null Hypothesis being rejected.

In the same way, one may want to test the Null Hypothesis that all the column effects β_j are zero. This is done in a similar way. With a $100\alpha\%$ level test one rejects the Null Hypothesis

$$H_0: \quad \beta_j = 0 \text{ for all } j$$

in favour of the Alternative Hypothesis

$$H_1: \quad \text{Some } \beta_j \text{ is not zero}$$

if for

$$F_C = \frac{rS_C^2}{S^2}$$

$$F_C \geq F_{\alpha,(c-1),(r-1)(c-1)}.$$

Example 9.3. Three varieties of potatoes are being compared for yield. The experiment was carried out by assigning each variety at random to four of twelve equal size plots, one being chosen in each of four locations. The following yields in bushels[7] per plot resulted:

[7] *A bushel is about 36.4 litres.*

	Potato		
Location	A	B	C
1	18	13	12
2	20	23	21
3	14	12	9
4	11	17	10

Test the hypothesis that there is no difference in the yielding capabilities of the three varieties.

Answer

This is a two-way classification, and we will assume that the yields are normally distributed with the same population variance. We assume that the population mean μ_{ij} for variety j in location i is of the form

$$\mu + \alpha_i + \beta_j.$$

Checking for no difference between yielding capabilities of the varieties is checking that the effects of Varieties 1, 2 and 3 are all the same. The classification by location helps to improve precision of estimation of variety effects by removing variation between locations.

The row means and overall mean are

$$\begin{aligned}
\bar{x}_{1.} &= 14.333 \\
\bar{x}_{2.} &= 21.333 \\
\bar{x}_{3.} &= 11.667 \\
\bar{x}_{4.} &= 12.667 \\
\bar{x}_{..} &= 15.000.
\end{aligned}$$

The estimated row effects $\hat{\beta}_j$ are the differences between the row means and the overall mean. So,

$$\begin{aligned}
\hat{\alpha}_1 &= -0.667 \\
\hat{\alpha}_2 &= 6.333 \\
\hat{\alpha}_3 &= -3.333 \\
\hat{\alpha}_4 &= -2.333.
\end{aligned}$$

A hand calculator routine will soon calculate from the row means the sample variance $s_R^2 = 19.037$, leading to the Mean Between Rows Sum of Squares $cs_R^2 = 3 \times 19.037 = 57.111$, and the Between Rows Sum of Squares $(r-1)cs_R^2 = 3 \times 57.111 = 171.333$. The Between Rows Sum of Squares is also the sum of the squares of the estimated row effects $\hat{\alpha}_i$ multiplied by c.

The column means and overall mean are

$$\begin{aligned}
\bar{x}_{.1} &= 15.75 \\
\bar{x}_{.2} &= 16.25 \\
\bar{x}_{.3} &= 13.00 \\
\bar{x}_{..} &= 15.00.
\end{aligned}$$

The estimated column effects $\hat{\beta}_j$ are the differences between the column means and the overall mean. So,

$$\hat{\beta}_1 = 0.75$$
$$\hat{\beta}_2 = 1.25$$
$$\hat{\beta}_3 = -2.00.$$

A hand calculator routine will soon calculate from the column means the sample variance $s_C^2 = 3.0625$, leading to the Mean Between Columns Sum of Squares $rs_C^2 = 4 \times 3.0625 = 12.25$, and the Between Columns Sum of Squares $(c-1)rs_R^2 = 2 \times 12.25 = 24.5$. The Between Columns Sum of Squares is also the sum of the squares of the estimated column effects $\hat{\beta}_j$ multiplied by r.

The sample variance of all the observations is $s_T^2 = 21.6364$, leading to the Total Sum of Squares as $(rc-1)s_T^2 = 11 \times 21.6364 = 238.000$. Subtracting the Sums of Squares Between Columns and Between Rows gives the Residual Sum of Squares $238.000 - 24.5 - 171.333 = 42.167$, and the Mean Residual Sum of Squares is $s^2 = 42.167/[(r-1)(c-1)] = 42.167/[3 \times 2] = 42.167/6 = 7.028$.

The test statistic to test the Null Hypothesis (that the variety effects are all zero) against the Alternative Hypothesis (that the variety effects are not all zero) is

$$F_C = \frac{rs_C^2}{s^2} = 12.25/7.028 = 1.74.$$

From Table 12(b) of the *New Cambridge Statistical Tables*, with $\nu_1 = (c-1) = 2$ and $\nu_2 = (r-1)(c-1) = 6$, we get $F_{.05,2,6} = 5.14$. Since the observed value for the test statistic does not exceed this critical value we do not reject the Null Hypothesis with a test at the 5% level of significance. There is no evidence of a difference between the varieties.

The ANOVA table is

Source	Sum of Squares	d.f.	Mean Sum of Squares	F
Between Locations	171.33	3	57.11	8.12
Between Varieties	24.50	2	12.25	1.74
Residual	42.17	6	7.03	
Total	238.00	11		

■

Confidence intervals

Single intervals

The principal interest for a two-way analysis is in the pairwise differences between row effects or column effects. Confidence intervals for differences of population row effects can be easily written down as in Chapter 7, using the result that for $i \neq j$

$$\frac{(\bar{X}_{i.} - \bar{X}_{j.}) - (\alpha_i - \alpha_j)}{S\sqrt{2/c}}$$

has a Student's t distribution with $(r-1)(c-1)$ degrees-of-freedom. The last result generalises those in Chapter 7 for two matched samples. It is easy to set intervals for differences in column effects in a symmetric fashion, just interchanging the roles of rows and columns.

Simultaneous intervals

We may want to think about setting intervals for all possible pairs of differences between any two α_is.

We can again use Scheffé's method. A $100(1-\alpha)$% set of simultaneous intervals for every linear combination $\sum_{i=1}^{k} d_i\alpha_i$, where $\sum_{i=1}^{k} d_i = 0$, is given by

$$\sum_{i=1}^{k} d_i\bar{X}_{i.} \pm S\sqrt{(r-1)F_{\alpha,r-1,(r-1)(c-1)}\sum d_i^2/c}.$$

In the particular case of a difference in population group means $\alpha_i - \alpha_j$ we have all $d_i = 0$ except for $d_i = -d_j = 1$, and the confidence interval is

$$\bar{X}_{i.} - \bar{X}_{j.} \pm S\sqrt{(r-1)F_{\alpha,r-1,(r-1)(c-1)}(2/c)}.$$

In a symmetric way one can set intervals for differences of column effects β_j.

Example 9.4. Continuing Example 9.3, a 95% confidence interval for $\beta_1 - \beta_2$ is

$$15.75 - 16.25 \pm \sqrt{7.03}t_{0.025,(6)}\sqrt{2/4}$$

where from Table 10 of the *New Cambridge Statistical Tables* with $P = 2.5$ and $\nu = 6$ we get $t_{0.025,(6)} = 2.447$. So the interval is

$$-0.5 \pm 4.69.$$

On the other hand, if we want a 95% confidence simultaneous set of intervals of the Scheffé type, then the corresponding interval is

$$15.75 - 16.25 \pm \sqrt{7.03}\sqrt{2F_{0.05,2,6}(2/4)}.$$

From Table 12(b) of the *New Cambridge Statistical Tables* with $\nu_1 = 2$, $\nu_2 = 6$ we get $F_{0.05,2,6} = 5.143$ and so the interval is

$$-0.5 \pm 6.01.$$

This is longer than the previous one, and since there are only three pairwise differences to look at, perhaps one does not need the Scheffé method here. Since we did not reject at the 5% level of significance the Null Hypothesis of no difference between variety effects in Example 9.3, we expect to see that 95% confidence interval of the Scheffé type always include zero.
■

Fitted values and residuals

In this section we briefly consider for two-way ANOVA how to see whether the model fits the data by directly comparing the observations with the estimated cell means. The model for two-way ANOVA postulates that the mean for the cell in the ith column and the jth row is

$$\mu + \alpha_i + \beta_j$$

where the sum of the row effects α_i and of the column effects β_j is zero. The overall population mean μ is estimated by the average of all the observations in the table. (For a simple notation, we will put 'hats' $\hat{}$ on a parameter to denote its estimator.)

$$\hat{\mu} = \bar{X}_{..} = \sum_{i=1}^{r} \sum_{j=1}^{c} X_{ij}/(rc).$$

The estimator of α_i is

$$\hat{\alpha}_i = \bar{X}_{i.} - \bar{X}_{..} = \sum_{j=1}^{c} X_{ij}/c - \bar{X}_{..}.$$

Notice that $\sum_i \hat{\alpha}_i = 0$. The estimator of β_j is

$$\hat{\beta}_j = \bar{X}_{.j} - \bar{X}_{..} = \sum_{i=1}^{r} X_{ij}/r - \bar{X}_{..}.$$

Notice that $\sum_j \hat{\beta}_j = 0$. The estimator of the cell mean μ_{ij} is

$$\begin{aligned} \hat{\mu}_{ij} &= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j \\ &= \bar{X}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) + (\bar{X}_{.j} - \bar{X}_{..}) \\ &= \bar{X}_{i.} + \bar{X}_{.j} - \bar{X}_{..}. \end{aligned} \tag{9.5}$$

The difference $\hat{\epsilon}_{ij} = X_{ij} - \hat{\mu}_{ij}$ is called the *residual*. It shows how far the observed value in cell (ij) is away from the estimated cell mean. A very large value for $\hat{\epsilon}_{ij}$

might show that the observation in cell (ij) was not consistent with the model for the two-way analysis.

We have

$$\hat{\epsilon}_{ij} = X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}$$

and one can see that $\hat{\epsilon}_{ij}$ sums to zero over i and over j.

Example 9.5. Continuing Example 9.4, the fitted values from (9.5) are

	Variety of Potato		
Location	A	B	C
1	15.08	15.58	12.33
2	22.08	22.58	19.33
3	12.42	12.92	9.67
4	13.42	13.92	10.67

For instance, the fitted value for variety B in location 2 is
Overall mean + Effect for Variety B + Effect for Location 2

$$= 15.00 + (16.25 - 15) + (21.33 - 15) = 22.58.$$

The residuals are found by taking the fitted values from the original observations to give

	Variety of Potato		
Location	A	B	C
1	2.92	−2.58	−0.33
2	−2.08	0.42	1.67
3	1.58	−0.92	−0.67
4	−2.42	3.08	−0.67

Notice that, within rounding errors, the rows and the columns of the residual table sum to 0. The sum of the squared residuals is 42.17, which is also[8] the Residual Sum of Squares in the analysis of variance table. There is no particular pattern or spectacularly large residual in the table here.

We can display the structure of the table of yields that we have found by fitting the

[8] *One would hope so.*

two-way model. We have

$$\begin{bmatrix} 1 & 18 & 13 & 12 \\ 2 & 20 & 23 & 21 \\ 3 & 14 & 12 & 9 \\ 4 & 11 & 17 & 10 \end{bmatrix} = \begin{bmatrix} 15.0 & 15.0 & 15.0 \\ 15.0 & 15.0 & 15.0 \\ 15.0 & 15.0 & 15.0 \\ 15.0 & 15.0 & 15.0 \end{bmatrix} + \begin{bmatrix} -0.67 & -0.67 & -0.67 \\ 6.33 & 6.33 & 6.33 \\ -3.33 & -3.33 & -3.33 \\ -2.33 & -2.33 & -2.33 \end{bmatrix} + \begin{bmatrix} 0.75 & 1.25 & -2.00 \\ 0.75 & 1.25 & -2.00 \\ 0.75 & 1.25 & -2.00 \\ 0.75 & 1.25 & -2.00 \end{bmatrix} + \begin{bmatrix} 2.92 & -2.58 & -0.33 \\ -2.08 & 0.42 & 1.67 \\ 1.58 & -0.92 & -0.67 \\ -2.42 & 3.08 & -0.67 \end{bmatrix}.$$

The original table of observations is displayed as the sum of an overall mean plus row effects plus column effects plus residuals.

■

Sum of squares identity

The fact that the sum of squared residuals is the Residual Sum of Squares in the ANOVA table implies a sum of squares identity. We can prove that directly as follows:

$$\begin{aligned} \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\hat{\mu})^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c}(\hat{\alpha}_i+\hat{\beta}_j+\hat{\epsilon}_{ij})^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c}(\hat{\alpha}_i^2+\hat{\beta}_j^2+\hat{\epsilon}_{ij}^2) \\ &= c\sum_{i=1}^{r}\hat{\alpha}_i^2 + r\sum_{j=1}^{c}\hat{\beta}_j^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}\hat{\epsilon}_{ij}^2. \end{aligned}$$

(After squaring all the cross-product terms sum to zero.) This sum of squares identity shows the addition property for the sums of squares in the analysis of variance table. One can think of the Between Rows Sum of Squares as $c\sum_{i=1}^{r}\hat{\alpha}_i^2$, the Between Columns Sum of Squares as $r\sum_{j=1}^{c}\hat{\beta}_j^2$ and the Residual Sum of Squares as $\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{\epsilon}_{ij}^2$.

Activity 9.7. What happens to all the sums of squares if one entry in a data table, for instance the data table in Example 9.3 on page 102, is allowed to become very large? This is the effect of a gross error in recording the data.
■

Learning outcomes

After working through this chapter you should be able to:

1. explain the purpose of analysis of variance

2. write down and interpret the models for one-way and two-way analysis of variance

3. carry out small examples of one-way and two-way analysis of variance with a hand calculator, presenting the results in an ANOVA table

4. carry out tests of hypotheses, and to write down confidence intervals as in this chapter

5. explain how to look at residuals from a two-way analysis of variance

6. derive sums of squares identities for one-way and for two-way analysis of variance.

Sample examination questions

1. (a) Explain and discuss the difference between one-way and two-way analysis of variance.

   (b) Explain qualitatively why in a one-way analysis of variance one rejects the null hypothesis of no differences between group means if the mean sum of squares between groups is large compared to the mean sum of squares within groups.

   (c) The table below shows measurements of sections taken from five European larch trees of the same age. Each section gives rise to 4 measurements of the trachoid length from each of the four aspects North, South, East and West.

	Aspect			
Tree	East	South	West	North
1	3.4	3.5	3.1	3.5
2	2.8	3.1	3.0	3.0
3	3.0	3.2	3.3	3.3
4	3.0	3.0	2.5	2.8
5	3.3	3.5	3.7	3.6

i. Give the analysis of variance table for a two-way analysis of variance for these data, using the classification by aspects and by tree number.

ii. Test the hypothesis that there is no difference between the trachoid lengths from different aspects.

(Elements of Statistics 2001, Zone A)

2. (a) Sometimes it is suggested that one carries out an analysis of variance on the logarithms of the original data. Why might this be a sensible transformation?

(b) The table below shows the percentage vote for the Democratic Party in US presidential elections of several different campaigns for different counties of Connecticut.

	Lich	Fairf	Middx	Toll
1920	32.5	30.9	33.1	31.0
1924	30.0	24.5	29.9	30.3
1928	36.0	43.7	39.7	39.6
1932	41.9	47.1	46.3	46.0

i. Give the analysis of variance table for a two-way analysis of variance for these data, using the classification by counties and by years.

ii. Are some year effects significantly different from 0?

iii. Are these data suitable for this form of analysis?

(Elements of Statistics 2001, Zone B)

3. (a) Give a model for the two-way analysis of variance, specifying the distribution of any random variables included in your model.

(b) Explain what is meant by interaction in a two-way analysis of variance.

(c) The table below shows the values of price index numbers for glasshouse fruit and vegetables (with base January 1969 at 100).

	Jan	Feb	March	April	May
1970	261	276	193	160	147
1971	214	239	193	2210	138
1972	332	248	208	164	128
1973	173	232	199	211	145
1974	328	314	259	209	121

i. Give the analysis of variance table for a two-way analysis of variance for these data, using the classification by years and by months.

ii. Give a set of 90% simultaneous confidence intervals for the differences between the first three years.

(Elements of Statistics 2000, Zone A)

Chapter 10

How to fit a straight line to a scatter plot of points.

Least squares

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], parts of chapter 12.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], parts of chapter 10.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods.* (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], parts of chapter 11.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction.* (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], parts of chapter 14.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics.* (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], 2.3, 2.4, parts of chapter 10.
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], parts of chapter 12.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 11.

Introduction

This chapter deals with fitting a straight line to a scatter plot by using the principle of least squares. There is some algebra to think about. Least squares is a principle linked to the normal distribution, and lies behind almost all the methodology taught in first year undergraduate statistics courses, though this is only obvious in the treatments of regression.

Response variable and explanatory variable

In some simple applications we are concerned to model the behaviour of one variable, the *response* variable[1] for fixed values of another variable, the *explanatory*[2] variable. One wants to predict or forecast or explain the values of the response variable from the values of the explanatory variable. One might for instance want to explain the amount of Local Government expenditure in a region by the proportion of poor households living there, or one might want to explain a measure of lung function for each of a collection of individuals by their ages. It is usual to label the response variable Y and the explanatory variable x. A simple model for a continuous measurement Y on a random individual chosen at random from those with explanatory variable value x is that

$$Y = \alpha + \beta x + \epsilon, \tag{10.1}$$

where ϵ is a random error[3] of measurement. This model says that the Y values for a particular choice x will be scattered above and below the straight line $y = \alpha + \beta x$. We will have for our investigation measurements $Y_1, Y_2, \ldots, Y_n$ of Y on[4] n individuals with corresponding known x values $x_1, x_2, \ldots, x_n$, and completely unknown errors of measurement $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$. The simplest assumptions are that the ϵs are random variables all with mean zero, the same variance σ^2, and uncorrelated. All these simple assumptions may be incorrect for a particular application, as we shall discuss later. With these assumptions the Y_is are random variables with means $\alpha + \beta x_i$, variance σ^2 and uncorrelated[5].

Equation (10.1) is called the population regression line. The parameter α is called the *intercept* of the population regression line, and the parameter β is called the *slope* of the population regression line.[6] Figure 10.1 shows a population regression line and nine examples of a sample of 15 observations y_i at the same chosen values x_i. From Figure 10.1 you can get some idea of how the observations fall around a population regression line.

Activity 10.1. Looking at Figure 10.1, would you expect that estimates of the regression based purely on the observed points would be close to the population line, or far away?
■

There is in the whole of regression methodology an asymmetry between the roles of the response and explanatory variables which is not readily apparent from such scatter diagrams alone.

Estimation of α and β

We do not know the population regression line, and must estimate it from (x_i, Y_i) for $i = 1, 2, \ldots, n$. One very useful idea is to find the intercept[7] A and slope B such

[1] *Also called the dependent, endogenous or y variable.*

[2] *Also called the independent, exogenous, regressor or x variable.*

[3] *We do not use an upper case Roman letter for this random variable for historical reasons.*

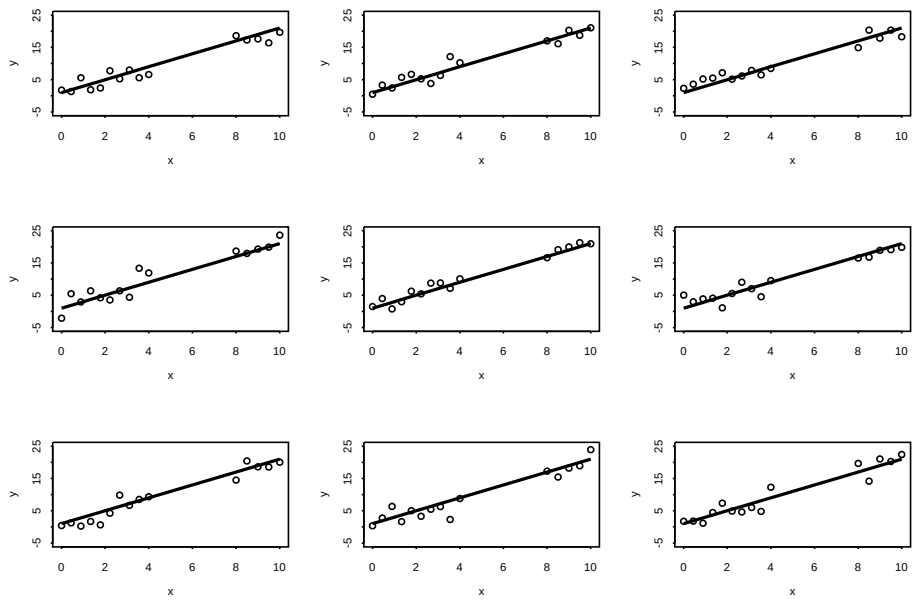
[4] *Notice the number of observation is n, not $2n$.*

[5] *The assumption of a common variance has been seen before.*

[6] *The word 'regression' is used for historical reasons.*

[7] *Both A and B are random variables.*

Figure 10.1: Nine examples of 15 random observations from the population with the population regression line shown.



that

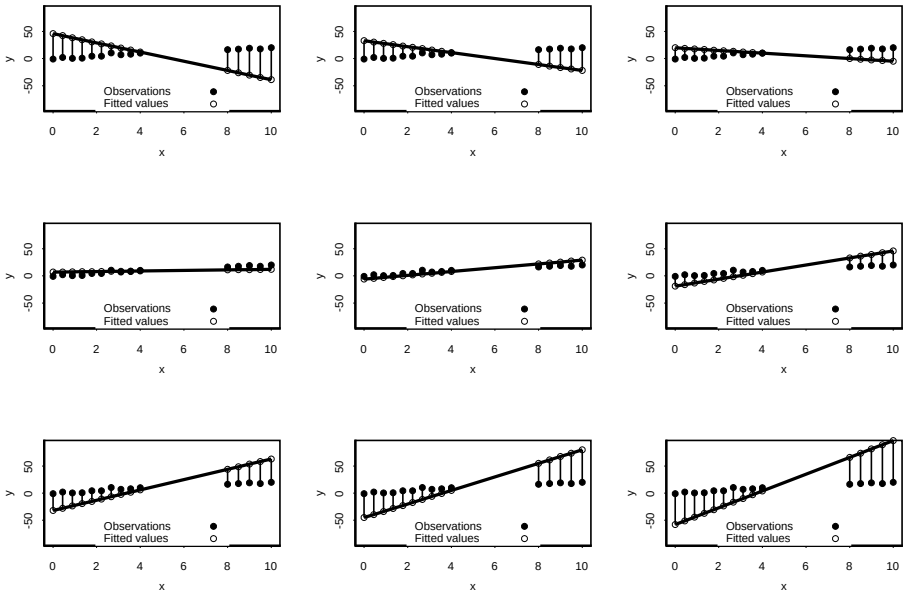
$$\sum_{i=1}^{n}(Y_i - A - Bx_i)^2 \leq \sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2, \tag{10.2}$$

for **all** possible α, β. The line $y = A + Bx$ is the *least squares fit* of the regression model. The line is chosen so that the sum of the squared distances of the Y_i from their estimated mean values $\hat{Y}_i = A + Bx_i$ is as small as possible, because it does at least as well as every one of the other possible lines $y = \alpha + \beta x$ on the right hand side of (10.2). We use A as a point estimate of α and B as a point estimate of β. The estimated mean values $\hat{Y}_i$ are called the *fitted values*.

Activity 10.2. Figures 10.2 on page 116 and 10.3 on page 117 show the same set of 15 points fitted by a variety of straight lines. In the first figure the lines go through the centre of gravity but have varying slopes. In the second figure the lines have something not too far off the right slope, but they don't all go through the centre of gravity of the observations. The line that fits best according to the least squares principal is the one that minimises the sum of the squared lengths of the little black

vertical lines that join the observations to the fitted values. Working by eye, choose the best line among those in Figure 10.2 and then among those in Figure 10.3.

Figure 10.2: Scatter plots for Activity 10.2



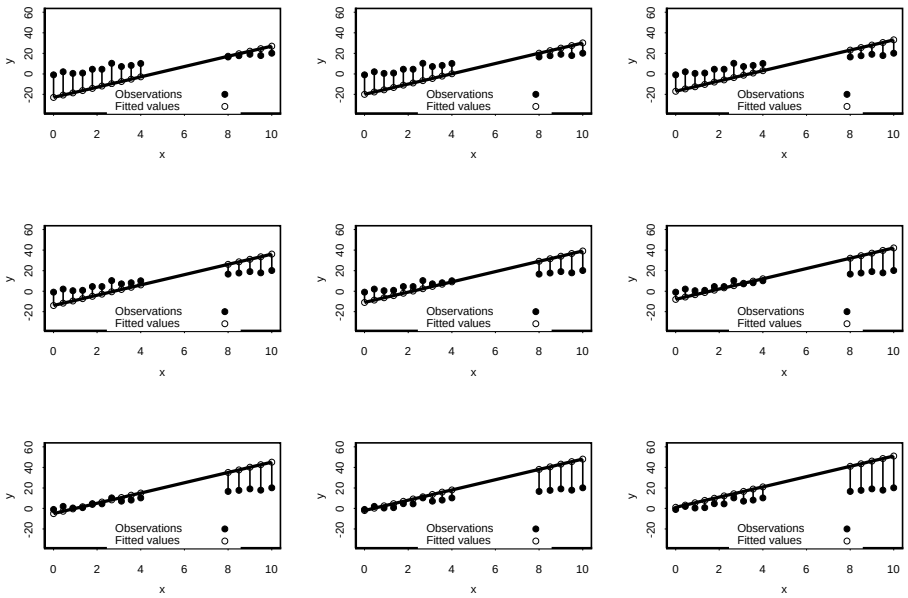
■

We now look at two special cases of regression models, using the least squares principle of (10.2) to find estimates.

Example 10.1. A very special case of the regression model is when $x_i = 0$ for all i. Then β plays no part in the model, which just says that the Y_is are uncorrelated random observations with mean α and variance σ^2. The least squares fit is obtained by choosing A so that

$$\sum_{i=1}^{n}(Y_i - A)^2 \leq \sum_{i=1}^{n}(Y_i - \alpha)^2$$

Figure 10.3: Scatter plots for Activity 10.2



for all possible α. Since

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\alpha)^2 &= \sum_{i=1}^{n}[(Y_i-\bar{Y})-(\bar{Y}-\alpha)]^2\\ &= \sum_{i=1}^{n}[(Y_i-\bar{Y})^2+(\bar{Y}-\alpha)^2]\\ &\geq \sum_{i=1}^{n}(Y_i-\bar{Y})^2,\end{aligned}$$

(the cross-products on squaring vanish after summing). It follows that $A=\bar{Y}$. The least squares estimator of α, which for this model is the population mean of the Y_is, is the sample mean $\bar{Y}$.

■

Example 10.2. Another very special case of the regression model is when we know that $\alpha=0$. Then the model is a straight line through the origin

$$Y_i=\beta x_i+\epsilon_i.$$

and the least squares fit chooses B so that

$$\sum_{i=1}^{n}(Y_i - Bx_i)^2 \leq \sum_{i=1}^{n}(Y_i - \beta x_i)^2$$

for all possible β.

If we make sure that

$$\sum_{i=1}^{n} x_i(Y_i - Bx_i) = 0,$$

which is true if

$$B = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

then

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \beta x_i)^2 &= \sum_{i=1}^{n}[(Y_i - Bx_i) - (B-\beta)x_i]^2 \\
&= \sum_{i=1}^{n}[(Y_i - Bx_i)^2 + (B-\beta)^2 x_i^2 + 2x_i(Y_i - Bx_i)(B-\beta)] \\
&= \sum_{i=1}^{n}[(Y_i - Bx_i)^2 + (B-\beta)^2 x_i^2] \\
&\geq \sum_{i=1}^{n}(Y_i - Bx_i)^2.
\end{aligned}$$

because the cross-products on squaring vanish after summing. It follows that the least squares estimator of β, the slope of the straight line through the origin is

$$B = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

One might wonder if there is anything intuitive about this estimator. To some extent there is, because an observation in the scatter diagram (x_i, Y_i) lies on a line through the origin with slope Y_i/x_i. The estimator B is just a weighted average of these slopes for all the observations (x_i, Y_i) with weights x_i^2,

$$B = \frac{\sum_{i=1}^{n} x_i^2 \frac{Y_i}{x_i}}{\sum_{i=1}^{n} x_i^2}.$$

It is not so obvious why one should use those weights x_i^2, though it's clear that observations with large x give better information about slope than points with small x.

Amusingly, if we choose $x_i = 1$ for all i we get back to Example 10.1.

■

Activity 10.3. Suppose that in Example 10.2 we miss out the observation (x_1, Y_1), and recalculate the slope estimate B. Find a formula for the difference between the old and new slope estimates. Check it works for $x_i = 1$, all i.
■

Finding A and B in the general case

Looking to the general model (10.1), we now find the least squares estimators A and B in the general case.

The *residual* $\hat{\epsilon}_i$ at x_i is the difference between the observation Y_i and the fitted value $\hat{Y}_i = A + Bx_i$. We will use these notations in what follows:

$$\hat{Y}_i = A + Bx_i,$$
$$\hat{\epsilon}_i = Y_i - \hat{Y}_i.$$

The least squares estimators A and B must be chosen so that (10.2) holds. Suppose we have chosen A and B so that

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0 \tag{10.3}$$
$$\sum_{i=1}^{n} x_i \hat{\epsilon}_i = 0. \tag{10.4}$$

Then we can find a sums of squares identity for the regression, since

$$\begin{aligned}
\sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2 &= \sum_{i=1}^{n} [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \alpha - \beta x_i)]^2 \\
&= \sum_{i=1}^{n} [\hat{\epsilon}_i + (A - \alpha) + (B - \beta) x_i]^2 \\
&= \sum_{i=1}^{n} [\hat{\epsilon}_i^2 + \{(A - \alpha) + (B - \beta) x_i\}^2] \\
&= \sum_{i=1}^{n} \hat{\epsilon}_i^2 + \sum_{i=1}^{n} [\hat{Y}_i - \alpha - \beta x_i]^2 \qquad (10.5)
\end{aligned}$$

because from the assumptions (10.3) and (10.4) the sums of cross-products vanish after squaring. We have found a sums of squares identity for regression, which is true for **all** choices of α, β provided A and B satisfy (10.3) and (10.4). It follows from (10.5) that when those conditions are satisfied

$$\sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2 \geq \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

for all α, β, and so A and B are least squares estimators of the regression parameters α, β.

Equations (10.3) and (10.4) can be written

$$\sum_{i=1}^{n}(Y_i - A - Bx_i) = 0$$
$$\sum_{i=1}^{n} x_i(Y_i - A - Bx_i) = 0.$$

The first reduces to

$$\bar{Y} - A - B\bar{x} = 0 \tag{10.6}$$

which is

$$A = \bar{Y} - B\bar{x}. \tag{10.7}$$

Substituting for A in the second equation

$$\sum_{i=1}^{n} x_i(Y_i - \bar{Y} + B\bar{x} - Bx_i) = 0$$
$$\sum_{i=1}^{n} x_i(Y_i - \bar{Y}) = B\sum_{i=1}^{n} x_i(x_i - \bar{x})$$
$$B = \frac{\sum_{i=1}^{n} x_i(Y_i - \bar{Y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \tag{10.8}$$

Very often one rewrites (10.8) as

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{10.9}$$

or as

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

Activity 10.4. Show that $\sum_{i=1}^{n} x_i(Y_i - \bar{Y}) = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$ and that $\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$.
■

Notice that (10.6) says that the fitted line $y = A + Bx$ goes through the point $(\bar{X}, \bar{Y})$. The result (10.9) shows that one can think of B as a weighted average of estimators of the slope formed from each pair $(x_i - \bar{x}, Y_i - \bar{Y})$ with weights $(x_i - \bar{x})^2$, where

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

The intuitive content of this is that we have decided that the estimated regression line will go through the point $(\bar{X}, \bar{Y})$, and so each pair $(x_i - \bar{x}, Y_i - \bar{Y})$ will suggest a slope of

$$\frac{Y_i - \bar{Y}}{x_i - \bar{x}}$$

The final estimate of slope is a weighted combination of all of these, with more weight being given to those points far from $(\bar{X}, \bar{Y})$ because they give more accurate information about the slope.

Fortunately, most calculators will compute A and B if the observations are entered, so one need not bother with complicated calculations based on (10.7) and (10.8). Those formulae help to understand the properties of the estimates, but are not needed for calculation.

Sums of squares identity

We found in (10.5) that for **all** values of α, β we have an identity

$$\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2 + \sum_{i=1}^{n}[\hat{Y}_i - \alpha - \beta x_i]^2.$$

Choosing $\alpha = \bar{Y}$ and $\beta = 0$ and using (10.6) we get a useful special case.

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}\hat{\epsilon}_i^2 + \sum_{i=1}^{n}[\hat{Y}_i - \bar{Y}]^2 \\ &= \sum_{i=1}^{n}\hat{\epsilon}_i^2 + \sum_{i=1}^{n}[A + Bx_i - A - B\bar{x}]^2 \\ &= \sum_{i=1}^{n}\hat{\epsilon}_i^2 + B^2\sum_{i=1}^{n}(x_i - \bar{x})^2\end{aligned}$$

This shows a Total Sum of Squares $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = (n-1)S_y^2$ (where S_y^2 is the sample variance of all the Y_is) broken into a Sum of Squares for the Slope

$$\begin{aligned}S_\beta^2 &= B^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= (n-1)B^2 s_x^2 \qquad (10.10)\end{aligned}$$

where s_x^2 is the sample variance of the x_is, and the Residual Sum of Squares

$$(n-2)S^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2.$$

The Mean Residual Sum of Squares is S^2, and the Mean Sum of Squares for the slope is the same as the Sum of Squares for the Slope.

With a hand calculator the Residual Sum of Squares can be found from

$$(n-2)S^2 = (n-1)S_y^2 - (n-1)B^2 s_x^2 = (n-1)[S_y^2 - B^2 s_x^2]. \qquad (10.11)$$

Example 10.3. The following table shows the proportions of non-manual workers and the Labour share of the vote in the 1970 General Election for constituencies in Kent (excluding the area transferred to Greater London).

Constituency	% of non-manual workers	% Labour share of vote
Ashford*	28.9	32.8
Canterbury*	25.3	41.0
Dartford*	45.0	36.6
Dover	48.6	32.4
Faversham	46.6	26.6
Folkestone & Hythe	32.8	39.6
Gillingham	41.2	33.9
Gravesend*	45.0	32.4
Isle of Thanet*	33.7	39.0
Maidstone	30.3	38.8
Rochester & Chatham	45.2	22.9
Sevenoaks*	25.5	44.4
Tonbridge*	29.9	45.5

Note: * Denotes a Liberal candidate in the constituency.

Source: Butler & Pinto-Duchinsky The British General Election of 1970

Plot a scatter diagram and then fit a straight line by least squares, treating the % Labour vote as the response variable and the % non-manual workers as the regressor variable.

Answer

Entering the data into a calculator operating in Regression Mode will quickly lead to the estimates

$$a = 56.86$$

and

$$b = -0.5717.$$

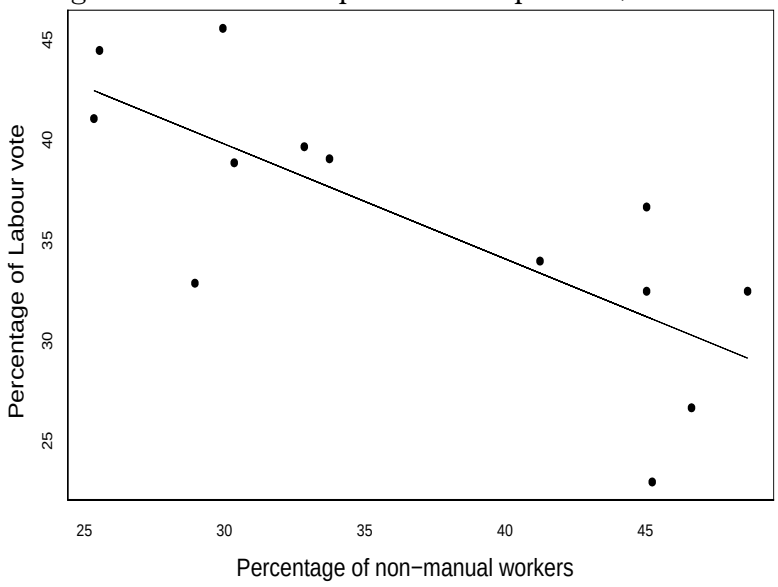
This gives the fitted line[8]

$$\text{Labour \% share of vote } = 56.86 - 0.5717 \text{ \% of non-manual workers.}$$

[8] *It is fairly easy to get the response variable and the explanatory variable confused when calculating.*

Figure 10.4 shows the observations plotted along with this fitted line. Notice how the observations lie on either side of the line, which goes through the centre of gravity of the points. The slope is negative, since the more non-manual workers in a constituency, the smaller is the percentage Labour vote.

Figure 10.4: A scatter plot for Example 10.3, and the fitted line



One might consider whether the assumptions that make least squares appropriate are true for this application. One assumption is that the variances are the same for each observation of the Labour percentage share of vote. If some of the percentages were too extreme, say close to 100% or to 0%, then this assumption could hardly be true. Intuitively, a very high or a very low percentage has less variability than one towards 50%. However, the percentages here are not too extreme.

Are the percentages in any way random values of uncorrelated random variables? It is a bit hard to think of the % Labour votes as being subject to random measurement error. As in many practical applications of regression, the theoretical framework is idealised.

■

Sample covariance and sample correlation

We have so far thought of the explanatory variable as having values that are fixed. Sometimes this is more a matter of convenience than truth, for the values of the explanatory variables may themselves come from measurements on individuals randomly chosen without regard to the value of the explanatory variable. This fluidity in the origins of observations to which lines are fitted leads to links with the ideas of covariance and correlation between two random variables in Chapter 5. On page

46 the covariance of two random variables X and Y was defined. There is a similar quantity defined for a sample of n pairs of observations (X_i, Y_i), $i = 1, 2, \ldots, n$ in a random sample. We define the *sample covariance* as

$$S_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}. \tag{10.12}$$

The *sample correlation coefficient* is defined as

$$R_{xy} = \frac{S_{xy}}{S_x S_y}. \tag{10.13}$$

Like its population counterpart, a sample correlation coefficient lies between -1 and $+1$. The sample covariance and correlation coefficient have similar interpretations to those for the population analogues.

Using these definitions, one can recast (10.10) for the sum of squares for the slope as

$$S_\beta^2 = B^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-1)R_{xy}^2 S_y^2.$$

The Residual Sum of Squares in (10.11) can be written

$$(n-2)S^2 = (n-1)S_y^2(1 - R_{xy}^2). \tag{10.14}$$

It is (10.14) that allows the easiest calculation of the Residual Sum of Squares, because the regression routine on a hand calculator will supply the value r_{xy} of the sample correlation coefficient at the same time as the intercept a and the slope b.

Example 10.4. Continuing Example 10.3, let us write down the values in the sum of squares identity. The Total Sum of Squares is $(n-1)s_y^2 = (13-1) \times 6.5585^2 = 516.167$. The Correlation coefficient is $r_{xy} = -0.75391$, so that the Residual Sum of Squares is $516.167[1 - (-0.75391)^2] = 222.787$. The Sum of Squares for the slope is

$$s_\beta^2 = (n-1)b^2 s_x^2 = (13-1)(-0.57174)^2 8.6482^2 = 293.380.$$

One can see that $516.167 = 222.787 + 293.380$. The Mean Residual Sum of Squares is $222.787/(13-2) = 20.253$. This is an estimate of the error variance σ^2.
■

Learning outcomes

After working through this chapter you should be able to:

1. explain the difference between a response variable and an explanatory variable
2. discuss the idea of minimising sums of squared residuals to obtain a fitted straight line

3. derive the algebraic formulae for A and B from first principles, and prove that the sums of squares identities hold
4. fit a simple linear regression with a hand-calculator, and find the residual sum of squares
5. give the definitions of sample covariance and correlation and show how to use them in regression.

Sample examination questions

1. (a) Derive from first principles the least squares estimator of slope for a simple linear regression.

   (b) The table below shows the population of England and Wales in millions for years in the 19th century.

Year	1801	1811	1821	1831	1841	1851	1861	1871
Popn.	8.89	10.16	12.00	13.90	15.91	17.93	20.07	22.71

   i. Find the least squares fit of a regression model for response variable population and explanatory variable year. Give the intercept and slope of the fitted line.

   ii. Should you fit a straight line through (0,0) to these data rather than allowing an arbitrary intercept?

   iii. How would your fitted regression line change if the population were measured in thousands?

   (Part of a question from Elements of Statistics 2001, Zone B)

2. (a) Find from first principles the least squares estimator for the slope of a line through the origin fitted to n pairs of values (x_i, Y_i).

   (b) The table below shows Regional Manufacturing Capital Stock Estimates in millions of pounds sterling at 1970 prices in the Wales and in the Scotland.

Year	1950	1951	1952	1953	1954	1955	1956	1957	1958
Wales	1116	1162	1219	1256	1316	1381	1426	1500	1563
Scotland	1746	1815	1868	1918	1958	2011	2066	2110	2153

i. Find the least squares fit of a regression model for response variable Scotland Capital Stock and explanatory variable Wales Capital Stock.

ii. Interpret your regression line.

(Part of a question from Elements of Statistics 2000, Zone A)

3. (a) Derive from first principles the least squares estimators of intercept and slope for a simple linear regression model.

(b) The following table shows the proportions of part-time women employees in Great Britain according to the New Earnings Survey (NElements of Statistics) and the Labour Force Survey (LFS), over several recent years.

Year	NES	LFS
1985	32.6	44.6
1986	32.9	45.0
1987	32.6	45.0
1988	32.6	44.5
1989	31.9	43.7
1990	32.8	43.3
1991	33.0	43.4
1992	33.9	43.8
1993	34.7	43.7

i. Make a scatter diagram for these data.

ii. Fit a regression model with response variable the LFS percentages, and explanatory variable the NES percentages.

iii. Is your fitted model sensible?

(Part of a question from Elements of Statistics 1999, Zone A)

Chapter 11

Tests and confidence intervals for regression models.

Simple linear regression

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 12.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], chapter 10.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods.* (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], chapter 11.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], chapter 14.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], 10.1 and 10.2.
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], 12.3, 12.4, 12.5.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 12.

Introduction

We carry through the ideas of least squares fitting, using further assumptions that allow the use of confidence intervals and tests.

The model for linear regression

As in Chapter 10 we have n observations $Y_1, Y_2, \ldots, Y_n$ on a response variable y. Each observation Y_i is associated with a value x_i of the explanatory variable x. We now assume that the observations Y_i are independent, and with distribution $N(\alpha + \beta x_i, \sigma^2)$. The additional assumptions over those in Chapter 10 are the normal distributions for, and independence of, the Y_is rather than just their zero correlation.

The point estimators for α and β are, as before, the least squares estimators

$$A = \bar{Y} - B\bar{x}, \tag{11.1}$$
$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Means and variances of A and B

The expected value of B under repeated sampling of Y_is for the same fixed x_is is

$$\begin{aligned} E[B] &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})E[Y_i]}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\beta\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\beta\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \beta. \end{aligned}$$

So B is an unbiased estimator of β. The variance of B is

$$\begin{aligned} \mathrm{var}[B] &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\,\mathrm{var}[Y_i]}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \sigma^2/[(n-1)s_x^2]. \end{aligned}$$

The expected value of A under repeated sampling for fixed x_is is

$$\begin{aligned} E[A] &= E[\bar{Y}] - E[B]\bar{x} \\ &= \alpha + \beta\bar{x} - \beta\bar{x} \\ &= \alpha. \end{aligned}$$

So A is an unbiased estimator of α. The variance of A under repeated sampling for fixed x_is is

$$\mathrm{var}[A] = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right]. \tag{11.2}$$

We shall not prove the last result.

Activity 11.1. Find the mean and variance of the estimator of slope for the model of a regression line through the origin, as in Example 10.2 on page 117.
■

If one knew the variance σ^2 then these results would be enough to allow the construction of confidence intervals and tests using the standard normal distribution, because with the assumptions of this chapter, both A and B have normal distributions. For instance, one could write down a 95% confidence interval for β as

$$B \pm 1.96\sqrt{\sigma^2/[(n-1)s_x^2]}. \tag{11.3}$$

In practice we will not know σ^2, and will estimate this by the Mean Residual Sum of Squares S^2 from (10.11). It is in fact true that $(n-2)S^2/\sigma^2$ has a χ^2 distribution with $(n-2)$ degrees-of-freedom, and that it is independent of A and of B. These distributional results, not proved here, allow the formation of tests and intervals using Student's t distribution. For instance, instead of (11.3) we get

$$B \pm t_{0.05,(n-2)}\sqrt{S^2/[(n-1)s_x^2]}. \tag{11.4}$$

There is no need to give much further detail on confidence intervals and tests, since they all use Student's t distribution and follow the patterns seen before in other chapters.

Example 11.1. Continuing Example 10.4 on page 124, a 90% confidence interval for β is given from (11.4) as

$$-0.5717 \pm t_{0.05,(11)}\sqrt{20.253/(12 \times 8.6482^2)}$$

where from Table 10 of the *New Cambridge Statistical Tables* with $\nu = 11$ and $P = 5$, $t_{0.05,(11)} = 1.796$. The 90% interval is

$$-0.5717 \pm 1.796 \times 0.1502$$

which is

$$-0.5717 \pm 0.270.$$

■

Interval estimates for fitted values

The fitted value $\hat{Y}$ for a value x of the explanatory variable is the best estimate of the mean of the response variable Y for that value x. It is given by

$$\hat{Y} = A + Bx.$$

This quantity can usually be found directly with a regression routine on a hand-calculator for any given x without any need to type in A and B. The expected value of $\hat{Y}$ is

$$E[A] + E[B]x = \alpha + \beta x,$$

which is the expected value of Y at x. So $\hat{Y}$ is an unbiased estimator of the expected value of Y at x. The variance of $\hat{Y}$ is

$$\sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right]. \tag{11.5}$$

Since $\hat{Y}$ has a normal distribution, and is independent of the Residual Sum of Squares $(n-2)S^2$, we can form confidence intervals for $\alpha + \beta x$ using Student's t distribution.

Activity 11.2. Why is (11.5) the same as (11.2) when $x = 0$?
■

Activity 11.3. Find the variance of $\hat{Y}$ for the model in in Example 10.2 on page 117.
■

Example 11.2. Continuing Example 11.1 we can find an 80% confidence interval for the expected percentage Labour vote if the percentage of non-manual workers is $x = 35\%$.

The fitted value is

$$56.86 - 0.5717 \times 35 = 36.85.$$

The estimated standard deviation is, from (11.5),

$$\sqrt{20.253 \left[\frac{1}{13} + \frac{(35 - 36.769)^2}{(12 \times 8.6482^2)} \right]} = 1.276.$$

The upper 10% point of the Student's t distribution with 11 degrees-of-freedom is from Table 10 of the *New Cambridge Statistical Tables* with $\nu = 11$ and $P = 10$, $t_{0.1,(11)} = 1.363$. The 80% confidence interval for the mean percentage Labour vote when the percentage of non-manual workers is 35% is

$$36.85 \pm 1.363 \times 1.276$$

which is

$$36.85 \pm 1.739.$$

■

Sometimes one wants to find an interval not for the mean of the response variable for a given x as in Example 11.2, but which covers a new observation Y at that x-value. This is more variable than $\hat{Y}$, and so the interval that covers it (usually called a *prediction* interval) is longer. The variance in (11.5) is increased to

$$\sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right]. \tag{11.6}$$

Example 11.3. If we want to cover the percentage Labour vote in another constituency, not included in the original regression, with a percentage of non-manual workers of 35%, an 80% interval becomes

$$36.85 \pm 1.363 \times \sqrt{20.253 \left[1 + \frac{1}{13} + \frac{(35 - 36.769)^2}{(12 \times 8.6482^2)} \right]}$$

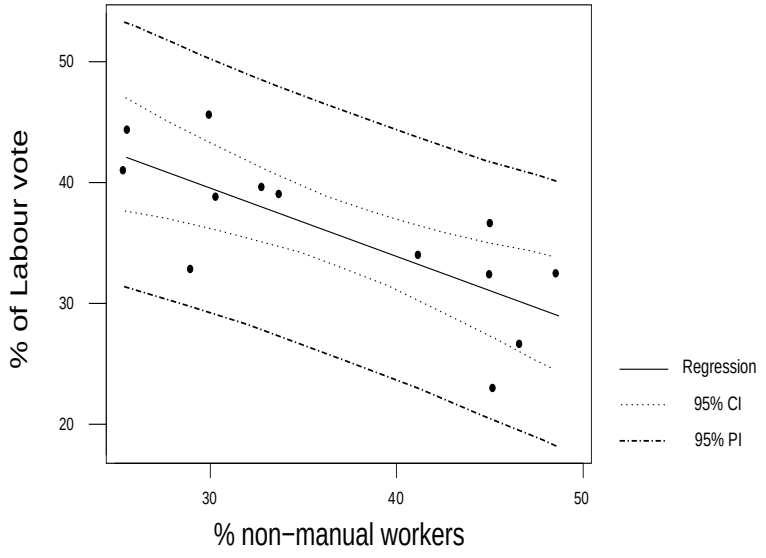
which is

$$36.85 \pm 6.376.$$

■

We can show the prediction intervals and the confidence intervals for the mean values on the scatter diagram. They both get wider as the value x moves away from the mean $\bar{x}$. Figure 11.1 shows 95% intervals on the scatter diagram for Example 11.3.

Figure 11.1: Scatter plot, fitted line and intervals for Example 11.3

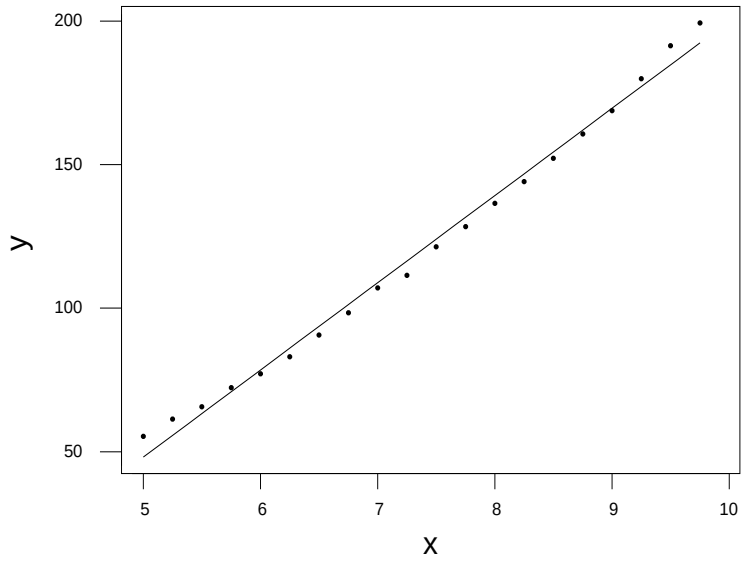


Spotting difficulties

The scatter diagram will allow us to see if the model is not likely to be correct. One should check:

- That the points lie roughly on a straight line (not a curve). Figure 11.2 shows a case where a quadratic curve would fit much better than the straight line.

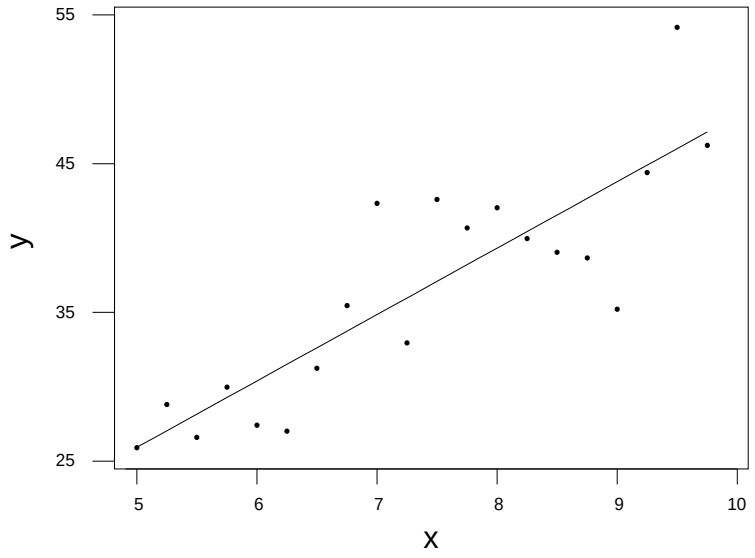
Figure 11.2: Scatter plot with fitted regression; quadratic curve needed



- That the scatter of points around the fitted line does not show any obvious pattern such as alternating above and below, or wider scatter with increasing x. Figure 11.3 shows the scatter around the line becoming larger as x increases.
- That there is no 'wild' outlying point far from the others, and not following the same regression line. Such a point can distort the least squares fitting. Figure 11.4 shows how the fitted line can be thrown out by a single very bad observation.

Activity 11.4. Which assumptions go wrong in each of the cases just mentioned? ■

Figure 11.3: Scatter plot with fitted regression; variation increasing with x



Example 11.4. The number of pages and the price in dollars of 15 books reviewed in the February 1982 issue of the journal *Technometrics* are given below:

Pages	302	425	526	532	145	556	426	359	465	246
Price	30	24	35	42	25	27	64	59	55	25
Pages	143	557	372	320	178					
Price	15	29	30	25	26					

Fit a straight line to these data using price as the response variable and number of pages as the regressor variable. Test the hypothesis that the population slope is zero against the alternative that it is greater than zero. Find a 95% confidence interval for the population slope. Find a 90% confidence interval for the mean price of a book with 250 pages, and a 90% prediction interval for the price of a book with 250 pages.

Answer

The first thing to do is to look at a scatter plot, to get some idea of what one should find. This gives Figure 11.5.

It is evident that there is no very close relation between the number of pages and the price. The three most expensive books seem to lie away from the line that best fits

Figure 11.4: Scatter plot with fitted regression; showing one very bad observation

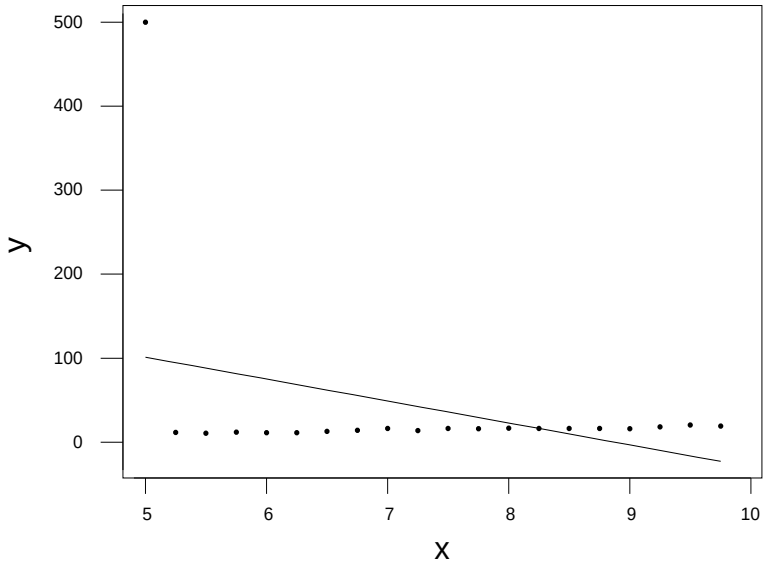
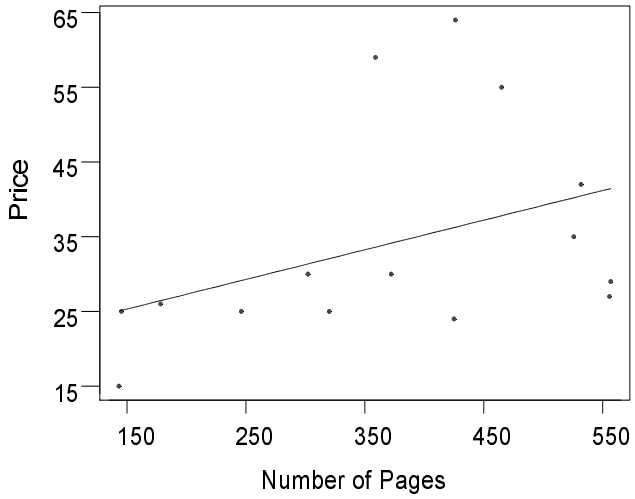


Figure 11.5: Scatter plot for page numbers and prices



the others, so there is good reason to think that a simple linear regression model may not best describe these data. If all the data are included, there seems to be a positive slope in the scatter diagram, though the association does not seem strong. There is a lot of variation around the best fitting line. Most statisticians would see little point in fitting a straight line to these data, and would expect the results to be of very little use.[1]

[1] *Think where one can see in the model-fitting carried out below that the model is not very good.*

A calculator regression routine will, after inputting the $n = 15$ pairs of observations, give

$$\begin{aligned} a &= 19.399 \\ b &= 0.03963 \\ s_x &= 145.6448 \\ s_y &= 14.4047 \\ r_{xy} &= 0.40069 \\ \bar{x} &= 370.133. \end{aligned}$$

The regression line is

$$\text{Price } = 19.399 + 0.03963 \times \text{Number of Pages}.$$

The estimate of the residual variance is

$$\begin{aligned} s^2 &= (n-1)s_y^2(1-r_{xy}^2)/(n-2) = 14(14.4047)^2(1-0.40069^2)/13 \\ &= 187.5 = 13.69^2. \end{aligned}$$

The Null Hypothesis is
H_0: $\beta = 0$
and the Alternative Hypothesis is
H_1: $\beta > 0$
The estimated standard deviation of B is $13.69/(145.6448\sqrt{14}) = 0.02511$. The test statistic is

$$\frac{B - 0}{\text{estimated standard deviation}}$$

which has value

$$\frac{0.03963}{0.02511} = 1.578.$$

The Null Hypothesis will be rejected at the 5% level of significance if the value of the test statistic is greater than $t_{0.05,(13)}$. From Table 10 of the *New Cambridge Statistical Tables* with $\nu = 13$ and $P = 5$, $t_{0.05,(13)} = 1.771$.

The Null Hypothesis that $\beta = 0$ is not rejected at the 5% level of significance. This is a one-tailed test for a one-sided Alternative Hypothesis.

The 95% confidence interval for the slope β is, from (11.4)

$$0.03963 \pm t_{0.025,(13)} 13.69/(\sqrt{(14)145.6448})$$
$$\text{ie } 0.03963 \pm 2.16 \times 13.69/(\sqrt{(14)145.6448})$$
$$\text{ie } 0.03963 \pm 2.16 \times 0.02511$$
$$\text{ie } 0.03963 \pm 0.0543.$$

The value 2.16 is from Table 10 of the *New Cambridge Statistical Tables* for $\nu = 13$ and $P = 2.5$.

Notice that this interval includes negative values, so we can't even be sure that the relation between number of pages and price is in the right direction. We should also think carefully whether the intercept $a = 19.399$ is needed for this model. It would correspond to a fixed charge for each book irrespective of its length, though such a charge is not entirely plausible.[2] Since the slope of the line is so imprecisely available, the confidence intervals for fitted values and the prediction intervals are likely also to be very wide.

[2] *It is not plausible for a book of 0 pages*

The estimated mean price of a book with 250 pages is

$$19.399 + 0.03963 \times 250 = 29.31.$$

The estimated standard error of this fitted value is, see (11.5),

$$13.69\sqrt{1/15 + (250 - 370.133)^2/(14 \times 145.6448^2)}$$
$$= 13.69 \times 0.3395$$
$$= 4.65.$$

The 90% confidence interval for the mean price of a 250 page book is

$$29.31 \pm 1.771 \times 4.65 = 29.31 \pm 8.23.$$

This interval is perhaps a bit shorter than we might have expected, but only a few cheap and a few expensive books in the original data are outside it. The figure 1.771 is from Table 10 of the *New Cambridge Statistical Tables* with $P = 5$ and $\nu = 13$.

The 90% prediction interval for an individual price of a book with 250 pages follows from (11.6). The estimated standard error is now

$$13.69\sqrt{1 + 1/15 + (250 - 370.133)^2/(14 \times 145.6448^2)} = 13.69 \times 1.056 = 14.46$$

The 90% prediction interval is

$$29.31 \pm 1.771 \times 14.46 = 29.31 \pm 25.60.$$

An interval this long is of very little use, and reflects the large variation in the data. Routine use of statistical methods avoids overly optimistic interpretation of data, though

it does not lead to good relations between a statistician and an over-eager client. The underlying problem is that there is too much variation in the data for a model to give good predictions.
■

Learning outcomes

After working through this chapter you should be able to:

1. derive from first principles the means of A and B and the variance of B
2. explain how to set confidence intervals and carry out tests about α and β from a small collection of data
3. demonstrate how to set confidence intervals for $\hat{Y}$, and know how to calculate a prediction interval and explain the difference between the two
4. discuss ways in which the regression model may not be appropriate, and how to spot them.

Sample examination questions

1. (a) Derive from first principles the least squares estimator of slope for a simple linear regression.

   (b) The table below shows the population of England and Wales in millions for years in the 19th century.

Year	1801	1811	1821	1831	1841	1851	1861	1871
Popn.	8.89	10.16	12.00	13.90	15.91	17.93	20.07	22.71

   i. Test the null hypothesis that the population regression slope is 0.21.

   (Part of a question from Elements of Statistics 2001, Zone B)

2. (a) Show that the least squares estimators of intercept and slope are unbiased estimators of the corresponding population parameters.

   (b) The table below shows heights in cm of male children on their fourth and fifth birthdays.

Child	1	2	3	4	5
Fourth Birthday	100.0	95.1	103.3	98.2	98.8
Fifth Birthday	105.5	101.5	110.0	104.5	104.8
Child	6	7	8	9	10
Fourth Birthday	103.0	98.6	97.5	95.3	97.7
Fifth Birthday	109.0	105.5	102.5	100.4	103.6

i. Find the least squares fit (ie intercept and slope) of a regression model for response variable height at fifth birthday and explanatory variable height at fourth birthday, and interpret your fitted line.

ii. Give a 90% confidence interval for the mean height on the fifth birthday for a height on fourth birthday of 98 cm.

iii. Test the null hypothesis that the population regression slope is 0.1.

3. (a) Derive from first principles the variance of the estimator of slope for a simple linear regression.

(b) The table below shows Regional Manufacturing Capital Stock Estimates in millions of pounds sterling at 1970 prices in the West Midlands and in the East Midlands.

Year	1950	1951	1952	1953	1954	1955	1956	1957	1958
West Midlands	2649	2742	2834	2918	3001	3114	3246	3385	3495
East Midlands	1748	1810	1854	1903	1944	1982	1991	2012	2028

i. Find the least squares fit of a regression model for response variable East Midlands Capital Stock and explanatory variable West Midlands Capital Stock.

ii. Give the analysis of variance table for this regression.

iii. Test the null hypothesis that the population regression slope is 0.

iv. Are the usual assumptions for inference on a regression model satisfied in this case?

(Elements of Statistics 2000, Zone B)

How to measure association between two continuous random variables.

Chapter 12

Correlation

Essential reading

Newbold, P. *Statistics for Business and Economics*. (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 12.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], sections 11.1, 11.2.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], 3.5, 3.6, chapter 11.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], chapter 13.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], chapter 2.
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], section 12.6.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 15.

Introduction

This chapter follows on from the work on regression in chapters 10, 11 and allows the understanding of some alternative ways of presenting the results. A connection between regression and analysis of variance becomes apparent. Correlations are very important for all work with many variables.

Correlation between two random variables

We saw on page 46 how to define the covariance between two random variables X and Y by using

$$\mathrm{cov}(X, Y) = E[(X - E(X))(Y - E(Y)] = E[XY] - E[X]E[Y],$$

and that the correlation coefficient of X and Y as

$$\rho_{xy} = \mathrm{cov}(X, Y)/\sqrt{\mathrm{var}\, X\, \mathrm{var}\, Y}.$$

For a random sample of n pairs of observations from a joint distribution of (X, Y), we can define S_{xy}, the sample covariance of X and Y where, as in (10.12),

$$S_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}.$$

The *sample correlation coefficient* is defined as in (10.13) by

$$R_{xy} = \frac{S_{xy}}{S_x S_y}.$$

The sample covariance and the sample correlation coefficient are used as estimators of the population covariance and correlation.

Notice that we can write the sample covariance as $R_{xy}S_xS_y$. As is the case for ρ_{xy}, it is true that

$$-1 \leq R_{xy} \leq 1.$$

This is easy to see, for since

$$\sum_{i=1}^{n}[(X_i - \bar{X}) - S_{xy}(Y_i - \bar{Y})/S_y^2]^2/(n-1) \geq 0, \tag{12.1}$$

on expanding the squared term we get

$$S_x^2 - 2S_{xy}^2/S_y^2 + S_{xy}^2/S_y^2 \geq 0$$

so that

$$S_x^2[1 - 2R_{xy}^2 + R_{xy}^2] = S_x^2[1 - R_{xy}^2] \geq 0.$$

It follows that $R_{xy}^2 \leq 1$, which implies $-1 \leq R_{xy} \leq 1$.

In fact, the extreme values of -1, +1 can only be reached if (X_i, Y_i) all lie exactly on a straight line with negative or positive slope respectively.

Activity 12.1. Show, by changing (12.1) to an equality, that if $R_{xy}^2 = 1$, then (X_i, Y_i) all lie exactly on a straight line.
■

This strongly suggests that the correlation coefficient measures the degree to which (X_i, Y_i) lie on a straight line. Other types of association[1] may not give a correlation coefficient much different from 0.

[1] *Association is not meant to be precisely defined*

Figures 12.1, 12.2 and 12.3 show some random samples from *bivariate normal* distributions with different population correlation coefficients. In each case the sample correlation coefficient is also given.

Figure 12.1: Scatter plot to show correlations

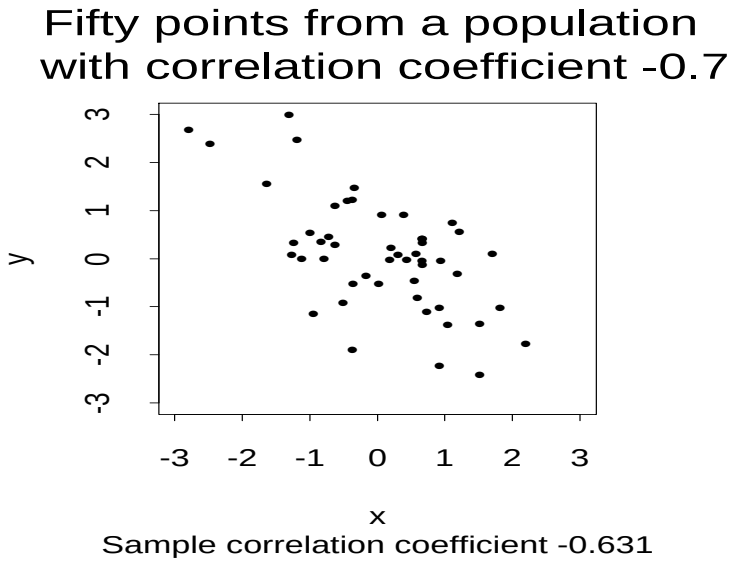


Figure 12.2: Scatter plot to show correlations

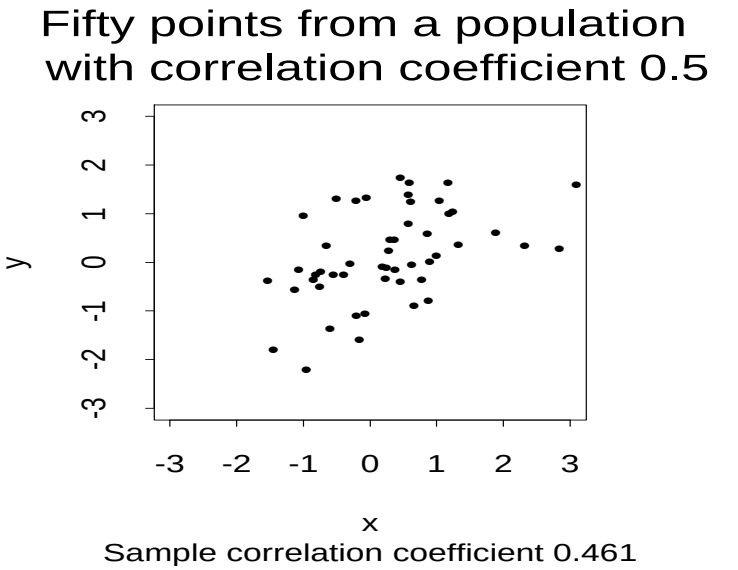
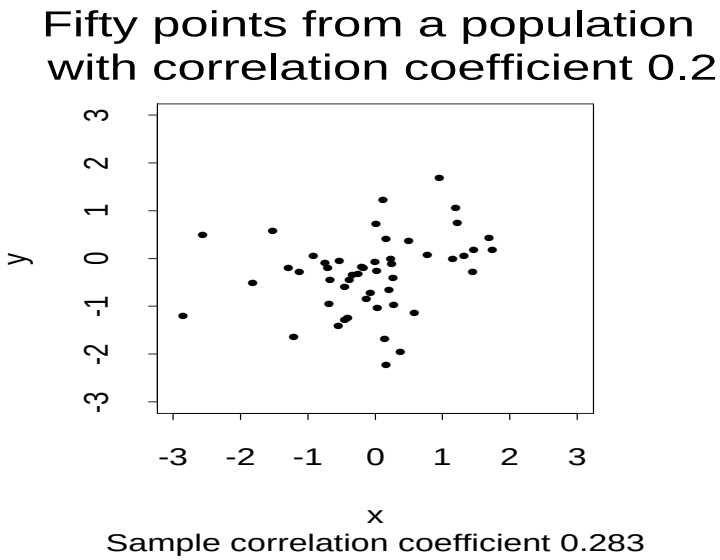


Figure 12.3: Scatter plot to show correlations



It will be seen from these diagrams that there can be quite large differences between a population correlation coefficient and the sample correlation coefficient of a fairly large random sample from that population. It is also clear that there may not be an enormous difference between the samples from a population with correlation coefficient 0.5 and one with correlation coefficient 0.2. You will find a little about formal tests of $\rho = 0$ towards the end of this chapter. Caution is needed when looking for association because of sample variation.[2] The bivariate normal distribution is a population that is often used when the distributions of both X and Y are normal distributions, and they are correlated.

[2] *One must also beware of the effect of the scales of the axes. A correlation coefficient is scale-free, but it does not always seem so to the eye.*

Regression and the coefficient of determination R^2

It is fairly common for the results of a regression fit to be summarised in an analysis of variance table.[3] The form of the table is

[3] *Don't be confused by ANOVA reappearing in regression.*

Source of variation	Degrees of freedom	Sums of Squares	Mean Sums of Squares	F-ratio
Explained by slope	1	S_β^2	S_β^2	S_β^2/S^2
Residual	$n-2$	$(n-2)S^2$	S^2	
Total	$n-1$	$(n-1)S_y^2$		

The ratio $R^2 = S_\beta^2/[(n-1)S_y^2]$ is called the *coefficient of determination*. It gives an intuitive guide to the proportion of variation of Y around its mean that is explained by variations in X. It is also the squared sample correlation coefficient between the y_is and x_is, which is defined as r_{xy}^2 in (10.13). It is also the squared sample correlation coefficient between the y_is and $\hat{y}_i$s.

Activity 12.2. Why is the squared sample correlation coefficient between the y_is and x_is the same as the squared sample correlation coefficient between the y_is and $\hat{y}_i$s? No algebra is needed for this.
■

One can test the Null Hypothesis H_0: $\beta = 0$ against the alternative H_1: $\beta \neq 0$ at the $100\gamma\%$ level by rejecting the null hypothesis if the F-ratio

$$S_\beta^2/S^2 > F_{\gamma,1,n-2}. \tag{12.2}$$

This test is equivalent to a Student's t test of H_0:$\beta = 0$ against the alternative H_1: $\beta \neq 0$ at the $100\gamma\%$ level because the distribution of the square of a random variable with a Student's t distribution with ν degrees of freedom has an F distribution with 1 degree of freedom for the numerator and ν for the denominator. Notice that this is a one-tailed test for a two-sided Alternative Hypothesis.

Example 12.1. Continuing Example 11.4 on page 133, the analysis of variance table for the regression of price of books on pages is:

Source of Variation	Degrees of freedom	Sums of Squares	Mean Sums of Squares	F-ratio
Explained by slope	1	467.2	467.2	2.49
Residual	13	2437.7	187.5	
Total	14	2904.9		

Notice that the degrees of freedom add to the total, as do the sums of squares. The upper 5% point of the F distribution with 1 degree of freedom for the numerator and 13 for the denominator is from Table 12(b) of the *New Cambridge Statistical Tables* with

$\nu_1 = 1$ and $\nu_2 = 13$, $F_{.05,1,13} = 4.667$. The observed ratio is 2.49, which is not greater than 4.667. So we do not reject the hypothesis that $\beta = 0$ at the 5% level of significance. Notice that the tabular value 4.667 is the square of the upper 2.5% point of the Student's t distribution, which from Table 10 of the *New Cambridge Statistical Tables* with $P = 2.5$ and $\nu = 13$ is 2.160.

■

Testing $\rho = 0$ for a bivariate normal distribution

If you are given a random sample of n pairs of observations (X_i, Y_i) from a bivariate normal distribution with correlation ρ, then you can, by a happy coincidence, use the test suggested in (12.2) to test the Null Hypothesis H_0: $\rho = 0$ against the Alternative Hypothesis H_1: $\rho \neq 0$. In a similar way one can think of the Student's t test of H_0:$\beta = 0$ against the alternative H_1: $\beta > 0$ as a test of H_0:$\rho = 0$ against the alternative H_1: $\rho > 0$. An analogous comment may be made about the test against the alternative H_1: $\rho < 0$.

Learning outcomes

After working through this chapter you should be able to:

1. define the sample correlation coefficient and link it to the appearance of scatter diagrams
2. construct and use an analysis of variance table for a regression, including the F-test for $\beta = 0$
3. show that a correlation coefficient is between -1 and +1.

Sample examination question

1. Suppose that we have 10 random observations (x_i, y_i), $i = 1, \ldots, 10$ of random variables (X, Y) that have a bivariate normal distribution.

x_i	0.33	1.01	-0.41	2.10	-0.29	-1.27	-0.48	0.31	-1.56	-0.81
y_i	1.60	1.47	-1.82	2.50	0.34	-1.37	-0.09	0.20	-1.16	-2.16

(a) Find the sample correlation coefficient between X and Y.

(b) Test the null hypothesis that the population correlation coefficient is equal to zero against the alternative that it is greater than zero.

Regression with more than one explanatory variable.

Chapter 13

Multiple Regression

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 13.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], 14.6.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods.* (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], chapter 12.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], chapter 15.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], chapter 11.
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], chapter 13.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 13.

Introduction

In practical application of regression models, there is almost always more than one explanatory variable. In this course we can only give a glimpse of the several new important matters that arise with such models, which are the most used in the whole of statistics.

The model for linear regression

As in Chapter 11 we have n observations, $Y_1, Y_2, \ldots, Y_n$, on a response variable y. Each observation Y_i is associated with the fixed values $x_{i1}, \ldots, x_{ip}$ of p explanatory variables $x_1, \ldots, x_p$. We assume that the observations Y_i are independent, and have

normal distributions with means

$$\alpha + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{13.1}$$

and in each case a variance σ^2.

We usually make sure that p is substantially less than n. There are now p slope parameters β_j and an intercept α to estimate from the n observations, so we must certainly have $n \geq p+1$. The interpretation of the slope parameter β_j is that it shows the change in Y for a unit change in x_j given that all other explanatory variables $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$ are held constant.

Least squares fitting

The usual method of estimation is by least squares. The least squares estimators $(A, B_1, \ldots, B_p)$ satisfy

$$\sum_{i=1}^{n} (Y_i - A - \sum_{j=1}^{p} B_j x_{ij})^2 \leq \sum_{i=1}^{n} (Y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

for all possible choices of $(\alpha, \beta_1, \ldots, \beta_p)$. The fitted values are

$$\hat{Y}_i = A + \sum_{j=1}^{p} B_j x_{ij}$$

and the residuals are

$$\hat{\epsilon}_i = Y_i - A - \sum_{j=1}^{p} B_j x_{ij}.$$

The estimators are found by solving the $p+1$ *least squares equations*

$$\begin{aligned} \sum_{i=1}^{n} \hat{\epsilon}_i &= 0, \\ \sum_{i=1}^{n} x_{i1} \hat{\epsilon}_i &= 0, \\ &\cdots \\ \sum_{i=1}^{n} x_{ip} \hat{\epsilon}_i &= 0. \end{aligned} \tag{13.2}$$

Notice that the first of these equations implies that $(\bar{Y}, \bar{x}_1, \ldots, \bar{x}_p)$ lies on the fitted regression

$$Y = A + \sum_{j=1}^{p} B_j x_j.$$

Activity 13.1. Justify the last statement.
■

Sum of squares identity

In just the same way as we obtain the sum of squares identity (10.5) on page 119, we can obtain, using (13.2), the more general identity, true for **any** choice of $\alpha, \beta_1, \ldots, \beta_p$

$$\sum_{i=1}^{n}(Y_i - \alpha - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2 + \sum_{i=1}^{n}[\hat{Y}_i - \alpha - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}]^2 \tag{13.3}$$

Equation (13.3) shows that solving the equations (13.2) is sure to give the least squares estimators.

Activity 13.2. Prove (13.3.)
■

By choosing in (13.3) to take $\alpha = \bar{Y}$ and $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, we obtain the simpler sum of squares identity

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}\hat{\epsilon}_i^2 + \sum_{i=1}^{n}[\hat{Y}_i - \bar{Y}]^2 \\ &= \sum_{i=1}^{n}\hat{\epsilon}_i^2 + \sum_{i=1}^{n}[B_1(x_{i1} - \bar{x}_1) + B_2(x_{i2} - \bar{x}_2) + \cdots + B_p(x_{ip} - \bar{x}_p)]^2.\end{aligned}$$

This shows a Total Sum of Squares $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = (n-1)S_y^2$ (where S_y^2 is the sample variance of all the Y_is) broken into a Sum of Squares for the Slopes

$$pS_\beta^2 = \sum_{i=1}^{n}[B_1(x_{i1} - \bar{x}_1) + B_2(x_{i2} - \bar{x}_2) + \cdots + B_p(x_{ip} - \bar{x}_p)]^2 \tag{13.4}$$

and a Residual Sum of Squares $(n-p-1)S^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2$. The Mean Residual Sum of Squares is S^2, and the Mean Sum of Squares for the slope is S_β^2.

Coefficient of Determination

One can measure how well the model succeeds in explaining the variation in the response by the *Coefficient of Determination* R^2, which is defined by the ratio of the Sum of squares for the slopes to the Total sum of squares.

$$R^2 = \frac{pS_\beta^2}{(n-1)S_y^2}.$$

R^2 is usually thought of as the proportion of the variation in the response variable explained by the regression. Often, one would look for R^2 over 60% before thinking that a model was useful, but in social science applications one is often content with much less. R^2 is the square of the correlation coefficient between the values of the response variable and the fitted values from the model.

[1] *In peculiar cases it might stay the same*

If the number of explanatory variables is increased, then R^2 always decreases[1]. So if one wants to choose how many explanatory variables to include in the model, one can't depend entirely on R^2, because that would always suggest putting in every explanatory variable available. One should look also at Student's t values for the estimated slopes, or perhaps at the *Adjusted Coefficient of Determination*, which makes an attempt to correct for the number of explanatory variables used. Most packages give an F-test for the Null Hypothesis that the population analogue of R^2 is 0 against the alternative that it is greater than 0. This is a test of whether there is any point in fitting the regression at all.

Computation

There is little point in trying to fit a regression with more than one explanatory variable using a hand calculator, though the more expensive ones can do it; a software package on a computer is better for this purpose. We shall therefore concentrate here on the interpretation of regression models, and on problems that arise from their use.

Example 13.1. We shall use data on the taste of cheese, suggested in *Introduction to the Practice of Statistics* by D.S. Moore and G.P. McCabe, published in 1998 by Freeman, another good textbook for some parts of this subject. The data give scores for the taste of a cheese (Taste) from 30 different formulations which caused variation in the concentration in the cheese of acetic acid (Acetic), hydrogen sulphide (H_2S) and lactic acid (Lactic). One would wish to model the dependence of the taste score on the concentrations of those three constituents, using the $n = 30$ observations that are given in Table 13.1 on page 149. A first check is to look at the scatter plots of the response variable 'Taste' against each of the explanatory variables. The plots are shown in Figure 13.1 on page 150.

The plots show that the taste score tends to increase with an increase in acetic acid, hydrogen sulphide or lactic acid. A more precise interpretation (of the first plot, for instance) is that the taste score tends to increase if acetic acid concentration increases, **if** one **ignores** all other sources of variability including that in hydrogen sulphide and lactic acid.

If the model

$$\text{Taste} = \alpha + \beta_1 \text{Acetic} + \beta_2 \text{H}_2\text{S} + \beta_3 \text{Lactic}$$

is fitted by least squares, the fitted model is:

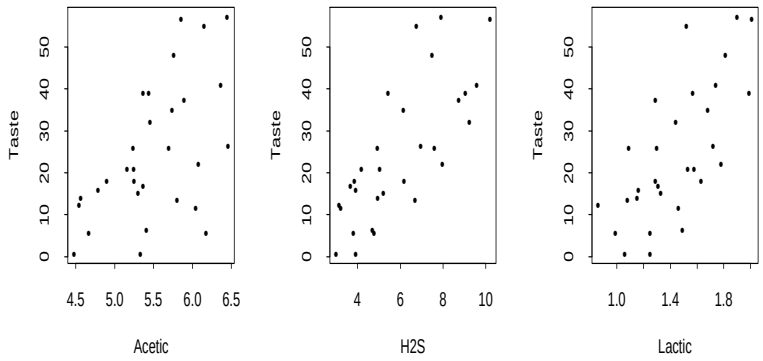
$$\text{Taste} = -28.9 + 0.33 \text{Acetic} + 3.91 \text{H}_2\text{S} + 19.67 \text{Lactic}$$

Table 13.1: Taste of cheese Data

Case	Taste	Acetic	H_2S	Lactic
01	12.3	4.543	3.135	0.86
02	20.9	5.159	5.043	1.53
03	39.0	5.366	5.438	1.57
04	47.9	5.759	7.496	1.81
05	5.6	4.663	3.807	0.99
06	25.9	5.697	7.601	1.09
07	37.3	5.892	8.726	1.29
08	21.9	6.078	7.966	1.78
09	18.1	4.898	3.850	1.29
10	21.0	5.242	4.174	1.58
11	34.9	5.740	6.142	1.68
12	57.2	6.446	7.908	1.90
13	0.7	4.477	2.996	1.06
14	25.9	5.236	4.942	1.30
15	54.9	6.151	6.752	1.52
16	40.9	6.365	9.588	1.74
17	15.9	4.787	3.912	1.16
18	6.4	5.412	4.700	1.49
19	18.0	5.247	6.174	1.63
20	38.9	5.438	9.064	1.99
21	14.0	4.564	4.949	1.15
22	15.2	5.298	5.220	1.33
23	32.0	5.455	9.242	1.44
24	56.7	5.855	10.199	2.01
25	16.8	5.366	3.664	1.31
26	11.6	6.043	3.219	1.46
27	26.5	6.458	6.962	1.72
28	0.7	5.328	3.912	1.25
29	13.4	5.802	6.685	1.08
30	5.5	6.176	4.787	1.25

This shows that we estimate that the Taste score increases with an increase in each of the explanatory variables **holding the other two explanatory variables fixed.** For instance, each increase of one unit in the concentration of lactic acid is estimated to lead to an increase of 19.67 in the score for taste, if we hold the concentration of acetic acid and hydrogen sulphide fixed. It may, of course, be very difficult when cheese making to increase Acetic acid concentration while holding Hydrogen sulphide and Lactic acid at fixed levels of concentration, so one must be very careful when

Figure 13.1: Scatter plots for Taste against explanatory variables



implementing policy changes using the fitted model.

We can check if all three explanatory variables are important in the fitted model by looking at the Student's t values for each of the slope coefficients. We have, from the software package,

```
              Value Std. Error  t value   p-value
(Intercept) -28.8768  19.7354    -1.4632   0.1554
     Acetic   0.3277   4.4598     0.0735   0.9420
        H2S   3.9118   1.2484     3.1334   0.0042
    Lactic   19.6705   8.6291     2.2796   0.0311
```

The small t-value, corresponding to a large p-value, for acetic acid suggests that this explanatory variable is of no importance when the model includes hydrogen sulphide and lactic acid. The package also gives $R^2 = 0.6518$, so that the model might be reasonably useful. The F-test for the population value of R^2 is given by the package as

```
F-statistic: 16.22 on 3 and 26 degrees of freedom, the p-value is
3.81e-006
```

The large F-statistic, and the small p-value show that it is worth fitting the model.

If we drop acetic acid from the model, and refit, the fitted model becomes:

$$\text{Taste} = -27.6 + 3.95\text{H}_2\text{S} + 19.89\text{Lactic}$$

which gives slopes for H_2S and Lactic only slightly different from before. In this model, we are not attempting to correct for acetic acid. For instance, the effect of an increase of one unit in lactic acid is estimated to give rise to a change of 19.89 in the taste score, when hydrogen sulphide is held constant and acetic acid is ignored altogether. So, the interpretation of the 19.89 in this model is different from that of the 19.61 in the previous model. Now, the Student's t values are

```
              Value Std. Error  t value   p-value
(Intercept) -27.5918   8.9818    -3.0720   0.0048
        H2S   3.9463   1.1357     3.4748   0.0017
     Lactic  19.8872   7.9590     2.4987   0.0188
```

All the explanatory variables now have small p-values, so we are not tempted to drop any of them. Notice that the intercept now also has a small p-value. It had a large value before because Acetic acid acted very much like an intercept term as well. R^2 is 0.6517 for this model, so it provides an explanation almost exactly as good as the one with three explanatory variables. The F-test for the population value of R^2 is given by the package as

```
F-statistic: 25.26 on 2 and 27 degrees of freedom, the p-value is
6.551e -007
```

The large F-statistic, and the small p-value show that it is worth fitting the model. As an alternative to the Student's t values, one can cast the results into an analysis of variance table.

```
Response: Taste

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
      H2S  1  4376.746 4376.746 44.27639 0.00000039
   Lactic  1   617.175  617.175  6.24352 0.01884987
Residuals 27  2668.965   98.851
```

This table shows $S^2 = 98.81$ and $S^2_\beta = \frac{4376.746+617.175}{1+1} = 2496.96$. Notice that the F-statistic given above is $2496.96/98.81 = 25.26$ (apart from rounding errors). The degrees-of-freedom for the residual sum of squares are $n-p-1 = 30-2-1 = 27$. The terms are added sequentially, so 4376.746 is the variation in the taste scores explained by hydrogen sulphide (ignoring lactic acid and acetic acid). Then 617.175 is the variation in the taste scores left after fitting hydrogen sulphide that is removed by fitting lactic acid (still ignoring acetic acid). The 617.175 is sometimes called the sum of squares for lactic acid **adjusted** for hydrogen sulphide. If we fit the explanatory variables in the other order then the analysis of variance becomes

```
Response: Taste

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
   Lactic  1  3800.398 3800.398 38.44589 0.000001250
      H2S  1  1193.523 1193.523 12.07402 0.001742869
Residuals 27  2668.965   98.851
```

This shows that lactic acid on its own does not explain as much as hydrogen sulphide on its own, because 3800.398 is smaller than 4376.746.
■

Extrapolation

It is generally unwise to use a regression model to predict values of the response for explanatory variable values dissimilar from those used to fit the model. There is no guarantee that the model will fit outside the ranges of values for the explanatory variables that have been used to construct it (see Example 11.4, page 136). When there are several explanatory variables it becomes much more difficult to see which values for the explanatory variables are close to those used to fit the model, and so one should pay careful attention to the length of the confidence intervals for the fitted values.

Example 13.2. Continuing Example 13.1, we can ask for the fitted model

$$\text{Taste } = -27.6 + 3.95\text{H}_2\text{S} + 19.89\text{Lactic}$$

what would be the fitted value for hydrogen sulphide at 20.0 and lactic acid at 6.0. A quick glance at the data will show that these values are extreme. The software package will give the fitted value 170.7 and the 95% confidence interval for the expected value of the response as $(121.7, 219.7)$. The very long interval suggests that extrapolation is taking place.
■

Collinearity

What happens if we make the values of two of the explanatory variables exactly the same (though they have different names)? Then some of the software packages break down, while some will produce rather different estimates from each other. From an intuitive point of view, the problem is that if there is no difference between two explanatory variables then it's hard to say how one of them affects the response when the other is held fixed, that being what a fitted model produces. So the results become fairly arbitrary.

Even worse, suppose that two variables, while not taking exactly the same values, have very similar values. Then one may well find some of the same problems. Whenever any one of the explanatory variables has values that can be predicted well from

those of the others, the slope coefficients in the model are likely to be ill-defined and hard to interpret. This is called the problem of *collinearity*. It arises because there are many different estimators of the parameters that give rise to the same minimum sum of squared residuals.

Activity 13.3. Suppose that $n = 4$ the values for $p = 3$ explanatory variables are in the columns

Variables		
V_1	V_2	V_3
1	2	3
2	3	5
3	4	7
4	5	9

Show that fitted values $V_1+V_2+V_3$ are the same as fitted values $0.5V_1+0.5V_2+1.5V_3$.
■

Diagnostic Plots

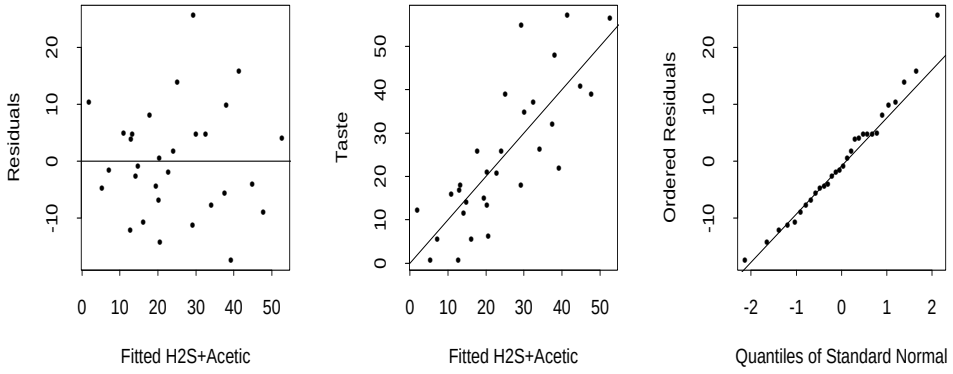
Most statistics packages provide several plots designed to allow identification of problems in the model. It is not so easy to use the simple plots on page 132 when there are several explanatory variables, but there are others that can be tried. One can plot the fitted values against the responses, or the residuals against the fitted values. Both these plots might show up outliers or the need to change to a non-linear model. One can also plot the residuals against each of the explanatory variables to check whether the error variance changes with the values of the explanatory variables.

Activity 13.4. If the model fits, then the fitted values and the residuals from the model are independent of each other. What do you expect to see if the model fits when you plot residuals against fitted values?
■

Example 13.3. Continuing Example 13.1, we look at some diagnostic plots. In Figure 13.2, the first diagram shows a plot of the residuals against the fitted values. No unusual structure is present. The points are scattered around evenly around the line for zero residuals. The second diagram shows observed values of taste against the fitted values. The diagram shows a points scattered around a straight line through the points with exactly equal values for taste and the fitted values for taste. The third diagram shows a *Normal Plot* of ordered values of the residuals against standard normal quantiles. The roughly straight line plot suggests that there are no obvious outliers, and that the measurement errors have approximately a normal distribution.

To see what happens if there is a very wild outlier, let us change the hydrogen sulphide reading on case 12 from 7.91 to 79.1, as might easily happen by error. This changes the fitted model very greatly:

Figure 13.2: Diagnostic Plots




```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -25.0166  10.4983    -2.3829   0.0245
        H2S   0.3064   0.1670     1.8352   0.0775
     Lactic  32.5952   7.4441     4.3787   0.0002

Residual standard error: 11.28 on 27 degrees of freedom Multiple
R-Squared: 0.5519 F-statistic: 16.62 on 2 and 27 degrees of
freedom, the p-value is 0.00001969
```

The R^2 value of 55% is not quite as good as before, but one might well believe from this output alone that one had a good model, and never notice the wild observation that has distorted it. None of the diagnostic plots we used above show much wrong either. This wild observation has pulled the whole model out towards it, and more sophisticated diagnostic plots (or perhaps the simple checks like those in Figure 13.1) are needed to detect the problem. ■

Learning outcomes

After working through this chapter you should be able to:

1. write down the model for linear regression with several explanatory variables, and explain its interpretation
2. give the assumptions on which the model is based
3. write down the least squares equations and obtain from them the sum of squares identity; use the sum of squares identity to establish that the solution of the least

squares equations is indeed a least squares estimator

4. interpret typical output from a computer package fitting of a regression model
5. understand the use of the coefficient of determination
6. describe the nature of problems of collinearity and outliers
7. interpret simple diagnostic plots.

Sample examination question

1. (a) Write down a sum of squares identity for a multiple regression model, and show how it implies that the solution of the least squares equations is a least squares estimator.

   (b) The following output is from a regression of record times in hours for Scottish Hill Races on the explanatory variables of distance run in miles and height climbed in feet. These were discussed by A.C. Atkinson in a paper in *Statistical Science* in 1986.

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -8.9920   4.3027    -2.0898   0.0447
       dist   6.2180   0.6011    10.3435   0.0000
      climb   0.0110   0.0021     5.3869   0.0000

Residual standard error: 14.68 on 32 degrees of freedom Multiple
R-Squared: 0.9191 F-statistic: 181.7 on 2 and 32 degrees of
freedom, the p-value is 0


Analysis of Variance Table

Response: time

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value         Pr(F)
     dist  1  71996.89 71996.89 334.2926 0.000000e+000
    climb  1   6249.74  6249.74  29.0185 6.445183e-006
Residuals 32   6891.87   215.37
```

   i. What is the fitted model? Interpret the model.

   ii. What is the estimated value of the record time in hours for The Goatfell Hill Race which has distance 8.0 miles and height climbed 2866 feet?

iii. How would you interpret the value of R^2?

iv. What diagnostic plots would you suggest for these data?

Seeing if counted data are consistent with a distribution.

Chapter 14

Tests for goodness-of-fit

Essential reading

Newbold, P. *Statistics for Business and Economics.* (London: Prentice-Hall, 1995) fourth edition [ISBN 0-13-855549-0], chapter 11.

Further reading

Daly, F., D.J. Hand, M.C. Jones, A.D. Lunn and K.J. McConway *Elements of Statistics*. (London: Open University and Addison-Wesley, 1995) [ISBN 0-201-42278-6], a small amount in 11.4.
Johnson, R.A. and G.K. Bhattacharyya *Statistics: Principles and Methods*. (New York: John Wiley and Sons, 2000) fourth edition [ISBN 0-471-388971], chapter 13.
Mason, R.D., D.A. Lind and W.A. Marchal *Statistics; an Introduction*. (New York: Duxberry Press, 1998) fifth edition [ISBN 0-534-35379-7], chapter 16.
Moore, D.S. and G.P. McCabe *Introduction to the Practice of Statistics*. (New York: W.H. Freeman and Company, 1998) third edition [ISBN 0-7167-3502-4], chapter 9.
Triola, M.F. and L.A. Franklin *Business Statistics: understanding populations and processes*. (New York: Addison-Wesley, 1994) [ISBN 0-201-58990-7], chapter 10.
Wonnacott, T.H. and R.J. Wonnacott *Introductory Statistics* (New York: John Wiley and Sons, 1990) fifth edition [ISBN 0-471-615188], chapter 17.

Introduction

Tests of goodness-of-fit are very often used in applications, and all statisticians must know when they can be applied, and what their limitations are. These tests are not concerned with the fit of a straight line to data, but with counted data. They are used when one wants to check whether a sample comes from some type of population, or when one wants to check that two samples are from the same population.

Basic counting model

The model underlying all the applications in this chapter is that of a *multinomial* distribution. There are k different classes indexed by $i = 1, \ldots, k$, with associated

probabilities π_i. There are n independent trials X_s, for $s = 1, 2, \ldots, n$. Each trial results in a choice of one of the classes, where $P[X_s = i] = \pi_i$ is the probability that class i is chosen. This is a generalisation of the binomial distribution on page 25.[1]

[1] *There is no overlap of classes. Each trial results in allocation to one class only.*

For instance, suppose that we count the number of people that were born on each day of the week in a random sample of size n from a large human population. Here there would be $k = 7$ distinct classes, Sunday to Saturday, and each person in the sample would have been born on one of those days. After classifying the sample, we will know the number of people in the sample born on Sunday, the number born on Monday, and so on. One might think that the probabilities π_i would in this case be equal, but that is not necessarily true. Even if there were equal proportions in the population born on each day of the week, so that all π_i are equal to $1/7$, the sample would not usually show equal frequencies because of sampling variation.

We will use the notation O_i for the number of the trials X_s, $s = 1, \ldots, n$ for which $X_s = i$. From the sample we will have the set of frequencies

$$O_1, O_2, \ldots, O_k,$$

where $\sum_i O_i = n$. The joint distribution of $O_1, O_2, \ldots, O_k$ is the multinomial distribution.

Since for fixed choice of i, O_i has a binomial distribution Binomial(n, π_i), we can see from page 57 that $E[O_i] = n\pi_i$ and $\mathrm{var}[O_i] = n\pi_i(1 - \pi_i)$. As a rough approximation, particularly if π_i is small, one could take

$$\frac{O_i - n\pi_i}{\sqrt{n\pi_i}}$$

as having mean zero and variance something like 1.

A goodness-of-fit statistic

One measure of the difference between the observed frequencies O_i and their means $n\pi_i$ is the X^2 statistic

$$X^2 = \sum \frac{(O_i - n\pi_i)^2}{n\pi_i}.$$

Writing $E_i = n\pi_i$, the usual way for the X^2 statistic to be written in the textbooks is

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}.$$

Another form offering fewer insights into the reasons for calculating it is

$$X^2 = \sum \frac{O_i^2}{E_i} - n.$$

Activity 14.1. Show that the two forms for X^2 are equivalent.
■

The X^2 statistic gives an overall measure of how close the observed frequencies are to their expected values. If the probabilities for the classes were actually different from the π_is, one would expect that X^2 would be found to take a larger value than when the probabilities for the classes were the π_is. For this reason X^2 is called a *goodness-of-fit statistic*. It has expected value $(k-1)$ when the Null Hypothesis is true.

Activity 14.2. Show from results for the binomial distribution that $EX^2 = k-1$. This is exactly the mean of a $\chi^2_{(k-1)}$ distribution.
■

For large n, the distribution[2] of X^2 is approximately $\chi^2_{(k-1)}$. This is used as the basis of a test of goodness-of-fit. To test

[2] *The approximation is poor if any E_i are too small.*

$$H_0\text{: The class probabilities are } \pi_1, \pi_2, \ldots, \pi_k$$

against the alternative

$$H_1\text{: the class probabilities are not } \pi_1, \pi_2, \ldots, \pi_k$$

with a test at the $100\alpha\%$ level, reject the Null Hypothesis if

$$X^2 > \chi^2_{\alpha,(k-1)}.$$

This is a one-tailed test for a two-sided Alternative Hypothesis.

Example 14.1. Twenty-five machine operators produce items that are sometimes defective. Random samples of the production of each operator give the number of defective items each produces out of 10 items sampled. These numbers are as below:

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

Do these results give enough evidence to show that the operators do not all have a 10% rate of producing defective items?

Answer

We suppose that each operator is independent of the others, and that items are produced independently of each other, and that the rate of production of defective items is constant for each operator. The Null Hypothesis is that the rate of production of defective items is 10% for each operator. Then each of the numbers in the table

above is a value for a binomial random variable with number of trials $n = 10$ and probability of 'success' $\pi = 0.1$.

We can classify each of the 25 operators by the number of defective items produced. This gives a table of frequencies:

Class number i	1	2	3	4	5
Number of defective items	0	1	2	3	≥ 4
Frequency O_i	12	8	3	2	0

We need to use a goodness-of-fit test to check if this table of frequencies is consistent with the Null Hypothesis. On the Null Hypothesis the class probabilities are those for a binomial distribution with $n = 10$ and probability of 'success' $\pi = 0.1$, which can be found from Table 1 of the *New Cambridge Statistical Tables* on page 9, with $n = 10$ and $p = 0.10$. The numbers given in that table are left-tail probabilities, so some differencing is necessary. The probabilities are in the table below, with the expected frequencies E_i obtained by multiplying the probabilities by $n = 25$.

Number of defective items	0	1	≥ 2
Probability π_i	0.3487	0.3874	0.2639
Expected frequency $E_i = n\pi_i$	8.71696	9.68551	6.59775

Notice that the last class of ≥ 2 defective items has been put in so that the class probabilities sum to 1, and to make sure all the E_i are greater than 3. Otherwise the X^2 statistics would not have a distribution approximated by the χ^2 distribution.

We have now constructed all the quantities for calculating X^2. There are $k = 3$ classes, and the O_i and E_i are as below:

i	1	2	3
O_i	12	8	5
E_i	8.71696	9.68551	6.59775

Notice that the sum of the E_i is the same as the sum of the O_i and both are equal to 25, the number n of operators classified for this example by the number of defective items they produced.

We have

$$\begin{aligned} x^2 &= \frac{(12 - 8.7196)^2}{8.7196} + \frac{(8 - 9.68551)^2}{9.68551} + \frac{(5 - 6.59775)^2}{6.59775} \\ &= 1.2341 + 0.2933 + 0.3869 \\ &= 1.9144. \end{aligned}$$

For a test at the 5% level of significance we use the upper 5% point of the χ^2 distribution with $k - 1 = 3 - 1 = 2$ degrees-of-freedom. From Table 8 of the *New Cambridge Statistical Tables* with $P = 5$ and $\nu = 2$ we get $\chi^2_{0.05,(2)} = 5.991$. Since

our observed value of X^2 is $x^2 = 1.9144$, which is lower than 5.991, we do not at the 5% level of significance reject the hypothesis that the rate of production of defective items is equal to 10% for all the operators.
■

Testing when there are unknown parameters

Sometimes the Null Hypothesis for a goodness-of-fit test does not completely specify the probabilities π_i for the classes, because some parameters are not known, but must be estimated from the observed frequencies O_i. If m distinct parameters are estimated from the O_is in an efficient way, then the test using the X^2 statistic stays the same, but with degrees-of-freedom $k - m - 1$ instead of $k - 1$.

Example 14.2. Let us think of a set-up as for Example 14.1, but with 100 machine operators. The frequencies are those in the table below

Class number i	1	2	3	4	5	6	7
Number of defective items	0	1	2	3	4	5	6
Frequency O_i	14	21	36	20	5	2	2

Let us now suppose that the null hypothesis is that the rate of production of defective items is the same for all the operators but is **unknown**. If that null hypothesis is true, the usual efficient estimator of that rate is the total number of failures, 195, divided by the total number of items tested, which is 1000. The estimated failure rate is therefore 0.195. We get the estimated class probabilities $\hat{\pi}_i$ from the binomial distribution with 10 trials and probability of a defective item 0.195. The statistical tables are not much use, so from a hand calculator we get

Number of defective items	0	1	2	3	≥ 4
Probability $\hat{\pi}_i$	0.1143	0.2768	0.3018	0.1949	0.1122
Expected freqy. E_i	11.43	27.68	30.18	19.49	11.22

The table below shows the observed and expected frequencies. There are $k = 5$ classes with one parameter (the rate of defective items) estimated from the data.

i	1	2	3	4	5
O_i	14	21	36	20	9
E_i	11.4277	27.6820	30.1751	19.4920	11.2230

The goodness-of-fit statistic takes the value

$$\begin{aligned} x^2 &= \frac{(14-11.4277)^2}{11.4277} + \frac{(21-27.6820)^2}{27.6820} + \frac{(36-30.1751)^2}{30.1751} + \frac{(20-19.4920)^2}{19.4920} \\ &\quad + \frac{(9-11.2230)^2}{11.2230} \\ &= 0.5790 + 1.4481 + 1.1244 + .01324 + 0.4403 \\ &= 3.6051. \end{aligned}$$

To test the Null Hypothesis at the 10% level, we check to see if x^2 is larger than $\chi^2_{.1,(3)}$, which from Table 8 of the *New Cambridge Statistical Tables* with $P = 10$ and $\nu = 3$ is 6.251. There are 3 degrees-of-freedom because $k = 5$ and $m = 1$ parameter is estimated giving $5 - 1 - 1 = 3$ degrees-of-freedom. Since the observed x^2 is not greater than 6.251, we do not reject the hypothesis that all machine operators have the same failure rate.

The test helps to put in perspective the wide variation from 0 to 6 in the number of defective items produced by the 100 operators. There is little evidence here that this variation is not simply sampling variation from equal population rates of defective items. The p-value is

$$P[\chi^2_{(3)} \geq 3.6051] = 0.3074.$$

which can be approximately found from Table 7 of the *New Cambridge Statistical Tables* with $\nu = 3$, $x = 3.6$ giving the approximate complementary probability 0.6920. This is a fairly large p-value. It suggests that X^2 would be at least as large as the value 3.6051 in over 30% of repeated observations on the 100 operators, even if they all have the same rate of producing defective items.

■

Testing for association in two-way tables

One specialised use for the χ^2 goodness-of-fit test is to check whether there is evidence of association between the row and column classifications of a two-way table of counts (often called a *contingency table*).

In this application we have a rectangular table of counts with r rows and c columns. In order to see association in a two-way table it is however best to present the table as a collection of row profiles or a collection of column profiles, rather than as a collection of counts. Each of the n trials now specifies a cell in a particular row and column. The Null Hypothesis is that information about the row classification of a trial gives no information about its column classification. That is, the Null Hypothesis says that the probability π_{ij} of being classified into cell (i, j) in row i and column j is the product of $\pi_{\text{row } i}$ and $\pi_{\text{col } j}$, which are the probabilities of being in classified in row i and column j respectively. The observed data are the counts O_{ij} in cells (i, j).

This Null Hypothesis does not specify $\pi_{\text{row } i}$ or $\pi_{\text{col } j}$, so they must be estimated from the data, and the degrees-of-freedom for the test reduced from one less than the total number of classes $(rc-1)$ by the number of distinct parameters estimated. Since both the $\pi_{\text{row } i}$ and the $\pi_{\text{col } j}$ add to 1, we have $(r-1)+(c-1)$ distinct parameters to estimate, so the degrees-of-freedom are $[rc-1-(r-1)-(c-1)] = (r-1)(c-1)$.

A suitable estimator of $\pi_{\text{row } i}$ is $\hat{\pi}_{ri} = c\bar{O}_{i.}/n$, which is the sample proportion of trials classified into the ith row. (We use the same notation here as in analysis of variance, so $\bar{O}_{i.}$ is the average frequency of row i, and $c\bar{O}_{i.}$ is the total frequency in row i.) Similarly, a suitable estimator of $\pi_{\text{col } j}$ is $\hat{\pi}_{cj} = r\bar{O}_{.j}/n$, which is the sample

proportion of trials classified into the jth column. These lead to the estimate of p_{ij} when the Null Hypothesis is true of

$$\hat{p}_{ij} = \frac{c\bar{O}_{i.}}{n}\frac{r\bar{O}_{.j}}{n},$$

giving expected frequencies

$$\begin{aligned} E_{ij} &= n\frac{r\bar{O}_{i.}}{n}\frac{c\bar{O}_{.j}}{n} \\ &= nrc\bar{O}_{i.}\bar{O}_{.j}/n^2 \\ &= rc\bar{O}_{i.}\bar{O}_{.j}/n. \end{aligned}$$

The test is carried out just as before.

Example 14.3. In a survey the samples from five towns were examined for the number of partly skilled or unskilled workers. The data are in Table 14.1: Does the

Table 14.1: Numbers of workers

Town	Partly skilled and unskilled workers	Other workers
A	80	184
B	58	147
C	114	276
D	55	196
E	83	229

population proportion of partly skilled and other workers vary with the area?

Answer

In order to see any association in the table, one can look at the column percentages (i.e. the column profiles). These are in Table 14.2.

The association that **might** be there is seen as a difference in the percentages as one moves from the column for *Partly skilled and unskilled workers* to the column for *Other workers*. If there were no association, the percentages would be the same in each column. The null hypothesis of no association says that the percentages in each column in the population are the same.

The χ^2 test will help to decide if the differences in the percentages we see in Table 14.2 are just due to random variation or are showing real differences in the population between the two profiles.

In Table 14.3 expected counts are shown in parenthesis below observed counts.

The expected values E_{ij} are easily found from the row and column totals. For instance, for the bottom right-hand corner of the table, the expected value is

$$312 \times 1032/1422 = 226.4.$$

Table 14.2: Column Profiles

Town	Partly skilled and unskilled workers	Other workers
A	20.5%	17.8%
B	14.9%	14.2%
C	29.2%	26.7%
D	14.1%	19.0%
E	21.3%	22.2%

Table 14.3: Expected and observed counts

Town	Partly skilled and unskilled workers	Other workers	Row total
A	80.0	184.0	264
	(72.4)	(191.6)	
B	58.0	147.0	205
	(56.2)	(148.8)	
C	114.0	276.0	390
	(107.0)	(283.0)	
D	55.0	196.0	251
	(68.8)	()182.2)	
E	83.0	229.0	312
	(85.6)	(226.4)	
Column total	390	1032	1422

The value of the X^2 statistic is

$$x^2 = (80 - 72.4)^2/72.4 + (184 - 191.6)^2/191.6 + \ldots$$

$$\begin{aligned} =&0.797 + 0.301+ \\ &0.056 + 0.021+ \\ &0.463 + 0.175+ \\ &2.782 + 1.051+ \\ &0.077 + 0.029 = 5.753 \end{aligned}$$

with df = (5 - 1)(2 - 1) = 4.

The Null Hypothesis is that the population proportion of partly skilled and unskilled workers does not vary with area. We will test this hypothesis at the 10% level

of significance. Looking in Table 8 of the *New Cambridge Statistical Tables* with $P = 10$ and $\nu = 4$ we get $\chi^2_{0.1,(4)} = 7.779$. Since the observed value $x^2 = 5.753$ is less than this, we do not reject the hypothesis that there is no variation with area of the population proportion of partly skilled and unskilled workers. From Table 7 of the with $\nu = 4$ and $x = 5.5$, $x = 6.0$ we get a p-value between 0.2 and 0.24 (more exactly 0.2184). One can see that the main contribution to the value x^2 is from area D, so, to the extent that there is a difference between areas, it looks as if area D is different from the others in its proportion of partly skilled and unskilled workers.
■

Learning outcomes

After working through this chapter you should be able to:

1. discuss the model that underlies all the work with χ^2 goodness-of-fit tests

2. carry out tests for small tables with or without parameter estimation

3. explain the idea of association in two-way tables, and be able to carry out tests of it for small tables.

Sample examination questions

1. (a) Why are the degrees of freedom for a test of independence of row and column classifications in an $r \times c$ contingency table equal to $(r-1)(c-1)$?

   (b) The table below shows the number of units sold by three sales operatives for three different products.

	Product		
Sales Operative	A	B	C
Alpha	14	12	4
Beta	21	16	8
Gamma	15	5	10

   i. Is there any difference in the patterns of sales for different Sales Operatives?

   ii. Display the information in the table in column profile form, and comment on any association displayed.

(Elements of Statistics 2001 Zone B)

2. (a) Why would it usually be unwise to carry out both a chi-squared test for independence of the row and column classifications of a table and a two-way analysis of variance for the same table.

   (b) The table below shows the numbers of piston ring failures in each of three legs of four compressors.

	Compressor legs		
Compressor	North	Centre	South
1	17	17	12
2	11	9	13
3	11	8	19
4	14	7	28

   i. Is there any difference in the pattern of failures over different legs for different compressors?

   ii. By looking at the contributions to χ^2, or profiles, give a qualitative description of any difference that you find.

(Elements of Statistics 2001 Zone A)

3. (a) Explain why the fitted values for a χ^2 test of association in a two-way table take the form that they do.

   (b) The table below shows the number of employees of a manufacturer of animal feeds and soap classified by gender, year of entrance, and length of service in months before resignation. Only those employees with lengths of service of less than 15 months are included.

	1950 Entrants		1951 Entrants	
Length of service	Male	Female	Male	Female
< 3	182	25	147	38
$> 3, < 6$	103	26	54	29
$> 6, < 9$	60	22	47	15
$> 9, < 12$	29	13	21	9
$> 12, < 15$	31	15	12	5

   i. Is there any difference in the patterns of length of service over the different columns of the table?

   ii. Would there be more or less association in the table if the results for female employees were excluded completely? Explain your answer.

(Elements of Statistics 2000 Zone B)

Appendix A

Sample examination paper

UNIVERSITY OF LONDON

279 0007ZB
996 D007ZB
990 0007ZB

BSc degrees in Economics, Management, Finance and the Social Sciences, the Diploma in Economics and Access Route for Students in the External Programme

Statistics 2 (Half-unit)

Wednesday, ?th May 200?: 10.00am to 12.00pm

Candidates should answer **THREE** of the following **FIVE** questions: **QUESTION 1** of Section A (40 marks in total) and **TWO** questions from Section B (30 marks each). **Candidates are strongly advised to divide their time accordingly**

A list of formulae is given at the end of the paper.

Graph paper is provided If this is used it must be securely fastened inside the answer book.

New Cambridge Statistical Tables (second edition) are provided.

A hand held non-programmable calculator may be used when answering questions on this paper. The make and type of machine must be stated clearly on the front of the answer book.

1

UL /10

SECTION A

Answer **all** parts of Question 1 (40 marks).

1. (a) For each of i to iv below, say whether the statement is true or false and briefly give your reasons:

   i. An event of probability 0 can never happen.

   ii. $4\,\mathrm{cov}(X, Y) = \mathrm{var}(X + Y) - \mathrm{var}(X - Y)$.

   iii. It is always better to have a shorter confidence interval for a parameter.

   iv. One should always use an unbiased estimator.

   (8 marks)

   (b) Write brief notes on each of the topics below. Explain in what way each topic is part of statistics, and why it is important.

   i. The Central Limit theorem

   ii. The Poisson distribution.

   iii. Error Sum of Squares.

   iv. Coefficient of Determination.

   (6 marks)

   (c) It is reported that one kg of fishmeal will provide 200g of fish protein, with a standard deviation of 4g.

   (d) Two independent samples of 1kg of fishmeal are tested for protein content. The sample protein content is 195g for one sample and 192g for the other. Are these results consistent with the reported fish protein content of 200g? **(5 marks)**

   (e) X is a random variable with expected value zero and $P(X = 1) = 0.3$ and $P(X = 2) = 0.5$. X takes just one other value besides 1 and 2.

   i. What is the probability that X is negative?

   ii. What is the other value that X takes?

   iii. What is the variance of X?

   (4 marks)

   (f) i. From your tables find the probability that there are more than 5 successes for the binomial distribution with 10 trials and probability of success 0.1.

   ii. From your tables find the probability that $X \leq 5$ when X has a Poisson distribution with mean 10.

   (4 marks)

   (g) A box has 3 red balls and 2 green balls. One ball is taken out of the box at random. If it is red, then two red balls are put back in the box. If the green

ball is taken out, then no ball is put back in. Another ball is then taken at random from the box. What is the probability that the first ball taken was red if the second ball taken is red?

(5 marks)

(h) i. 100 observations are thought to be independent realisations x of a Poisson random variable X with mean 15. They are tabulated below:

$x \leq 9$	$9 < x \leq 14$	$14 < x \leq 19$	$19 < x \leq 24$	$x > 24$
10	33	42	12	3

Test the hypothesis that the observation on X are from a Poisson distribution with mean 15.

(8 marks)

UL /10

SECTION B Answer **two** questions from this section (30 marks each).

2. (a) Mathematicians A, B and C are discussing the proof of a new theorem by one of their colleagues. A will say the proof is false if and only if either B or C say they have found a mistake in the proof. Neither of B, C have had time to read the proof, but B will randomly with a 50% probability say that he has found a mistake. C will wait to hear what B says, but if B does not say he has found a mistake, then C will randomly with a 10% probability say that he has found a mistake in the proof. What is the probability that B said he found a mistake if A says the proof is false? **(12 marks)**

   (b) X is a random variable with density function

$$f_X(x) = 3x^2$$

   over the range $(0, 1)$. Using calculus, find the mean and the variance of X. **(18 marks)**

3. (a) What is an unbiased estimator? How is mean squared error related to bias and variance? **(5 marks)**

   (b) The mean of a Poisson distribution is known to be either 1 or 2. One random observation X is available from the Poisson distribution. Estimators S and T are defined as below:

	$X = 0$	$X = 1$	$X \geq 2$
S	1	1.5	2
T	0	1	2

      i. What are the mean and variance for each of S and T? (Remember to work with both possible parameter values.)

      ii. Which of S, T is the best estimator?

   (12 marks)

   (c) Twenty-four hospitals are randomly divided into two groups of 12. A different expert rates the performance of each group of hospitals on a scale between 50 and 100.

Expert	
1	2
56.79	68.34
62.70	71.50
58.83	72.20
59.03	65.53
61.93	70.13
58.87	69.37
62.65	67.52
58.25	68.06
56.60	67.41
57.16	68.98
61.59	67.63
59.50	69.11

i. Find a 90% confidence interval for the difference between mean ratings for the two experts.

(9 marks)

ii. Test the null hypothesis that Expert 2 gives ratings that are on average 8 greater than Expert 1.

(4 marks)

4. (a) Derive the sum of squares identity for a two-way analysis of variance.

(8 marks)

(b) Suppose that the numbers in the cells of a two-way table are each the mean of several observations appropriate for that cell. If the numbers of observations for each mean are different, is it still appropriate to use a two-way analysis of variance of the means?

(6 marks)

(c) The table below shows scores of four types used for constructing an index of deprivation in a few small areas of England (Wards).

	Types of Score—			
Wards	Employment	Health	Education	Housing
Ampthill	3.80	-1.11	-0.54	-0.77
Arlesey	4.60	-0.90	-0.10	-0.23
Aspley	2.97	-1.56	-0.39	-1.23
Biggleswade Ivel	6.10	-0.58	0.45	0.38
Biggleswade Stratton	4.11	-0.87	0.40	-1.49
Blunham	4.25	-0.86	-0.36	-0.38

i. Give a two-way analysis of variance table for these data.

(10 marks)

ii. Is there significant evidence that Ward effects are different from zero?

(4 marks)

iii. Does this look like a table suitable for analysis of variance?

(2 mark)

5. (a) i. Derive the least squares estimator of slope for a regression model for a line through the origin.

(6 marks)

ii. Show that the slope estimator for the regression through the origin is unbiased.

(6 marks)

(b) The table below shows the relative risk of lung cancer (adjusted for age and sex) in several areas of Cornwall and Devon, and the mean concentration of radon gas (Bq m^{-3}) for those areas.

Area	Relative Risk	Radon
Penrith	1.21	107
Kerrier	1.11	159
Restormel	1.06	89
Carrick	1.04	94
Caradon	0.82	72
North Cornwall	0.82	58
West Devon	0.88	69

i. Fit a straight line to these data using relative risk of lung cancer as the response variable and mean concentration of radon gas as the explanatory variable.

(6 marks)

ii. Give an 85% confidence interval for the expected value of the relative risk of lung cancer if the mean radon value is 95.

(6 marks)

iii. It would be possible to try to improve the regression model by using an additional explanatory variable equal to the square of the radon concentration for each area. Make a scatter plot of these data, and use it to say whether you think it would be better to fit this quadratic term in the regression in addition to the linear one. Sketch on your scatter plot a rough estimate of the fitted curve you would get from such a model.

(6 marks)

Formulae for Statistics

Discrete Distributions

Binomial Distribution

The probability of x successes in n trials is

$$\binom{n}{x}\pi^x(1-\pi)^{n-x}$$

for $x = 0, 1, \ldots, n$ The mean number of successes is $n\pi$ and the variance is $n\pi(1-\pi)$.

Poisson Distribution

The probability of x is

$$e^{-\mu}\frac{\mu^x}{x!}.$$

The mean number of successes is μ and the variance is μ.

Hypergeometric Distribution

The probability of x successes in a sample of size n from a population of size N with M successes is

$$\binom{M}{x}\binom{N-M}{n-x}/\binom{N}{n}.$$

The mean number of successes is nM/N and the variance is
$n(M/N)(1-M/N)(N-n)/(N-1)$.

Sample Quantities

Sample Variance	$s^2 = \sum(x_i - \bar{x})^2/(n-1) = (\sum x_i^2 - n\bar{x}^2)/(n-1)$
Sample Covariance	$\sum(x_i - \bar{x})(y_i - \bar{y})/(n-1) = (\sum x_i y_i - n\bar{x}\bar{y})/(n-1)$
Sample Correlation	$(\sum x_i y_i - n\bar{x}\bar{y})/\sqrt{(\sum y_i^2 - n\bar{y}^2)(\sum x_i^2 - n\bar{x}^2)}$

Inference

Variance of Sample Mean	σ^2/n
One-sample t statistic	$\sqrt{n}(\bar{x}-\mu)/s$ with $(n-1)$ degrees of freedom
Two-sample t statistic	$\dfrac{(\bar{x_1}-\bar{x_2})-(\mu_1-\mu_2)}{\sqrt{[1/n_1+1/n_2]\{[(n_1-1)s_1^2+(n_2-1)s_2^2]/(n_1+n_2-2)\}}}$
Variances for differences of binomial proportions	
Pooled	$\left[\dfrac{(n_1p_1+n_2p_2)}{(n_1+n_2)}\right]\left[1-\dfrac{(n_1p_1+n_2p_2)}{(n_1+n_2)}\right]\left[\dfrac{1}{n_1}+\dfrac{1}{n_2}\right]$
Separate	$p_1(1-p_1)/n_1+p_2(1-p_2)/n_2$
Least Squares	Estimates for $y=\alpha+\beta x$ fitted to (y_i,x_i) for $i=1,2,\ldots,n$ are $a=\bar{y}-b\bar{x}$ and $b=\sum(x_i-\bar{x})(y_i-\bar{y})/\sum(x_i-\bar{x})^2$. The estimate of variance is $[\sum(y_i-\bar{y})^2-b^2\sum(x_i-\bar{x})^2]/(n-2)$. The variance of b is $\sigma^2/\sum(x_i-\bar{x})^2$
Chi-squared Statistic	$\sum(\text{Observed}-\text{Expected})^2/\text{Expected}$, with degrees of freedom depending on the hypothesis tested.

END OF PAPER

Postscript

Statistics is a rewarding subject which needs powers of abstract reasoning, profound knowledge of the field of application and unflagging attention to detail. Few who study this introductory course will ever need to call themselves statisticians, but you should have obtained, by this point, some impression of the scope and importance of a statistician's work. Look at a newspaper, and you will probably see several examples of downright foolish conclusions based on data. One should try to avoid joining the ignorant.