CS1

Actuarial Statistics

Combined Materials Pack
for exams in 2019

The Actuarial Education Company
on behalf of the Institute and Faculty of Actuaries

Subject CS1

2019 Study Guide

Introduction

This Study Guide has been created to help guide you through Subject CS1. It contains all the information that you will need before starting to study Subject CS1 for the 2019 exams and you may also find it useful to refer to throughout your Subject CS1 journey.

The guide is split into two parts:

- Part 1 contains general information about the Core Principles subjects

- Part 2 contains specific information about Subject CS1.

Please read this Study Guide carefully before reading the Course Notes, even if you have studied for some actuarial exams before.

1.1 Before you start

When studying for the UK actuarial exams, you will need:

- a copy of the **Formulae and Tables for Examinations of the Faculty of Actuaries and the Institute of Actuaries, 2nd Edition (2002)** – these are often referred to as simply the *Yellow Tables* or the *Tables*

- a 'permitted' **scientific calculator** – you will find the list of permitted calculators on the profession's website. Please check the list carefully, since it is reviewed each year.

These are both available from the Institute and Faculty of Actuaries' eShop. Please visit **www.actuaries.org.uk**.

The CS1 course assumes that students do have a certain level of statistical knowledge before they start. More detail on this is given in the CS1 Syllabus (see pages 19-26 in this document).

If you feel that you do not have this level of background, you may want to consider ordering the ActEd course 'Pure Maths and Statistics for Actuarial Studies'. More information on this is given later (see page 31 of this document).

Alternatively, a good A-level statistics textbook would help to fill any gaps.

1.2   Core study material

This section explains the role of the Syllabus, Core Reading and supplementary ActEd text. It also gives guidance on how to use these materials most effectively in order to pass the exam.

Some of the information below is also contained in the introduction to the Core Reading produced by the Institute and Faculty of Actuaries.

Syllabus

The Syllabus for Subject CS1 has been produced by the Institute and Faculty of Actuaries. The relevant individual Syllabus Objectives are included at the start of each course chapter and a complete copy of the Syllabus is included in Section 2.2 of this Study Guide. We recommend that you use the Syllabus as an important part of your study.

Core Reading

The Core Reading has been produced by the Institute and Faculty of Actuaries. The purpose of the Core Reading is to ensure that tutors, students and examiners understand the requirements of the syllabus for the qualification examinations for Fellowship of the Institute and Faculty of Actuaries.

It is therefore important that students have a good understanding of the concepts covered by the Core Reading.

The examinations require students to demonstrate their understanding of the concepts given in the syllabus and described in the Core Reading; this will be based on the legislation, professional guidance *etc* that are in force when the Core Reading is published, *ie* on 31 May in the year preceding the examinations.

Therefore the exams in April and September 2019 will be based on the Syllabus and Core Reading as at 31 May 2018. We recommend that you always use the up-to-date Core Reading to prepare for the exams.

Examiners will have this Core Reading when setting the papers. In preparing for examinations, students are advised to work through past examination questions and may find additional tuition helpful. The Core Reading will be updated each year to reflect changes in the syllabus and current practice, and in the interest of clarity.

Accreditation

The Institute and Faculty of Actuaries would like to thank the numerous people who have helped in the development of the material contained in this Core Reading.

ActEd text

Core Reading deals with each syllabus objective and covers what is needed to pass the exam. However, the tuition material that has been written by ActEd enhances it by giving examples and further explanation of key points. Here is an excerpt from some ActEd Course Notes to show you how to identify Core Reading and the ActEd material. **Core Reading is shown in this bold font.**

Note that in the example given above, the index *will* fall if the actual share price goes below the theoretical ex-rights share price. Again, this is consistent with what would happen to an underlying portfolio.

After allowing for chain-linking, **the formula for the investment index then becomes:**

$$I(t) = \frac{\sum_i N_{i,t} P_{i,t}}{B(t)}$$

where **$N_{i,t}$ is the number of shares issued for the *i*th constituent at time *t*;**

$B(t)$ is the base value, or divisor, at time *t*.

This is ActEd text

This is Core Reading

Here is an excerpt from some ActEd Course Notes to show you how to identify Core Reading for R code.

The R code to draw a scatterplot for a bivariate data frame, `<data>`**, is:**

```
plot(<data>)
```

Further explanation on the use of R will not be provided in the Course Notes, but instead be picked up in the Paper B Online Resources (PBOR). We recommend that you refer to and use PBOR at the end of each chapter, or couple of chapters, that contains a significant number of R references.

Copyright

1.3   ActEd study support

This section gives a description of the products offered by ActEd.

Successful students tend to undertake three main study activities:

1.   *Learning* – initial study and understanding of subject material

2.   *Revision* – learning subject material and preparing to tackle exam-style questions

3.   *Rehearsal* – answering exam-style questions, culminating in answering questions at exam speed without notes.

Different approaches suit different people. For example, you may like to learn material gradually over the months running up to the exams or you may do your revision in a shorter period just before the exams. Also, these three activities will almost certainly overlap.

We offer a flexible range of products to suit you and let you control your own learning and exam preparation. The following table shows the products that we produce. Note that not all products are available for all subjects.

LEARNING	LEARNING & REVISION	REVISION	REVISION & REHEARSAL	REHEARSAL
Course Notes PBOR	Assignments Combined Materials Pack (CMP) Assignment Marking Tutorials Online Classroom	Flashcards	Revision Notes ASET	Mock Exam Mock Marking

The products and services are described in more detail below.

'Learning' products

Course Notes

The Course Notes will help you develop the basic knowledge and understanding of principles needed to pass the exam. They incorporate the complete Core Reading and include full explanation of all the syllabus objectives, with worked examples and questions (including some past exam questions) to test your understanding.

Each chapter includes:

- the relevant syllabus objectives

- a chapter summary

- a page of important formulae or definitions (where appropriate)

- practice questions with full solutions.

Paper B Online Resources (PBOR)

The Paper B Online Resources (PBOR) will help you prepare for the computer-based paper. Delivered through a virtual learning environment (VLE), you will have access to worked examples and practice questions. PBOR will also include the Y Assignments, which are two exam-style assessments.

'Learning & revision' products

X Assignments

The Series X Assignments are written assessments that cover the material in each part of the course in turn. They can be used to both develop and test your understanding of the material.

Combined Materials Pack (CMP)

The Combined Materials Pack (CMP) comprises the Course Notes, PBOR and the Series X Assignments.

The CMP is available in **eBook** format for viewing on a range of electronic devices. eBooks can be ordered separately or as an addition to paper products. Visit **www.ActEd.co.uk** for full details about the eBooks that are available, compatibility with different devices, software requirements and printing restrictions.

X / Y Assignment Marking

We are happy to mark your attempts at the X and/or Y assignments. Marking is not included with the Assignments or the CMP and you need to order both Series X and Series Y Marking separately. You should submit your script as an attachment to an email, in the format detailed in your assignment instructions. You will be able to download your marker's feedback via a secure link on the internet.

Don't underestimate the benefits of doing and submitting assignments:

- Question practice during this phase of your study gives an early focus on the end goal of answering exam-style questions.

- You're incentivised to keep up with your study plan and get a regular, realistic assessment of your progress.

- Objective, personalised feedback from a high quality marker will highlight areas on which to work and help with exam technique.

In a recent study, we found that students who attempt more than half the assignments have significantly higher pass rates.

There are two different types of marking product: Series Marking and Marking Vouchers.

Series Marking

Series Marking applies to a specified subject, session and student. If you purchase Series Marking, you will **not** be able to defer the marking to a future exam sitting or transfer it to a different subject or student.

We typically provide full solutions with the Series Assignments. However, if you order Series Marking at the same time as you order the Series Assignments, you can choose whether or not to receive a copy of the solutions in advance. If you choose not to receive them with the study material, you will be able to download the solutions via a secure link on the internet when your marked script is returned (or following the final deadline date if you do not submit a script).

If you are having your attempts at the assignments marked by ActEd, you should submit your scripts regularly throughout the session, in accordance with the schedule of recommended dates set out in information provided with the assignments. This will help you to pace your study throughout the session and leave an adequate amount of time for revision and question practice.

The recommended submission dates are realistic targets for the majority of students. Your scripts will be returned more quickly if you submit them well before the final deadline dates.

Any script submitted *after* the relevant final deadline date will not be marked. It is your responsibility to ensure that we receive scripts in good time.

Marking Vouchers

Marking Vouchers give the holder the right to submit a script for marking at any time, irrespective of the individual assignment deadlines, study session, subject or person.

Marking Vouchers can be used for any assignment. They are valid for four years from the date of purchase and can be refunded at any time up to the expiry date.

Although you may submit your script with a Marking Voucher at any time, you will need to adhere to the explicit Marking Voucher deadline dates to ensure that your script is returned before the date of the exam. The deadline dates are provided with the assignments.

Tutorials

Our tutorials are specifically designed to develop the knowledge that you will acquire from the course material into the higher-level understanding that is needed to pass the exam.

We run a range of different tutorials including face-to-face tutorials at various locations, and Live Online tutorials. Full details are set out in our *Tuition Bulletin*, which is available on our website at **www.ActEd.co.uk**.

Regular and Block Tutorials

In preparation for these tutorials, we expect you to have read the relevant part(s) of the Course Notes before attending the tutorial so that the group can spend time on exam questions and discussion to develop understanding rather than basic bookwork.

You can choose *one* of the following types of tutorial:

- **Regular Tutorials** spread over the session.

- **A Block Tutorial** held two to eight weeks before the exam.

The tutorials outlined above will focus on and develop the skills required for the written Paper A examination. Students wishing for some additional tutor support working through exam-style questions for Paper B may wish to attend a Preparation Day. These will be available Live Online or face-to-face, where students will need to provide their own device capable of running Excel or R as required.

Online Classroom

The Online Classroom acts as either a valuable add-on or a great alternative to a face-to-face or Live Online tutorial, focussing on the written Paper A examination.

At the heart of the Online Classroom in each subject is a comprehensive, easily-searched collection of tutorial units. These are a mix of:

- **teaching** units, helping you to really get to grips with the course material, and

- guided **questions**, enabling you to learn the most efficient ways to answer questions and avoid common exam pitfalls.

The best way to discover the Online Classroom is to see it in action. You can watch a sample of the Online Classroom tutorial units on our website at **www.ActEd.co.uk**.

'Revision' products

Flashcards

For most subjects, there is *a lot of material* to revise. Finding a way to fit revision into your routine as painlessly as possible has got to be a good strategy. Flashcards are a relatively inexpensive option that can provide a massive boost. They can also provide a variation in activities during a study day, and so help you to maintain concentration and effectiveness.

Flashcards are a set of A6-sized cards that cover the key points of the subject that most students want to commit to memory. Each flashcard has questions on one side and the answers on the reverse. We recommend that you use the cards actively and test yourself as you go.

Flashcards are available in **eBook** format for viewing on a range of electronic devices. eBooks can be ordered separately or as an addition to paper products. Visit **www.ActEd.co.uk** for full details about the eBooks that are available, compatibility with different devices, software requirements and printing restrictions.

The following questions and comments might help you to decide if flashcards are suitable for you:

- Do you have a regular train or bus journey?

    Flashcards are ideal for regular bursts of revision on the move.

- Do you want to fit more study into your routine?

    Flashcards are a good option for 'dead time', eg using flashcards on your phone or sticking them on the wall in your study.

- Do you **find** yourself cramming for exams (even if that's not your original plan)?

    Flashcards are an extremely efficient way to do your pre-exam memorising.

If you are retaking a subject, then you might consider using flashcards if you didn't use them on a previous attempt.

'Revision & rehearsal' products

Revision Notes

Our Revision Notes have been designed with input from students to help you revise efficiently. They are suitable for first-time sitters who have worked through the ActEd Course Notes or for retakers (who should find them much more useful and challenging than simply reading through the course again).

The Revision Notes are a set of A5 booklets – perfect for revising on the train or tube to work. Each booklet covers one main theme or a set of related topics from the course and includes:

- Core **Reading** with a set of integrated short questions to develop your bookwork knowledge

- relevant **past** exam questions with concise solutions from the last ten years

- other **useful** revision aids.

ActEd Solutions with Exam Technique (ASET)

The ActEd Solutions with Exam Technique (ASET) contains our solutions to eight past exam papers, plus comment and explanation. In particular, it highlights how questions might have been analysed and interpreted so as to produce a good solution with a wide range of relevant points. This will be valuable in approaching questions in subsequent examinations.

'Rehearsal' products

Mock Exam

The Mock Exam consists of two papers. There is a 100-mark mock exam for the written Paper A examination and a separate mock exam for the computer-based Paper B exam. These provide a realistic test of your exam readiness.

Mock Marking

We are happy to mark your attempts at the mock exams. The same general principles apply as for the Assignment Marking. In particular:

- Mock Exam Marking applies to a specified subject, session and student. In this subject it covers the marking of both papers.

- Marking Vouchers can be used for each mock exam paper. Note that you will need two marking vouchers in order to have the two mock papers marked.

Recall that:

- marking is not included with the products themselves and you need to order it separately

- you should submit your script via email in the format detailed in the mock exam instructions

- you will be able to download the feedback on your marked script via a secure link on the internet.

1.4   Skills

Technical skills

The Core Reading and exam papers for these subjects tend to be very technical. The exams themselves have many calculation and manipulation questions. The emphasis in the exam will therefore be on *understanding* the mathematical techniques and applying them to various, frequently unfamiliar, situations. It is important to have a feel for what the numerical answer should be by having a deep understanding of the material and by doing reasonableness checks.

As a high level of pure mathematics and statistics is generally required for the Core Principles subjects, it is important that your mathematical skills are extremely good. If you are a little rusty you may wish to consider purchasing additional material to help you get up to speed. The course 'Pure Maths and Statistics for Actuarial Studies' is available from ActEd and it covers the mathematical techniques that are required for the Core Principles subjects, some of which are beyond A-Level (or Higher) standard. You do not need to work through the whole course in order – you can just refer to it when you need help on a particular topic. An initial assessment to test your mathematical skills and further details regarding the course can be found on our website at www.ActEd.co.uk.

Study skills

Overall study plan

We suggest that you develop a realistic study plan, building in time for relaxation and allowing some time for contingencies. Be aware of busy times at work, when you may not be able to take as much study leave as you would like. Once you have set your plan, be determined to stick to it. You don't have to be too prescriptive at this stage about what precisely you do on each study day. The main thing is to be clear that you will cover all the important activities in an appropriate manner and leave plenty of time for revision and question practice.

Aim to manage your study so as to allow plenty of time for the concepts you meet in these courses to 'bed down' in your mind. Most successful students will probably aim to complete the courses at least a month before the exam, thereby leaving a sufficient amount of time for revision. By finishing the courses as quickly as possible, you will have a much clearer view of the big picture. It will also allow you to structure your revision so that you can concentrate on the important and difficult areas.

You can also try looking at our discussion forum on the internet, which can be accessed at **www.ActEd.co.uk/forums** (or use the link from our home page at **www.ActEd.co.uk**). There are some good suggestions from students on how to study.

Study sessions

Only do activities that will increase your chance of passing. Try to avoid including activities for the sake of it and don't spend time reviewing material that you already understand. You will only improve your chances of passing the exam by getting on top of the material that you currently find difficult.

Ideally, each study session should have a specific purpose and be based on a specific task, *eg 'Finish reading Chapter 3 and attempt Practice Questions 1.4, 1.7 and 1.12 '*, as opposed to a specific amount of time, *eg 'Three hours studying the material in Chapter 3'*.

Try to study somewhere quiet and free from distractions (*eg* a library or a desk at home dedicated to study). Find out when you operate at your peak, and endeavour to study at those times of the day. This might be between 8*am* and 10*am* or could be in the evening. Take short breaks during your study to remain focused – it's definitely time for a short break if you find that your brain is tired and that your concentration has started to drift from the information in front of you.

Order of study

We suggest that you work through each of the chapters in turn. To get the maximum benefit from each chapter you should proceed in the following order:

1.   Read the Syllabus Objectives. These are set out in the box at the start of each chapter.

2.   Read the Chapter Summary at the end of each chapter. This will give you a useful overview of the material that you are about to study and help you to appreciate the context of the ideas that you meet.

3.   Study the Course Notes in detail, annotating them and possibly making your own notes. Try the self-assessment questions as you come to them. As you study, pay particular attention to the listing of the Syllabus Objectives and to the Core Reading.

4.   Read the Chapter Summary again carefully. If there are any ideas that you can't remember covering in the Course Notes, read the relevant section of the notes again to refresh your memory.

5.   Attempt (at least some of) the Practice Questions that appear at the end of the chapter.

6.   Where relevant, work through the relevant Paper B Online Resources for the chapter(s). You will need to have a good understanding of the relevant section of the paper-based course before you attempt the corresponding section of PBOR.

It's a fact that people are more likely to remember something if they review it several times. So, do look over the chapters you have studied so far from time to time. It is useful to re-read the Chapter Summaries or to try the Practice Questions again a few days after reading the chapter itself. It's a good idea to annotate the questions with details of when you attempted each one. This makes it easier to ensure that you try all of the questions as part of your revision without repeating any that you got right first time.

Once you've read the relevant part of the notes and tried a selection of questions from the Practice Questions (and attended a tutorial, if appropriate) you should attempt the corresponding assignment. If you submit your assignment for marking, spend some time looking through it carefully when it is returned. It can seem a bit depressing to analyse the errors you made, but you will increase your chances of passing the exam by learning from your mistakes. The markers will try their best to provide practical comments to help you to improve.

To be really prepared for the exam, you should not only know and understand the Core Reading but also be aware of what the examiners will expect. Your revision programme should include plenty of question practice so that you are aware of the typical style, content and marking structure of exam questions. You should attempt as many past exam questions as you can.

Active study

Here are some techniques that may help you to study actively.

1.   Don't believe everything you read. Good students tend to question everything that they read. They will ask 'why, how, what for, when?' when confronted with a new concept, and they will apply their own judgement. This contrasts with those who unquestioningly believe what they are told, learn it thoroughly, and reproduce it (unquestioningly?) in response to exam questions.

2.   Another useful technique as you read the Course Notes is to think of possible questions that the examiners could ask. This will help you to understand the examiners' point of view and should mean that there are fewer nasty surprises in the exam room. Use the Syllabus to help you make up questions.

3.   Annotate your notes with your own ideas and questions. This will make you study more actively and will help when you come to review and revise the material. Do not simply copy out the notes without thinking about the issues.

4.   Attempt the questions in the notes as you work through the course. Write down your answer before you refer to the solution.

5.   Attempt other questions and assignments on a similar basis, *ie* write down your answer before looking at the solution provided. Attempting the assignments under exam conditions has some particular benefits:

    •   It forces you to think and act in a way that is similar to how you will behave in the exam.

    •   When you have your assignments marked it is *much* more useful if the marker's comments can show you how to improve your performance under exam conditions than your performance when you have access to the notes and are under no time pressure.

    •   The knowledge that you are going to do an assignment under exam conditions and then submit it (however good or bad) for marking can act as a powerful incentive to make you study each part as well as possible.

    •   It is also quicker than trying to write perfect answers.

6.   Sit a mock exam four to six weeks before the real exam to identify your weaknesses and work to improve them. You could use a mock exam written by ActEd or a past exam paper.

You can find further information on how to study in the profession's Student Handbook, which you can download from their website at:

   www.actuaries.org.uk/studying

Revision and exam skills

Revision skills

You will have sat many exams before and will have mastered the exam and revision techniques that suit you. However it is important to note that due to the high volume of work involved in the Core Principles subjects it is not possible to leave all your revision to the last minute. Students who prepare well in advance have a better chance of passing their exams on the first sitting.

Unprepared students find that they are under time pressure in the exam. Therefore it is important to find ways of maximising your score in the shortest possible time. Part of your preparation should be to practise a large number of exam-style questions under timed exam conditions as soon as possible. This will:

- help you to develop the necessary understanding of the techniques required

- highlight the key topics, which crop up regularly in many different contexts and questions

- help you to practise the specific skills that you will need to pass the exam.

There are many sources of exam-style questions. You can use past exam papers, the Practice Questions at the end of each chapter (which include many past exam questions), assignments, mock exams, the Revision Notes and ASET.

Exam question skill levels

Exam questions are not designed to be of similar difficulty. The Institute and Faculty of Actuaries specifies different skill levels that questions may be set with reference to.

Questions may be set at any skill level:

- Knowledge – demonstration of a detailed knowledge and understanding of the topic

- Application – demonstration of an ability to apply the principles underlying the topic within a given context

- Higher Order – demonstration of an ability to perform deeper analysis and assessment of situations, including forming judgements, taking into account different points of view, comparing and contrasting situations, suggesting possible solutions and actions, and making recommendations.

Command verbs

The Institute and Faculty of Actuaries use command verbs (such as 'Define', 'Discuss' and 'Explain') to help students to identify what the question requires. The profession has produced a document, 'Command verbs used in the Associate and Fellowship written examinations', to help students to understand what each command verb is asking them to do.

It also gives the following advice:

- The use of a specific command verb within a syllabus objective does not indicate that this is the only form of question which can be asked on the topic covered by that objective.

- The Examiners may ask a question on any syllabus topic using any of the agreed command verbs, as are defined in the document.

You can find the relevant document on the profession's website at:

https://www.actuaries.org.uk/studying/prepare-your-exams

1.5   The examination

What to take to the exam

IMPORTANT NOTE: The following information was correct at the time of printing, however it is important to keep up-to-date with any changes. See the profession's website for the latest guidance.

For the written exams the examination room will be equipped with:

- the question paper

- an answer booklet

- rough paper

- a copy of the Yellow Tables.

Remember to take with you:

- black pens

- a permitted scientific calculator – please refer to **www.actuaries.org.uk** for the latest advice.

Please also refer to the profession's website and your examination instructions for details about what you will need for the computer-based Paper B exam.

Past exam papers

You can download some past exam papers and Examiners' Reports from the profession's website at **www.actuaries.org.uk**. However, please be aware that these exam papers are for the pre-2019 syllabus and not all questions will be relevant.

1.6   Queries and feedback

Questions and queries

From time to time you may come across something in the study material that is unclear to you. The easiest way to solve such problems is often through discussion with friends, colleagues and peers – they will probably have had similar experiences whilst studying. If there's no-one at work to talk to then use our discussion forum at **www.ActEd.co.uk/forums** (or use the link from our home page at **www.ActEd.co.uk**).

Our online forum is dedicated to actuarial students so that you can get help from fellow students on any aspect of your studies from technical issues to study advice. You could also use it to get ideas for revision or for further reading around the subject that you are studying. ActEd tutors will visit the site from time to time to ensure that you are not being led astray and we also post other frequently asked questions from students on the forum as they arise.

If you are still stuck, then you can send queries by email to the relevant subject email address (see Section 2.6), but we recommend that you try the forum first. We will endeavour to contact you as soon as possible after receiving your query but you should be aware that it may take some time to reply to queries, particularly when tutors are away from the office running tutorials. At the busiest teaching times of year, it may take us more than a week to get back to you.

If you have many queries on the course material, you should raise them at a tutorial or book a personal tuition session with an ActEd tutor. Information about personal tuition is set out in our current brochure. Please email **ActEd@bpp.com** for more details.

Feedback

If you find an error in the course, please check the corrections page of our website (**www.ActEd.co.uk/paper_corrections.html**) to see if the correction has already been dealt with. Otherwise please send details via email to the relevant subject email address (see Section 2.6).

Each year our tutors work hard to improve the quality of the study material and to ensure that the courses are as clear as possible and free from errors. We are always happy to receive feedback from students, particularly details concerning any errors, contradictions or unclear statements in the courses. If you have any comments on this course please email them to the relevant subject email address (see Section 2.6).

Our tutors also work with the profession to suggest developments and improvements to the Syllabus and Core Reading. If you have any comments or concerns about the Syllabus or Core Reading, these can be passed on via ActEd. Alternatively, you can send them directly to the Institute and Faculty of Actuaries' Examination Team by email to **education.services@actuaries.org.uk**.

2.1   Subject CS1 – background

History

The Actuarial Statistics subjects (Subjects CS1 and CS2) are new subjects in the Institute and Faculty of Actuaries 2019 Curriculum.

Subject CS1 is *Actuarial Statistics*.

Predecessors

The topics covered in the Actuarial Statistics subjects (Subjects CS1 and CS2) cover content previously in Subjects CT3, CT4, CT6 and a small amount from Subject ST9:

- Subject CS1 contains material from Subjects CT3 and CT6.

- Subject CS2 contains material from Subjects CT4, CT6 and ST9.

Exemptions

You will need to have passed or been granted an exemption from Subject CT3 to be eligible for a pass in Subject CS1 during the transfer process.

Links to other subjects

- Subject CS2 – Risk Modelling and Survival Analysis builds directly on the material in this subject.

- Subjects CM1 and CM2 – Actuarial Mathematics 1 and Financial Engineering and Loss Reserving apply the material in this subject to actuarial and financial modelling.

2.2   Subject CS1 – Syllabus and Core Reading

Syllabus

The Syllabus for Subject CS1 is given here. To the right of each objective are the chapter numbers in which the objective is covered in the ActEd course.

Aim

The aim of the Actuarial Statistics 1 subject is to provide a grounding in mathematical and statistical techniques that are of particular relevance to actuarial work.

Competences

On successful completion of this subject, a student will be able to:

1.   describe the essential features of statistical distributions

2.   summarise data using appropriate statistical analysis, descriptive statistics and graphical presentation

3.   describe and apply the principles of statistical inference

4.   describe, apply and interpret the results of the linear regression model and generalised linear models

5.   explain the fundamental concepts of Bayesian statistics and use them to compute Bayesian estimators.

Syllabus topics

1.	Random variables and distributions	(20%)
2.	Data analysis	(15%)
3.	Statistical inference	(20%)
4.	Regression theory and applications	(30%)
5.	Bayesian statistics	(15%)

The weightings are indicative of the approximate balance of the assessment of this subject between the main syllabus topics, averaged over a number of examination sessions.

The weightings also have a correspondence with the amount of learning material underlying each syllabus topic. However, this will also reflect aspects such as:

- the relative complexity of each topic, and hence the amount of explanation and support required for it

- the need to provide thorough foundation understanding on which to build the other objectives

- the extent of prior knowledge which is expected

- the degree to which each topic area is more knowledge or application based.

Assumed knowledge

This subject assumes that a student will be competent in the following elements of foundational mathematics and basic statistics:

1   Summarise the main features of a data set (exploratory data analysis)

   1.1   Summarise a set of data using a table or frequency distribution, and display it graphically using a line plot, a box plot, a bar chart, histogram, stem and leaf plot, or other appropriate elementary device.

   1.2   Describe the level/location of a set of data using the mean, median, mode, as appropriate.

   1.3   Describe the spread/variability of a set of data using the standard deviation, range, interquartile range, as appropriate.

   1.4   Explain what is meant by symmetry and skewness for the distribution of a set of data.

2   Probability

   2.1   Set functions and sample spaces for an experiment and an event.

   2.2   Probability as a set function on a collection of events and its basic properties.

   2.3   Calculate probabilities of events in simple situations.

   2.4   Derive and use the addition rule for the probability of the union of two events.

   2.5   Define and calculate the conditional probability of one event given the occurrence of another event.

   2.6   Derive and use Bayes' Theorem for events.

   2.7   Define independence for two events, and calculate probabilities in situations involving independence.

3   Random variables

   3.1   Explain what is meant by a discrete random variable, define the distribution function and the probability function of such a variable, and use these functions to calculate probabilities.

   3.2   Explain what is meant by a continuous random variable, define the distribution function and the probability density function of such a variable, and use these functions to calculate probabilities.

   3.3   Define the expected value of a function of a random variable, the mean, the variance, the standard deviation, the coefficient of skewness and the moments of a random variable, and calculate such quantities.

3.4   Evaluate probabilities associated with distributions (by calculation or by referring to tables as appropriate).

3.5   Derive the distribution of a function of a random variable from the distribution of the random variable.

Detailed syllabus objectives

1   Random variables and distributions   (20%)

1.1   Define basic univariate distributions and use them to calculate probabilities, quantiles and moments.   (Chapter 1)

   1.1.1   Define and explain the key characteristics of the discrete distributions: geometric, binomial, negative binomial, hypergeometric, Poisson and uniform on a finite set.

   1.1.2   Define and explain the key characteristics of the continuous distributions: normal, lognormal, exponential, gamma, chi-square, t, F, beta and uniform on an interval.

   1.1.3   Evaluate probabilities and quantiles associated with distributions (by calculation or using statistical software as appropriate).

   1.1.4   Define and explain the key characteristics of the Poisson process and explain the connection between the Poisson process and the Poisson distribution.

   1.1.5   Generate basic discrete and continuous random variables using the inverse transform method.

   1.1.6   Generate discrete and continuous random variables using statistical software.

1.2   Independence, joint and conditional distributions, linear combinations of random variables   (Chapter 3)

   1.2.1   Explain what is meant by jointly distributed random variables, marginal distributions and conditional distributions.

   1.2.2   Define the probability function/density function of a marginal distribution and of a conditional distribution.

   1.2.3   Specify the conditions under which random variables are independent.

   1.2.4   Define the expected value of a function of two jointly distributed random variables, the covariance and correlation coefficient between two variables, and calculate such quantities.

   1.2.5   Define the probability function/density function of the sum of two independent random variables as the convolution of two functions.

   1.2.6   Derive the mean and variance of linear combinations of random variables.

   1.2.7   Use generating functions to establish the distribution of linear combinations of independent random variables.

1.3   Expectations, conditional expectations   (Chapter 4)

   1.3.1   Define the conditional expectation of one random variable given the value of
           another random variable, and calculate such a quantity.

   1.3.2   Show how the mean and variance of a random variable can be obtained from
           expected values of conditional expected values, and apply this.

1.4   Generating functions   (Chapter 2)

   1.4.1   Define and determine the moment generating function of random variables.

   1.4.2   Define and determine the cumulant generating function of random variables.

   1.4.3   Use generating functions to determine the moments and cumulants of random
           variables, by expansion as a series or by differentiation, as appropriate.

   1.4.4   Identify the applications for which a moment generating function, a cumulant
           generating function and cumulants are used, and the reasons why they are used.

1.5   Central Limit Theorem – statement and application   (Chapter 5)

   1.5.1   State the Central Limit Theorem for a sequence of independent, identically
           distributed random variables.

   1.5.2   Generate simulated samples from a given distribution and compare the sampling
           distribution with the Normal.

2   Data analysis   (15%)

2.1   Exploratory data analysis   (Chapter 10)

   2.1.1   Describe the purpose of exploratory data analysis.

   2.1.2   Use appropriate tools to calculate suitable summary statistics and undertake
           exploratory data visualizations.

   2.1.3   Define and calculate Pearson's, Spearman's and Kendall's measures of correlation
           for bivariate data, explain their interpretation and perform statistical inference as
           appropriate.

   2.1.4   Use Principal Components Analysis to reduce the dimensionality of a complex
           data set.

2.2   Random sampling and sampling distributions   (Chapter 6)

    2.2.1   Explain what is meant by a sample, a population and statistical inference.

    2.2.2   Define a random sample from a distribution of a random variable.

    2.2.3   Explain what is meant by a statistic and its sampling distribution.

    2.2.4   Determine the mean and variance of a sample mean and the mean of a sample variance in terms of the population mean, variance and sample size.

    2.2.5   State and use the basic sampling distributions for the sample mean and the sample variance for random samples from a normal distribution.

    2.2.6   State and use the distribution of the t-statistic for random samples from a normal distribution.

    2.2.7   State and use the F distribution for the ratio of two sample variances from independent samples taken from normal distributions.

3   Statistical inference   (20%)

3.1   Estimation and estimators   (Chapter 7)

    3.1.1   Describe and apply the method of moments for constructing estimators of population parameters.

    3.1.2   Describe and apply the method of maximum likelihood for constructing estimators of population parameters.

    3.1.3   Define the terms: efficiency, bias, consistency and mean squared error.

    3.1.4   Define and apply the property of unbiasedness of an estimator.

    3.1.5   Define the mean square error of an estimator, and use it to compare estimators.

    3.1.6   Describe and apply the asymptotic distribution of maximum likelihood estimators.

    3.1.7   Use the bootstrap method to estimate properties of an estimator.

3.2 Confidence intervals   (Chapter 8)

3.2.1   Define in general terms a confidence interval for an unknown parameter of a distribution based on a random sample.

3.2.2   Derive a confidence interval for an unknown parameter using a given sampling distribution.

3.2.3   Calculate confidence intervals for the mean and the variance of a normal distribution.

3.2.4   Calculate confidence intervals for a binomial probability and a Poisson mean, including the use of the normal approximation in both cases.

3.2.5   Calculate confidence intervals for two-sample situations involving the normal distribution, and the binomial and Poisson distributions using the normal approximation.

3.2.6   Calculate confidence intervals for a difference between two means from paired data.

3.2.7   Use the bootstrap method to obtain confidence intervals.

3.3 Hypothesis testing and goodness of fit   (Chapter 9)

3.3.1   Explain what is meant by the terms null and alternative hypotheses, simple and composite hypotheses, type I and type II errors, test statistic, likelihood ratio, critical region, level of significance, probability-value and power of a test.

3.3.2   Apply basic tests for the one-sample and two-sample situations involving the normal, binomial and Poisson distributions, and apply basic tests for paired data.

3.3.3   Apply the permutation approach to non-parametric hypothesis tests.

3.3.4   Use a chi-square test to test the hypothesis that a random sample is from a particular distribution, including cases where parameters are unknown.

3.3.5   Explain what is meant by a contingency (or two-way) table, and use a chi-square test to test the independence of two classification criteria.

4   Regression theory and applications   (30%)

4.1   Linear regression   (Chapters 11, 11b)

    4.1.1   Explain what is meant by response and explanatory variables.

    4.1.2   State the simple regression model (with a single explanatory variable).

    4.1.3   Derive the least squares estimates of the slope and intercept parameters in a simple linear regression model.

    4.1.4   Use appropriate software to fit a simple linear regression model to a data set and interpret the output.

- Perform statistical inference on the slope parameter.

- Describe the use of measures of goodness of fit of a linear regression model.

- Use a fitted linear relationship to predict a mean response or an individual response with confidence limits.

- Use residuals to check the suitability and validity of a linear regression model.

    4.1.5   State the multiple linear regression model (with several explanatory variables).

    4.1.6   Use appropriate software to fit a multiple linear regression model to a data set and interpret the output.

    4.1.7   Use measures of model fit to select an appropriate set of explanatory variables.

4.2   Generalised linear models   (Chapter 12)

    4.2.1   Define an exponential family of distributions. Show that the following distributions may be written in this form: binomial, Poisson, exponential, gamma, normal.

    4.2.2   State the mean and variance for an exponential family, and define the variance function and the scale parameter. Derive these quantities for the distributions above.

    4.2.3   Explain what is meant by the link function and the canonical link function, referring to the distributions above.

    4.2.4   Explain what is meant by a variable, a factor taking categorical values and an interaction term. Define the linear predictor, illustrating its form for simple models, including polynomial models and models involving factors.

    4.2.5   Define the deviance and scaled deviance and state how the parameters of a generalised linear model may be estimated. Describe how a suitable model may be chosen by using an analysis of deviance and by examining the significance of the parameters.

    4.2.6   Define the Pearson and deviance residuals and describe how they may be used.

4.2.7   Apply statistical tests to determine the acceptability of a fitted model: Pearson's chi-square test and the likelihood ratio test.

4.2.8   Fit a generalised linear model to a data set and interpret the output.

5   Bayesian statistics   (15%)

(Chapters 13, 14 and 15)

5.1   Explain the fundamental concepts of Bayesian statistics and use these concepts to calculate Bayesian estimators.

5.1.1   Use Bayes' theorem to calculate simple conditional probabilities.

5.1.2   Explain what is meant by a prior distribution, a posterior distribution and a conjugate prior distribution.

5.1.3   Derive the posterior distribution for a parameter in simple cases.

5.1.4   Explain what is meant by a loss function.

5.1.5   Use simple loss functions to derive Bayesian estimates of parameters.

5.1.6   Explain what is meant by the credibility premium formula and describe the role played by the credibility factor.

5.1.7   Explain the Bayesian approach to credibility theory and use it to derive credibility premiums in simple cases.

5.1.8   Explain the empirical Bayes approach to credibility theory and use it to derive credibility premiums in simple cases.

5.1.9   Explain the differences between the two approaches and state the assumptions underlying each of them.

Core Reading

The Subject CS1 Course Notes include the Core Reading in full, integrated throughout the course.

Accreditation

The Institute and Faculty of Actuaries would like to thank the numerous people who have helped in the development of the material contained in this Core Reading.

Further reading

The exam will be based on the relevant Syllabus and Core Reading and the ActEd course material will be the main source of tuition for students.

2.3   Subject CS1 – the course structure

There are four parts to the Subject CS1 course. The parts cover related topics and have broadly equal marks in the paper-based exam. The parts are broken down into chapters.

The following table shows how the parts, the chapters and the syllabus items relate to each other. The end columns show how the chapters relate to the days of the regular tutorials. We have also given you a broad indication of the length of each chapter. This table should help you plan your progress across the study session.

Part	Chapter	Title	No of pages	Syllabus objectives	4 full days
1	1	Probability distributions	63	1.1	1
	2	Generating functions	30	1.4	
	3	Joint distributions	57	1.2	
	4	Conditional expectation	19	1.3	
2	5	Central limit theorem	27	1.5	2
	6	Sampling and statistical inference	35	2.2	
	7	Point estimation	63	3.1	
	8	Confidence intervals	46	3.2	
3	9	Hypothesis testing	86	3.3	3
	10	Data analysis	41	2.1	
	11	Linear regression	54	4.1.1-4.1.4	
	11b	Multiple linear regression	24	4.1.5-4.1.7	
4	12	Generalised linear models	74	4.2	4
	13	Bayesian statistics	38	5.1.1-5.1.5	
	14	Credibility theory	32	5.1.6, 5.1.7, 5.1.9	
	15	Empirical Bayes credibility theory	54	5.1.8, 5.1.9	

2.4   Subject CS1 – summary of ActEd products

The following products are available for Subject CS1:

- Course Notes

- PBOR (including the Y Assignments)

- X Assignments – four assignments:

    – X1, X2: 80-mark tests (you are allowed 2¾ hours to complete these)

    – X3, X4: 100-mark tests (you are allowed 3¼ hours to complete these)

- Series X Marking

- Series Y Marking

- Online Classroom – over 150 tutorial units

- Flashcards

- Revision Notes

- ASET – four years' exam papers, *ie* eight papers, with full worked solutions, covering the period April 2014 to September 2017

- Mock Exam

- Mock Exam Marking

- Marking Vouchers.

We will endeavour to release as much material as possible but unfortunately some revision products may not be available until the September 2019 or even April 2020 exam sessions. Please check the ActEd website or email ActEd@bpp.com for more information.

The following tutorials are typically available for Subject CS1:

- Regular Tutorials (four days)

- Block Tutorials (four days)

- a Preparation Day for the computer-based exam.

Full details are set out in our *Tuition Bulletin*, which is available on our website at **www.ActEd.co.uk**.

2.5   Subject CS1 – skills and assessment

Technical skills

The *Actuarial Statistics* subjects (Subjects CS1 and CS2) are very mathematical and have relatively few questions requiring wordy answers.

Exam skills

Exam question skill levels

In the CS subjects, the approximate split of assessment across the three skill types is:

- Knowledge – 20%

- Application – 65%

- Higher Order skills – 15%.

Assessment

Assessment consists of a combination of a 3¾-hour written examination and a 1¾-hour computer-based practical examination.

2.6   Subject CS1 – frequently asked questions

Q: ***What knowledge of earlier subjects should I have?***

A: No knowledge of earlier subjects is required.

Q: ***What level of mathematics is required?***

A: The level of maths you need for this course is broadly A-level standard. However, there may be some symbols (*eg* the gamma function) that are not usually included on A-level syllabuses. You will find the course (and the exam) much easier if you feel comfortable with the mathematical techniques (*eg* integration by parts) used in the course and you feel confident in applying them yourself.

If your maths or statistics is a little rusty you may wish to consider purchasing additional material to help you get up to speed. The course 'Pure Maths and Statistics for Actuarial Studies' is available from ActEd and it covers the mathematical techniques that are required for the Core Principles subjects, some of which are beyond A-Level (or Higher) standard. You do not need to work through the whole course in order – you can just refer to it when you need help on a particular topic. An initial assessment to test your mathematical skills and further details regarding the course can be found on our website.

Q: ***What should I do if I discover an error in the course?***

A: If you find an error in the course, please check our website at:

www.ActEd.co.uk/paper_corrections.html

to see if the correction has already been dealt with. Otherwise please send details via email to **CS1@bpp.com**.

Q: ***Who should I send feedback to?***

A: We are always happy to receive feedback from students, particularly details concerning any errors, contradictions or unclear statements in the courses.

If you have any comments on this course in general, please email **CS1@bpp.com**.

If you have any comments or concerns about the Syllabus or Core Reading, these can be passed on to the profession via ActEd. Alternatively, you can send them directly to the Institute and Faculty of Actuaries' Examination Team by email to **education.services@actuaries.org.uk**.

1

Probability distributions

Syllabus objectives

1.1   Define basic univariate distributions and use them to calculate probabilities, quantiles and moments.

   1.1.1   Define and explain the key characteristics of the discrete distributions: geometric, binomial, negative binomial, hypergeometric, Poisson and uniform on a finite set.

   1.1.2   Define and explain the key characteristics of the continuous distributions: normal, lognormal, exponential, gamma, chi-square, *t*, *F*, beta and uniform on an interval.

   1.1.3   Evaluate probabilities and quantiles associated with distributions (by calculation or using statistical software as appropriate).

   1.1.4   Define and explain the key characteristics of the Poisson process and explain the connection between the Poisson process and the Poisson distribution.

   1.1.5   Generate basic discrete and continuous random variables using the inverse transform method.

   1.1.6   Generate discrete and continuous random variables using statistical software.

0   Introduction

This unit introduces the standard distributions that are used in actuarial work.

We look in this chapter at all the standard probability distributions used in Subject CS1.

This chapter does assume that you have some basic knowledge of statistics and probability. If your knowledge in this area is rusty, you can purchase additional ActEd materials to remind you of those statistical ideas. Please see the ActEd website for further details.

There is a book called Formulae and Tables for Examinations (simply denoted the *Tables* in this course) available in the exam, which contains many relevant formulae for the distributions in this chapter as well as probability tables. This is available from the Profession and you should purchase a copy as soon as possible (if you have not already done so) as it is essential to your studying of the Subject CS1 course.

Many of the formulae in this course are contained in the *Tables*. So you should concentrate on being able to *apply* them to calculate means, variances, coefficients of skewness and probabilities, rather than memorising them.

If you have studied statistics to A-Level standard or equivalent you should find this chapter straightforward. However, some of the standard distributions (*eg* lognormal and gamma) that are used frequently in statistical work in finance and insurance, may be new to you. Since we will be using the properties of these distributions in the rest of the course, it is vital that you feel confident with them.

1   Important discrete distributions

In this section we will look at the standard discrete distributions that we will use in actuarial modelling work.

Remember all of these results are given in the *Tables* – concentrate on understanding and applying them, particularly to calculating probabilities, rather than memorising them.

The distributions considered here are all models for the number of something – *eg* number of 'successes', number of 'trials', number of deaths, number of claims. The values assumed by the variables are integers from the set {0, 1, 2, 3, …} – such variables are often referred to as counting variables.

1.1   Uniform distribution

Sample space $S = \{1,\ 2,\ 3,\ldots,k\}$.

Probability measure: equal assignment (1/k) to all outcomes, *ie* all outcomes are equally likely.

Random variable X defined by $X(i) = i$, $(i = 1,\ 2,\ 3,\ldots,k)$.

Distribution: $P(X = x) = \dfrac{1}{k}$ $(x = 1,\ 2,\ 3,\ldots,k)$

Moments:

$$\mu = E[X] = \frac{(1 + 2 + \cdots + k)}{k} = \frac{\frac{1}{2}k(k+1)}{k} = \frac{k+1}{2}$$

$$E[X^2] = \frac{(1^2 + 2^2 + \cdots + k^2)}{k} = \frac{\frac{1}{6}k(k+1)(2k+1)}{k} = \frac{(k+1)(2k+1)}{6}$$

$$\Rightarrow \sigma^2 = \frac{k^2 - 1}{12}$$

For example, if X is the score on a fair die, $P(X = x) = \frac{1}{6}$ for $x = 1,2,\ldots,6$.

Question

Verify that $\sigma^2 = \dfrac{k^2 - 1}{12}$ for the uniform distribution.

Solution

σ is the standard deviation, and σ^2 is the variance, which is calculated as:

$$\sigma^2 = E(X^2) - [E(X)]^2 = \frac{(k+1)(2k+1)}{6} - \frac{(k+1)^2}{4}$$

$$= \frac{2(2k^2 + 3k + 1) - 3(k^2 + 2k + 1)}{12}$$

$$= \frac{k^2 - 1}{12}$$

R code for simulating a random sample from the discrete uniform distribution.

Generate a vector for sample space S = {1, 2, 3, ... , 20}:

```
S = 1:20
```

Simulate 100 values:

```
sample(S, 100, replace = TRUE)
```

1.2 Bernoulli distribution

A Bernoulli trial is an experiment which has (or can be regarded as having) only two possible outcomes – s ('success') and f ('failure').

Sample space $S = \{s, f\}$. The words 'success' and 'failure' are merely labels – they do not necessarily carry with them the ordinary meanings of the words.

For example in life insurance, a success could mean a death.

Probability measure: $P(\{s\}) = p$, $P(\{f\}) = 1 - p$

Random variable X defined by $X(s) = 1$, $X(f) = 0$. X is the number of successes that occur (0 or 1).

Distribution: $P(X = x) = p^x(1-p)^{1-x}$, x = 0,1; $0 < p < 1$

Moments: $\mu = p$

$\sigma^2 = p - p^2 = p(1 - p)$

A Bernoulli variable is also called an 'indicator' variable – its value can be used to indicate whether or not some specified event, for example A, occurs. Set $X = 1$ if A occurs, 0 if A does not occur. If $P(A) = p$ then X has the above Bernoulli distribution.

The event A could, for example, be the survival of an assured life over one year.

An assured life is a person with an insurance policy that makes a payment on death.

Another example of a Bernoulli random variable occurs when a fair die is thrown once. If X is the number of sixes obtained, $p = \frac{1}{6}$, $(1-p) = \frac{5}{6}$ and $P(X = 0) = \frac{5}{6}$ and $P(X = 1) = \frac{1}{6}$.

R code. See R code for Binomial distribution.

1.3   Binomial distribution

Consider a sequence of n Bernoulli trials as above such that:

(i)   the trials are independent of one another, *ie* the outcome of any trial does not depend on the outcomes of any other trials

and:

(ii)   the trials are identical, *ie* at each trial $P(\{s\}) = p$.

Such a sequence is called a 'sequence of n independent, identical, Bernoulli (p) trials' or, for short, a 'sequence of n Bernoulli (p) trials'.

A quick way of saying independent and identically distributed is IID. We will need this idea later.

The independence allows the probability of a joint outcome involving two or more trials to be expressed as the product of the probabilities of the outcomes associated with each separate trial concerned.

Sample space S: the joint set of outcomes of all n trials

Probability measure: as above for each trial

Random variable X is the number of successes that occur in the n trials.

Distribution: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0,\ 1,\ 2,...,n$; $0 < p < 1$

The coefficients here are the same as in the binomial expansion that can be obtained using the numbers from Pascal's triangle, *ie* $\binom{n}{x} = {}^nC_x = \dfrac{n!}{(n-x)!\,x!}$. We can work out these quantities using the *nCr* function on a calculator.

If X is distributed binomially with parameters n and p, then we can write $X \sim Bin(n,p)$.

The fact that a $Bin(n,p)$ distribution arises from the sum of n independent and identical Bernoulli (p) trials is important and will be used later to prove some important results.

Moments: $\mu = np$

$$\sigma^2 = np(1-p)$$

Very often when using the binomial distribution we will write $1 - p = q$.

As an example of the binomial distribution, suppose that X is the number of sixes obtained when a fair die is thrown 10 times. Then $P(X = x) = {}^{10}C_x \left(\frac{1}{6} \right)^x \left(\frac{5}{6} \right)^{10-x}$ and the probability of exactly one 'six' in ten throws is ${}^{10}C_1 \left(\frac{1}{6} \right)^1 \left(\frac{5}{6} \right)^9 = 0.3230$. There are 10 $\left(= {}^{10}C_1 \right)$ ways of obtaining exactly one 'six', *ie* the 'six' could be on the first throw, the second throw, …. or the tenth throw.

Question

Calculate the probability that at least 9 out of a group of 10 people who have been infected by a serious disease will survive, if the survival probability for the disease is 70%.

Solution

The number of survivors is distributed binomially with parameters $n = 10$, and $p = 0.7$. If X is the number of survivors, then:

$$P(X \geq 9) = P(X = 9 \text{ or } 10) = \binom{10}{9} \times 0.7^9 \times 0.3 + \binom{10}{10} \times 0.7^{10} = 0.1493$$

Alternatively, we could use the cumulative binomial probabilities given on page 187 of the Tables.

The R code for simulating values and calculating probabilities and quantiles from the binomial distribution uses the R functions `rbinom`, `dbinom`, `pbinom` **and** `qbinom`**. The prefixes** `r`, `d`, `p`, **and** `q` **stand for random generation, density, distribution and quantile functions respectively.**

R code for simulating a random sample of 100 values from the binomial distribution with $n = 20$ **and** $p = 0.3$ **:**

```
n = 20
p = 0.3
rbinom(100, n, p)
```

Calculate $P(X = 2)$ **:**

```
dbinom(2, n, p)
```

Similarly, the cumulative distribution function (CDF) and quantiles can be calculated with `pbinom` **and** `qbinom.`

For a Bernoulli distribution the parameter n **is set to** $n = 1$**.**

1.4 Geometric distribution

Consider again a sequence of independent, identical Bernoulli trials with $P(\{s\}) = p$. The variable of interest now is the number of trials that has to be performed until the first success occurs. Because trials are performed one after the other and a success is awaited, this distribution is one of a class of distributions called *waiting-time distributions*.

Random variable X : Number of the trial on which the first success occurs

Distribution: For $X = x$ there must be a run of $(x - 1)$ failures followed by a
                        success, so $P(X = x) = p(1-p)^{x-1}$, $x = 1, 2, 3,\ldots$ $(0 < p < 1)$

Moments: $$\mu = \frac{1}{p}$$

                        $$\sigma^2 = \frac{(1-p)}{p^2}$$

For example, if the probability that a phone call leads to a sale is $\frac{1}{4}$ and X is the number of

phone calls required to make the first sale, then $P(X = 3) = \frac{1}{4} \times \left(\frac{3}{4}\right)^2 = 0.140625$.

Question

If the probability of having a male or female child is equal, calculate the probability that a woman's fourth child is her first son.

Solution

The probability is $\left(\dfrac{1}{2}\right)^3 \times \dfrac{1}{2} = 0.0625$.

Consider the conditional probability $P(X > x + n \mid X > n)$.

Given that there have already been n trials without a success, what is the probability that more than x additional trials are required to get a success?

To answer this, we will need the conditional probability formula $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$.

The intersection of the events '$X > n$' and '$X > x + n$' is just '$X > x + n$', so:

$$P(X > x + n \mid X > n) = \frac{P(X > x + n)}{P(X > n)} = \frac{(1-p)^{x+n}}{(1-p)^n} = (1-p)^x = P(X > x)$$

ie **just the same as the original probability that more than x trials are required.**

The lack of success on the first n trials is irrelevant – under this model the chances of success are not any better because there has been a run of bad luck.

This characteristic – a reflection of the 'independent, identical trials' structure – is important, and is referred to as the 'memoryless' property.

Question

The probability of having a male or female child is equal. A woman has two boys and a girl. Calculate the probability that her next two children are girls.

Solution

Due to the memoryless property, the children she has so far are irrelevant when it comes to

working out the probability that the next two are girls. So the probability is $\left(\dfrac{1}{2}\right)^2 = 0.25$.

Another formulation of the geometric distribution is sometimes used. Let Y be the number of failures before the first success. Then $P(Y = y) = p(1-p)^y$, $y = 0, 1, 2, 3,\dots$ with mean

$$\mu = \frac{1-p}{p}.$$

$Y = X - 1$, where X is defined as above.

Question

Determine the variance for this formulation.

Solution

Since $Y = X - 1$:

$$\text{var}(Y) = \text{var}(X) = \frac{1-p}{p^2}$$

since subtracting a constant from a random variable does not change the spread of the distribution.

The R code for simulating values and calculating probabilities and quantiles from the geometric distribution is similar to the R code used for the binomial distribution using the R functions `rgeom,` `dgeom,` `pgeom` **and** `qgeom`**.**

For example:

```
dgeom(10, 0.3)
```

calculates the probability $P(Y = 10)$ for $p = 0.3$.

1.5 Negative binomial distribution

This is a generalisation of the geometric distribution.

The random variable X is the number of the trial on which the k th success occurs, where k is a positive integer.

For example, in a telesales company, X might be the number of phone calls required to make the fifth sale.

Distribution: $P(X = x) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} p^k (1-p)^{x-k}$ $x = k,\ k+1,\dots;\ 0 < p < 1$

We say that X has a Type 1 negative binomial (k, p) distribution.

The probabilities satisfy the recurrence relationship:

$$P(X = x) = \frac{x-1}{x-k}(1-p)P(X = x-1)$$

Note that in applying this model, the value of k is known.

Moments: $\qquad \mu = \dfrac{k}{p}$ $\qquad$ **and:** $\qquad\qquad \sigma^2 = \dfrac{k(1-p)}{p^2}$

Note: The mean and variance are just k times those for the geometric (p) variable, which is itself a special case of this random variable (with $k = 1$). Further, the negative binomial variable can be expressed as the sum of k geometric variables (the number of trials to the first success, plus the number of additional trials to the second success, plus … to the ($k-1$)th success, plus the number of additional trials to the k th success.)

Question

If the probability that a person will believe a rumour about a scandal in politics is 0.8, calculate the probability that the ninth person to hear the rumour will be the fourth person to believe it.

Solution

Let X be the 'position' of the fourth person who believes it. Then $p = 0.8$, $X = 9$ and $k = 4$, and we have:

$$P(X = 9) = \begin{pmatrix} 8 \\ 3 \end{pmatrix} \times 0.8^4 \times 0.2^5 = 0.00734$$

Another formulation of the negative binomial distribution is sometimes used.

Let Y be the number of failures before the k th success.

Then $P(Y = y) = \dbinom{k + y - 1}{y} p^k (1 - p)^y$, $y = 0, 1, 2, 3,...$, with mean $\mu = \dfrac{k(1 - p)}{p}$. $Y = X - k$,

where X is defined as above.

This formulation is called the Type 2 negative binomial distribution and can be found on page 9 of the *Tables*. It should be noted that in the *Tables* the combinatorial factor has been rewritten in terms of the gamma function (defined later in this chapter).

The previous formulation is known as the Type 1 negative binomial distribution. The formulae for this version are given on page 8 of the *Tables*.

The R code for simulating values and calculating probabilities and quantiles from the negative binomial distribution is similar to the R code used for the binomial distribution using the R functions `rnbinom, dnbinom, pnbinom` **and** `qnbinom`**.**

For example:

```
dnbinom(15, 10, 0.3)
```

calculates the probability $P(Y = 15) = 0.0366544$ for $p = 0.3$ and $k = 10$.

1.6   Hypergeometric distribution

This is the 'finite population' equivalent of the binomial distribution, in the following sense. Suppose objects are selected at random, one after another, without replacement, from a finite population consisting of k 'successes' and $N - k$ 'failures'. The trials are not independent, since the result of one trial (the selection of a success or a failure) affects the make-up of the population from which the next selection is made.

Random variable X : is the number of 'successes' in a sample of size n from a population of size N that has k 'successes' and $N - k$ 'failures' .

$$P(X = x) = \frac{\dbinom{k}{x}\dbinom{N - k}{n - x}}{\dbinom{N}{n}} , \; x = 1, 2, 3,... ; \; 0 < p < 1.$$

Moments: $\mu = \dfrac{nk}{N}$

$$\sigma^2 = \frac{nk(N - k)(N - n)}{N^2(N - 1)}$$

(The details of the derivation of the mean and variance of the number of successes are not required by the syllabus).

Note that the mean is given by $\mu = \dfrac{nk}{N}$, which parallels the '$\mu = np$' result for the binomial

distribution – the initial proportion of successes here being $\dfrac{k}{N}$.

In the above context, the binomial is the model appropriate to selecting *with* replacement, which is equivalent to selecting from an infinite population $N \to \infty$ for which:

$$P(\text{success}) = p = \frac{k}{N}$$

is kept fixed. Hence, the binomial, with $p = k/N$, provides a good approximation to the hypergeometric when N is large compared to n.

The hypergeometric distribution is used in the grouping of signs test in Subject CS2.

R

The R code for simulating values and calculating probabilities and quantiles from the hypergeometric distribution is similar to the R code used for other distributions using the R functions `rhyper, dhyper, phyper` and `qhyper`.

For example:

```
rhyper(20, 15, 10, 5)
```

simulates 20 values from samples of size 5 from a population in which *k* = 15 and *N – k* = 10.

Question

Among the 58 people applying for a job, only 30 have a particular qualification. If 5 of the group are randomly selected for a survey about the job application procedure, determine the probability that none of the group selected have the qualification.

Calculate the answer:

(i)   exactly

(ii)   using the binomial approximation.

Solution

(i)   Let X denote the number of applicants from the group of 5 that have the qualification. Using the probability function of the hypergeometric distribution with $N = 58$, $k = 30$, and $n = 5$:

$$P(X = 0) = \frac{\dbinom{30}{0}\dbinom{28}{5}}{\dbinom{58}{5}} = 0.0214$$

Alternatively we could consider in turn the probabilities that each candidate is unqualified, and multiply the probabilities together:

$$\frac{28}{58} \times \frac{27}{57} \times \cdots \times \frac{24}{54} = 0.0214$$

(ii)   Using the hypergeometric distribution, $N = 58$, and $k = 30$, so we will use a binomial approximation with $n = 5$, and $p = \frac{30}{58}$:

$$P(X = 0) \approx \binom{5}{0} p^5 q^0 = \left(\frac{28}{58}\right)^5 = 0.0262$$

1.7   Poisson distribution

This distribution models the number of events that occur in a specified interval of time, when the events occur one after another in time in a well-defined manner. This manner presumes that the events occur singly, at a constant rate, and that the numbers of events that occur in separate (*ie* non-overlapping) time intervals are independent of one another. These conditions can be described loosely by saying that the events occur 'randomly, at a rate of .. per ..', and such events are said to occur according to a Poisson process. We will formally define this later in this chapter.

Another approach to the Poisson distribution uses arguments which appear at first sight to be unrelated to the above. Consider a sequence of binomial (n, p) distributions as $n \to \infty$ and $p \to 0$ together, such that the mean np is held constant at the value λ. The limit leads to the distribution of the Poisson variable, with parameter λ.

Here $\lambda = np$.

Distribution: $P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, 2, 3, \ldots$; $\lambda > 0$

The probabilities satisfy the recurrence relationship:

$$P(X = x) = \frac{\lambda}{x} P(X = x - 1)$$

If X has a Poisson distribution with parameter λ, then we can write $X \sim Poi(\lambda)$.

Moments:

Since the binomial mean is held constant at λ through the limiting process, it is reasonable to suggest that the distribution of X (the limiting distribution) also has mean λ. This is in fact the case.

The binomial variance is:

$$np(1 - p) = n\left(\frac{\lambda}{n}\right)\left(1 - \frac{\lambda}{n}\right) = \lambda\left(1 - \frac{\lambda}{n}\right) \to \lambda \text{ as } n \to \infty$$

This suggests that X has variance λ. This is in fact also the case. So $\mu = \sigma^2 = \lambda$.

Question

Using the probability function for the Poisson distribution, prove the formulae for the mean and variance. Hint: for the variance, consider $E[X(X-1)]$.

Solution

The mean is:

$$E(X) = \sum_x x P(X=x) = \lambda e^{-\lambda} + 2\frac{\lambda^2}{2!}e^{-\lambda} + 3\frac{\lambda^3}{3!}e^{-\lambda} + 4\frac{\lambda^4}{4!}e^{-\lambda} + \cdots$$

$$= \lambda e^{-\lambda} + \lambda^2 e^{-\lambda} + \frac{\lambda^3}{2!}e^{-\lambda} + \frac{\lambda^4}{3!}e^{-\lambda} + \cdots$$

$$= \lambda e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right)$$

Since $e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots$, we obtain:

$$E(X) = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

For the variance we need to work out $E(X^2)$. However, the easiest way to work out the variance is actually to consider $E[X(X-1)]$:

$$E[X(X-1)] = \sum_x x(x-1)P(X=x) = 2\times1\frac{\lambda^2}{2}e^{-\lambda} + 3\times2\frac{\lambda^3}{3!}e^{-\lambda} + 4\times3\frac{\lambda^4}{4!}e^{-\lambda} + \cdots$$

$$= \lambda^2 e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \cdots\right)$$

$$= \lambda^2 e^{-\lambda} e^{\lambda}$$

$$= \lambda^2$$

$$E[X(X-1)] = E(X^2) - E(X) = \lambda^2 \quad \Rightarrow \quad E(X^2) = \lambda^2 + E(X) = \lambda^2 + \lambda$$

$$\text{var}(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

We can calculate Poisson probabilities in the usual way, using the probability function or the cumulative probabilities given in the *Tables*.

Question

If goals are scored randomly in a game of football at a constant rate of three per match, calculate the probability that more than 5 goals are scored in a match.

Solution

The number of goals in a match can be modelled as a Poisson distribution with mean $\lambda = 3$.

$$P(X > 5) = 1 - P(X \leq 5)$$

We can use the recurrence relationship given:

$$P(X = 0) = e^{-3} = 0.0498$$

$$P(X = 1) = \frac{3}{1} \times 0.0498 = 0.1494$$

$$P(X = 2) = \frac{3}{2} \times 0.1494 = 0.2240$$

$$P(X = 3) = \frac{3}{3} \times 0.2240 = 0.2240$$

$$P(X = 4) = \frac{3}{4} \times 0.2240 = 0.1680$$

$$P(X = 5) = \frac{3}{5} \times 0.1680 = 0.1008$$

So we have $P(X > 5) = 1 - 0.9161 = 0.0839$.

Alternatively, we could obtain this directly using the cumulative Poisson probabilities given on page 176 of the Tables.

The Poisson distribution provides a very good approximation to the binomial when *n* is large and *p* is small – typical applications have $n = 100$ or more and $p = 0.05$ or less. The approximation depends only on the product np ($= \lambda$) – the individual values of *n* and *p* are irrelevant. So, for example, the value of $P(X = x)$ in the case $n = 200$ and $p = 0.02$ is effectively the same as the value of $P(X = x)$ in the case $n = 400$ and $p = 0.01$. When dealing with large numbers of opportunities for the occurrence of 'rare' events (under 'binomial assumptions'), the distribution of the number that occurs depends only on the expected number.

We will look at other approximations in Chapter 5.

Question

If each of the 55 million people in the UK independently has probability 1×10^{-8} of being killed by a falling meteorite in a given year, use an approximation to calculate the probability of exactly 2 such deaths occurring in a given year.

Solution

If X is the number of people killed by a meteorite in a year then X has a binomial distribution with $n = 55,000,000$ and $p = 1 \times 10^{-8}$. We can approximate this by using a Poisson distribution with:

$$\lambda = np = 55,000,000 \times 1 \times 10^{-8} = 0.55$$

Hence:

$$P(X = 2) = \frac{0.55^2}{2!} e^{-0.55} = 0.0873$$

The Poisson distribution is often used to model the number of claims that an insurance company receives per unit of time. It is also used to model the number of accidents along a particular stretch of road.

When events are described as occurring 'as a Poisson process with rate λ' or 'randomly, at a rate of λ per unit time' then the number of events that occur in a time period of length *t* has a Poisson distribution with mean λt.

The Poisson process is discussed in more detail in Section 3.

The R code for simulating values and calculating probabilities and quantiles from the Poisson distribution is similar to the R code used for other distributions using the R functions `rpois, dpois, ppois` **and** `qpois`**.**

For example, to calculate $P(X \le 5) = 0.9432683$ for $\lambda = 2.7$ use the R code:

```
ppois(5, 2.7)
```

Question

The number of home insurance claims a company receives in a month is distributed as a Poisson random variable with mean 2. Calculate the probability that the company receives exactly 30 claims in a year. Treat all months as if they are of equal length.

Solution

Let X denote the number of home insurance claims received in a year. Since the number of claims in a month has a $Poi(2)$ distribution; $X \sim Poi(24)$. The required probability is:

$$P(X = 30) = \frac{24^{30}}{30!} e^{-24} = 0.0363$$

Alternatively, we could use the cumulative Poisson probabilities given on page 184 of the Tables:

$$P(X = 30) = P(X \le 30) - P(X \le 29) = 0.90415 - 0.86788 = 0.0363$$

2   Important continuous distributions

2.1   Uniform distribution

X takes values between two specified numbers α and β say.

Probability density function: $f_X(x) = \dfrac{1}{\beta - \alpha}$ $\alpha < x < \beta$

$X \sim U(\alpha, \beta)$ is often written as shorthand for 'the random variable X has a continuous uniform distribution over the interval (α, β).

Moments: $\mu = \dfrac{\alpha + \beta}{2}$, by symmetry, the mid-point of the range of possible values

$$\sigma^2 = \frac{(\beta - \alpha)^2}{12}$$

Question

Prove the variance result, by considering $E[(X - \mu)^2]$ directly.

Solution

The variance is:

$$\text{var}[X] = E[(X - \mu)^2] = \int_x (x - \mu)^2 f(x)\,dx = \int_\alpha^\beta \left(x - \tfrac{1}{2}(\alpha + \beta)\right)^2 \frac{1}{\beta - \alpha}\,dx$$

$$= \left[\frac{\left(x - \tfrac{1}{2}(\alpha + \beta)\right)^3}{3(\beta - \alpha)} \right]_\alpha^\beta$$

$$= \frac{\left(\beta - \tfrac{1}{2}(\alpha + \beta)\right)^3}{3(\beta - \alpha)} - \frac{\left(\alpha - \tfrac{1}{2}(\alpha + \beta)\right)^3}{3(\beta - \alpha)}$$

$$= \frac{\left(\tfrac{1}{2}(\beta - \alpha)\right)^3}{3(\beta - \alpha)} - \frac{\left(-\tfrac{1}{2}(\beta - \alpha)\right)^3}{3(\beta - \alpha)}$$

$$= \frac{1}{24}(\beta - \alpha)^2 + \frac{1}{24}(\beta - \alpha)^2 = \frac{1}{12}(\beta - \alpha)^2$$

In this model, the total probability of 1 is spread 'evenly' between the two limits, so that subintervals of the same length have the same probability.

Question

If $Y \sim U(50,150)$, calculate $P(Y > 74)$ and $P(50 < Y < 126)$.

Solution

The PDF is given by $f(y) = \dfrac{1}{150 - 50} = \dfrac{1}{100}$ for $50 < y < 150$. This gives:

$$P(Y > 74) = \frac{76}{100} = 0.76$$

Similarly, $P(50 < Y < 126) = 0.76$. This probability is the same since the two subintervals have the same length.

The R code for simulating 100 values from a $U(0,3)$ distribution is given by:

```
runif(100, min=0, max=3)
```

The PDF is obtained by `dunif(x, min=0, max=3)` **and is useful for graphing.**

To calculate probabilities for a continuous distribution we use the CDF which is obtained by `punif`. **For example, to calculate $P(X \leq 1.8) = 0.6$ for $U(0,3)$ use the R code:**

```
punif(1.8, min=0, max=3)
```

Similarly, the quantiles can be calculated with `qunif`.

Although there are not many real-life examples of the continuous uniform distribution, it is nevertheless an important distribution. A sample of random numbers from $U(0,1)$ is often used to generate random samples from other distributions. We will do this in Section 4.

2.2 Gamma (including exponential and chi-square) distributions

The gamma family of distributions has 2 positive parameters and is a versatile family. The PDF can take different shapes depending on the values of the parameters. The range of the variable is $\{x: x > 0\}$.

The parameter α changes the shape of the graph of the PDF, and the parameter λ changes the x-scale. The gamma distribution may be written in shorthand as $Gamma(\alpha, \lambda)$, or $Ga(\alpha, \lambda)$.

First note that the gamma function $\Gamma(\alpha)$ is defined for $\alpha > 0$ as follows:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

Note in particular that $\Gamma(1) = 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for $\alpha > 1$ (*ie* if α is an integer $\Gamma(\alpha) = (\alpha - 1)!$), and $\Gamma\left(\dfrac{1}{2}\right) = \sqrt{\pi}$.

These results are given on page 5 of the *Tables* and are all that is required to answer examination questions.

The R code for the gamma function $\Gamma(n)$ is `gamma(n)`.

The PDF of the gamma distribution with parameters α and λ is defined by:

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad \text{for } x > 0$$

Moments: $\quad \mu = \dfrac{\alpha}{\lambda}, \ \sigma^2 = \dfrac{\alpha}{\lambda^2}$

Question

Prove the formulae given for the mean and variance.

Solution

Remembering that the formulae for mean and variance are $E(X) = \int\limits_x x f(x)\, dx$, and

$\text{var}(X) = \int\limits_x x^2 f(x)\, dx - [E(X)]^2$, using appropriate limits, we have:

$$E(X) = \int\limits_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x}\, dx$$

Using integration by parts with $u = \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha$ and $\dfrac{dv}{dx} = e^{-\lambda x}$, we obtain:

$$E(X) = \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha \times -\frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty - \int\limits_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} \alpha x^{\alpha-1} \times -\frac{1}{\lambda} e^{-\lambda x}\, dx$$

$$= \frac{\alpha}{\lambda} \int\limits_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx$$

The integral is the integral of the PDF over the whole range which is 1, giving:

$$E(X) = \frac{\alpha}{\lambda}$$

For the variance we need $E(X^2)$:

$$E(X^2) = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha+1} e^{-\lambda x} \, dx$$

Using integration by parts with $u = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha+1}$, we obtain:

$$E(X^2) = \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha+1} \times -\frac{1}{\lambda} e^{-\lambda x} \right]_0^\infty - \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} (\alpha+1) x^\alpha \times -\frac{1}{\lambda} e^{-\lambda x} \, dx$$

$$= \frac{\alpha+1}{\lambda} \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x} \, dx$$

The integral is $E(X)$, so we have $E(X^2) = \frac{\alpha+1}{\lambda} \times \frac{\alpha}{\lambda}$, hence:

$$\text{var}(X) = \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda} \right)^2 = \frac{\alpha}{\lambda^2}$$

We shall see in a later chapter that these results can be proved far more easily using moment generating functions.

We can calculate gamma probabilities in simple cases by integrating the PDF.

Question

If $X \sim Gamma(2, 1.5)$, calculate $P(X > 4)$.

Solution

Using integration by parts:

$$P(X > 4) = \int_4^\infty \frac{1.5^2}{\Gamma(2)} x e^{-1.5x} dx$$

$$= 2.25 \left\{ \left[-\frac{x}{1.5} e^{-1.5x} \right]_4^\infty + \int_4^\infty \frac{1}{1.5} e^{-1.5x} dx \right\}$$

$$= 2.25 \left\{ \frac{4}{1.5} e^{-6} + \left[-\frac{1}{1.5^2} e^{-1.5x} \right]_4^\infty \right\} = 2.25 \left\{ \frac{4}{1.5} e^{-6} + \frac{1}{1.5^2} e^{-6} \right\} = 0.0174$$

We will see a quicker way to do this question later in the chapter.

The R code for simulating a random sample of 100 values from the gamma distribution with $\alpha = 2$ and $\lambda = 0.25$ is:

```
rgamma(100, 2, 0.25)
```

Similarly, the PDF, cumulative distribution function (CDF) and quantiles can be obtained using the R functions `dgamma`**,** `pgamma` **and** `qgamma`**.**

Special case 1: exponential distribution

Gamma with $\alpha = 1$.

PDF: $f_X(x) = \lambda e^{-\lambda x}$, $x > 0$

$X \sim Exp(\lambda)$ is often written as shorthand for 'the random variable X has an exponential distribution with parameter λ'.

Moments: $\mu = \dfrac{1}{\lambda}$, $\sigma^2 = \dfrac{1}{\lambda^2}$

$$F_X(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}$$

For many of the continuous distributions, the CDF is given in the *Tables*.

Question

Determine the median of the $Exp(\lambda)$ distribution. (The median is the value of *m* such that $P(X \le m) = \frac{1}{2}$.)

Solution

Since $P(X \le m) = F_X(m)$, we have:

$$1 - e^{-\lambda m} = 0.5 \;\Rightarrow\; 0.5 = e^{-\lambda m} \;\Rightarrow\; -\lambda m = \ln 0.5 \;\Rightarrow\; m = -\frac{1}{\lambda}\ln 0.5$$

Since $\ln 0.5 = -\ln 2$, we can say $m = \dfrac{\ln 2}{\lambda}$.

The exponential distribution is used as a simple model for the lifetimes of certain types of equipment. Very importantly, it also gives the distribution of the waiting time, *T*, from one event to the next in a Poisson process with rate λ. This is proved in Section 3 of this chapter.

The R code for simulating values and obtaining the PDF, CDF and quantiles from the exponential distribution is similar to the R code used for other continuous distributions using the R functions `rexp`, `dexp`, `pexp` and `qexp`.

Question

Claims to a general insurance company's 24 hour call centre occur according to a Poisson process with a rate of 3 per hour. Calculate the probability that the next call arrives after more than ½ hour.

Solution

The number of claims, X, in an hour can be modelled as a Poisson distribution with mean $\lambda = 3$. Hence, the waiting time, T, between claims can be modelled as an exponential distribution with $\lambda = 3$. So:

$$P(T > \tfrac{1}{2}) = \int_{\frac{1}{2}}^{\infty} 3e^{-3x}\, dx = [-e^{-3x}]_{\frac{1}{2}}^{\infty} = 0 - (-e^{-1\frac{1}{2}}) = 0.2231$$

In fact the time from any specified starting point (not necessarily the time at which the last event occurred) to the next event occurring has this exponential distribution. This property can also be expressed as the memoryless property.

Recall that the geometric distribution in Section 1.4 had the memoryless property. For the exponential distribution we can also show that:

$$P(X > x + n \mid X > n) = P(X > x)$$

For example, the probability that we wait at least a further 10 minutes given that we have already waited 20 minutes is equal to the unconditional probability of waiting at least 10 minutes.

Question

Prove that if $X \sim Exp(\lambda)$ then $P(X > x + n \mid X > n) = P(X > x)$.

Solution

$$P(X > x + n \mid X > n) = \frac{P(X > x + n, X > n)}{P(X > n)}$$

$$= \frac{P(X > x + n)}{P(X > n)}$$

$$= \frac{e^{-\lambda(x+n)}}{e^{-\lambda n}} = e^{-\lambda x} = P(X > x)$$

Note: A gamma variable with parameters $\alpha = k$ (a positive integer) and λ can be expressed as the sum of k exponential variables, each with parameter λ. This gamma distribution is in fact the model for the time from any specified starting point to the occurrence of the kth event in a Poisson process with rate λ.

The fact that a *Gamma*(α, λ) random variable can be thought of as the sum of α independent and identical *Exp*(λ) random variables is important and will be used in a later chapter to prove some important results.

Special case 2: chi-square (χ^2) distribution with parameter 'degrees of freedom' υ

Gamma with $\alpha = \dfrac{\upsilon}{2}$ where υ is a positive integer, and $\lambda = \dfrac{1}{2}$.

So the PDF of a χ^2 distribution is:

$$f_X(x) = \frac{(\frac{1}{2})^{\frac{1}{2}\upsilon}}{\Gamma(\frac{1}{2}\upsilon)} x^{\frac{1}{2}\upsilon - 1} e^{-\frac{1}{2}x} \quad \text{for } x > 0.$$

Moments: $\mu = \upsilon, \ \sigma^2 = 2\upsilon$

Note: A χ^2 variable with $\upsilon = 2$ is the same as an exponential variable with mean 2.

Since integrating the PDF isn't straightforward, extensive probability tables for the chi-square distribution are given in the *Tables*. These can be found on pages 164-166.

Another result that we will use in later work is:

If $W \sim Gamma(\alpha, \lambda)$, then $2\lambda W$ has a $\chi^2_{2\alpha}$ distribution (*ie* a chi-square distribution with 2α degrees of freedom). This result is also in the *Tables* (on page 12).

We can prove this result using moment generating functions which we will meet in a later chapter.

This is an important result as it is the only practical way we can calculate probabilities for a gamma distribution in an exam. We can look up probabilities associated with the χ^2 distribution, for certain degrees of freedom, in the *Tables*.

> **The R code for simulating values and obtaining the PDF, CDF and quantiles from the chi-square distribution is similar to the R code used for other continuous distributions using the R functions `rchisq`, `dchisq`, `pchisq` and `qchisq`.**

Question

If the random variable X has a χ_5^2 distribution, calculate:

(a) $P(X > 6.5)$

(b) $P(X < 11.8)$.

Solution

Using the χ^2 probabilities given on pages 164–166 of the *Tables*, we obtain:

(a) $1 - 0.7394 = 0.2606$

(b)   Here we need to interpolate between the two closest probabilities given,
           ie $P(X < 11.5) = 0.9577$ and $P(X < 12) = 0.9652$, so:

$$P(X < 11.8) \approx 0.9577 + \frac{11.8 - 11.5}{12 - 11.5} \times (0.9652 - 0.9577) = 0.9622$$

Alternatively, we could use interpolation on the χ^2 percentage points tables given on page 168-169 of the Tables. These give the approximate answers of 0.2644 and 0.9604.

We now repeat an earlier question using the χ^2 result.

Question

If $X \sim Gamma(2, 1.5)$, calculate $P(X > 4)$, by using the chi square tables.

Solution

Since $X \sim Gamma(2, 1.5)$, we know that $3X \sim \chi_4^2$. So:

$$P(X > 4) = P(3X > 12) = P(\chi_4^2 > 12) = 1 - 0.9826 = 0.0174$$

using the χ^2 probability tables given on page 165 of the *Tables*.

This gives us the same answer as we obtained earlier, but without the integration by parts.

2.3   Beta distribution

This is another versatile family of distributions with two positive parameters. The range of the variable is $\{x :\ 0 < x < 1\}$.

First note that the beta function $\mathrm{B}(\alpha, \beta)$ is defined by:

$$\mathrm{B}(\alpha, \beta) = \int_{0}^{1} x^{\alpha-1}(1-x)^{\beta-1}dx$$

The relationship between beta functions and gamma functions is:

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

The R code for the beta function $\mathrm{B}(a,b)$ is `beta(a,b)`.

The PDF of a beta distribution is defined by:

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0 < x < 1$$

Moments:

$$\mu = \frac{\alpha}{\alpha + \beta}, \ \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

The (continuous) **uniform distribution on (0,1) is a special case (with $\alpha = \beta = 1$).**

The beta distribution is a useful distribution because it can be rescaled and shifted to create a wide range of shapes − from straight lines to curves, and from symmetrical distributions to skewed distributions. Since the random variable can only take values between 0 and 1, it is often used to model proportions, such as the proportion of a batch that is defective or the percentage of claims that are over £1,000.

Question

The random variable X has PDF $f_X(x) = kx^3(1-x)^2$, $0 < x < 1$, where k is a constant. Determine the value of k.

Solution

Comparing the PDF directly with that of the beta distribution, we can see that $\alpha = 4$ and $\beta = 3$. So:

$$k = \frac{\Gamma(7)}{\Gamma(4)\Gamma(3)} = 60$$

k can also be found directly from $\int_0^1 kx^3(1-x)^2 dx = 1$ by multiplying out the bracket first and then integrating.

The R code for simulating values and obtaining the PDF, CDF and quantiles from the beta distribution is similar to the R code used for other continuous distributions using the R functions `rbeta, dbeta, pbeta` **and** `qbeta`**.**

2.4 Normal distribution

This distribution, with its symmetrical 'bell-shaped' density curve is of fundamental importance in both statistical theory and practice. Its roles include the following:

(i)   **it is a good model for the distribution of measurements that occur in practice in a wide variety of different situations,** for example heights, weights, IQ scores or exam scores.

(ii)   **it provides good approximations to various other distributions – in particular it is a limiting form of the binomial (n, p).**

It is also used to approximate the Poisson distribution. Both of these approximations are covered in Chapter 5.

(iii)   **it provides a model for the sampling distributions of various statistics – see Chapter 6.**

(iv)   **much of large sample statistical inference is based on it, and some procedures require an assumption that a variable is normally distributed.**

We will look at this in Chapters 8 and 9.

(v)   **it is a 'building block' for many other distributions.**

The distribution has two parameters, which can conveniently be expressed directly as the mean μ and the standard deviation σ of the distribution. The distribution is symmetrical about μ.

The notation used for the Normal distribution is $X \sim N(\mu, \sigma^2)$.

The PDF of the normal distribution is defined by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty$$

A linear function of a normal variable is also a normal variable, *ie* if X is normally distributed, so is $Y = aX + b$.

This result can be proved using moment generating functions which we will meet in the next chapter.

It is not possible to find an explicit expression for $F_X(x) = P(X \le x)$, so tables have to be used. These are provided for the distribution of $Z = \dfrac{X - \mu}{\sigma}$, which is the standard normal variable – it has mean 0 and standard deviation 1. The distribution is symmetrical about 0.

We can also prove this result using moment generating functions.

The *x*-values μ, $\mu + \sigma$, $\mu + 2\sigma$, $\mu + 3\sigma$ correspond to the *z*-values 0, 1, 2, 3 respectively, and so on. The *z*-value measures how many standard deviations the corresponding *x* value is above or below the mean. For example the value $x = 30$ from a normal distribution with mean 20 and standard deviation 5 has *z*-value $+2$ (30 is 2 standard deviations above the mean of 20).

The calculation of a probability for a normal variable is always done in the same way – transform to the standard normal via $z = \dfrac{x - \mu}{\sigma}$ and look up in the tables.

Standard normal probabilities are given on pages 160-161 of the *Tables*.

The probabilities in the table are 'left hand' probabilities, in other words they give $P(Z < z)$, the cumulative distribution function of Z. We sometimes use $\Phi(z)$ for the CDF of Z.

Since Z is symmetrical about zero, it follows that:

$$P(Z < -z) = P(Z > z) = 1 - P(Z < z)$$

$$P(Z > -z) = P(Z < z)$$

Question

If $X \sim N(25, 36)$, calculate:

(i) $P(X < 28)$

(ii) $P(X > 30)$

(iii) $P(X < 20)$

(iv) $P(|X - 25| < 4)$.

Solution

(i) $P(X < 28) = P\left(Z < \dfrac{28 - 25}{\sqrt{36}} \right) = P(Z < 0.5) = 0.69146$

The following answers use interpolation between tabulated values:

(ii) $P(X > 30) = P(Z > 0.833) = 1 - P(Z < 0.833) = 1 - 0.7976 = 0.2024$

(iii) $P(X < 20) = P(Z < -0.833) = 1 - P(Z < 0.833) = 1 - 0.7976 = 0.2024$

(iv)   We need to simplify the expression involving the absolute value:

$$P(|X-25|<4)=P(-4<X-25<4)$$

$$=P(21<X<29)$$

$$=P(X<29)-P(X<21)$$

$$=P(Z<0.667)-P(Z<-0.667)$$

$$=P(Z<0.667)-\left[1-P(Z<0.667)\right]$$

$$=0.4952$$

The normal distribution is used in many areas of statistics and often we need to find values of the standard normal distribution connected to certain probabilities, for example the value of a such that $P(-a<Z<a)=0.99$. Common examples of this type of calculation are now given.

95% and 99% intervals:

$P(Z<1.96)=0.97500$ so $P(0<Z<1.96)=0.97500-0.5=0.47500$

$\therefore\ P(-1.96<Z<1.96)=2\times0.47500=0.95$

Similarly $\therefore\ P(-2.5758<Z<2.5758)=0.99$. So (approximately):

95% of a normal distribution is contained in the interval '1.96 standard deviations on either side of the mean', and 99% is contained in the interval '2.5758 standard deviations on either side of the mean'.

Note: All but 0.3% of the distribution is contained in the interval $(\mu-3\sigma,\mu+3\sigma)$ – the so-called '3σ limits'. (The range of a large set of observations from a normal distribution is usually about 6 or 7 standard deviations).

Finally, we note that, if X has the standard normal distribution, then X^2 has the chi-squared distribution (the special case of the gamma distribution given above).

In fact X^2 is χ_1^2 here. This result can be used to find $E(Z^2)$ and $\text{var}(Z^2)$.

The R code for simulating values and obtaining the PDF, CDF and quantiles from the normal distribution is similar to the R code used for other continuous distributions using the R functions `rnorm`, `dnorm`, `pnorm` and `qnorm`.

2.5   Lognormal distribution

If X represents, for example, claim size and $Y=\log X$ has a normal distribution, then X is said to have a *lognormal* distribution.

$\log X$ here refers to natural log, or log to base e, ie $\ln X$.

If X has a lognormal distribution with parameters μ and σ, then we write $X\sim\log N(\mu,\sigma^2)$.

The PDF of the lognormal distribution is defined by:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2} \quad \text{for } 0 < x < \infty$$

The lower limit for x is 0 and not $-\infty$, as it is for the normal distribution. This is because $\log x$ is not defined for $x \leq 0$.

Question

If $W \sim \log N(5,6)$, calculate $P(W > 3,000)$.

Solution

If $W \sim \log N(5,6)$, then $\ln W \sim N(5,6)$. This gives:

$$P(W > 3,000) = P(\ln W > 8.006) = P(Z > 1.227) = 1 - 0.8901 = 0.1099$$

The mean and variance of the lognormal distribution are *not* μ and σ^2 but are given by

$E[X] = e^{\mu + \frac{1}{2}\sigma^2}$, and $\text{var}[X] = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$.

Question

If the mean of the lognormal distribution is 9.97 and the variance is 635.61, calculate the parameters μ and σ^2.

Solution

$e^{\mu + \frac{1}{2}\sigma^2} = 9.97$ and $e^{2\mu + \sigma^2}(e^{\sigma^2} - 1) = 635.61$, so $9.97^2(e^{\sigma^2} - 1) = 635.61$.

This can be rearranged to give $\sigma^2 = 2$.

Substituting into the equation for the mean, we get $e^{\mu + 1} = 9.97$. Taking logs gives us $\mu = 1.3$.

The lognormal distribution is positively skewed and is therefore a good model for the distribution of claim sizes. We also use the lognormal distribution in Subject CM2 to calculate the probabilities associated with accumulating funds.

The R code for simulating values and obtaining the PDF, CDF and quantiles from the lognormal distribution is similar to the R code used for other continuous distributions using the R functions `rlnorm, dlnorm, plnorm` and `qlnorm`.

2.6 *t* distribution

If the variable **X** has a χ^2_ν distribution and another independent variable **Z** has the standard normal distribution of the form **N(0,1)** then the function:

$$\frac{Z}{\sqrt{X/\nu}}$$

is said to have a *t* distribution with parameter 'degrees of freedom' ν.

The *t* distribution, like the normal, is symmetrical about 0.

You do not need to know the PDF of the *t* distribution for the exam. It is in fact given on page 163 of the *Tables*.

Calculating probabilities by integrating this PDF is not easy. Fortunately, we will only be expected to look up probabilities using page 163 of the *Tables*.

Question

Use the *t* tables to calculate:

(i) $P(t_{15} < 1.341)$

(ii)   the value of *a* such that $P(t_8 > a) = 0.01$

(iii) $P(t_{24} < -0.5314)$.

Solution

From the *Tables*:

(i) $P(t_{15} > 1.341) = 10\%$, so $P(t_{15} < 1.341) = 90\%$.

(ii) $a = 2.896$

(iii)   By symmetry:

$$P(t_{24} < -0.5314) = P(t_{24} > 0.5314) = 30\%$$

The **t** distribution is used to find confidence intervals and carry out hypothesis tests on the mean of a distribution. We will meet it again in Chapters 6, 8 and 9.

The R code for simulating values and obtaining the PDF, CDF and quantiles from the *t* distribution is similar to the R code used for other continuous distributions using the R functions `rt, dt, pt` **and** `qt`.

2.7 *F* distribution

If two independent random variables, *X* and *Y* have χ^2 distributions with parameters n_1 and n_2 respectively, then the function:

$$\frac{X / n_1}{Y / n_2}$$

is said to have an *F* distribution with parameters (degrees of freedom) n_1 and n_2.

Once again, it is not necessary to know the PDF of this distribution. We find probabilities by using the *F* tables given on pages 170-174 of the *Tables*.

The *F* distribution is *not* symmetrical. Given that only upper tail probabilities are given in the *Tables*, we will need to know the fact that $P(F_{a,b} > c) = P\left(\frac{1}{F_{a,b}} < \frac{1}{c}\right) = P\left(F_{b,a} < \frac{1}{c}\right)$ to find lower tail probabilities. This will be covered in greater detail in Chapter 6.

Question

Use the *F* tables to calculate:

(i) $P(F_{5,12} < 3.106)$

(ii)   the value of *a* such that $P(F_{7,4} > a) = 0.01$.

Solution

From the *Tables*:

(i) $P(F_{5,12} > 3.106) = 5\%$, so $P(F_{5,12} < 3.106) = 95\%$

(ii) $a = 14.98$.

This distribution is used to find confidence intervals and carry out hypothesis tests on the variances of two distributions. We will meet it again in Chapters 6, 8, 9, and 11.

> **The R code for simulating values and obtaining the PDF, CDF and quantiles from the *F* distribution is similar to the R code used for other continuous distributions using the R functions** `rf, df, pf` **and** `qf`**.**

3   The Poisson process

Earlier, in Section 1.7, we met the Poisson distribution, $X \sim Poi(\lambda)$ with probability function (PF):

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, 2, \ldots$$

This is useful for modelling the number of events (*eg* claims or deaths) occurring per unit time. For $X \sim Poi(\lambda)$, we have events occurring at a rate of λ per unit time.

A Poisson *process* occurs when we let the time period vary. So instead of looking at the number of events occurring per unit time, we now look at the number of events occurring up to time *t*.

Question

An insurer receives car claims at a rate of 8 per calendar week. Write down the distribution of the number of claims received:

(i)   per day

(ii)   per year.

Solution

The number of car claims per week has a *Poi*(8) distribution, therefore the number of:

(i)   car claims per day has a $Poi\left(\frac{8}{7}\right)$ distribution

(ii)   car claims per year has a *Poi*(416) distribution (using 52 weeks in a year).

From the previous question it should be clear to see that if we have $X \sim Poi(\lambda)$ modelling the number of claims per unit time, then $X(t) \sim Poi(\lambda t)$ will model the number of claims up to time *t*.

$$P(X(t) = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \qquad x = 0, 1, 2, \ldots$$

Question

The number of deaths amongst retired members of a pension scheme occurs at a rate of 3 per calendar month. Calculate the probability of:

(i)   5 deaths in January to March inclusive

(ii)   12 deaths in June to October inclusive.

Solution

Let $X(t)$ be the number of deaths in a t month period. The number of deaths per calendar month has a $Poi(3)$ distribution, therefore:

(i)   the number of deaths in January to March inclusive has a $Poi(9)$ distribution:

$$P(X(3)=5)=\frac{9^5}{5!}e^{-9}=0.0607$$

(ii)   the number of deaths in June to October inclusive has a $Poi(15)$ distribution:

$$P(X(5)=12)=\frac{15^{12}}{12!}e^{-15}=0.0829$$

3.1   Deriving Poisson process formulae

In Section 1.7, we stated that the Poisson distribution could be used to model events occurring randomly one after another in time at a constant rate and that the numbers of events that occur in separate (*ie* non-overlapping) time intervals are independent of one another. However, we derived the distribution from the binomial distribution.

In this section, we shall look at the Poisson process, $N(t)$, by considering events occurring in a small interval of time. To start with, we shall define mathematically the properties of a counting process and a Poisson process.

The Poisson process is an example of a counting process. Here the number of events occurring is of interest. Since the number of events is being counted over time, the event number process $\{N(t)\}_{t\geq 0}$ must satisfy the following conditions.

(i) $N(0)=0$, *ie* no events have occurred at time 0.

 (ii)   for any $t>0$, $N(t)$ must be integer valued

         ie we can't have 2.3 claims.

(iii)   When $s<t$, $N(s)\leq N(t)$, *ie* the number of events over time is non-decreasing.

         ie if we have counted, say, 5 deaths in 2 months, then the number of deaths counted in a 3 month period which includes the 2 month period *must* be *at least* 5.

(iv)   When $s<t$, $N(t)-N(s)$ represents the number of events occurring in the time interval (s,t).

         ie we have counted $N(t)$ events up to time t and $N(s)$ events up to time s, so there were $N(t)-N(s)$ events counted between time s and time t.

These are the mathematical properties of *any* counting process; we will now define the mathematical properties for a Poisson process.

The event number process $\{N(t)\}_{t \geq 0}$ is defined to be a Poisson process with parameter λ if the following three conditions are satisfied:

(i) $N(0) = 0$, and $N(s) \leq N(t)$ when $s < t$

These are just properties (i) and (iii) from above for any counting process.

(ii) $P(N(t+h) = r \mid N(t) = r) = 1 - \lambda h + o(h)$

$P(N(t+h) = r + 1 \mid N(t) = r) = \lambda h + o(h)$ (1.1)

$P(N(t+h) > r + 1 \mid N(t) = r) = o(h)$

(Note that a function $f(h)$ is described as $o(h)$ if $\lim\limits_{h \to 0} \dfrac{f(h)}{h} = 0$.)

(iii)   when $s < t$, the number of events in the time interval $(s, t]$ is independent of the number of events up to time s.

ie the numbers of events that occur in separate (*ie* non-overlapping) time intervals are independent of one another.

Condition (ii) states that in a very short time interval of length h, the only possible numbers of events are zero or one. Condition (ii) also implies that the number of events in a time interval of length h does not depend on when that time interval starts.

Question

Explain how motor insurance claims could be represented by a Poisson process.

Solution

The events in this case are occurrences of claim events (*ie* accidents, fires, thefts, *etc*) reported to the insurer. The parameter λ represents the average rate of occurrence of claims (*eg* 50 per day), which we are assuming remains constant throughout the year and at different times of day. The assumption that, in a sufficiently short time interval, there can be at most one claim is satisfied if we assume that claim events cannot lead to multiple claims (*ie* no motorway pile-ups, *etc*).

The reason why a process satisfying conditions (i) to (iii) is called a Poisson process is that for a fixed value of *t*, the random variable $N(t)$ has a Poisson distribution with parameter λt. This is proved as follows.

First we need a little shorthand:

Let $p_n(t) = P(N(t) = n)$.

So if $N(t)$ satisfies conditions (i) to (iii) given above then $N(t) \sim Poi(\lambda t)$ with probability function:

$$p_n(t) = \exp\{-\lambda t\} \frac{(\lambda t)^n}{n!} \tag{1.2}$$

This will be proved by deriving 'differential-difference' equations from the conditions and then showing that 1.2 is their solution.

Recall that for a partition $B_1, \dots, B_k$, the probability of any event *A* is:

$$P(A) = P(A \mid B_1)P(B_1) + \dots + P(A \mid B_k)P(B_k)$$

For a fixed value of $t > 0$ and a small positive value of *h*, condition on the number of events at time *t* and write:

$$P(n \text{ by time } t + h) = P(n \text{ by time } t + h \mid n \text{ by time } t)P(n \text{ by time } t)$$

$$+ P(n \text{ by time } t + h \mid n - 1 \text{ by time } t)P(n - 1 \text{ by time } t)$$

$$+ \cdots$$

Hence using (1.1) and the $p_n(t)$ notation, we obtain:

$$p_n(t + h) = p_{n-1}(t)[\lambda h + o(h)] + p_n(t)[1 - \lambda h + o(h)] + o(h)$$

$$= \lambda h p_{n-1}(t) + [1 - \lambda h]p_n(t) + o(h)$$

Thus:

$$p_n(t + h) - p_n(t) = \lambda h[p_{n-1}(t) - p_n(t)] + o(h) \tag{1.3}$$

and this identity holds for $n = 1, 2, 3, \dots$.

Now recall the formal definition of differentiation:

$$\frac{d}{dt}f(t) = \lim_{h \to 0}\left(\frac{f(t+h)-f(t)}{h}\right)$$

Now divide (1.3) by h, and let h go to zero from above to get the differential-difference equation:

$$\lim_{h \to 0^+}\frac{p_n(t+h)-p_n(t)}{h} = \lim_{h \to 0^+}\frac{\lambda h[p_{n-1}(t)-p_n(t)]+o(h)}{h}$$

$$\frac{d}{dt}p_n(t) = \lambda[p_{n-1}(t)-p_n(t)] \tag{1.4}$$

By definition $\lim_{h \to 0^+}\frac{o(h)}{h} \to 0$.

We will now consider the special case when $n = 0$.

There is only one possibility when $n = 0$:

$$P(0 \text{ by time } t+h) = P(0 \text{ by time } t+h \mid 0 \text{ by time } t)P(0 \text{ by time } t)$$

ie $p_0(t+h) = p_0(t)[1-\lambda h + o(h)]$

So:

$$p_0(t+h) - p_0(t) = -\lambda h p_0(t) + o(h)$$

and therefore:

$$\lim_{h \to 0^+}\frac{p_0(t+h)-p_0(t)}{h} = \lim_{h \to 0^+}\frac{-\lambda h p_0(t)+o(h)}{h}$$

When $n = 0$, an identical analysis yields:

$$\frac{d}{dt}p_0(t) = -\lambda p_0(t) \tag{1.5}$$

with initial condition $p_0(0) = 1$.

It is now straightforward to verify that the suggested solution (1.2) satisfies both the differential equations (1.4) and (1.5) as well as the initial conditions..

Question

Show that $p_n(t) = e^{-\lambda t} \dfrac{(\lambda t)^n}{n!}$ satisfies the differential equations:

$$\frac{d}{dt} p_n(t) = \lambda[p_{n-1}(t) - p_n(t)]$$

$$\frac{d}{dt} p_0(t) = -\lambda p_0(t)$$

Solution

We have $p_n(t) = e^{-\lambda t} \dfrac{(\lambda t)^n}{n!}$. Calculating the derivative using the product rule gives:

$$\frac{d}{dt} p_n(t) = \frac{d}{dt}\left\{ e^{-\lambda t} \frac{(\lambda t)^n}{n!} \right\}$$

$$= -\lambda e^{-\lambda t} \frac{(\lambda t)^n}{n!} + e^{-\lambda t} \frac{n\lambda^n t^{n-1}}{n!}$$

$$= -\lambda e^{-\lambda t} \frac{(\lambda t)^n}{n!} + \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}$$

$$= \lambda\left[p_{n-1}(t) - p_n(t) \right]$$

Similarly, $p_0(t) = e^{-\lambda t}$ which gives a derivative of:

$$\frac{d}{dt} p_0(t) = \frac{d}{dt}\left\{ e^{-\lambda t} \right\} = -\lambda e^{-\lambda t} = -\lambda p_0(t)$$

3.2   Waiting times between events in a Poisson process

This study of the Poisson process concludes by considering the distribution of the time to the first event, T_1, and the times between events, $T_2, T_3, \ldots$. These inter-event times are often called the waiting times or holding times.

In Section 2.2, we said that the waiting time between consecutive events in a Poisson distribution has an $Exp(\lambda)$ distribution.

$P(T_1 > t)$ is the probability that no events occur between time 0 and time t. Hence

$$P(T_1 > t) = P\big(N(t) = 0\big) = \exp\{-\lambda t\}$$

So the distribution function of T_1 is

$$F(t) = P(T_1 \leq t) = 1 - \exp\{-\lambda t\}$$

so that T_1 has an exponential distribution with parameter λ.

Now we consider the distribution of the time between the first and the second event.

Consider the conditional distribution of T_2 given the value of T_1.

$$P(T_2 > t \mid T_1 = r) = P(T_1 + T_2 > t + r \mid T_1 = r)$$
$$= P(N(t+r) = 1 \mid N(r) = 1)$$
$$= P(N(t+r) - N(r) = 0 \mid N(r) = 1)$$

Because the number of events in the time interval is independent of the number of events up to the start of that time interval (condition (iii) above):

$$P(N(t+r) - N(r) = 0 \mid N(r) = 1) = P(N(t+r) - N(r) = 0)$$

Since the number of events in a time interval of length r does not depend on when that time interval starts (condition (ii) above, equations (1.4.1)) we have:

$$P(N(t+r) - N(r) = 0) = P(N(t) = 0) = \exp\{-\lambda t\}$$

Hence, T_2 has an exponential distribution with parameter λ and T_2 is independent of T_1. This calculation can be repeated for T_2, T_3, ...

So we have now shown that each waiting time has an $Exp(\lambda)$ distribution.

The inter-event time is independent of the absolute time. In other words the time until the next event has the same distribution, irrespective of the time since the last event or the number of events that have already occurred.

Question

If reported claims follow a Poisson process with rate 5 per day (and the insurer has a 24 hour hotline), calculate the probability that:

(i)   there will be fewer than 2 claims reported on a given day

(ii)   the time until the next reported claim is less than an hour.

Solution

(i)   The number of claims per day, X, has a $Poi(5)$ distribution, so:

$$P(X < 2) = P(X = 0) + P(X = 1)$$

$$= e^{-5} + 5e^{-5}$$

$$= 0.0404$$

Alternatively, we can read the value of $P(X \leq 1)$ from the cumulative Poisson tables listed on page 176 of the Tables.

(ii)   The number of claims per hour, Y, has a $Poi\left(\frac{5}{24}\right)$ distribution, so the waiting time (in hours), T, has an $Exp\left(\frac{5}{24}\right)$ distribution. Hence:

$$P(T < 1) = \int_0^1 \frac{5}{24} e^{-\frac{5}{24}t} \, dt = \left[-e^{-\frac{5}{24}t} \right]_0^1 = 1 - e^{-\frac{5}{24}} = 0.188$$

Alternatively, we could just use the cumulative distribution function for the exponential distribution given on page 11 of the Tables.

4   Monte Carlo simulation

With the advent of high-speed personal computers Monte Carlo simulations have become one of the most valuable tools of the actuarial profession. This is because the vast majority of the practically important problems are not amenable to analytical solution.

> **We have already seen that we can simulate samples from distributions listed in Sections 2 and 3 using the R functions** `rbinom, rgeom, rnbinom, rhyper, rpois, runif, rgamma, rexp, rchisq, rbeta, rnorm, rlnorm, rt` **and** `rf`**.**

Below we outline one basic simulation technique that can be used to simulate values from most of these distributions.

This is known as the inverse transform method. It can be applied to both continuous and discrete distributions.

4.1   Inverse Transform method for continuous distributions

The method works by first generating a random number from a uniform distribution on the interval $(0,1)$. We then use the cumulative distribution function of the distribution we are trying to simulate to obtain a random value from that distribution.

First we generate a random number, U, from the $U(0,1)$ distribution. We can use this to simulate a random variate X with PDF $f(x)$ by using the CDF, $F(x)$.

Let U be the probability that X takes on a value less than or equal to x, *ie*:
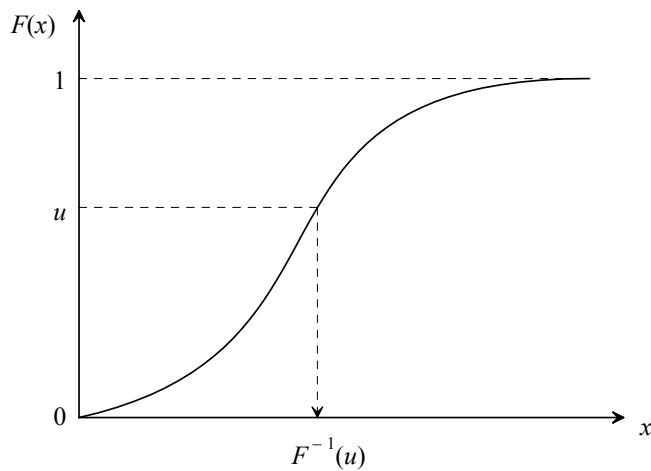
$U = P(X \leq x) = F(x)$. Then x can be derived as:

$$x = F^{-1}(u)$$

Hence, the following two-step algorithm is used to generate a random variate x from a continuous distribution with CDF $F(x)$:

1.   generate a random number u from $U(0, 1)$,

2.   return $x = F^{-1}(u)$.

We can represent this on a diagram as follows. We have a random value, u, between 0 and 1. Recall that the cumulative distribution, $F(x)$, increases from 0 to 1 as x increases:

If we set $u = F(x)$ we can obtain a random value, x, by inverting the cumulative distribution, $x = F^{-1}(u)$. Hence this method is called the inverse transform method.

This method requires that our distribution has a cumulative distribution function, $F(x)$, in the first place. This rules out the gamma, normal, lognormal and beta distributions.

Formally, we can prove that the random variable $X = F^{-1}(U)$ has the CDF $F(x)$, as follows:

$$P(X \leq x) = P[F^{-1}(U) \leq x] = P[U \leq F(x)] = F(x)$$

Example

Generate a random variate from the exponential distribution with parameter λ.

The distribution function of X is given by

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

Hence

$$x = F^{-1}(u) = -\log(1-u)/\lambda$$

Thus, to generate a random variate x from an exponential distribution we can use the following algorithm:

1.   **generate a random variate u from $U(0, 1)$**

2.   **return $x = -\log(1-u)/\lambda$.**

The main disadvantage of the inverse transform method is the necessity to have an explicit expression for the inverse of the distribution function $F(x)$. For instance, to generate a random variate from the standard normal distribution using the inverse transform method we need the inverse of the distribution function

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \ .$$

However, no explicit solution to the equation $u = F(x)$ can be found in this case.

However, it is possible to generate simulated values from a standard normal distribution. The procedure is as follows.

1.   Generate a random number u from a $U(0,1)$ distribution.

2.   If $u > 0.5$, use the tables directly to find z such that $P(Z \le z) = u$. In this case our simulated value from $N(0,1)$ is z.

3.   If $u < 0.5$, use the tables to find z such that $P(Z \le z) = 1 - u$. In this case our simulated value from $N(0,1)$ is $-z$.

We can generalise this method to generate a value from any normal distribution $X \sim N(\mu, \sigma^2)$ by using the transformation $x = \mu + \sigma z$.

Question

Simulate three values from an *Exp*(0.1) distribution using the values 0.113, 0.608 and 0.003 from $U(0,1)$.

Solution

Using the inverse transform method, we have:

$$x = -\frac{1}{0.1} \ln(1-u) = -10\ln(1-u)$$

This gives:

$$x = -10\ln(1-0.113) = 1.20$$
$$x = -10\ln(1-0.608) = 9.36$$
$$x = -10\ln(1-0.003) = 0.03$$

We can also generate random samples from other distributions, for example the uniform distribution.

Question

Generate three random values from a $U(-1,4)$ using the following random values from $U(0,1)$:

     0.07   0.628   0.461

Solution

The distribution function for a $U(-1,4)$ distribution is:

$$F(x) = \frac{x+1}{5}$$

We now set our random value, u, equal to this and rearrange:

$$u = \frac{x+1}{5} \quad \Rightarrow \quad x = 5u - 1$$

Substituting, we obtain:

$$x = 5 \times 0.07 - 1 = -0.65$$

$$x = 5 \times 0.628 - 1 = 2.14$$

$$x = 5 \times 0.461 - 1 = 1.305$$

We can also see intuitively that if we start by generating a random number from $U(0,1)$, then if we multiply it by 5 it will become a random number from $U(0,5)$. If we then subtract 1, it will become a random number from $U(-1,4)$.

Example

Generate a random variate X from the double exponential distribution with density function

$$f(x) = \tfrac{1}{2}\, \theta\, e^{-\theta |x|}, \quad x \in \mathbb{R}$$

It is possible in this case to find the distribution function F corresponding to f and to use the inverse transform method, but an alternative method is presented here. The density f is symmetric about 0; we can therefore generate a variate Y having the same distribution as $|X|$ and set $X = +Y$ or $X = -Y$ with equal probability.

The density of | X | is

$$f_{|X|}(y) = \theta e^{-\theta y}, \quad y > 0,$$

easily recognised as the density of the exponential distribution. The following algorithm therefore generates a value for X.
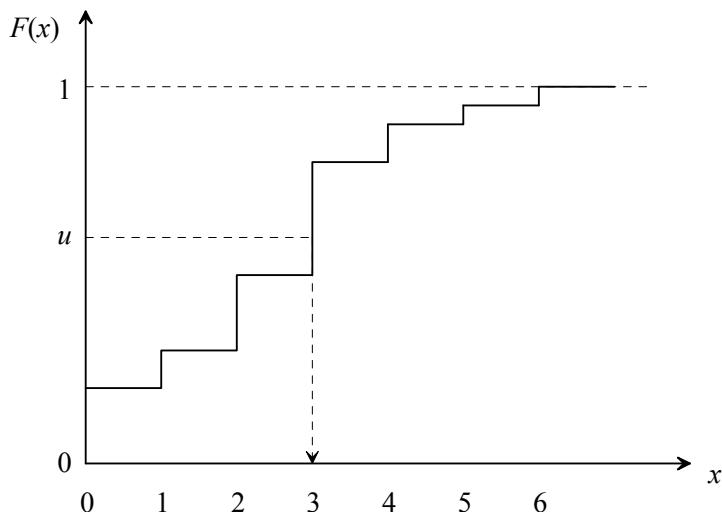
1. **generate u_1 and u_2 from $U(0, 1)$,**

2. **if $u_1 < 0.5$ return $y = -\ln(1 - u_2)/\theta$, otherwise return $y = \ln(1 - u_2)/\theta$.**

4.2 Discrete distributions

We cannot invert algebraically the distribution function of a discrete random variable, as it is a step function. The distribution function, $F(x)$, is the sum of the probabilities so far, *eg*:

$$F(5) = P(X \le 5) = P(X = 0) + P(X = 1) + \cdots + P(X = 5)$$

Given a random value, u, from $U(0,1)$ we can read off the x value from the distribution function graph as follows:

From the graph, we can see that in this particular case our value of u lies between $F(2)$ and $F(3)$. This gives $x = 3$ as our simulated value.

So in general, if our value u lies between $F(x_{j-1})$ and $F(x_j)$ then our simulated value is x_j. If the value of u corresponds exactly to the position of a step, then by convention we use the lower of the x values, *ie* the point corresponding to the left hand end of the step.

Let X be a discrete random variable which can take only the values $x_1, x_2, \ldots, x_N$, where $x_1 < x_2 < \cdots < x_N$.

The first step is to generate a random number, U, from the $U(0,1)$ distribution. We can use this to simulate a random variate X with PDF $f(x)$ by using the CDF, $F(x)$.

Let U be the probability that X takes on a value less than or equal to x. Then $X = x_j$ if:

$$F(x_{j-1}) < U \le F(x_j)$$

ie $$P(X = x_1) + P(X = x_2) + \cdots + P(X = x_{j-1}) < U \le P(X = x_1) + P(X = x_2) + \cdots + P(X = x_j)$$

Note that for $x < x_1$ we have $F(x) = 0$.

Hence, the following three-step algorithm is used to generate a random variate x from a discrete distribution with CDF $F(x)$:

1.   generate a random number u from $U(0,1)$.

2.   find the positive integer i such that $F(x_{i-1}) < u \le F(x_i)$.

3.   return $x = x_i$.

We can see that the algorithm can return only variates x from the range $\{x_1, x_2, \ldots, x_N\}$ and that the probability that a particular value $x = x_i$ is returned is given by:

$$P(\text{value returned is } x_i) = P[F(x_{i-1}) < U \le F(x_i)] = F(x_i) - F(x_{i-1}) = P(X = x_i)$$

Question

Simulate two random values from a *Poi*(2) distribution using the random values 0.721 and 0.128.

Solution

$$P(X = x) = e^{-\mu} \frac{\mu^x}{x!}, \quad x = 0, 1, 2, \ldots$$

$P(X = 0) = e^{-2} = 0.1353$ $\Rightarrow$ $F(0) = 0.1353$

$P(X = 1) = 2e^{-2} = 0.2707$ $\Rightarrow$ $F(1) = 0.4060$

$P(X = 2) = \frac{2^2}{2!} e^{-2} = 0.2707$ $\Rightarrow$ $F(2) = 0.6767$

$P(X = 3) = \frac{2^3}{3!} e^{-2} = 0.1804$ $\Rightarrow$ $F(3) = 0.8571$, *etc*.

Since $F(2) < 0.721 < F(3)$, our first simulated value is 3.

Since $0.128 < F(0)$, our second simulated value is 0.

Alternatively, we could use the cumulative Poisson tables on page 175 of the Tables instead of calculating the values.

We can use a similar approach for the binomial distribution.

Question

Generate three random values from a $Bin(4, 0.6)$ distribution using the following random values from $U(0,1)$:

0.588	0.222	0.906

Solution

The probability function for a $Bin(4, 0.6)$ distribution is:

$$P(X = x) = \binom{4}{x} 0.6^x 0.4^{4-x}, \qquad x = 0, 1, 2, 3, 4$$

Calculating the probabilities and the cumulative distribution function:

$$P(X = 0) = 0.4^4 = 0.0256 \qquad\qquad \Rightarrow\quad F(0) = 0.0256$$

$$P(X = 1) = 4 \times 0.6 \times 0.4^3 = 0.1536 \quad \Rightarrow\quad F(1) = 0.1792$$

$$P(X = 2) = 6 \times 0.6^2 \times 0.4^2 = 0.3456 \Rightarrow\quad F(2) = 0.5248$$

$$P(X = 3) = 4 \times 0.6^3 \times 0.4 = 0.3456 \quad\; \Rightarrow\quad F(3) = 0.8704$$

$$P(X = 4) = 0.6^4 = 0.1296 \qquad\qquad \Rightarrow\quad F(4) = 1$$

Since $F(2) < 0.588 < F(3)$, our first simulated value is 3.

Since $F(1) < 0.222 < F(2)$, our second simulated value is 2.

Since $F(3) < 0.906 < F(4)$, our third simulated value is 4.

Alternatively, it is much quicker to use the cumulative binomial probabilities given on page 186 of the Tables.

Chapter 1 Summary

Standard discrete distributions covered in this course are the discrete uniform, Bernoulli, binomial, geometric, negative binomial, hypergeometric and Poisson.

Standard continuous distributions covered in this course are the continuous uniform, gamma, exponential, chi-square, normal, lognormal, beta, t and F.

The geometric and exponential distributions have the memoryless property:

$$P(X > x + n \mid X > n) = P(X > x)$$

The properties of the distributions are summarised on the next page.

The t distribution with k degrees of freedom is defined as:

$$t_k = \frac{N(0,1)}{\sqrt{\chi_k^2 / k}}$$

The F distribution with m, n degrees of freedom is defined as:

$$F_{m,n} = \frac{\chi_m^2 / m}{\chi_n^2 / n}$$

The Poisson process counts events occurring up to and including time t:

$$N(t) \sim Poi(\lambda t)$$

To calculate probabilities we consider events occurring in a small time interval h. The waiting times between events in a Poisson process have exponential distributions.

Random variables can be simulated by using the inverse transform method. First we take a random number, u, from $U(0,1)$ then we set:

continuous	$x = F^{-1}(u)$		
discrete	$x = x_j$	where	$F(x_{j-1}) < u \le F(x_j)$

	Distribution	PF or PDF	Mean	Variance
Discrete Distributions	Discrete uniform	$\dfrac{1}{k}$	$\dfrac{k+1}{2}$	$\dfrac{k^2-1}{12}$
	Bernouilli	$p^x(1-p)^{1-x}$	p	$p(1-p)$
	Binomial	$\dbinom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
	Geometric	$p(1-p)^{x-1}$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
	Negative binomial	$\dbinom{x-1}{k-1}p^k(1-p)^{x-k}$	$\dfrac{k}{p}$	$\dfrac{k(1-p)}{p^2}$
	Poisson	$\dfrac{\lambda^x e^{-\lambda}}{x!}$	λ	λ
Continuous Distributions	Continuous uniform	$\dfrac{1}{\beta-\alpha}$	$\dfrac{1}{2}(\alpha+\beta)$	$\dfrac{1}{12}(\beta-\alpha)^2$
	Gamma	$\dfrac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$	$\dfrac{\alpha}{\lambda}$	$\dfrac{\alpha}{\lambda^2}$
	Exponential	$\lambda e^{-\lambda x}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
	Chi-square	$\dfrac{\left(\frac{1}{2}\right)^{\frac{\upsilon}{2}}}{\Gamma(\frac{\upsilon}{2})}x^{\frac{\upsilon}{2}-1}e^{-\frac{1}{2}x}$	υ	2υ
	Beta	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
	Lognormal	$\dfrac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2}$	$e^{\mu+\frac{1}{2}\sigma^2}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$
	Normal	$\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2

Chapter 1 Practice Questions

1.1   If $X \sim N(14, 20)$, calculate:

(i) $P(X < 14)$

(ii) $P(X > 20)$

(iii) $P(X < 9)$

(iv) r such that $P(X > r) = 0.41294$.

1.2   Determine the third non-central moment of the normal distribution with mean 10 and variance 25.

1.3   Calculate $P(X < 8)$ if:

(i) $X \sim U(5, 10)$

(ii) $X \sim N(10, 5)$

(iii) $X \sim Exp(0.5)$

(iv) $X \sim \chi_5^2$

(v) $X \sim Gamma(8, 2)$

(vi) $X \sim \log N(2, 5)$.

1.4   A random variable X has a $Poi(3.6)$ distribution.

(i)   Calculate the mode of the probability distribution.

(ii)   Calculate the standard deviation of the distribution.

(iii)   State, with reasons, whether the distribution is positively or negatively skewed.

1.5   If U denotes a continuous random variable that is uniformly distributed over the range $(-1, 1)$ and V denotes a discrete random variable that is equally likely to take any of the values $\{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$, calculate $var(U)$ and $var(V)$. Comment on your answers.

1.6   An analyst is interested in using a gamma distribution with parameters $\alpha = 2$ and $\lambda = \frac{1}{2}$, that is,
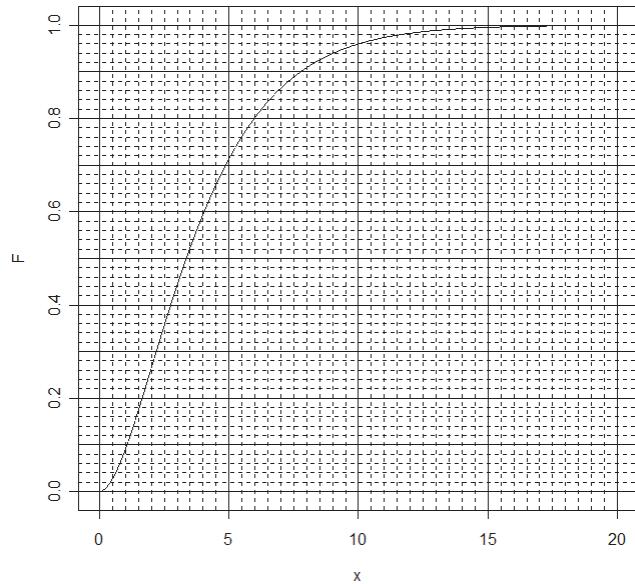
Exam style   with density function $f(x) = \frac{1}{4} x e^{-\frac{1}{2}x}$, $0 < x < \infty$.

(i)   (a)   State the mean and standard deviation of this distribution.

       (b)   Hence comment briefly on its shape.   [2]

(ii)   Show that the cumulative distribution function is given by

$$F(x) = 1 - (1 + \frac{1}{2}x)e^{-\frac{1}{2}x}, \quad 0 < x < \infty \quad \text{(zero otherwise).}$$ [3]

The analyst wishes to simulate values x from this gamma distribution and is able to generate random numbers u from a uniform distribution on (0,1).

(iii)   (a)   Specify an equation involving x and u, the solution of which will yield the simulated value x.

        (b)   Comment briefly on how this equation might be solved.

        (c)   The graph below gives $F(x)$ plotted against x. Use this graph to determine the simulated value of x corresponding to the random number $u = 0.66$.

[3]
[Total 8]

1.7   Calculate $P(X < 8)$ if:

(i) X is the number of claims reported in a year by 20 policyholders. Claims reporting from each policyholder occurs randomly at a rate of 0.2 per year independently of the other policyholders.

(ii) X is the number of claims examined up to and including the fourth claim that exceeds £20,000. The probability that any claim received exceeds £20,000 is 0.3 independently of any other claim.

(iii) X is the number of deaths in the coming year amongst a group of 500 policyholders. Each policyholder has a 0.01 probability of dying in the coming year independently of any other policyholder.

(iv) X is the number of phone calls made *before* an agent makes the first sale. The probability that any phone call leads to a sale is 0.01 independently of any other call.

1.8   A random variable has a lognormal distribution with mean 10 and variance 4. Calculate the probability that the variable will take a value between 7.5 and 12.5.

1.9   The random variable N has a Poisson distribution with parameter λ and $P(N = 1 | N \geq 1) = 0.4$. Calculate the value of λ to 2 decimal places.

1.10   Simulate two observations from the distribution with probability density function:

$$f(x) = \frac{50}{(5+x)^3}, \qquad x > 0$$

using the random numbers 0.863 and 0.447 selected from the uniform distribution on the interval (0,1).

1.11   Claim amounts are modelled as an exponential random variable with mean £1,000.

Exam style

(i)   Calculate the probability that a randomly selected claim amount is greater than £5,000. [1]

(ii)   Calculate the probability that a randomly selected claim amount is greater than £5,000 given that it is greater than £1,000.   [2]

[Total 3]

1.12   The ratio of the standard deviation to the mean of a random variable is called the *coefficient of variation*.

Exam style

For each of the following distributions, decide whether increasing the mean of the random variable increases, decreases, or has no effect on the value of the coefficient of variation:

(a)   Poisson with mean λ

(b)   exponential with mean μ

(c)   chi-square with n degrees of freedom.   [6]

1.13   Consider the following simple model for the number of claims, N, which occur in a year on a policy:

n	0	1	2	3
$P(N=n)$	0.55	0.25	0.15	0.05

(a)   Explain how you would simulate an observation of N using a number r, an observation of a random variable that is uniformly distributed on $(0,1)$.

(b)   Illustrate your method described in (a) by simulating three observations of N using the following random numbers between 0 and 1:

0.6221, 0.1472, 0.9862   [4]

1.14   It is assumed that claims arising on an industrial policy can be modelled as a Poisson process at a rate of 0.5 per year.

(i)   Determine the probability that no claims arise in a single year.   [1]

(ii)   Determine the probability that, in three consecutive years, there is one or more claims in one of the years and no claims in each of the other two years.   [2]

(iii) Suppose a claim has just occurred. Determine the probability that more than two years will elapse before the next claim occurs.   [2]
                                                                       [Total 5]

1.15   Consider the following three probability statements concerning an F variable with 6 and 12 degrees of freedom.

(a) $P(F_{6,12} > 0.250) = 0.95$

(b) $P(F_{6,12} < 4.821) = 0.99$

(c) $P(F_{6,12} < 0.13) = 0.01$

State, with reasons, whether each of these statements is true.   [3]

Chapter 1 Solutions

1.1 (i)   Since 14 is the mean, the probability is 0.5.

(ii) $P(X > 20) = P\left(Z > \dfrac{20-14}{\sqrt{20}}\right) = P(Z > 1.342) \approx 1 - 0.91020 = 0.0898$

(iii) $P(X < 9) = P\left(Z < \dfrac{9-14}{\sqrt{20}}\right) = P(Z < -1.118) \approx 1 - 0.86821 = 0.1318$

(iv) $P(X > r) = P\left(Z > \dfrac{r-14}{\sqrt{20}}\right) = 0.41294$, which gives:

$$P\left(Z < \frac{r-14}{\sqrt{20}}\right) = 0.58706 \Rightarrow \frac{r-14}{\sqrt{20}} = 0.22 \Rightarrow r = 14.98$$

1.2   The third non-central moment is $E[X^3]$. The formula for the skewness is:

$$E[(X-\mu)^3] = E[X^3] - 3\mu E[X^2] + 2\mu^3$$

We also know that the skewness of the normal distribution is zero, so:

$$0 = E[X^3] - 3 \times 10 \times (25 + 10^2) + 2 \times 10^3 \Rightarrow E[X^3] = 1,750$$

We have worked out $E[X^2]$ here by turning around the relationship $var(X) = E[X^2] - E^2[X]$.

1.3 (i)   *Uniform*

$$P(X < 8) = \int_5^8 0.2\,dx = \left[0.2x\right]_5^8 = 0.6$$

Alternatively, we could use the DF given on page 13 of the Tables. $P(X < 8) = F(8) = \frac{8-5}{10-5} = 0.6$.

(ii)   *Normal*

$$P(X < 8) = P\left(Z < \frac{8-10}{\sqrt{5}}\right) = P(Z < -0.894)$$

$$= 1 - P(Z < 0.894)$$

$$\approx 1 - 0.81434$$

$$= 0.1857$$

(iii)   **Exponential**

$$P(X < 8) = \int_0^8 0.5 e^{-0.5x}\, dx = \left[-e^{-0.5x} \right]_0^8 = 1 - e^{-4} = 0.98168$$

Alternatively, we could use the DF given on page 11. $P(X < 8) = F(8) = 1 - e^{-0.5 \times 8} = 0.98168$.

(iv)   **Chi-square**

Using the χ^2 tables on page 165 of the *Tables* gives $P(X < 8) = 0.8438$.

(v)   **Gamma**

The only practical way in a written exam to calculate probabilities of an $X \sim Gamma(\alpha, \lambda)$ distribution is to use the relationship $2\lambda X \sim \chi^2_{2\alpha}$ and then read off the probability from the χ^2 tables.

$$P(X < 8) = P(2\lambda X < 16\lambda) = P(4X < 32) = P(\chi^2_{16} < 32) = 0.9900$$

(vi)   **Lognormal**

Using the fact that if $X \sim \log N(\mu, \sigma^2)$ then $\ln X \sim N(\mu, \sigma^2)$:

$$P(X < 8) = P(\ln X < \ln 8) = P\left(Z < \frac{\ln 8 - 2}{\sqrt{5}} \right)$$

$$= P(Z < 0.036) \approx 0.5144$$

1.4   (i)   **Mode**

We can find the mode by calculating probabilities and seeing which value has the highest probability.

$$P(X = 0) = e^{-3.6} = 0.02732$$

Using the iterative formula for the Poisson distribution gives:

$$P(X = 1) = \tfrac{3.6}{1} \times 0.02732 = 0.09837$$

$$P(X = 2) = \tfrac{3.6}{2} \times 0.09837 = 0.17706$$

$$P(X = 3) = \tfrac{3.6}{3} \times 0.17706 = 0.21247$$

$$P(X = 4) = \tfrac{3.6}{4} \times 0.21247 = 0.19122$$

$$P(X = 5) = \tfrac{3.6}{5} \times 0.19122 = 0.13768$$

etc
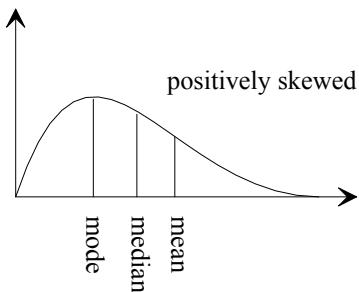
We can see that 3 is the mode.

(ii)   Standard deviation

The variance of the $Poi(\lambda)$ distribution is λ. So the standard deviation of the $Poisson(3.6)$ distribution is $\sqrt{3.6} = 1.8974$.

(iii)   Skewness

The Poisson distribution is positively skewed as the mode of 3 is lower than the mean of 3.6. In fact the Poisson distribution is *always* positively skewed. For most positively skewed distributions, we find that mode < median < mean.

1.5   The probability density function of U is constant, ie $f_U(u) = \frac{1}{2}$, $-1 < x < 1$.

The probability function of V is constant ie $f_V(v) = \frac{1}{5}$, $x = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$.

By symmetry the mean value of both variables is zero.

Alternatively:

$$E(U) = \int_u uf(u)\,du = \int_{-1}^{1} \frac{1}{2}u\,du = \left[\frac{1}{4}u^2\right]_{-1}^{1} = \frac{1}{4} - \frac{1}{4} = 0$$

$$E(V) = \sum vP(V = v)$$
$$= (-1 \times \tfrac{1}{5}) + (-\tfrac{1}{2} \times \tfrac{1}{5}) + (0 \times \tfrac{1}{5}) + (\tfrac{1}{2} \times \tfrac{1}{5}) + (1 \times \tfrac{1}{5}) = 0$$

So the variance of U is calculated from:

$$E(U^2) = \int_{-1}^{1} \frac{1}{2}u^2\,du = \left[\frac{1}{6}u^3\right]_{-1}^{1} = \frac{1}{3}$$

$$\Rightarrow \quad var(U) = \frac{1}{3} - 0^2 = \frac{1}{3}$$

Alternatively, you could use the formula $\frac{1}{12}(b-a)^2$ from page 13 of the Tables.

So the variance of V is:

$$E\left(V^2\right)=\sum v^2 P(V=v)=\tfrac{1}{5}\left[(-1)^2+(\tfrac{1}{2})^2+0^2+(\tfrac{1}{2})^2+1^2\right]=\tfrac{1}{2}$$

$$\Rightarrow \quad var(V)=\tfrac{1}{2}-0^2=\tfrac{1}{2}$$

The variance is a measure of the spread of values. Both distributions take values in the range from -1 to $+1$ and are centred around zero. However, the variance of V is greater than the variance of U because there is a greater probability of obtaining the extreme values -1 and $+1$.

1.6 **(i)(a)** *The mean and standard deviation of the distribution*

For a gamma distribution with $\alpha=2$ and $\lambda=0.5$:

$$E(X)=\frac{\alpha}{\lambda}=\frac{2}{0.5}=4 \qquad sd(X)=\sqrt{\frac{\alpha}{\lambda^2}}=\sqrt{\frac{2}{0.5^2}}=\sqrt{8}=2.828 \tag{1}$$

(i)(b) *The shape of the distribution*

Since X cannot take negative values and the standard deviation is large relative to the mean, the gamma distribution with $\alpha=2$ and $\lambda=0.5$ is positively skewed. [1]

(ii) *Cumulative distribution function*

The cumulative distribution function, $F_X(x)$, is:

$$F_X(x)=P(X\le x)=\int_{t=0}^{x}\frac{1}{4}te^{-\frac{1}{2}t}\,dt \qquad x>0 \tag{1}$$

Using integration by parts, with $u=\tfrac{1}{4}t$ and $\frac{dv}{dt}=e^{-\frac{1}{2}t}$:

$$F_X(x)=\int_{t=0}^{x}\tfrac{1}{4}te^{-\frac{1}{2}t}\,dt$$

$$=\left[\tfrac{1}{4}t\times\frac{e^{-\frac{1}{2}t}}{-\tfrac{1}{2}}\right]_0^x-\int_0^x\frac{e^{-\frac{1}{2}t}}{-\tfrac{1}{2}}\times\tfrac{1}{4}\,dt$$

$$=\left[-\tfrac{1}{2}te^{-\frac{1}{2}t}\right]_0^x+\int_0^x\tfrac{1}{2}e^{-\frac{1}{2}t}dt=-\tfrac{1}{2}xe^{-\frac{1}{2}x}-\left[e^{-\frac{1}{2}t}\right]_0^x=1-e^{-\frac{1}{2}x}\left(1+\tfrac{1}{2}x\right) \qquad x>0 \tag{2}$$

(iii)(a) *Equation to simulate values of x*

We equate the random number u to the cumulative distribution function:
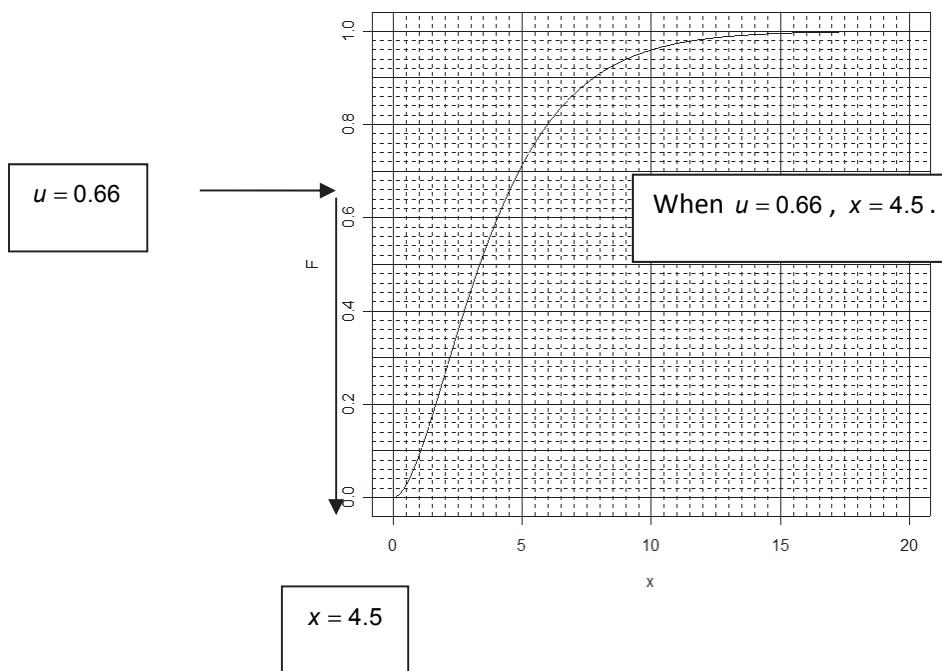
$$u = 1 - e^{-\frac{1}{2}x}\left(1 + \frac{1}{2}x\right) \quad x > 0 \qquad [1]$$

(iii)(b) *Solving the equation*

For a given random number, *ie* a given value of u, we could solve for x by:

- trial and error

- using Table Mode on the calculator

- using the Newton-Raphson method or some other iterative approach.   [1]

Alternatively, the function for u could be plotted against x, and then used to determine the x value corresponding to a u value.

(iii)(c) *Using the graph*

So, the simulated value of x is 4.5.   [1]

1.7   (i)   *Poisson*

Each policyholder has a $Poi(0.2)$ distribution for the number of claims. Therefore the number of claims for the 20 policyholders has a $Poi(4)$ distribution.

Since the Poisson distribution only takes integer values $P(X < 8) = P(X \leq 7)$. Using the Poisson cumulative probability tables gives 0.94887.

Alternatively, we could use $P(X=x) = \frac{\lambda^x}{x!}e^{-\lambda}$ to calculate the values of

$P(X=0), P(X=1), \ldots, P(X=7)$ (the iterative formula would speed up this process), and then add them up.

(ii)   *Negative binomial*

We are counting the number of trials up to and including the 4th success. This describes a Type 1 negative binomial with $k=4$ and $p=0.3$.

$$P(X=x) = \binom{x-1}{3} 0.3^4 \, 0.7^{x-4} \qquad\qquad x = 4, 5, 6, \ldots$$

So $P(X<8) = P(X=4) + \cdots + P(X=7)$.

$$P(X=4) = \binom{3}{3} 0.3^4 = 0.0081$$

Now using the iterative formula $P(X=x) = \frac{x-1}{x-4} q P(X=x-1)$, we get:

$$P(X=5) = \frac{4}{1} \times 0.7 \times 0.0081 = 0.02268$$

$$P(X=6) = \frac{5}{2} \times 0.7 \times 0.02268 = 0.03969$$

$$P(X=7) = \frac{6}{3} \times 0.7 \times 0.03969 = 0.05557$$

Hence, $P(X<8) = 0.0081 + 0.02268 + 0.03969 + 0.05557 = 0.12604$.

Alternatively, we could have calculated each of the probabilities using the probability function.

(iii)   *Binomial*

Here we have a binomial distribution with $n=500$ and $p=0.01$. Since *n* is large and *p* is small we could use a Poisson approximation and then use the cumulative Poisson tables (as we did in part (i)).

$$Bin(500, 0.01) \sim Poi(5) \text{ (approximately)}$$

Using the cumulative Poisson tables gives $P(X<8) = P(X \le 7) = 0.86663$.

Alternatively, we could calculate this accurately, starting with the probability of no deaths:

$$P(X=0) = \binom{500}{0} \times 0.99^{500} = 0.00657$$

Now using the iterative formula $P(X = x) = \dfrac{n-x+1}{x} \times \dfrac{p}{q} \times P(X = x-1)$:

$$P(X = 1) = \frac{500}{1} \times \frac{0.01}{0.99} \times 0.00657 = 0.03318$$

$$P(X = 2) = \frac{499}{2} \times \frac{0.01}{0.99} \times 0.03318 = 0.08363$$

$$P(X = 3) = \frac{498}{3} \times \frac{0.01}{0.99} \times 0.08363 = 0.14023$$

$$P(X = 4) = \frac{497}{4} \times \frac{0.01}{0.99} \times 0.14023 = 0.17600$$

$$P(X = 5) = \frac{496}{5} \times \frac{0.01}{0.99} \times 0.17600 = 0.17635$$

$$P(X = 6) = \frac{495}{6} \times \frac{0.01}{0.99} \times 0.17635 = 0.14696$$

$$P(X = 7) = \frac{494}{7} \times \frac{0.01}{0.99} \times 0.14696 = 0.10476$$

Hence, $P(X = 8) = P(X = 0) + \cdots + P(X = 7) = 0.86768$.

(iv)   *Geometric*

We are counting the number of trials up to, but not including, the 1st success. This describes a Type 2 geometric distribution with $p = 0.01$.

$$P(X = x) = 0.01 \times 0.99^x \quad x = 0, 1, 2, \ldots$$

Now:

$$P(X < 8) = P(X \le 7)$$

$$= P(X = 0) + \cdots + P(X = 7)$$

$$= 0.01 + 0.01 \times 0.99 \times 0.01 \times 0.99^2 + \cdots + 0.01 \times 0.99^7$$

This is a geometric series, so the quickest way to add this up is to use the formula for the sum of a geometric series $S_n = \dfrac{a(1-r^n)}{1-r}$. This gives:

$$P(X < 8) = \frac{0.01 \times (1 - 0.99^8)}{1 - 0.99} = 0.07726$$

1.8   Let X denote the random variable.

Using the formulae for the mean and variance of a lognormal distribution:

$$E[X] = e^{\mu + \frac{1}{2}\sigma^2} = 10 \qquad\qquad (1)$$

$$\text{var}(X) = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1) = 4 \qquad\qquad (2)$$

Squaring equation (1) and substituting into equation (2):

$$\text{var}(X) = 10^2\left(e^{\sigma^2} - 1\right) = 4$$

$$\Rightarrow \quad e^{\sigma^2} - 1 = 0.04$$

$$\Rightarrow \quad \sigma^2 = \log 1.04 = 0.03922$$

Substituting this into equation (1) gives:

$$\mu = \log 10 - \frac{1}{2}\sigma^2 = 2.2830$$

So the required probability is:

$$P(7.5 < X < 12.5) = P(X < 12.5) - P(X < 7.5)$$

$$= P(\ln X < \ln 12.5) - P(\ln X < \ln 7.5)$$

$$= P\left(Z < \frac{\log 7.5 - 2.2830}{\sqrt{0.03922}}\right) - P\left(Z < \frac{\log 12.5 - 2.2830}{\sqrt{0.03922}}\right)$$

$$= \Phi(1.226) - \Phi(-1.354)$$

$$\approx 0.88990 - 0.08787$$

$$= 0.8020$$

1.9   The conditional probability is:

$$P(N = 1 \mid N \geq 1) = \frac{P(N=1)}{P(N \geq 1)} = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} = \frac{\lambda}{e^{\lambda} - 1}$$

Trial and error gives $\dfrac{1.62}{e^{1.62} - 1} = 0.3997$. So $\lambda \approx 1.62$.

1.10   To simulate a random variable we require the distribution function, $F(x)$:

$$F(x) = P(X \leq x) = \int_0^x 50(5+t)^{-3}\, dt = \left[-25(5+t)^{-2}\right]_0^x = 1 - \frac{25}{(5+x)^2}$$

We can now use the inverse transform method:

$$u = 1 - \frac{25}{(5+x)^2} \quad \Rightarrow \quad x = \sqrt{\frac{25}{1-u}} - 5$$

Substituting in our values of u, we obtain:

$$x_1 = \sqrt{\frac{25}{1-0.863}} - 5 = 8.51$$

$$x_2 = \sqrt{\frac{25}{1-0.447}} - 5 = 1.72$$

1.11 (i) **Probability**

$$P(X > 5,000) = 1 - F(5,000) = e^{-0.001 \times 5,000} = e^{-5}$$

[1]

$$= 0.00674$$

(ii) **Conditional probability**

$$P(X > 5,000 \,|\, X > 1,000) = \frac{P(X > 5,000 \cap X > 1,000)}{P(X > 1,000)}$$

$$= \frac{P(X > 5,000)}{P(X > 1,000)}$$

[1]

We have already found the numerator, we just need to find the denominator:

$$P(X > 1,000) = 1 - F(1,000) = e^{-0.001 \times 1,000} = e^{-1}$$

So the required probability is:

$$P(X > 5,000 \,|\, X > 1,000) = \frac{e^{-5}}{e^{-1}} = e^{-4} = 0.0183$$

[1]

1.12 (a) **Poisson**

The $Poi(\lambda)$ distribution has mean λ and variance λ, so:

$$\text{coefficient of variation} = \sqrt{\lambda} \big/ \lambda = 1 \big/ \sqrt{\lambda}$$

[1]

As the mean, λ, increases the coefficient of variation, $1\big/\sqrt{\lambda}$, decreases.   [1]

(b) **Exponential**

We are given the mean of the exponential distribution, which is $\mu = 1/\lambda$. So working in terms of μ the mean is μ and the variance is μ^2. Hence:

$$\text{coefficient of variation} = \sqrt{\mu^2} \big/ \mu = 1$$

[1]

As the mean, μ, increases the coefficient of variation, 1, is unchanged, *ie* changing the mean has no effect on the coefficient.   [1]

(c)   ***Chi-square***

The χ^2_n distribution has a mean of n and a variance of $2n$. Hence:

$$\text{coefficient of variation} = \sqrt{2n}\big/n = \sqrt{2/n} \qquad\qquad [1]$$

As the mean, n, increases the coefficient of variation, $\sqrt{2/n}$, decreases.   [1]

1.13   (a)   ***Method for simulating an observation***

To simulate a value from a discrete distribution, we follow these two steps:

1.   Calculate the DF, $F(n)$

2.   If $F(n-1) < r \le F(n)$, then the simulated value is n.   [1]

The CDF is:

n	0	1	2	3
$P(N \le n)$	0.55	0.8	0.95	1

So the simulated value is given by:

$$n = \begin{cases} 0 & 0 \le r < 0.55 \\ 1 & 0.55 \le r < 0.8 \\ 2 & 0.8 \le r < 0.95 \\ 3 & 0.95 \le r < 1 \end{cases} \qquad\qquad [1]$$

(b)   ***Simulating three values***

Since $0.55 < 0.6221 < 0.8$, the first simulated value is 1. Since $0 < 0.1472 < 0.55$, the second simulated value is 0. Since $0.95 < 0.9862 < 1$, the third simulated value is 3.   [2]

1.14   (i)   ***Probability of no claims in one year***

The distribution of the number of claims, N, in one year is *Poi*(0.5). Hence the probability of no claims in one year is:

$$P(N = 0) = \frac{0.5^0}{0!} e^{-0.5} = 0.60653 \qquad\qquad [1]$$

(ii)   ***Probability of no claims in two of three years***

Using our result from part (i), the probability of one or more claims in one year is:

$$P(N \ge 1) = 1 - P(N = 0) = 1 - 0.60653 = 0.39347$$

If X is the number of years with one or more claim, then:

$$X \sim Bin(3, 0.39347)$$

So we have:

$$P(X = 1) = {}^3C_1 \times 0.39347 \times 0.60653^2 = 0.43425 \qquad [2]$$

(iii) Probability that more than two years will elapse before the next claim

The waiting time, T, in years has an $Exp(0.5)$ distribution.

$$P(T > 2) = 1 - F(2) = e^{-0.5 \times 2} = e^{-1} = 0.36788 \qquad [2]$$

1.15 In this question we will use the notation $F_{a,b,\alpha}$ to be the upper α% point of the $F_{a,b}$ distribution.

(a) Statement (a), true or false?

We know that $F_{6,12,95} = \dfrac{1}{F_{12,6,5}}$. From the *Tables*, $\dfrac{1}{F_{12,6,5}} = \dfrac{1}{4.0} = 0.250$, so (a) is true. $\qquad [1]$

(b) Statement (b), true or false?

From the *Tables*, $F_{6,12,1} = 4.821$, so (b) is true. $\qquad [1]$

(c) Statement (c), true or false?

We know that $F_{6,12,99} = \dfrac{1}{F_{12,6,1}}$. From the *Tables*, $\dfrac{1}{F_{12,6,1}} = \dfrac{1}{7.718} = 0.13$, so (c) is true. $\qquad [1]$

2

Generating functions

Syllabus objectives

1.4   Generating functions

     1.4.1   Define and determine the moment generating function of random variables.

     1.4.2   Define and determine the cumulant generating function of random variables.

     1.4.3   Use generating functions to determine the moments and cumulants of random variables, by expansion as a series or by differentiation, as appropriate.

     1.4.4   Identify the applications for which a moment generating function, a cumulant generating function and cumulants are used, and the reasons why they are used.

0   Introduction

Generating functions provide a neat way of working out various properties of probability distributions without having to use integration repeatedly. For example they can be used to:

(a)   Find the mean, variance and higher moments of a probability distribution. This will recap and build upon the work of the previous chapter.

(b)   Find the distribution of a linear combination of independent random variables, *eg* $X + Y$ where $X \sim Poi(\lambda)$ and $Y \sim Poi(\mu)$. This will be covered in a later chapter.

(c)   Determine the properties of compound distributions. We will meet these in Subject CS2.

In this chapter we will introduce two types of generating functions: moment generating functions (MGFs) and cumulant generating functions (CGFs). We will use them to derive formulae for the moments of statistical distributions. MGFs are used to generate moments (and so are the most useful to us at this point) and CGFs are used to generate cumulants. For our present purposes, all we need to know is that the first three cumulants are the mean, variance and skewness.

A lot of students get confused in this chapter, as they want to know 'where these definitions come from'. Basically, they were invented to make calculations of means and variances easier. We saw some examples in the previous chapter of how to calculate the mean and variance for different well-known probability distributions. In this chapter we will see how we can use generating functions to derive many of these results.

The syllabus says 'define and determine', so make sure you know the definitions of MGFs and CGFs and can find them (where they exist) for all the distributions met in the previous chapter. In addition, the syllabus requires us to 'determine the moments and cumulants', so ensure you can calculate $E(X)$ and $\text{var}(X)$ for each of these distributions.

1   Moment generating functions

1.1   General formula

A moment generating function (MGF) can be used to generate moments (about the origin) of the distribution of a random variable (discrete or continuous), ie $E(X), E(X^2), E(X^3), \ldots.$

Although the moments of most distributions can be determined directly by evaluation using the necessary integrals or summation, utilising moment generating functions sometimes provides considerable simplifications.

Definition

The moment generating function, $M_X(t)$, of a random variable X is given by:

$$M_X(t) = E\left[e^{tX} \right]$$

for all values of t for which the expectation exists.

MGFs can be defined for both discrete and continuous random variables.

Question

Write down the value of $M_X(0)$.

Solution

$$M_X(0) = E\left[e^0 \right] = E\left[1 \right] = 1$$

This is true for any random variable X.

This can be a useful check in the exam – make sure that the expression you obtain for the MGF gives 1 when $t = 0$.

We have defined the expectation of a function of a random variable, $g(X)$, to be

$E[g(X)] = \sum_x g(x)P(X = x)$ or $\int_x g(x)f_X(x)\,dx$. So the MGF is given by:

$$M_X(t) = E[e^{tX}] = \sum_x e^{tx} P(X = x) \quad \text{or} \quad \int_x e^{tx} f_X(x)\,dx$$

Question

Derive the MGF of the random variable X with probability function:

$$P(X = x) = \frac{3}{4^x} \qquad x = 1, 2, 3, \ldots$$

Solution

The MGF is:

$$M_X(t) = E(e^{tX}) = \sum_{x=1}^{\infty} e^{tx} \frac{3}{4^x} = \frac{3}{4} e^t + \frac{3}{16} e^{2t} + \frac{3}{64} e^{3t} + \cdots$$

This is an infinite geometric series with first term $a = \frac{3}{4} e^t$ and common ratio $r = \frac{1}{4} e^t$. Summing

this geometric series to infinity using the formula $S_\infty = \dfrac{a}{1-r} \quad -1 < r < 1$ gives:

$$M_X(t) = \frac{\frac{3}{4} e^t}{1 - \frac{1}{4} e^t} = \frac{3e^t}{4 - e^t}$$

where $-1 < \frac{1}{4} e^t < 1 \;\Rightarrow\; -4 < e^t < 4 \;\Rightarrow\; t < \ln 4$.

Now let's derive the MGF of a continuous distribution.

Question

Derive the MGF of the random variable X with probability density function:

$$f(x) = \tfrac{1}{2}(1-x) \qquad -1 \le x \le 1$$

Solution

The MGF is:

$$M_X(t) = E(e^{tX}) = \int_x e^{tx} f(x)\, dx$$

$$= \int_{-1}^{1} \tfrac{1}{2}(1-x) e^{tx}\, dx$$

$$= \int_{-1}^{1} \tfrac{1}{2} e^{tx}\, dx - \int_{-1}^{1} \tfrac{1}{2} x e^{tx}\, dx$$

Using integration by parts on the second integral we obtain:

$$M_X(t) = \left[\tfrac{1}{2t} e^{tx} \right]_{-1}^{1} - \left\{ \left[\tfrac{1}{2t} x e^{tx} \right]_{-1}^{1} - \tfrac{1}{2t} \int_{-1}^{1} e^{tx} \, dx \right\}$$

$$= \left[\tfrac{1}{2t} e^{tx} \right]_{-1}^{1} - \left[\tfrac{1}{2t} x e^{tx} \right]_{-1}^{1} + \tfrac{1}{2t} \left[\tfrac{1}{t} e^{tx} \right]_{-1}^{1}$$

$$= \tfrac{1}{2t} e^{t} - \tfrac{1}{2t} e^{-t} - \left(\tfrac{1}{2t} e^{t} + \tfrac{1}{2t} e^{-t} \right) + \tfrac{1}{2t} \left(\tfrac{1}{t} e^{t} - \tfrac{1}{t} e^{-t} \right)$$

$$= -\tfrac{1}{t} e^{-t} + \tfrac{1}{2t^2} e^{t} - \tfrac{1}{2t^2} e^{-t}$$

This is known as the triangular distribution (draw a sketch of the PDF and you will see why). You will meet this distribution again in Subject CS2.

In a moment we'll look at how to obtain the MGFs of the standard distributions given in the previous chapter, but first let's find out how we can use MGFs to calculate moments.

Calculating moments

The method is to differentiate the MGF with respect to t and then set $t = 0$, the r th derivative giving the r th moment about the origin.

For example, $M_X'(t) = E[X e^{tX}]$ so $M_X'(0) = E(X)$.

Similarly:

$$M_X''(t) = E[X^2 e^{tX}] \implies M_X''(0) = E[X^2]$$

$$M_X'''(t) = E[X^3 e^{tX}] \implies M_X'''(0) = E[X^3]$$

$$etc$$

Question

Calculate the mean and variance of a random variable, X, with MGF given by:

$$M_X(t) = \left(1 - \tfrac{t}{5} \right)^{-1} \qquad t < 5$$

Solution

Differentiating the MGF and substituting $t = 0$ into the resulting expressions gives:

$$M_X'(t) = \tfrac{1}{5} \left(1 - \tfrac{t}{5} \right)^{-2} \implies E(X) = M_X'(0) = \tfrac{1}{5}$$

$$M_X''(t) = \frac{2}{25}\left(1 - \frac{t}{5}\right)^{-3} \quad \Rightarrow \quad E(X^2) = M_X''(0) = \frac{2}{25}$$

$$\Rightarrow \quad \text{var}(X) = E(X^2) - E^2(X) = \frac{2}{25} - \left(\frac{1}{5}\right)^2 = \frac{1}{25}$$

We now look at an alternative method that uses a series expansion of the MGF. Although it might, at first glance, appear to be long-winded it can be useful if differentiation is particularly complicated.

Expanding the exponential function and taking expected values throughout (a procedure which is justifiable for the distributions here) gives:

$$\boldsymbol{M_X(t)} = E(e^{tX}) = E\left(1 + tX + \frac{t^2}{2!}X^2 + \frac{t^3}{3!}X^3 + \cdots\right)$$

$$= \boldsymbol{1 + tE[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \cdots}$$

from which it is seen that the r th moment of the distribution about the origin, $\boldsymbol{E[X^r]}$, is obtainable as the coefficient of $\dfrac{t^r}{r!}$ in the power series expansion of the MGF.

To use this method to find moments, we need to obtain a series expansion of the MGF. We then equate the coefficients of the powers of t with the above expression.

Question

Use a series expansion to derive $E(X), E(X^2)$ and $E(X^3)$, where the MGF of X is given by:

$$M_X(t) = \left(1 - \frac{t}{5}\right)^{-1} \qquad t < 5$$

Solution

Using the binomial expansion given on page 2 of the *Tables*:

$$M_X(t) = \left(1 - \frac{t}{5}\right)^{-1}$$

$$= 1 + (-1) \times \left(-\frac{t}{5}\right) + \frac{-1 \times -2}{2!}\left(-\frac{t}{5}\right)^2 + \frac{-1 \times -2 \times -3}{3!}\left(-\frac{t}{5}\right)^3 + \cdots$$

$$= 1 + \frac{1}{5}t + \frac{1}{25}t^2 + \frac{1}{125}t^3 + \cdots$$

Now the MGF can also be written as:

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots$$

Equating the coefficients gives:

$$E(X) = \frac{1}{5}$$

$$\frac{1}{2!}E(X^2) = \frac{1}{25} \implies E(X^2) = \frac{2}{25}$$

$$\frac{1}{3!}E(X^3) = \frac{1}{125} \implies E(X^3) = \frac{6}{125}$$

If we differentiate the series expansion for the MGF with respect to t and then substitute $t = 0$ this gives $M_X'(0) = E(X), M_X''(0) = E(X^2), \ldots$ as before.

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots$$

$$\implies M_X'(t) = E(X) + tE(X^2) + \frac{t^2}{2!}E(X^3) + \cdots \implies M_X'(0) = E(X)$$

$$\implies M_X''(t) = E(X^2) + tE(X^3) + \cdots \implies M_X''(0) = E(X^2) \qquad etc$$

The Uniqueness property of MGFs

If the distribution of a random variable X is known, in theory at least, all moments of the distribution that exist can be calculated. If the moments are specified, then the distribution can be identified.

Without going deeply into mathematical rigour, it can in fact be said that if all moments of a random variable exist (and if they satisfy a certain convergence condition) then the sequence of moments uniquely determines the distribution of X.

Further, if a moment generating function has been found, then there is a unique distribution with that MGF. Thus an MGF can be recognised as the MGF of a particular distribution. (There is a one-to-one correspondence between MGFs and distributions with MGFs).

This 'uniqueness property' will be used in a number of proofs in future chapters.

Question

A random variable, X, has MGF given by $M_X(t) = \exp\{5t + 3t^2\}$.

Use the MGFs listed in the *Tables* and the 'uniqueness property' to identify the distribution of X.

Solution

Examining the MGFs given in the *Tables* we want one that involves an exponential term. The normal distribution has the following MGF:

$$M(t) = \exp\{\mu t + \tfrac{1}{2}\sigma^2 t^2\}$$

Equating coefficients, we see that $\mu = 5$ and $\sigma^2 = 6$. By the uniqueness property since X has the same MGF as a $N(5,6)$ distribution, it means that X has the normal distribution with mean 5 and variance 6.

We can also identify a distribution by the series expansion of its MGF.

Question

Identify the continuous distribution for which $E[X^k] = \dfrac{k!}{\lambda^k}$ where $k = 1, 2, 3, \ldots$, and $\lambda > 0$.

Solution

The moment generating function of X is:

$$M_X(t) = 1 + tE[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \cdots$$

Substituting in the values of the moments given:

$$M_X(t) = 1 + \frac{1}{\lambda}t + \frac{2!}{\lambda^2}\frac{t^2}{2!} + \frac{3!}{\lambda^3}\frac{t^3}{3!} + \cdots = 1 + \frac{t}{\lambda} + \frac{t^2}{\lambda^2} + \frac{t^3}{\lambda^3} + \cdots$$

This is $(1 - t/\lambda)^{-1}$. By comparing this to standard MGFs we can see that the distribution is the exponential distribution with parameter λ.

1.2   Important examples – discrete distributions

The MGFs for some of the distributions introduced earlier are found as follows.

Discrete Uniform

The probability function for the discrete uniform distribution on the integers 1, 2, ..., k is:

$$P(X = x) = 1/k\,, \quad x = 1, 2, 3, \dots, k$$

So the MGF is:

$$M_X(t) = E(e^{tX}) = (1/k)(e^t + e^{2t} + \cdots + e^{kt})$$
$$= (e^t/k)(1 - e^{kt})/(1 - e^t) \qquad \text{for } t \neq 0$$

Binomial (n, p) (including Bernoulli, for which $n = 1$)

The probability function for the *Bin*(n, p) distribution is:

$$P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}\,, \quad x = 1, 2, \dots, n$$

So the moment generating function is:

$$M_X(t) = \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x q^{n-x} = (q + pe^t)^n$$

Negative binomial (k, p) (including geometric, for which $k = 1$)

The probability function is:

$$P(X = x) = \binom{x-1}{k-1}p^k q^{x-k}\,, \quad x = k, k+1, k+2, \dots$$

So the MGF is:

$$M_X(t) = \sum_{x=k}^{\infty} \binom{x-1}{k-1} e^{tx} p^k q^{x-k}$$

$$= (pe^t)^k \sum_{x=k}^{\infty} \binom{x-1}{k-1} (qe^t)^{x-k}$$

$$= (pe^t)^k (1 - qe^t)^{-k}$$

$$= [pe^t / (1 - qe^t)]^k$$

Note: The summation is valid for $\left| qe^t \right| < 1$, *ie* for $t < \ln(1/q)$.

Hypergeometric

MGF not used.

Poisson (λ)

The probability function is:

$$P(X = x) = \lambda^x \exp(-\lambda) / x!, \quad x = 0, 1, 2, 3, \dots$$

So the MGF is:

$$M_X(t) = e^{-\lambda} \sum_{x=0}^{\infty} (\lambda e^t)^x / x! = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

1.3   Important examples – continuous variables

We will now look at how to calculate the MGF of some standard continuous distributions. So here we will be integrating to obtain the MGF.

Uniform (*a*, *b*)

Multiplying the PDF by e^{tx} and integrating:

$$M_X(t) = \int_a^b e^{tx} \frac{1}{b-a} \, dx = \frac{e^{bt} - e^{at}}{t(b-a)}$$

Gamma (α, λ)

Integrate $e^{tx} f(x)$ from 0 to ∞.

This gives:

$$M_X(t) = \int_0^{\infty} e^{tx} \frac{\lambda^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} dx = \int_0^{\infty} \frac{\lambda^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)} e^{-(\lambda-t)x} dx$$

Writing out the integral and substituting $y = (\lambda - t)x$ **, so that** $\dfrac{dy}{dx} = \lambda - t$ **, we have:**

$$M_X(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{1}{\lambda - t}\right)^\alpha y^{\alpha-1} e^{-y} \ dy$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\lambda - t}\right)^\alpha \Gamma(\alpha)$$

$$= \left(\frac{\lambda}{\lambda - t}\right)^\alpha$$

In the second line we've used the definition of the gamma function. This is given on page 5 of the *Tables*.

So $M_X(t) = \left(\dfrac{\lambda}{\lambda - t}\right)^\alpha = \left(\dfrac{1}{1 - t/\lambda}\right)^\alpha = \left(1 - \dfrac{t}{\lambda}\right)^{-\alpha}$.

This formula only holds when $t < \lambda$ and is given on page 12 of the *Tables*.

Question

Describe what happens if we try to evaluate $E\left(e^{tX}\right)$ for the gamma distribution when $t \geq \lambda$.

Solution

$$M_X(t) = E\left(e^{tX}\right) = \int_0^\infty \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-(\lambda - t)x} dx \ .$$

If $t \geq \lambda$, then the power in the exponential factor in the integral is positive and therefore the answer is infinite. So the MGF does not exist in this case.

From this:

$$M_X'(t) = \alpha \lambda^\alpha (\lambda - t)^{-\alpha-1} \text{ so } E[X] = M_X'(0) = \frac{\alpha}{\lambda}$$

$$M_X''(t) = \alpha(\alpha + 1)\lambda^\alpha (\lambda - t)^{-\alpha-2} \text{ so } E[X^2] = M_X''(0) = \frac{\alpha(\alpha + 1)}{\lambda^2}$$

Hence, $\mu = \dfrac{\alpha}{\lambda}$ **,** $\sigma^2 = \dfrac{\alpha(\alpha + 1)}{\lambda^2} - \left(\dfrac{\alpha}{\lambda}\right)^2 = \dfrac{\alpha}{\lambda^2}$.

It follows that the MGF of the exponential distribution with mean θ is given by:

$$M_X(t) = (1 - \theta t)^{-1}$$

Remember that the exponential distribution is a special case of the gamma distribution when $\alpha = 1$. The mean is $\theta = \dfrac{1}{\lambda}$.

Note: The MGF of the chi-square υ distribution is given by $M_X(t) = (1 - 2t)^{-\upsilon/2}$.

Question

Show that this is true.

Solution

χ_υ^2 is gamma with $\alpha = \dfrac{\upsilon}{2}$ and $\lambda = \dfrac{1}{2}$. So it has moment generating function:

$$M_X(t) = \frac{\left(\dfrac{1}{2}\right)^{\frac{\upsilon}{2}}}{\left(\dfrac{1}{2} - t\right)^{\frac{\upsilon}{2}}} = \left(\frac{\dfrac{1}{2}}{\dfrac{1}{2} - t}\right)^{\frac{\upsilon}{2}} = \left(\frac{1}{1 - 2t}\right)^{\frac{\upsilon}{2}} = \left(1 - 2t\right)^{-\frac{\upsilon}{2}}$$

Normal (μ, σ^2)

The two crucial steps in evaluating the integral to obtain the MGF for the normal distribution are (i) completing the square in the exponent, and (ii) recognising that the resulting integral is simply that of a normal density and hence equal to 1. The derivation is not given in the Core Reading, but is detailed in the following question.

The result is:

$$M_X(t) = \exp(\mu t + \tfrac{1}{2}\sigma^2 t^2)$$

Question

Prove this result.

Solution

The moment generating function of the $N(\mu, \sigma^2)$ distribution is given by:

$$\int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] dx$$

First we need to complete the square:

$$M_X(t) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[tx - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right] dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left(x^2 - 2\mu x - 2tx\sigma^2 + \mu^2 \right) \right] dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left(x^2 - 2x(\mu + t\sigma^2) + \mu^2 \right) \right] dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left((x - (\mu + t\sigma^2))^2 + \mu^2 - (\mu + t\sigma^2)^2 \right) \right] dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x - (\mu + t\sigma^2))^2 \right] \exp\left[-\frac{1}{2\sigma^2}(-2\mu t\sigma^2 - t^2\sigma^4) \right] dx$$

Since the second factor in the integral does not depend on x, we can take it outside the integral:

$$M_X(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(-2\mu t\sigma^2 - t^2\sigma^4) \right] \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x - (\mu + t\sigma^2))^2 \right] dx$$

$$= \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2 \right] \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - (\mu + t\sigma^2)}{\sigma} \right)^2 \right] dx$$

The function now being integrated is the PDF of a normal distribution with mean $\mu + t\sigma^2$ and standard deviation σ, so the integral must be 1, giving us:

$$M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2 \right], \text{ as required}$$

We can also check the moments of the normal distribution.

Expanding:

$$M_X(t) = 1 + \left(\mu t + \frac{1}{2}\sigma^2 t^2 \right) + \frac{\left(\mu t + \frac{1}{2}\sigma^2 t^2 \right)^2}{2!} + \cdots$$

$$E[X] = \textbf{coefficient of } t = \mu$$

(confirming that the parameter μ does indeed represent the mean).

$E[X^2]$ **= coefficient of** $\dfrac{t^2}{2!} = \sigma^2 + \mu^2$ **so** $\text{var}[X] = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$ **(confirming that the parameter σ does indeed represent the standard deviation).**

Alternatively, we could differentiate the MGF to obtain the mean and variance. However, the series method is actually quicker in this case.

By setting $\mu = 0$ and $\sigma^2 = 1$, we can see that:

The standard normal random variable Z has MGF:

$$M_Z(t) = \exp(\tfrac{1}{2}t^2) = 1 + \frac{1}{2}t^2 + \frac{\left(\frac{1}{2}t^2\right)^2}{2!} + \cdots$$

Hence $E[Z] = 0$, $E[Z^2] = 1$, $E[Z^3] = 0$, $E[Z^4] = 3$ (coefficient of $t^4/4!$), ...

Now $X = \sigma Z + \mu$, and it follows that $E\left[(X - \mu)^3\right] = 0$, $E\left[(X - \mu)^4\right] = 3\sigma^4$.

Remember that $E\left[(X - \mu)^3\right]$ is the skewness. We stated earlier that the normal distribution is symmetrical, hence we expect $E\left[(X - \mu)^3\right]$ to be zero. This has now been proved.

Question

In this last result, we have used the fact that if we standardise a normal random variable X by setting $Z = \dfrac{X - \mu}{\sigma}$, then Z has the standard normal distribution. Use moment generating functions to show that this is true.

Solution

The MGF of $Z = \dfrac{X - \mu}{\sigma}$ is:

$$M_Z(t) = E[e^{tz}] = E[e^{t\left(\frac{X-\mu}{\sigma}\right)}] = e^{-\frac{\mu t}{\sigma}} E[e^{\frac{t}{\sigma}X}] = e^{-\frac{\mu t}{\sigma}} M_X\left(\frac{t}{\sigma}\right)$$

Using the formula for the MGF of the $N(\mu, \sigma^2)$ distribution gives:

$$M_Z(t) = e^{-\frac{\mu t}{\sigma}} e^{\frac{\mu t}{\sigma} + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma}\right)^2} = e^{\frac{1}{2}t^2}$$

which we recognise as the MGF of $N(0,1)$. So, using the uniqueness property of MGFs, we can conclude that $\dfrac{X - \mu}{\sigma}$ has a standard normal distribution.

The MGFs do not exist in closed form for the Beta and lognormal distributions. Hence, they are excluded from this section.

2   Cumulant generating functions

For many random variables the cumulant generating function (CGF) is easier to use than the MGF in evaluating the mean and variance.

Definition

The cumulant generating function, $C_X(t)$, of a random variable X is given by:

$$C_X(t) = \ln M_X(t)$$

We can treat this as the definition of the CGF.

Question

The MGF of the $Bin(n, p)$ distribution is given by:

$$M(t) = \left(q + pe^t\right)^n$$

State the CGF of the $Bin(n, p)$ distribution.

Solution

$$C_X(t) = \ln M_X(t) = \ln(q + pe^t)^n = n\ln(q + pe^t)$$

As a result if $C_X(t)$ is known it is easy to determine $M_X(t)$.

We have $M_X(t) = e^{C_X(t)}$.

Calculating moments

The first three derivatives of $C_X(t)$ evaluated at $t = 0$ give the mean, variance and skewness of X directly.

These results can be proved as follows:

$$C_X'(t) = \frac{M_X'(t)}{M_X(t)}$$

$$C_X''(t) = \frac{M_X''(t)M_X(t) - (M_X'(t))^2}{(M_X(t))^2}$$

and $$C_X'''(t) = \frac{M_X'''(t)(M_X(t))^3 - 3(M_X(t))^2 M_X'(t)M_X''(t) + 2M_X(t)(M_X'(t))^3}{(M_X(t))^4}$$

Now $M_X(0) = 1$ so

$$C_X'(0) = \frac{M_X'(0)}{M_X(0)} = \frac{E[X]}{1}$$

$$C_X''(0) = \frac{M_X''(0)M_X(0) - (M_X'(0))^2}{M_X^2(0)} = \frac{E[X^2](1) - (E[X])^2}{1^2} = var[X] ;$$

and $C_X'''(0) = \dfrac{M_X'''(0)(M_X(0))^3 - 3(M_X(0))^2 M_X'(0)M_X''(0) + 2M_X(0)(M_X'(0))^3}{(M_X(0))^4}$

$$= \frac{E[X^3](1)^3 - 3(1)^2 E[X]E^2[X] + 2(1)E^3[X]}{1^4}$$

$$= skew(X)$$

Question

State the CGF of X where $X \sim Gamma(\alpha, \lambda)$. Hence prove that $E(X) = \dfrac{\alpha}{\lambda}$, $var(X) = \dfrac{\alpha}{\lambda^2}$ and

$skew(X) = \dfrac{2\alpha}{\lambda^3}$.

Solution

$$M_X(t) = \frac{\lambda^\alpha}{(\lambda - t)^\alpha} = \left(1 - \frac{t}{\lambda}\right)^{-\alpha} \quad \Rightarrow \quad C_X(t) = -\alpha \ln\left(1 - \frac{t}{\lambda}\right) \qquad t < \lambda$$

Differentiating with respect to t, and substituting $t = 0$, we obtain:

$$C_X'(t) = -\alpha \times \frac{-\frac{1}{\lambda}}{\left(1 - \frac{t}{\lambda}\right)} = \frac{\alpha}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-1} \qquad \Rightarrow \quad E(X) = C_X'(0) = \frac{\alpha}{\lambda}$$

$$C_X''(t) = -\frac{\alpha}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-2} \times -\frac{1}{\lambda} = \frac{\alpha}{\lambda^2}\left(1 - \frac{t}{\lambda}\right)^{-2} \quad \Rightarrow \quad var(X) = C_X''(0) = \frac{\alpha}{\lambda^2}$$

$$C_X'''(t) = -\frac{2\alpha}{\lambda^2}\left(1 - \frac{t}{\lambda}\right)^{-3} \times -\frac{1}{\lambda} = \frac{2\alpha}{\lambda^3}\left(1 - \frac{t}{\lambda}\right)^{-3} \quad \Rightarrow \quad skew(X) = C_X'''(0) = \frac{2\alpha}{\lambda^3}$$

The coefficient of $\dfrac{t^r}{r!}$ in the Maclaurin series of $C_X(t) = \ln M_X(t)$ is called the r th cumulant and is denoted by κ_r.

Another method for finding the cumulants is to differentiate the CGF with respect to t and then set $t = 0$. The r th derivative then gives the r th cumulant, κ_r.

So:

$$\kappa_1 = C_X'(0)$$

$$\kappa_2 = C_X''(0)$$

$$\kappa_3 = C_X'''(0) \qquad etc$$

Cumulants are similar to moments. The first three cumulants are the mean, the variance and the skewness.

Question

By using the CGF of the $Poi(\mu)$ distribution, derive the 2nd, 3rd and 4th cumulants.

Solution

For the Poisson distribution:

$$M_X(t) = e^{\mu(e^t - 1)} \quad \Rightarrow \quad C_X(t) = \ln M_X(t) = \mu(e^t - 1)$$

Differentiating and setting $t = 0$ we obtain:

$$C_X'(t) = \mu e^t$$

$$C_X''(t) = \mu e^t \quad \Rightarrow \quad \kappa_2 = C_X''(0) = \mu e^0 = \mu$$

$$C_X'''(t) = \mu e^t \quad \Rightarrow \quad \kappa_3 = C_X'''(0) = \mu$$

$$C_X''''(t) = \mu e^t \quad \Rightarrow \quad \kappa_4 = C_X''''(0) = \mu$$

So the 2nd, 3rd and 4th cumulants of the Poisson distribution are all equal to μ. In fact, all the cumulants are the same.

We can see that the CGF is particularly useful when the MGF is an exponential function, as it makes the differentiation a lot easier.

3 Linear functions

Suppose X has MGF $M_X(t)$ and the distribution of a linear function $Y = a + bX$ is of interest. The MGF of Y, $M_Y(t)$ say, can be obtained from that of X as follows:

$$M_Y(t) = E[e^{tY}] = E[e^{t(a+bX)}] = e^{at} E[e^{btX}] = e^{at} M_X(bt)$$

Question

Use MGFs to show that if X has a $Gamma(\alpha, \lambda)$ distribution, then $2\lambda X$ has a $\chi^2_{2\alpha}$ distribution. Hence, if X has the ($Gamma(20, 0.4)$) distribution, estimate the probability that $X > 75$.

Solution

The MGF of the $Gamma(\alpha, \lambda)$ distribution is $M_X(t) = \left(1 - \dfrac{t}{\lambda}\right)^{-\alpha}$.

If $Y = a + bX$, then $M_Y(t) = e^{at} M_X(bt)$.

In this question, $a = 0$, and $b = 2\lambda$, so:

$$M_Y(t) = M_X(2\lambda t) = \left(1 - \frac{2\lambda t}{\lambda}\right)^{-\alpha} = (1 - 2t)^{-\alpha}$$

This is the moment generating function of the chi-square distribution with 2α degrees of freedom, so by the uniqueness of MGFs, we can say that $2\lambda X$ has a $\chi^2_{2\alpha}$ distribution.

If X has the $Gamma(20, 0.4)$ distribution, then this result tells us that $0.8X \sim \chi^2_{40}$, which gives:

$$P(X > 75) = P(0.8X > 60) = P(\chi^2_{40} > 60)$$

From the percentage points table of the χ^2_{40} distribution, we see that this probability is just less than 0.025.

Question

If X has the gamma distribution with parameters $\alpha = 2$ and $\lambda = 0.4$, find by direct integration $P(X > 10)$.

Solution

The PDF of X is:

$$f_X(x) = \frac{1}{\Gamma(2)} 0.4^2 x e^{-0.4x} = 0.16 x e^{-0.4x}, \qquad x \geq 0$$

Integrating the PDF using integration by parts:

$$P(X \leq 10) = \int\limits_{0}^{10} 0.16 x e^{-0.4x} dx = \left[0.16x \left(\frac{e^{-0.4x}}{-0.4} \right) \right]_0^{10} - \int\limits_{0}^{10} 0.16 \frac{e^{-0.4x}}{-0.4} \, dx$$

$$= \frac{1.6e^{-4}}{-0.4} + \left[\frac{0.16}{0.4} \times \frac{e^{-0.4x}}{-0.4} \right]_0^{10} = -4e^{-4} + \left(-e^{-4} - (-1) \right) = 1 - 5e^{-4}$$

So we have:

$$P(X > 10) = 1 - \left(1 - 5e^{-4} \right) = 5e^{-4} = 0.09158$$

We can check this result by obtaining $P(\chi_4^2 > 8)$ using page 165 of the Tables. Alternatively we can obtain $P(X > 10)$ directly by integrating between the limits of 10 and ∞.

This method rapidly becomes tedious (or impossible) for values of α other than very small integers.

We can also obtain the CGF of a linear function.

Question

If $Y = a + bX$, derive and simplify an expression for $C_Y(t)$ in terms of $C_X(t)$.

Solution

Since $C_Y(t) = \ln M_Y(t)$, using the expression for $M_Y(t)$, we have:

$$C_Y(t) = \ln M_Y(t) = \ln[e^{at} M_X(bt)] = at + \ln M_X(bt) = at + C_X(bt)$$

4   Further applications of generating functions

Generating functions can be used to establish the distribution of linear combinations of random variables. This will be covered in detail in a later chapter.

A linear combination of the random variables $X_1, \dots, X_n$ is an expression of the form:

$$c_1 X_1 + \cdots + c_n X_n$$

where $c_1, \dots, c_n$ are constants.

We can use MGFs (or CGFs) to obtain the distribution of such a linear combination. For example, we can show that if $X_1 \sim Poi(\mu_1)$ and $X_2 \sim Poi(\mu_2)$, and X_1 and X_2 are independent, then $X_1 + X_2$ has a $Poi(\mu_1 + \mu_2)$ distribution. We will prove results such as this later in the course.

Moment generating functions can also be used to calculate moments for and specify compound distributions. This will be covered in detail in Subject CS2.

Chapter 2 Summary

Generating functions are used to make it easier to find moments of distributions.

The moment generating function (MGF) of a random variable is defined to be:

$$M_X(t) = E\left[e^{tX} \right]$$

The series expansion for MGFs is:

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots$$

The formulae for the mean and variance are:

$$E(X) = M_X'(0) \qquad\qquad var(X) = M_X''(0) - \left[M_X'(0) \right]^2$$

The cumulant generating function (CGF) of a random variable is defined to be:

$$C_X(t) = \ln M_X(t)$$

The formulae for the moments are:

$$E(X) = C_X'(0) \qquad var(X) = C_X''(0) \qquad skew(X) = C_X'''(0)$$

The uniqueness property means that if two variables have the same MGF or CGF then they have the same distribution.

If $Y = a + bX$, then:

$$M_Y(t) = e^{at}M_X(bt) \text{ and } C_Y(t) = at + C_X(bt)$$

The questions start on the next page so that you can keep
all the chapter summaries together for revision purposes.

Chapter 2 Practice Questions

2.1   (i)   Determine the moment generating function of the two parameter exponential random
              variable X, defined by the probability density function

Exam style

$$f(x) = \lambda e^{-\lambda(x-\alpha)}, \qquad x \geq \alpha \qquad \text{where } \lambda, \alpha > 0.$$ [3]

       (ii)   Hence, or otherwise, determine the mean and variance of the random variable X.   [4]
                                                                                              [Total 7]

2.2   Derive from first principles the moment generating function of a random variable X, where
       $P(X = x) = \theta(1-\theta)^{x-1} \quad x = 1, 2, 3, \ldots$.

2.3   Determine the cumulant generating function of the $N\left(\mu, \sigma^2\right)$ distribution, and hence determine
       the mean, variance and coefficient of skewness of this distribution.

2.4   (i)   If the moment generating function of X is $M_X(t)$, then derive an expression for the
              moment generating function of $2X + 3$ in terms of $M_X(t)$.   [2]

Exam style

       (ii)   Hence, if X is normally distributed with mean μ and variance σ^2, derive the distribution
              of $2X + 3$.   [2]
                                                                                              [Total 4]

2.5   The moment generating function, $M_Y(t)$, of a random variable, Y, is given by:

Exam style

$$M_Y(t) = \left(1 - 4t\right)^{-2} \qquad t < 0.25$$

       Calculate:

       (i) $E(Y)$ [1]

       (ii)   the standard deviation of Y [2]

       (iii) $E(Y^6)$.   [2]
                                                                                              [Total 5]

2.6   The random variable U has a geometric distribution with probability function:

Exam style

$$P(U = u) = pq^{u-1} \quad u = 1, 2, 3, \ldots \qquad \text{where} \quad p + q = 1$$

       (i)   Derive the moment generating function of U.   [2]

       (ii)   Write down the CGF of U, and hence show that $E(U) = 1/p$.   [3]
                                                                                              [Total 5]

2.7   A random variable X has probability density function:

$$f(x) = ke^{-2x} \qquad x > R$$

where R and k are positive constants.

(i)   (a)   Derive a formula for the moment generating function of X.

       (b)   State the values of t for which the formula in (i)(a) is valid.   [4]

(ii)   Hence determine the value of the constant k in terms of R.   [1]
                                                                              [Total 5]

2.8   (i)   Derive, from first principles, the moment generating function of a $Gamma(\alpha, \lambda)$

       distribution.   [3]

(ii)   Use this moment generating function to show that the mean and variance are α/λ and

       α/λ^2, respectively.   [2]
                                                                              [Total 5]

2.9 X is normally distributed with mean μ and variance σ^2. Use generating functions to determine

the fourth central moment of X.   [3]

2.10   The claim amount X in units of £1,000 for a certain type of industrial policy is modelled as a

gamma variable with parameters $\alpha = 3$ and $\lambda = 1/4$.

(i)   Use moment generating functions to show that $\frac{1}{2}X \sim \chi_6^2$.   [3]

(ii)   Hence use the *Tables* to calculate the probability that a claim amount exceeds £20,000. [2]
                                                                              [Total 5]

Chapter 2 Solutions

2.1   (i)   Using the definition of an MGF:

$$M_X(t) = E\left[e^{tX}\right] = \int_\alpha^\infty e^{tx} \lambda e^{-\lambda(x-\alpha)} \, dx$$

$$= \int_\alpha^\infty \lambda e^{\lambda\alpha} e^{-(\lambda-t)x} \, dx$$

$$= \left[-\frac{\lambda e^{\lambda\alpha}}{\lambda-t} e^{-(\lambda-t)x} \right]_\alpha^\infty$$

$$= \frac{\lambda}{\lambda-t} e^{t\alpha} \qquad \text{provided } t < \lambda \qquad [3]$$

(ii)   Re-writing the MGF to make it easier to differentiate:

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-1} e^{t\alpha}$$

$$M_X'(t) = \frac{1}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-2} e^{t\alpha} + \alpha\left(1 - \frac{t}{\lambda}\right)^{-1} e^{t\alpha}$$

$$\Rightarrow \quad E(X) = M_X'(0) = \frac{1}{\lambda} + \alpha \qquad [2]$$

$$M_X''(t) = \frac{2}{\lambda^2}\left(1 - \frac{t}{\lambda}\right)^{-3} e^{t\alpha} + \frac{2\alpha}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-2} e^{t\alpha} + \alpha^2\left(1 - \frac{t}{\lambda}\right)^{-1} e^{t\alpha}$$

$$\Rightarrow \quad E(X^2) = M_X''(0) = \frac{2}{\lambda^2} + \frac{2\alpha}{\lambda} + \alpha^2$$

$$\Rightarrow \quad \text{var}(X) = \frac{2}{\lambda^2} + \frac{2\alpha}{\lambda} + \alpha^2 - \left(\frac{1}{\lambda} + \alpha\right)^2 = \frac{1}{\lambda^2} \qquad [2]$$

2.2   The MGF is given by:

$$M_X(t) = E\left(e^{tX}\right) = \sum_{x=1}^\infty e^{tx} P(X = x) = \sum_{x=1}^\infty e^{tx} \theta(1-\theta)^{x-1}$$

$$= \theta\left(e^t + e^{2t}(1-\theta) + e^{3t}(1-\theta)^2 + \cdots\right)$$

The expression in the brackets is an infinite geometric series with $a = e^t$ and $r = e^t(1-\theta)$. Summing it gives:

$$M_X(t) = \frac{\theta e^t}{1-(1-\theta)e^t} \qquad \text{where } -1 < e^t(1-\theta) < 1 \implies t < \ln\left(\frac{1}{1-\theta}\right)$$

2.3   For the normal distribution:

$$M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]$$

$$\Rightarrow C_X(t) = \ln M_X(t) = \mu t + \frac{1}{2}\sigma^2 t^2$$

Differentiating and setting $t = 0$ gives:

$$C_X'(t) = \mu + \sigma^2 t \quad \Rightarrow \quad E(X) = C_X'(0) = \mu$$

$$C_X''(t) = \sigma^2 \qquad \Rightarrow \quad \text{var}(X) = C_X''(0) = \sigma^2$$

$$C_X'''(t) = 0 \qquad \Rightarrow \quad Skew(X) = C_X'''(0) = 0$$

Since the skewness is zero, the coefficient of skewness is also 0.

2.4   (i)   **MGF**

The MGF of X is:

$$M_X(t) = E(e^{tX})$$

So the MGF of $2X + 3$ is:

$$M_{2X+3}(t) = E[e^{t(2X+3)}] = E[e^{(2t)X+3t}] = e^{3t}E[e^{(2t)X}] = e^{3t}M_X(2t) \qquad [2]$$

(ii)   **Distribution**

The MGF for a $N(\mu, \sigma^2)$ is $e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. Using the formula derived in part (i):

$$M_{2X+3}(t) = e^{3t}M_X(2t) = e^{3t}e^{2\mu t + 2\sigma^2 t^2} = e^{(2\mu+3)t + \frac{1}{2}(4\sigma^2)t^2}$$

This is the MGF of a $N(2\mu+3, 4\sigma^2)$ distribution. Therefore by the uniqueness property of MGFs, $2X + 3$ has a $N(2\mu+3, 4\sigma^2)$ distribution. [2]

2.5   **(i)**   ***Expectation***

$$M_Y'(t) = 8(1-4t)^{-3} \implies E(Y) = M_X'(0) = 8 \qquad [1]$$

(ii)   ***Standard deviation***

$$M_Y''(t) = 96(1-4t)^{-4} \implies E(Y^2) = 96$$

$$\implies \text{var}(Y) = 96 - 8^2 = 32 \implies \text{standard deviation} = \sqrt{32} = 5.6569 \qquad [2]$$

(iii)   ***Sixth moment***

Recall that $E(Y^6)$ is the coefficient of $\dfrac{t^6}{6!}$ in the expansion of $M_Y(t)$. From the binomial

expansion of the MGF, $(1-4t)^{-2}$, (using the formula given on page 2 of the *Tables*) the term is:

$$\frac{-2 \times -3 \times -4 \times -5 \times -6 \times -7}{6!}(-4t)^6 \qquad [1]$$

Hence, $E(Y^6) = -2 \times -3 \times -4 \times -5 \times -6 \times -7 \times (-4)^6 = 20{,}643{,}840$. $\qquad [1]$

Alternatively, we can use $E(Y^6) = M_Y^{(6)}(0)$ but this requires us to differentiate the MGF six times.

2.6   **(i)**   ***MGF***

The MGF of U is:

$$M_U(t) = E\left[e^{tU}\right] = \sum_{u=1}^{\infty} e^{tu} P(U = u) = \sum_{u=1}^{\infty} e^{tu} pq^{u-1}$$

$$= pe^t + pqe^{2t} + pq^2 e^{3t} + \cdots \qquad [1]$$

This is an infinite geometric series with $a = pe^t$ and $r = qe^t$ so using the formula $S_\infty = \dfrac{a}{1-r}$ gives:

$$M_U(t) = \frac{pe^t}{1 - qe^t} \qquad [1]$$

(iii)   ***CGF and mean***

We have:

$$C_U(t) = \ln\left(\frac{pe^t}{1-qe^t}\right) = \ln p + t - \ln\left(1 - qe^t\right) \qquad [1]$$

Differentiating the CGF:

$$C_U'(t) = 1 - \left(\frac{-qe^t}{1-qe^t} \right) = 1 + \frac{qe^t}{1-qe^t}$$

[1]

Substituting in $t = 0$:

$$E(U) = C_U'(0) = 1 + \frac{q}{1-q} = \frac{1}{1-q} = \frac{1}{p}$$

[1]

2.7   (i)(a)   **MGF**

The MGF of X is:

$$M_X(t) = E[e^{tX}] = \int_R^\infty e^{tx} k e^{-2x} dx$$

[1]

$$= k \int_R^\infty e^{-(2-t)x} dx = k \left[\frac{e^{-(2-t)x}}{-(2-t)} \right]_R^\infty$$

[1]

$$= \frac{k e^{-(2-t)R}}{(2-t)}$$

[1]

(i)(b)   **Values of t for which valid**

The integral converges as $x \to \infty$ only if $2-t$ is positive. So the MGF is valid for $t < 2$.   [1]

(ii)   **Evaluate k**

Putting $t = 0$ gives $M_X(0) = \frac{1}{2} k e^{-2R}$.   [½]

Since $M_X(0)$ must equal 1, this tells us that $k = 2e^{2R}$.   [½]

2.8   (i)   **MGF**

$$M_X(t) = E\left(e^{tX}\right) = \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx$$

[1]

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x} dx$$

The integral looks like the PDF of a $Gamma(\alpha, \lambda - t)$, so putting in the appropriate constants:

$$M_X(t) = \frac{\lambda^\alpha}{(\lambda - t)^\alpha} \int_0^\infty \frac{(\lambda - t)^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-(\lambda - t)x} \, dx$$

$$= \frac{\lambda^\alpha}{(\lambda - t)^\alpha} \qquad \text{provided } t < \lambda$$

$$= \left(\frac{\lambda}{\lambda - t} \right)^\alpha = \left(\frac{\lambda - t}{\lambda} \right)^{-\alpha} = \left(1 - \frac{t}{\lambda} \right)^{-\alpha} \qquad [2]$$

Since the integral of a Gamma PDF over the whole range is 1.

Alternatively, we can use the substitution method.

(ii)   ***Mean and variance***

Using the results $E(X) = M_X'(0)$ and $E(X^2) = M_X''(0)$:

$$M_X'(t) = \frac{\alpha}{\lambda} \left(1 - \frac{t}{\lambda} \right)^{-\alpha - 1} \quad \Rightarrow \quad E(X) = M_X'(0) = \frac{\alpha}{\lambda} \qquad [1]$$

$$M_X''(t) = \frac{\alpha(\alpha + 1)}{\lambda^2} \left(1 - \frac{t}{\lambda} \right)^{-\alpha - 2} \quad \Rightarrow \quad E\left(X^2 \right) = M_X''(0) = \frac{\alpha(\alpha + 1)}{\lambda^2}$$

$$\text{var}(X) = \frac{\alpha(\alpha + 1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2} \qquad [1]$$

2.9   The MGF of $X - \mu$, which has a $N(0, \sigma^2)$ distribution, is:

$$M_{X-\mu}(t) = e^{\frac{1}{2}\sigma^2 t^2} = 1 + \frac{1}{2}\sigma^2 t^2 + \frac{1}{2}\left(\frac{1}{2}\sigma^2 t^2 \right)^2 + \cdots \qquad [1]$$

So $E[(X - \mu)^4]$ is the coefficient of $\frac{t^4}{4!}$ in this series, *ie*:

$$E[(X - \mu)^4] = 3\sigma^4 \qquad [2]$$

Alternatively, we can differentiate the MGF four times and substitute $t = 0$ each time to obtain $E(X), E(X^2), E(X^3)$ and $E(X^4)$. We then use the expansion of $E[(X - \mu)^4]$:

$$E[(X - \mu)^4] = E(X^4) - 4\mu E(X^3) + 6\mu^2 E(X^2) - 3\mu^4$$

It might be tempting to use the CGF − after all, it gives the second and third central moments $\text{var}(X) = C_X''(0)$ and $skew(X) = C_X'''(0)$. However, thereafter the CGF does **not** give central moments.

2.10 **(i)** *MGF*

We are given that $X \sim Gamma\left(3, \frac{1}{4}\right)$. From the *Tables*:

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-\alpha} = (1 - 4t)^{-3} \qquad \text{[1]}$$

Let $Y = \frac{1}{2}X$. Then:

$$M_Y(t) = E\left[e^{tY}\right] = E\left[e^{\frac{1}{2}(tX)}\right] = M_X\left(\frac{t}{2}\right)$$

$$= \left(1 - 4\left(\frac{t}{2}\right)\right)^{-3} = (1 - 2t)^{-3} \qquad \text{[1]}$$

So the moment generating function of $Y = \frac{1}{2}X$ is $(1 - 2t)^{-3}$.

By comparing this with the MGF of the gamma distribution in the *Tables*, we see that this is the MGF of a $Gamma(3, \frac{1}{2})$ distribution. Looking also at the definition of the chi-square distribution, we see that $Gamma(3, \frac{1}{2})$ is the definition of a chi-square distribution with 6 degrees of freedom.

By the uniqueness property of moment generating functions, therefore, we have shown that $\frac{1}{2}X \sim \chi_6^2$. $\qquad$ [1]

(ii) *The probability that a claim exceeds £20,000*

We are given that $X \sim Gamma\left(3, \frac{1}{4}\right)$ where X is the claim amount in units of £1,000.

We also know that $2\lambda X \sim \chi_{2\alpha}^2$, *ie* $\frac{1}{2}X \sim \chi_6^2$. Therefore, using the *Tables*:

$$P(X > 20) = P\left(\frac{1}{2}X > 10\right) = P\left(\chi_6^2 > 10\right) = 1 - 0.8753 = 0.1247 \qquad \text{[2]}$$

3

Joint distributions

Syllabus objectives

1.2   Independence, joint and conditional distributions, linear combinations of random variables

    1.2.1   Explain what is meant by jointly distributed random variables, marginal distributions and conditional distributions.

    1.2.2   Define the probability function/density function of a marginal distribution and of a conditional distribution.

    1.2.3   Specify the conditions under which random variables are independent.

    1.2.4   Define the expected value of a function of two jointly distributed random variables, the covariance and correlation coefficient between two variables, and calculate such quantities.

    1.2.5   Define the probability function/density function of the sum of two independent random variables as the convolution of two functions.

    1.2.6   Derive the mean and variance of linear combinations of random variables.

    1.2.7   Use generating functions to establish the distribution of linear combinations of independent random variables.

0   Introduction

As yet, we have only considered situations involving one random variable. In this chapter we will look at some general results involving two or more random variables.

This chapter is quite long, and contains a large amount of material. It may therefore be helpful to notice the parallels with the single random variable notation, in order to aid understanding of the overall structure of the chapter.

Firstly we will define a joint probability (density) function $P(X = x, Y = y)$ or $f(x, y)$. We will see how we can obtain a marginal distribution *ie* $P(X = x)$ or $f(x)$ from the joint distribution. Then we will look at conditional distributions $P(X = x | Y = y)$ or $f(x | y)$. The study of conditional distributions continues in the next chapter. It might be worth studying the next chapter with this one as the material in the two chapters is quite closely linked.

Given a joint distribution we are also able to work out the mean and variance of the marginal distributions, and the *covariance* of the joint distribution. We will also look at the *correlation coefficient*. This work will be continued in a later chapter, where we will attempt to estimate what the correlation is from a sample.

Finally, we will extend our work on MGFs from the previous chapter to combine distributions together. This will give us easier ways of obtaining results for the binomial, negative binomial and gamma distributions, amongst others.

1   Joint distributions

1.1   Joint probability (density) functions

Defining several random variables simultaneously on a sample space gives rise to a multivariate distribution. In the case of just two variables, it is a bivariate distribution.

Discrete case

To illustrate this for a pair of discrete variables, X and Y, the probabilities associated with the various values of (x, y) are as follows:

		x	
y	**1**	**2**	**3**
1	0.10	0.10	0.05
2	0.15	0.10	0.05
3	0.20	0.05	-
4	0.15	0.05	-

So, for example, $P(X = 3, Y = 1) = 0.05$, and $P(X = 1, Y = 3) = 0.20$. Note here that the comma means 'and', '&' or '$\cap$'.

The function $f(x, y) = P(X = x, Y = y)$ for all values of (x, y) is the (joint/bivariate) probability function of (X, Y) – it specifies how the total probability of 1 is divided up amongst the possible values of (x, y) and so gives the (joint/bivariate) probability distribution of (X, Y).

The requirements for a function to qualify as the probability function of a pair of discrete random variables are:

$f(x, y) \geq 0$ for all values of x and y in the domain

$$\sum_x \sum_y f(x, y) = 1$$

This parallels earlier results, where the probability function was $P(X = x)$ which had to satisfy $P(X = x) \geq 0$ for all values of x and $\sum_x P(X = x) = 1$.

For example, consider the discrete random variables M and N with joint probability function:

$$P(M=m, N=n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

Let's draw up a table showing the values of the joint probability function for M and N.

Starting with the smallest possible values of M and N, $P(M=1, N=1) = \frac{1}{35 \times 2^{-1}} = \frac{2}{35}$.

Calculating the joint probability for all combinations of M and N, we get the table shown below.

			M		
		1	2	3	4
	1	$\frac{2}{35}$	$\frac{4}{35}$	$\frac{6}{35}$	$\frac{8}{35}$
N	2	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{4}{35}$
	3	$\frac{1}{70}$	$\frac{1}{35}$	$\frac{3}{70}$	$\frac{2}{35}$

Question

Use the table of probabilities given above to calculate:

(i) $P(M=3, N=1 \text{ or } 2)$

(ii) $P(N=3)$

(iii) $P(M=2 \mid N=3)$.

Solution

(i)   Since the events $P(N=1)$ and $P(N=2)$ are mutually exclusive, we have:

$$P(M=3, N=1 \text{ or } 2) = P(M=3, N=1) + P(M=3, N=2) = \frac{6}{35} + \frac{3}{35} = \frac{9}{35}$$

(ii)   We require $P(N=3)$, and since this does not depend on the value of M it is the same as finding $P(N=3, M=1, 2, 3$ or $4)$, *ie* we are summing over all possible values of M:

$$P(N=3) = \frac{1}{70} + \frac{1}{35} + \frac{3}{70} + \frac{2}{35} = \frac{1}{7}$$

(iii)   Using the formula for conditional probability, $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$, gives:

$$P(M=2|N=3) = \frac{P(M=2, N=3)}{P(N=3)} = \frac{1/35}{1/7} = \frac{1}{5}$$

Continuous case

In the case of a pair of continuous variables, the distribution of probability over a specified area in the (x, y) plane is given by the (joint) probability density function $f(x, y)$. The probability that the pair (X, Y) takes values in some specified region A is obtained by integrating $f(x, y)$ over A – this integral is a double integral.

Thus:

$$P(x_1 < X < x_2, y_1 < Y < y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x, y) \, dx \, dy$$

The joint distribution function $F(x, y)$ is defined by:

$$F(x, y) = P(X \le x, Y \le y)$$

and it is related to the joint density function by:

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

The conditions for a function to qualify as a joint probability density function of a pair of continuous random variables **are:**

$f(x, y) \ge 0$ **for all values of *x* and *y* in the domain**

$$\int_x \int_y f(x, y) \, dx \, dy = 1$$

These results parallel those for a single random variable, where the probability density function was $f(x)$ which had to satisfy $f(x) \ge 0$ for all values of x and $\int_x f(x) \, dx = 1$. Recall also that probabilities were calculated using $P(a < X < b) = \int_{x=a}^{b} f(x) \, dx$.

The next question involves the use of double integrals.

Question

The continuous random variables U and V have the joint probability density function:

$$f_{U,V}(u,v) = \frac{2u+v}{3,000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

Calculate $P(10 < U < 15, V > 0)$.

Solution

From the formula for the joint probability function:

$$P(10 < U < 15, V > 0) = \int_{u=10}^{15} \int_{v=0}^{5} \frac{2u+v}{3,000} \, dv \, du$$

This can be integrated with respect to either u or v first. If we do v first, we get:

$$\int_{u=10}^{15} \int_{v=0}^{5} \frac{2u+v}{3,000} \, dv \, du = \int_{u=10}^{15} \left[\frac{2uv + \frac{1}{2}v^2}{3,000} \right]_{v=0}^{5} du = \int_{10}^{15} \frac{10u+12.5}{3,000} \, du$$

$$= \left[\frac{5u^2 + 12.5u}{3,000} \right]_{10}^{15}$$

$$= 0.229$$

If we integrate first with respect to u and then with respect to v, we obtain the same answer as before.

1.2 Marginal probability (density) functions

Discrete case

The marginal distribution of a discrete random variable X is defined to be:

$$f_X(x) = \sum_y f(x,y)$$

This is the distribution of X alone without considering the values that Y can take.

This is what we were doing in the first question in this chapter when we calculated the probability that $N = 3$. If we want the marginal distribution for X, we sum over all the values that Y can take.

Let X and Y have the joint probability function given in the Core Reading above:

		x		
		1	2	3
	1	0.10	0.10	0.05
	2	0.15	0.10	0.05
y	3	0.20	0.05	-
	4	0.15	0.05	-

Let's find the marginal probability distribution of *X*.

The marginal probabilities are:

$$P(X=1) = 0.1 + 0.15 + 0.2 + 0.15 = 0.6$$

$$P(X=2) = 0.1 + 0.1 + 0.05 + 0.05 = 0.3$$

$$P(X=3) = 0.05 + 0.05 = 0.1$$

So the probability distribution of *X* is:

x	1	2	3
$P(X=x)$	0.6	0.3	0.1

We are just adding up the numbers in each column. For the marginal distribution of Y we would just find the row totals.

We can also do this if we are given the joint distribution in the form of a function.

Question

Obtain the probability functions for the marginal distributions of M and N, where:

$$P(M=m, N=n) = \frac{m}{35 \times 2^{n-2}}, \text{ for } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

Solution

Summing over the values of N gives:

$$P(M=m) = \sum_{n=1}^{3} P(M=m, N=n) = \sum_{n=1}^{3} \frac{m}{35 \times 2^{n-2}} = \frac{m}{35}\left(2 + 1 + \frac{1}{2}\right) = \frac{m}{10}$$

Summing over the values of M gives:

$$P(N=n) = \sum_{m=1}^{4} P(M=m, N=n) = \sum_{m=1}^{4} \frac{m}{35 \times 2^{n-2}} = \frac{1}{35 \times 2^{n-2}} (1+2+3+4) = \frac{1}{7 \times 2^{n-3}}$$

Continuous case

In the case of continuous variables the marginal probability density function (PDF) of X, $f_X(x)$ is obtained by integrating over y (for the given value of x) the joint PDF $f(x,y)$.

This means that $f_X(x) = \int\limits_{y} f(x,y) \, dy$.

The resulting $f_X(x)$ is a proper PDF – it integrates to 1. Similarly for $f_Y(y)$, we obtain this **by integrating over x (for the given value of y).**

In some cases the region of definition of (X,Y) may be such that the limits of integration for one variable will involve the other variable.

Question

Determine the marginal probability density functions for U and V, where:

$$f_{U,V}(u,v) = \frac{2u+v}{3,000}, \text{ for } 10 < u < 20 \text{ and } -5 < v < 5$$

Solution

To find the PDF of the marginal distribution of U, we integrate out V:

$$f_U(u) = \int\limits_{v=-5}^{5} \frac{2u+v}{3,000} \, dv = \left[\frac{2uv + \frac{1}{2}v^2}{3,000} \right]_{v=-5}^{5} = \frac{(10u+12.5) - (-10u+12.5)}{3,000} = \frac{u}{150}$$

Therefore the marginal distribution of U is $f_U(u) = \dfrac{u}{150}$, $10 < u < 20$.

Similarly for V, we integrate out U:

$$f_V(v) = \int\limits_{u=10}^{20} \frac{2u+v}{3,000} \, du = \left[\frac{u^2 + uv}{3,000} \right]_{u=10}^{20} = \frac{(400+20v) - (100+10v)}{3,000} = \frac{30+v}{300}$$

Therefore the marginal distribution of V is $f_V(v) = \dfrac{30+v}{300}$, $-5 < v < 5$.

To check that these functions are PDFs, we can integrate them over the appropriate range. The answers should both be 1.

1.3 Conditional probability (density) functions

The distribution of X for a particular value of Y is called the conditional distribution of X given y.

Discrete case

The probability function $P_{X|Y=y}(x \mid y)$ for the conditional distribution of X given $Y = y$ for discrete random variables X and Y is:

$$P_{X|Y=y}(x, y) = P(X = x \mid Y = y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}$$

for all values x in the range of X.

This is what we were doing earlier when we calculated $P(M = 2 | N = 3)$ in a previous example.

Question

A bivariate distribution has the following probability function:

		X		
		0	1	2
	1	0.1	0.1	0
Y	2	0.1	0.1	0.2
	3	0.2	0.1	0.1

Determine:

(i)   the marginal distribution of X

(ii)   the conditional distribution of $X | Y = 2$.

Solution

(i)   The marginal distribution of X can be found by summing the columns in the table.

$$P(X = 0) = 0.4, \ P(X = 1) = 0.3, \ P(X = 2) = 0.3$$

(ii)   Using the definition of conditional probability:

$$P(X=0|Y=2) = \frac{P(X=0,Y=2)}{P(Y=2)} = \frac{0.1}{0.4} = 0.25$$

$$P(X=1|Y=2) = \frac{P(X=1,Y=2)}{P(Y=2)} = \frac{0.1}{0.4} = 0.25$$

$$P(X=2|Y=2) = \frac{P(X=2,Y=2)}{P(Y=2)} = \frac{0.2}{0.4} = 0.5$$

Alternatively here we could have scaled up the probabilities in the second row so that they

add to one eg $P(X=0|Y=2) = \dfrac{0.1}{0.1+0.1+0.2} = \dfrac{0.1}{0.4} = 0.25$.

Continuous case

The probability density function $f_{X|Y=y}(x\,|\,y)$ for the conditional distribution of X given $Y = y$ for the continuous variables X and Y is a function such that:

$$\int_{x=x_1}^{x_2} f_{X|Y=y}(x,y)\,dx = P(x_1 < X < x_2 \,|\, Y = y)$$

for all values x in the range of X .

This conditional distribution in both instances is only defined for those values of y for which $f_Y(y) > 0$.

We calculate the form of the conditional PDF similarly to the method we used in the discrete case – we just divide the joint PDF by the marginal PDF. So:

$$f_{X|Y=y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Question

Let X and Y have joint density function:

$$f(x,y) = \frac{1}{16}(x+3y) \qquad\qquad 0 < x < 2, 0 < y < 2$$

Determine the conditional density function of X given $Y = y$.

Solution

The marginal PDF of Y is:

$$f_Y(y) = \int\limits_{x=0}^{2} \frac{1}{16}(x+3y)\,dx = \frac{1}{16}\left[\frac{1}{2}x^2 + 3xy\right]_{x=0}^{2} = \frac{1}{16}(2+6y)$$

So:

$$f_{X|Y=y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{16}(x+3y)}{\frac{1}{16}(2+6y)} = \frac{x+3y}{2(1+3y)} \quad 0 < x < 2$$

1.4 Independence of random variables

Consider a pair of variables (X,Y), and suppose that the conditional distribution of Y given $X = x$ does not actually depend on x at all. It follows that the probability function/PDF $f(y|x)$ must be simply that of the marginal distribution of Y, $f_Y(y)$.

Here $f(y|x)$ is an abbreviation for $f_{Y|X=x}(y,x)$.

So, if conditional is equivalent to marginal, then:

$$f_Y(y) = f_{Y|X=x}(y,x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

ie $\qquad f_{X,Y}(x,y) = f_X(x)f_Y(y)$

so joint PF/PDF is the product of the marginals.

This motivates the definition, which is given here for two variables.

Definition

The random variables X and Y are independent if, and only if, the joint probability function/PDF is the product of the two marginal probability functions/PDFs for all (x,y) in the range of the variables, *ie*:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ for all } (x,y) \text{ in the range}$$

Discrete case

It follows that probability statements about values assumed by (X,Y) can be broken down into statements about X and Y separately. So if X and Y are independent discrete variables then:

$$P(X=x, Y=y) = P(X=x)P(Y=y)$$

Question

Determine whether the variables X and Y given below are independent.

		X		
		0	1	2
	1	0.1	0.1	0
Y	2	0.1	0.1	0.2
	3	0.2	0.1	0.1

Solution

They are not independent. For example, we saw earlier that the conditional distribution of X given $Y = 2$ is not the same as the marginal distribution of X.

Alternatively, we can see that, for example, $P(X = 0, Y = 1) = 0.1$. However, $P(X = 0) = 0.4$ and $P(Y = 1) = 0.2$. Since $0.4 \times 0.2 \neq 0.1$, the two random variables are not independent.

To show that the random variables are *not* independent, we only need to show that the joint probability is not equal to the product of the marginal probabilities in any one particular case. If we wish to show that they *are* independent, we need to show that the multiplication works for *all* possible values of x and y.

As a quick check in the discrete case, note that, for independence, the probabilities in each row in the table must be in the same ratios as the probabilities in every other row (and similarly for the columns). This is not the case here.

Question

Determine whether the variables M and N are independent. The joint probability function of M and N is:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

Solution

They are independent. We saw earlier that $P(M = m) = \frac{m}{10}$ and $P(N = n) = \frac{1}{7 \times 2^{n-3}}$. Hence the joint probability distribution is the product of the two marginal distributions. So the variables are independent.

Continuous case

If X and Y are continuous, the double integral required to evaluate a joint probability splits into the product of two separate integrals, one for X and one for Y, and we have:

$$P(x_1 < X < x_2, y_1 < Y < y_2) = P(x_1 < X < x_2)P(y_1 < Y < y_2)$$

This means that in the continuous case, if the two random variables are independent, we can factorise the joint PDF into two separate expressions, one of which will be a function of x only, and the other will be a function of y only.

Question

Determine whether the variables U and V are independent. The joint PDF of U and V is:

$$f_{U,V}(u,v) = \frac{2u + v}{3000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

Solution

They are not independent. If they were it would be possible to factorise $\dfrac{2u + v}{3,000}$ into two

functions of the form $g(u)h(v)$. As this is not possible, the random variables are not independent.

Functions of random variables

If the random variables X and Y are independent, then any functions $g(X)$ and $h(Y)$ are also independent.

This should be intuitively obvious if we think of independence as meaning that the quantities have no influence on each other.

Several variables

When considering three or more variables, the definition of independence involves the factorisation of the joint probability function into the product of all the individual marginal probability functions. For X, Y, and Z to be independent it is not sufficient that they are independent taken two at a time (pairwise independent).

Question

Consider the joint probability density function of X, Y and Z given by:

$$f(x,y,z) = \begin{cases} (x+y)e^{-z} & \text{for } 0 < x < 1,\ 0 < y < 1,\ z > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Verify that the random variables X, Y and Z are not independent, but that the two random variables X and Z are pairwise independent, and also that the two random variables Y and Z are pairwise independent.

Solution

We first need to find the joint density functions of X and Z, Y and Z, and the marginal distributions of X, Y and Z.

$$f_{X,Z}(x,z) = \int_0^1 (x+y)e^{-z}\ dy = \left[e^{-z}(xy + \tfrac{1}{2}y^2) \right]_0^1 = e^{-z}(x + \tfrac{1}{2})$$

$$f_{Y,Z}(y,z) = \int_0^1 (x+y)e^{-z}\ dx = \left[e^{-z}(\tfrac{1}{2}x^2 + yx) \right]_0^1 = e^{-z}(y + \tfrac{1}{2})$$

$$f_X(x) = \int_0^\infty e^{-z}(x + \tfrac{1}{2})\ dz = \left[-e^{-z}(x + \tfrac{1}{2}) \right]_0^\infty = x + \tfrac{1}{2}$$

$$f_Z(z) = \int_0^1 e^{-z}(x + \tfrac{1}{2})\ dx = \left[e^{-z}(\tfrac{1}{2}x^2 + \tfrac{1}{2}x) \right]_0^1 = e^{-z}$$

$$f_Y(y) = \int_0^\infty e^{-z}(y + \tfrac{1}{2})\ dz = \left[-e^{-z}(y + \tfrac{1}{2}) \right]_0^\infty = y + \tfrac{1}{2}$$

If we multiply together the marginal PDFs for X, Y and Z, we obtain:

$$f_X(x)f_Y(y)f_Z(z) = (x + \tfrac{1}{2})(y + \tfrac{1}{2})e^{-z}$$

Comparing this to the joint distribution PDF $f(x,y,z) = (x+y)e^{-z}$, we see that they are not the same. So X, Y and Z are not independent.

However the product of the marginal distribution PDFs for X and Z, and for Y and Z do give the respective joint PDFs, so X and Z, and Y and Z are pairwise independent.

2   Expectations of functions of two variables

2.1   Expectations

The expression for the expected value of a function $g(X,Y)$ of the random variables (X,Y) is found by summing (discrete case) or integrating (continuous case) the product:

value × probability of assuming that value

over all values (or combinations of) (x,y). The summation is a double summation, the integral a double integral.

Thus for discrete variables

$$E\big[g(X,Y)\big] = \sum_{x} \sum_{y} g(x,y) p_{X,Y}(x,y) = \sum_{x} \sum_{y} g(x,y) P(X = x, Y = y)$$

where the summation is over all possible values of x and y.

This result parallels that for single random variables, where the expected value of a function of a discrete random variable was defined to be $E[g(X)] = \sum_{x} g(x) P(X = x)$.

Question

Calculate the expected value of $\dfrac{N+1}{M}$, where the joint distribution of M and N is:

		\(M\)			
		1	2	3	4
\(N\)	1	$\dfrac{2}{35}$	$\dfrac{4}{35}$	$\dfrac{6}{35}$	$\dfrac{8}{35}$
	2	$\dfrac{1}{35}$	$\dfrac{2}{35}$	$\dfrac{3}{35}$	$\dfrac{4}{35}$
	3	$\dfrac{1}{70}$	$\dfrac{1}{35}$	$\dfrac{3}{70}$	$\dfrac{2}{35}$

ie $P(M = m, N = n) = \dfrac{m}{35 \times 2^{n-2}}$.

Solution

From the table of values, working across from the top left gives:

$$E\left[\frac{N+1}{M}\right] = 2 \times \frac{2}{35} + 1 \times \frac{4}{35} + \cdots + \frac{4}{3} \times \frac{3}{70} + 1 \times \frac{2}{35} = \frac{36}{35}$$

Alternatively, we could work from the formula:

$$P(M=m, N=n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m=1, 2, 3, 4 \text{ and } n=1, 2, 3$$

This gives:

$$E\left[\frac{N+1}{M}\right] = \sum_{m=1}^{4} \sum_{n=1}^{3} \frac{n+1}{m} P(M=m, N=n)$$

$$= \sum_{m=1}^{4} \sum_{n=1}^{3} \frac{n+1}{m} \times \frac{m}{35 \times 2^{n-2}}$$

$$= \frac{1}{35} \sum_{m=1}^{4} \sum_{n=1}^{3} \frac{n+1}{2^{n-2}}$$

$$= \frac{1}{35} \times 4 \times \left(\frac{1+1}{2^{-1}} + \frac{2+1}{2^{0}} + \frac{3+1}{2^{1}}\right) = \frac{36}{35}$$

For continuous variables

$$E\big[g(X,Y)\big] = \int_{x} \int_{y} g(x,y) f_{X,Y}(x,y)\, dy\, dx$$

where the integration is over all possible values of x and y.

This result parallels that for single random variables, where the expected value of a function of a continuous random variable was defined to be $E[g(X)] = \int_{x} g(x) f(x)\, dx$.

Question

U and V have the joint distribution:

$$f_{U,V}(u,v) = \frac{2u+v}{3,000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

(i)   Calculate $E(U)$ and $E(V)$:

    (a)   using $f_{U,V}(u,v)$ (b)   using $f_U(u)$ and $f_V(v)$.

(ii)   Comment on your answers.

Solution

(i)(a)   Integrating with respect to v first and then with respect to u, we obtain:

$$E(U) = \int_{u=10}^{20} \int_{v=-5}^{5} u\frac{(2u+v)}{3,000} dv\, du = \int_{u=10}^{20} \int_{v=-5}^{5} \frac{2u^2+uv}{3,000} dv\, du = \int_{u=10}^{20} \left[\frac{2u^2v+\frac{1}{2}uv^2}{3,000} \right]_{-5}^{5} du$$

$$= \int_{u=10}^{20} \frac{u^2}{150}\, du = \left[\frac{u^3}{450} \right]_{10}^{20} = \frac{140}{9}$$

Similarly:

$$E(V) = \int_{u=10}^{20} \int_{v=-5}^{5} v\frac{(2u+v)}{3,000} dv\, du = \int_{u=10}^{20} \int_{v=-5}^{5} \frac{2uv+v^2}{3,000} dv\, du = \int_{u=10}^{20} \left[\frac{uv^2+\frac{1}{3}v^3}{3,000} \right]_{-5}^{5} du$$

$$= \int_{u=10}^{20} \frac{1}{36}\, du = \left[\frac{1}{36}u \right]_{10}^{20} = \frac{5}{18}$$

(i)(b)   We have already found that $f_U(u) = \frac{u}{150}$ and $f_V(v) = \frac{30+v}{300}$. Hence:

$$E(U) = \int_{u=10}^{20} u\frac{u}{150}\, du = \int_{u=10}^{20} \frac{u^2}{150}\, du = \left[\frac{u^3}{450} \right]_{10}^{20} = \frac{140}{9}$$

$$E(V) = \int_{v=-5}^{5} v\frac{30+v}{300}\, dv = \int_{v=-5}^{5} \frac{30v+v^2}{300}\, dv = \left[\frac{15v^2+\frac{1}{3}v^3}{300} \right]_{v=-5}^{5} = \frac{5}{18}$$

(ii)   Both methods are equivalent.

2.2   Expectation of a sum

It follows that:

$$E\big[ag(X)+bh(Y)\big]=aE\big[g(X)\big]+bE\big[h(Y)\big]$$

where *a* **and** *b* **are constants, so handling the expected value of a linear combination of functions is no more difficult than handling the expected values of the individual functions.**

The definition of expected value and this last result (on the expected value of a sum of functions) extend to functions of more than 2 variables.

In particular $E\big[g(X)+h(Y)\big]=E\big[g(X)\big]+E\big[h(Y)\big]$ **– and these expected values will usually be easier to find from the respective marginal distributions, so there would be no need for double sums/integrals.**

So if we had the sum of a number of random variables:

$$S=X_1+X_2+\cdots+X_n$$

Then by extension of the result above:

$$E(S)=E(X_1)+E(X_2)+\cdots+E(X_n)$$

So the expectation of the sum is equal to the sum of the expectations.

This is true whether or not the random variables X_i are independent.

Question

Verify that $E\big[X^2+2Y\big]=E[X^2]+E[2Y]$, for the random variables X and Y given here:

		|	0	1	2
				X	
	1	|	0.1	0.1	0
Y	2	|	0.1	0.1	0.2
	3	|	0.2	0.1	0.1

Solution

Reading the values from the table, we have:

$$E\big[X^2+2Y\big]=(0^2+2\times1)\times0.1+(1^2+2\times1)\times0.1+\cdots+(2^2+2\times3)\times0.1=5.9$$

Looking at the terms on the right hand side:

$$E\left[X^2\right] = 0 \times 0.4 + 1 \times 0.3 + 4 \times 0.3 = 1.5$$

$$E[2Y] = 2 \times 0.2 + 4 \times 0.4 + 6 \times 0.4 = 4.4$$

Thus $E\left[X^2\right] + E[2Y] = 5.9$, and the result has been verified.

2.3 Expectation of a product

For independent random variables X and Y:

$$\boldsymbol{E\left[g(X)h(Y)\right] = E\left[g(X)\right]E\left[h(Y)\right]}$$

since the joint density function factorises into the two marginal density functions.

This result is true only for *INDEPENDENT* random variables.

Question

Verify that $E\left[\dfrac{N+1}{M}\right] = E\left[\dfrac{1}{M}\right]E[N+1]$, where the joint probability function of M and N is given

by:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

<div align="center">

M

		1	2	3	4
	1	$\dfrac{2}{35}$	$\dfrac{4}{35}$	$\dfrac{6}{35}$	$\dfrac{8}{35}$
N	2	$\dfrac{1}{35}$	$\dfrac{2}{35}$	$\dfrac{3}{35}$	$\dfrac{4}{35}$
	3	$\dfrac{1}{70}$	$\dfrac{1}{35}$	$\dfrac{3}{70}$	$\dfrac{2}{35}$

</div>

Solution

We have found previously that $E\left(\dfrac{N+1}{M}\right) = \dfrac{36}{35}$.

We can calculate $E\left(\dfrac{1}{M}\right)$ and $E(N+1)$ using the marginal probability functions:

$$E\left(\frac{1}{M}\right) = \sum_{m=1}^{4} \frac{1}{m} P(M=m) = \sum_{m=1}^{4} \frac{1}{m} \times \frac{m}{10} = \sum_{m=1}^{4} \frac{1}{10} = \frac{2}{5}$$

$$E(N+1) = \sum_{n=1}^{3} (n+1) P(N=n) = 2 \times \frac{4}{7} + 3 \times \frac{2}{7} + 4 \times \frac{1}{7} = \frac{18}{7}$$

This gives $E\left(\dfrac{1}{M}\right) E(N+1) = \dfrac{2}{5} \times \dfrac{18}{7} = \dfrac{36}{35}$, which verifies the result.

Note that $E\left(\dfrac{N+1}{M}\right) \neq \dfrac{E(N+1)}{E(M)}$

This should not be surprising, since we showed earlier that M and N are independent random variables here.

If we take the functions to be $g(X) = X$ and $h(Y) = Y$, these last two results give us some simple relationships between two random variables X and Y:

(a) $E[X+Y] = E[X] + E[Y]$

(b)   if X and Y are independent, $E[XY] = E[X]E[Y]$.

2.4   Covariance and correlation coefficient

The covariance $cov[X,Y]$ of two random variables X and Y is defined by:

$$cov[X,Y] = E\big[(X - E[X])(Y - E[Y])\big]$$

which simplifies to:

$$cov[X,Y] = E[XY] - E[X]E[Y]$$

Notice the similarity between the covariance defined here and the definition of the variance:

$$var(X) = E[(X - E(X))^2] = E(X^2) - E^2(X)$$

Question

Show that the simplification $cov[X,Y] = E[XY] - E[X]E[Y]$ is correct.

Solution

If we expand the definition of the covariance we obtain:

$$\text{cov}(X,Y) = E\big[(X - E[X])(Y - E[Y])\big]$$

$$= E\big(XY - XE[Y] - YE[X] + E[X]E[Y]\big)$$

$$= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y]$$

$$= E[XY] - E[X]E[Y]$$

It is often easier to use the formula $\text{cov}(X,Y) = E[XY] - E[X]E[Y]$ when calculating covariances, rather than using the definition.

If we rearrange this formula it tells us how to find $E[XY]$ for random variables that are not independent, *ie* $E[XY] = E[X]E[Y] + \text{cov}[X,Y]$. We will return to independence shortly.

Note: The units of $\text{cov}(X,Y)$ are the product of those of X and Y. So for example if X is a time in hours, and Y is a sum of money in £, then cov is in $£ \times hours$. Note also that $\text{cov}[X,X] = \text{var}[X]$.

Question

Calculate the covariance of the random variables X and Y whose joint distribution is as follows:

		|	0	1	2
	1	|	0.1	0.1	0
Y	2	|	0.1	0.1	0.2
	3	|	0.2	0.1	0.1

X is shown above the columns 0, 1, 2.

Solution

We will use the formula $\text{cov}(X,Y) = E[XY] - E[X]E[Y]$.

From the table of values:

$$E[XY] = 0 \times 1 \times 0.1 + \cdots + 2 \times 3 \times 0.1 = 2$$

The (marginal) probability distribution of X is:

X	0	1	2
$P(X = x)$	0.4	0.3	0.3

So: $E[X] = 0 \times 0.4 + 1 \times 0.3 + 2 \times 0.3 = 0.9$

The (marginal) probability distribution of Y is:

y	1	2	3
$P(Y = y)$	0.2	0.4	0.4

So: $E[Y] = 1 \times 0.2 + 2 \times 0.4 + 3 \times 0.4 = 2.2$

Hence $cov(X,Y) = 2 - 0.9 \times 2.2 = 0.02$.

Useful results on handling covariances

(a) $cov[aX + b, cY + d] = ac\, cov[X, Y]$

Proof:

$E[aX + b] = aE[X] + b$ and $E[cY + d] = cE[Y] + d$

so $aX + b - E[aX + b] = a(X - E[X])$ and $cY + d - E[cY + d] = c(Y - E[Y])$

$$\therefore cov[aX + b, cY + d] = E\big[a(X - E[X])c(Y - E[Y])\big] = ac\, cov[X, Y]$$

Note: The changes of origin (*b* and *d*) have no effect, because we are using deviations from means. The changes of scale (*a* and *c*) carry through.

This means that constants that are added or subtracted can be ignored and constants that are multiplied or divided are pulled out. Note the similarity between this result and that for the variance: $var[aX + b] = a^2\, var[X]$.

(b) $cov[X, Y + Z] = cov[X, Y] + cov[X, Z]$

Proof:

$E\big[X(Y + Z)\big] = E[XY] + E[XZ]$ and $E[Y + Z] = E[Y] + E[Z]$

$$\therefore cov[X, Y + Z] = E[XY] + E[XZ] - E[X]\big(E[Y] + E[Z]\big)$$
$$= E[XY] - E[X]E[Y] + E[XZ] - E[X]E[Z]$$
$$= cov[X, Y] + cov[X, Z]$$

These two results hold for any random variables X, Y and Z (whenever the covariances exist). This result is just like multiplying out brackets using the distributive law: $x(y+z) = xy + xz$.

Question

Write down the formula for $\text{cov}[X+Y, W+Z]$.

Solution

$$\text{cov}(X+Y, W+Z) = \text{cov}(X,W) + \text{cov}(X,Z) + \text{cov}(Y,W) + \text{cov}(Y,Z)$$

The next result concerns random variables that are independent.

(c) If X and Y are independent, $\text{cov}[X,Y] = 0$.

Proof:

$$\text{cov}[X,Y] = E[XY] - E[X]E[Y] = 0$$

The covariance of M and N used in earlier examples is zero as they are independent.

The result $E[XY] = E[X]E[Y]$ extends to the expected value of the product of any finite number of independent variables, *ie* $E[X_1 \ldots X_n] = E[X_1] \ldots E[X_n]$.

The covariance between X and Y is a measure of the strength of the linear association or linear relationship between the variables. However it suffers from the fact that its value is dependent on the units of measurement of the variables.

A related quantity to the covariance is the correlation coefficient which is a dimensionless quantity (*ie* it has no units).

The correlation coefficient (X,Y) (written as $\text{corr}(X,Y)$) or $\rho(X,Y)$ of two random variables X and Y is defined by:

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

Question

Calculate the correlation coefficient of U and V, where:

$$f_{U,V}(u,v) = \frac{2u+v}{3000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

We are given that $E(U) = \dfrac{140}{9}$ and $E(V) = \dfrac{5}{18}$.

Solution

First we need $E[UV]$:

$$E[UV] = \int\limits_{u=10}^{20} \int\limits_{v=-5}^{5} uv \frac{2u+v}{3,000} \, dv \, du$$

Integrating first with respect to u:

$$E[UV] = \int\limits_{v=-5}^{5} \left[\frac{\frac{2}{3}u^3 v + \frac{1}{2}u^2 v^2}{3,000} \right]_{u=10}^{u=20} dv = \int\limits_{v=-5}^{5} \left(\frac{14v}{9} + \frac{v^2}{20} \right) dv = \left[\frac{14v^2}{18} + \frac{v^3}{60} \right]_{-5}^{5} = \frac{25}{6}$$

So the covariance of U and V is:

$$\text{cov}(U,V) = \frac{25}{6} - \frac{140}{9} \times \frac{5}{18} = -\frac{25}{162}$$

We now need the variance of U and V:

$$\text{var}(U) = \int\limits_{10}^{20} u^2 \times \frac{u}{150} \, du - \left(\frac{140}{9} \right)^2 = \left[\frac{u^4}{600} \right]_{10}^{20} - \left(\frac{140}{9} \right)^2 = \frac{650}{81}$$

Similarly:

$$\text{var}(V) = \int\limits_{-5}^{5} v^2 \times \frac{30+v}{300} \, dv - \left(\frac{5}{18} \right)^2 = \left[\frac{v^3}{30} + \frac{v^4}{1,200} \right]_{-5}^{5} - \left(\frac{5}{18} \right)^2 = \frac{2,675}{324}$$

So the correlation coefficient is $\text{corr}(U,V) = \dfrac{-\dfrac{25}{162}}{\sqrt{\dfrac{650}{81} \times \dfrac{2,675}{324}}} = -\dfrac{1}{\sqrt{2,782}} = -0.019$.

The correlation coefficient takes a value in the range $-1 \le \rho \le 1$. It reflects the degree of association between the two variables.

Use this range to do a reasonableness check for any numerical answer. Any figure outside this range is automatically wrong.

A value for ρ of ± 1 indicates that the variables have 'perfect linear correlation' – what this means is that one variable is actually a linear function of the other (with probability 1).

If $\rho = 0$, the random variables are said to be uncorrelated.

Independent variables are uncorrelated (but not all uncorrelated variables are independent).

Note that this means that the converse of Result (c) given previously is not true. If X and Y are independent, their covariance is equal to zero. However, if the covariance of X and Y is zero, this does not necessarily mean that they are independent.

In simple terms, 'independent' means that 'probabilities factorise', and 'uncorrelated' means that 'expectations factorise'.

Question

A bivariate distribution has the following probability function:

		P		
		−1	0	1
	−1	0.1	0.6	0.1
Q	1	0.1	0	0.1

Show that P and Q are uncorrelated but not independent.

Solution

We have:

$$E[P]=0 \, , \; E[Q]=-0.6 \; \text{ and } \; E[PQ]=0$$

so the covariance is zero.

However the conditional distribution of, say, $P|Q=1$ takes the values −1 and 1 each with probability 0.5, whereas the marginal distribution of P takes the values −1, 0 and 1 with probabilities 0.2, 0.6 and 0.2 respectively. So the marginal distributions are different from the conditional distributions, and P and Q are not independent.

Alternatively, we could compare, for example, $P(P=-1, Q=1)$ with $P(P=-1) \times P(Q=1)$.

Note that this confirms the sentence given in the Core Reading previously. This is an example of a random variable that is uncorrelated, but not independent.

2.5 Variance of a sum

For any random variables X and Y:

$$\mathbf{var}[X+Y] = \mathbf{var}[X] + \mathbf{var}[Y] + 2\,\mathbf{cov}[X,Y]$$

This can be proved from the definitions of variance and covariance.

One possible proof is as follows:

$$\text{var}[X+Y] = E\left[([X+Y]-E[X+Y])^2\right]$$

$$= E\left[[(X-E[X])+(Y-E[Y])]^2\right]$$

$$= E\left[(X-E[X])^2\right] + E\left[(Y-E[Y])^2\right] + 2E\left[(X-E[X])(Y-E[Y])\right]$$

$$= \text{var}[X] + \text{var}[Y] + 2\text{cov}(X,Y)$$

Question

Set out an alternative proof of the above result starting from $\text{var}(X+Y) = \text{cov}(X+Y, X+Y)$.

Solution

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

$$= \text{cov}(X,X) + \text{cov}(X,Y) + \text{cov}(Y,X) + \text{cov}(Y,Y)$$

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

For independent random variables, this can be simplified:

$$\textbf{var}[X+Y] = \textbf{var}[X] + \textbf{var}[Y]$$

since $\textbf{cov}[X,Y] = 0$.

Question

Show from first principles that the random variables M, N, whose joint probability function is given below, satisfy $\text{var}[M+N] = \text{var}[M] + \text{var}[N]$.

		M			
		1	2	3	4
	1	$\dfrac{2}{35}$	$\dfrac{4}{35}$	$\dfrac{6}{35}$	$\dfrac{8}{35}$
N	2	$\dfrac{1}{35}$	$\dfrac{2}{35}$	$\dfrac{3}{35}$	$\dfrac{4}{35}$
	3	$\dfrac{1}{70}$	$\dfrac{1}{35}$	$\dfrac{3}{70}$	$\dfrac{2}{35}$

Solution

By adding up the probabilities from the table, the random variable $M+N$ has the distribution:

$m+n$	2	3	4	5	6	7
$P(M+N=m+n)$	$\dfrac{2}{35}$	$\dfrac{5}{35}$	$\dfrac{17}{70}$	$\dfrac{12}{35}$	$\dfrac{11}{70}$	$\dfrac{2}{35}$

The expectation of $M+N$ is:

$$E[M+N] = 2 \times \frac{2}{35} + \cdots + 7 \times \frac{2}{35} = \frac{32}{7}$$

The variance of $M+N$ is:

$$\text{var}(M+N) = 2^2 \times \frac{2}{35} + \cdots + 7^2 \times \frac{2}{35} - \left(\frac{32}{7}\right)^2 = \frac{75}{49}$$

Looking at the marginal distributions:

$$\text{var}(M) = 1^2 \times \frac{1}{10} + \cdots + 4^2 \times \frac{4}{10} - 3^2 = 1$$

$$\text{var}(N) = 1^2 \times \frac{4}{7} + \cdots + 3^2 \times \frac{1}{7} - \left(\frac{11}{7}\right)^2 = \frac{26}{49}$$

So M and N satisfy the given relationship, since $1 + \dfrac{26}{49} = \dfrac{75}{49}$.

Similarly, it can be shown that:

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

and so for independent random variables, we get:

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

Note that the variance of a difference is not equal to the difference in the variances. It is equal to the sum of the variances. This must be true, since the difference in the variances could easily be negative and variance is always a positive quantity.

3　Convolutions

3.1　Introduction

Much of statistical theory involves the distributions of sums of random variables. In particular the sum of a number of independent variables is especially important.

Discrete case

Consider the sum of two discrete random variables, so let $Z = X + Y$, where (X, Y) has joint probability function $P(x, y)$.

Then $P(Z = z)$ is found by summing $P(x, y)$ over all values of (x, y) such that $x + y = z$ ie

$$P_Z(z) = \sum_x P(x, z - x).$$

We did this when we calculated the distribution of $M + N$ in the previous question.

Now suppose that X and Y are independent variables. Then $P(x, y)$ is the product of the two marginal probability functions, so

$$P_Z(z) = \sum_x P_X(x) P_Y(z - x)$$

Definition

When a function P_Z can be expressed as a sum of this form, then P_Z is called the convolution of the functions P_X and P_Y. This is written symbolically as $P_Z = P_X * P_Y$. So here, the probability function of $Z = X + Y$ is the convolution of the (marginal) probability functions of X and Y.

Continuous case

In the case where X and Y are independent continuous variables with joint probability density function $f(x, y)$, the corresponding expression is:

$$f_Z(z) = \int_x f_X(x) f_Y(z - x) \, dx$$

Question

If $X \sim Poi(\lambda)$ and $Y \sim Poi(\mu)$ are independent random variables, obtain the probability function of $Z = X + Y$.

Solution

Using the convolution formula for discrete random variables:

$$P(Z = z) = \sum_{x=0}^{z} P(X = x)P(Y = z - x)$$

$$= \sum_{x=0}^{z} \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^{z-x} e^{-\mu}}{(z-x)!}$$

$$= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^{z} \frac{z!}{x!(z-x)!} \lambda^x \mu^{z-x}$$

$$= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^{z} \binom{z}{x} \lambda^x \mu^{z-x}$$

$$= \frac{e^{-(\lambda+\mu)}}{z!} (\lambda + \mu)^z$$

Since this matches the probability function for a $Poi(\lambda + \mu)$ distribution (and Z can take the values $Z = 0, 1, 2, \dots$), Z has a $Poi(\lambda + \mu)$ distribution.

We can also use a convolution approach to derive the sum of two continuous random variables.

Question

If $X \sim Exp(\lambda)$ and $Y \sim Exp(\mu)$ are independent random variables, obtain the PDF of $Z = X + Y$.

Solution

Using the formula for continuous random variables (and assuming that $\lambda \neq \mu$):

$$f_Z(z) = \int_0^z \lambda e^{-\lambda x} \mu e^{-\mu(z-x)} dx = \frac{\lambda \mu e^{-\mu z}}{\lambda - \mu} \int_0^z (\lambda - \mu) e^{-(\lambda-\mu)x} dx$$

$$= \frac{\lambda \mu}{\lambda - \mu} e^{-\mu z} \left[1 - e^{-(\lambda-\mu)z} \right] = \frac{\lambda \mu}{\lambda - \mu} \left(e^{-\mu z} - e^{-\lambda z} \right)$$

If $\lambda = \mu$, we get $\lambda^2 z e^{-\lambda z}$, which is the PDF of a $Gamma(2, \lambda)$ distribution.

We can also use MGFs to find the distribution of a sum of random variables. This will be dealt with in Section 4. The MGF method is much easier than the convolution method.

3.2 Moments of linear combinations of random variables

In the last section we looked at the properties of functions of two random variables. We can now extend these results to more than two variables.

Mean

If $X_1, X_2, ..., X_n$ are any random variables (not necessarily independent), then:

$$E(c_1 X_1 + c_2 X_2 + \cdots + c_n X_n) = c_1 E(X_1) + c_2 E(X_2) + \cdots + c_n E(X_n)$$

where $c_1, c_2, ..., c_n$ are any constants.

ie $$E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i E(X_i)$$

This is an extension of the result concerning the expectation of a function of two random variables that we saw earlier *ie* $E[ag(X) + bh(Y)] = aE[g(X)] + bE[h(Y)]$.

Variance

Let $Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$, where the variables are not necessarily independent, and let us now consider the variance:

$$\begin{aligned}
\text{var}(Y) &= \text{cov}(Y, Y) \\
&= \text{cov}(c_1 X_1 + c_2 X_2 + \cdots + c_n X_n, c_1 X_1 + c_2 X_2 + \cdots + c_n X_n) \\
&= \sum_i c_i^2 \,\text{cov}(X_i, X_i) + 2 \sum_{i<j} \sum_j c_i c_j \,\text{cov}(X_i, X_j)
\end{aligned}$$

This is an extension of the result $\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2) + 2\,\text{cov}(X_1, X_2)$.

If $X_1, X_2, ..., X_n$ are pairwise uncorrelated (and hence certainly if they are independent) random variables, then:

$$\text{var}(c_1 X_1 + c_2 X_2 + \cdots + c_n X_n) = c_1^2 \,\text{var}(X_1) + c_2^2 \,\text{var}(X_2) + \cdots + c_n^2 \,\text{var}(X_n)$$

ie $$\text{var}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 \,\text{var}(X_i)$$

Question

If the random variables X, Y, and Z have means and variances $\mu_X = 4$, $\mu_Y = -5$, $\mu_Z = 6$, $\sigma_X^2 = 1$, $\sigma_Y^2 = 4$ and $\sigma_Z^2 = 3$, and the covariances are $\text{cov}(X,Y) = -3$, $\text{cov}(X,Z) = -2$ and $\text{cov}(Y,Z) = 1$, calculate the mean and variance of $W = X - 2Y + 3Z$.

Solution

The mean is:

$$E[W] = E[X] - 2E[Y] + 3E[Z] = 4 - (2 \times -5) + (3 \times 6) = 32$$

Since the random variables X, Y and Z are *not* independent, we can see that:

$$\text{var}(W) \neq \text{var}(X) + 4\,\text{var}(Y) + 9\,\text{var}(Z)$$

Instead we have:

$$\text{var}(W) = \text{var}(X - 2Y + 3Z) = \text{cov}(X - 2Y + 3Z, X - 2Y + 3Z)$$

$$= \text{var}(X) + 4\,\text{var}(Y) + 9\,\text{var}(Z) - 4\,\text{cov}(X,Y) + 6\,\text{cov}(X,Z) - 12\,\text{cov}(Y,Z)$$

$$= 1 + (4 \times 4) + (9 \times 3) - (4 \times -3) + (6 \times -2) - (12 \times 1) = 32$$

It is important to note that there is a distinction between adding up random variables, and multiplying a random variable by a constant.

Question

If $X_1, X_2, \ldots, X_n$ are independent random variables with mean μ and variance σ^2, obtain the mean and variance of $S = X_1 + X_2 + \cdots + X_n$ and $T = nX_1$.

Solution

The mean and variance of S (which is a sum of random variables) are:

$$E[S] = E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n] = n\mu$$

$$\text{var}(S) = \text{var}(X_1 + X_2 + \cdots + X_n) = \text{var}[X_1] + \text{var}[X_2] + \cdots + \text{var}[X_n] = n\sigma^2$$

The mean and variance of T (which is a single random variable multiplied by a constant) are:

$$E[T] = E[nX_1] = nE[X_1] = n\mu$$

$$\text{var}(T) = \text{var}(nX_1) = n^2\,\text{var}(X_1) = n^2\sigma^2$$

The means are the same but that the variance of S is smaller.

4   Using generating functions to derive distributions of linear combinations of independent random variables

In the last section, we saw that we can find the distribution of a sum of a number of independent random variables using convolutions. In this section we look at an alternative and frankly much easier method.

In many cases generating functions may make it possible to specify the actual distribution of Y, where $Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$.

4.1   Moment generating functions

Suppose X_1 and X_2 are independent random variables with MGFs $M_{X_1}(t)$ and $M_{X_2}(t)$ respectively, and let $S = c_1 X_1 + c_2 X_2$.

Then:

$$M_S(t) = E\left[e^{(c_1 X_1 + c_2 X_2)t} \right]$$

$$= E\left[e^{c_1 X_1 t} \right] E\left[e^{c_2 X_2 t} \right]$$

$$= M_{X_1}(c_1 t) M_{X_2}(c_2 t)$$

In the case of a simple sum $Z = X + Y$, we have:

$$M_Z(t) = M_X(t) M_Y(t)$$

so the MGF of the sum of two independent variables is the product of the individual MGFs.

The result extends to the sum of more than two variables.

Let $Y = X_1 + X_2 + \cdots + X_n$ where the X_i are independent and X_i has MGF $M_i(t)$. Then:

$$M_Y(t) = M_1(t) M_2(t) ... M_n(t)$$

(And if X_i in the sum is replaced by $c X_i$ then $M_i(t)$ in the product is replaced by $M_i(ct)$.)

If, in addition, the X_i's are identically distributed, each with MGF $M(t)$, and $Y = X_1 + X_2 + \cdots + X_n$, then $M_Y(t) = \left[M(t) \right]^n$.

Both of the last two results are important to remember and are quotable in the Subject CS1 exam.

4.2   Using MGFs to derive relationships among variables

Bernoulli/binomial

We will now derive the MGF of a $Bin(n,p)$ distribution using an alternative method to that used in the previous chapter. This method uses the fact that a $Bin(n,p)$ is the sum of n independent $Bernoulli(p)$ trials.

Let X_i, $i = 1, 2, \ldots, n$, be independent Bernoulli (p) variables.

Then each has MGF $M(t) = q + pe^t$.

So $Y = X_1 + X_2 + \cdots + X_n$ has MGF $\left[q + pe^t\right]^n$ which is the MGF of a Bin (n,p) variable.

So the Bin (n,p) random variable is the sum of n independent Bernoulli (p) random variables.

Each Bernoulli variable has mean p and variance pq; hence the binomial has mean np and variance npq.

Physically, the number of successes in n trials is the sum of the numbers of successes (0 or 1) at each trial.

Further, the sum of two independent binomial variables, one (n,p) and the other (m,p), is a Bin $(n+m,p)$ variable.

Question

Show that if $X \sim Bin(m,p)$ and $Y \sim Bin(n,p)$ are independent random variables, then $X + Y$ also has a binomial distribution.

Solution

The moment generating functions for X and Y are:

$$M_X(t) = (q + pe^t)^m \qquad \text{and} \qquad M_Y(t) = (q + pe^t)^n$$

Since X and Y are independent, we can write:

$$M_{X+Y}(t) = (q + pe^t)^m (q + pe^t)^n = (q + pe^t)^{m+n}$$

which we recognise as the MGF of a $Bin(m+n,p)$ distribution. Hence, by uniqueness of MGFs, $X + Y$ has a binomial distribution with parameters $m+n$ and p.

This result should be obvious. If we toss a coin 10 times in the morning and count the number of heads, the number would be distributed as $Bin(10, \frac{1}{2})$. If we toss the coin a further 20 times in the afternoon, the number of heads will be distributed as $Bin(20, \frac{1}{2})$. Adding the totals together is obviously the same as the $Bin(30, \frac{1}{2})$ distribution that we would expect for the whole day.

The phrase 'by uniqueness of MGFs' is important here. What we are saying here is that it is not possible for two different distributions to have the same MGF. If it was, then, once we had found the MGF of the sum, it would not be possible to determine which of the two distributions having the same MGF was the one for the variable $X + Y$. Fortunately, MGFs do uniquely define the distribution to which they belong, and so we know the exact distribution of $X + Y$.

Geometric/negative binomial

We will now derive the MGF of a negative binomial distribution with parameters k and p using an alternative method to that used in the previous chapter. This method uses the fact that a negative binomial (k, p) random variable is the sum of k independent geometric random variables with parameter p.

Let X_i, $i = 1, 2, \ldots, k$, be independent geometric (p) variables.

Then each has MGF $M(t) = \dfrac{pe^t}{1 - qe^t}$.

So $Y = X_1 + X_2 + \cdots + X_k$ has MGF $\left(\dfrac{pe^t}{1 - qe^t} \right)^k$, which is the MGF of a negative binomial (k, p) variable.

So the negative binomial (k, p) random variable is the sum of k independent geometric (p) random variables.

Each geometric variable has mean $\dfrac{1}{p}$ and variance $\dfrac{q}{p^2}$; hence the negative binomial has mean $\dfrac{k}{p}$ and variance $\dfrac{kq}{p^2}$.

Physically, the number of trials up to the k th success is the sum of the number of trials to the first success, plus the additional number to the second success,…, plus the additional number to the k th success.

Further, the sum of two independent negative binomial variables, one (k, p) and the other (m, p), is a negative binomial $(k + m, p)$ variable.

This is straightforward to prove using MGFs.

Poisson

We will now find the distribution of the sum of two independent Poisson random variables using MGFs. This is an alternative method to the convolution method.

Let X and Z be independent Poi(λ) and Poi(γ) variables.

Then X has MGF $M_X(t) = \exp\left\{\lambda\left(e^t - 1\right)\right\}$, Z has MGF $M_Z(t) = \exp\left\{\gamma\left(e^t - 1\right)\right\}$.

So the sum $X + Z$ has MGF $\left[\exp\left\{\lambda\left(e^t - 1\right)\right\}\right]\left[\exp\left\{\gamma\left(e^t - 1\right)\right\}\right] = \exp\left\{(\lambda + \gamma)\left(e^t - 1\right)\right\}$, which is the MGF of a Poi$(\lambda + \gamma)$ variable.

So the sum of independent Poisson variables is a Poisson variable.

X has mean = variance = λ, Z has mean = variance = γ, and the sum has mean = variance = $\lambda + \gamma$.

This is an important result to remember and is quotable in the Subject CS1 exam. It can be extended (in an obvious way) to the sum of more than two Poisson random variables.

Question

A company has three telephone lines coming into its switchboard. The first line rings on average 3.5 times per half-hour, the second rings on average 3.9 times per half-hour, and the third line rings on average 2.1 times per half-hour. Assuming that the numbers of calls are independent random variables having Poisson distributions, calculate the probability that in half an hour the switchboard will receive:

(i)   at least 5 calls

(ii)   exactly 7 calls.

Solution

Summing the Poisson variables, the total number of telephone calls coming in has a Poisson distribution with mean $3.5 + 3.9 + 2.1 = 9.5$.

(i) $P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.04026 = 0.95974$

(ii) $P(X = 7) = P(X \leq 7) - P(X \leq 6) = 0.26866 - 0.16495 = 0.10371$

These figures are taken from the cumulative Poisson table on page 178 of the *Tables*.

Alternatively we could use the Poisson probability formula.

It's worth being aware of the probabilities given towards the back of the *Tables.* Some binomial probabilities are also given there.

Exponential/gamma

We will now derive the MGF of a $Gamma(\alpha, \lambda)$ distribution using an alternative method to that used earlier. This method uses the fact that a $Gamma(\alpha, \lambda)$ distribution can be regarded as the sum of α independent $Exp(\lambda)$ random variables.

Let $X_i, i = 1, 2, ..., k$, be independent $Exp(\lambda)$ variables.

Then each has MGF $M(t) = \lambda(\lambda - t)^{-1}$.

So $Y = X_1 + X_2 + \cdots + X_k$ has MGF $\left[\lambda(\lambda - t)^{-1}\right]^k$, which is the MGF of a Gamma (k, λ) variable.

So the Gamma (k, λ) random variable (for k a positive integer) is the sum of k independent $Exp(\lambda)$ random variables.

Each exponential variable has mean $\dfrac{1}{\lambda}$ and variance $\dfrac{1}{\lambda^2}$; hence the Gamma (k, λ) has mean $\dfrac{k}{\lambda}$ and variance $\dfrac{k}{\lambda^2}$.

Physically, the time to the k th event in a Poisson process with rate λ is the sum of k individual inter-event times.

Further, the sum of two independent gamma variables, one (α, λ) and the other (δ, λ), is a Gamma $(\alpha + \delta, \lambda)$ variable.

This result can also be proved using MGFs.

Question

If the number of minutes it takes for a mechanic to check a tyre is a random variable having an exponential distribution with mean 5, obtain the probability that the mechanic will take:

(i)   more than eight minutes to check two tyres

(ii)   at least fifteen minutes to check three tyres.

Solution

(i) The sum of two independent exponential random variables with mean 5, has a gamma distribution with parameters $\alpha = 2$ and $\lambda = \frac{1}{5}$. If we let X be the total time taken for the mechanic to check the tyres, then:

$$P(X > 8) = \int_8^\infty \frac{\left(\frac{1}{5}\right)^2}{\Gamma(2)} x e^{-\frac{1}{5}x} dx$$

Integrating by parts, using $u = x$, we obtain:

$$P(X > 8) = \frac{1}{25}\left[\left[-5xe^{-\frac{1}{5}x}\right]_8^\infty + 5\int_8^\infty e^{-\frac{1}{5}x} dx\right]$$

$$= \frac{1}{25}\left[40e^{-\frac{8}{5}} - 25\left[e^{-\frac{1}{5}x}\right]_8^\infty\right]$$

$$= \frac{1}{25}\left[40e^{-\frac{8}{5}} + 25e^{-\frac{8}{5}}\right]$$

$$= 0.525$$

Alternatively, we could use the Poisson process. If we let Y be the number of tyres checked in a time period of t minutes, then $Y \sim Poi(0.2t)$. The probability that it takes more than 8 minutes to check two tyres is equivalent to the probability that the number of tyres checked in 8 minutes is only 0 or 1. Using $Y \sim Poi(0.2 \times 8)$, the required probability is therefore:

$$P(Y = 0 \text{ or } 1) = e^{-1.6} + 1.6e^{-1.6} = 0.525$$

(ii) The sum of three independent exponential random variables with mean 5 has a gamma distribution with parameters $\alpha = 3$ and $\lambda = \frac{1}{5}$. If we let X be the total time taken for the mechanic to check the tyres, then we require:

$$P(X > 15)$$

We could solve this by integrating the PDF – but this would require integration by parts (twice).

The easier way to find this probability is to use the gamma-chi squared relationship proved earlier, that is, that if X is a gamma random variable with parameters α and λ, then $2\lambda X$ has a chi-squared distribution with 2α degrees of freedom:

$$P(X > 15) = P(2\lambda X > 30\lambda)$$

$$= P(\chi^2_{2\alpha} > 30\lambda)$$

Substituting $\alpha = 3$ and $\lambda = \frac{1}{5}$, and using the χ^2 values given on page 166 of the *Tables*, we obtain:

$$P(X > 15) = P(\chi^2_6 > 6) = 1 - 0.5768 = 0.4232$$

Alternatively, we could use the Poisson distribution with mean 0.2×15 and calculate the probability of 0, 1 or 2 tyres checked within 15 minutes.

Note that the difference in the wording in the two parts of the question – 'more than' versus 'at least' – is not significant here. Since we are working in continuous time, the probability that an event occurs at exactly time 8 (or time 15) is zero.

Chi-square

From the above result with $\lambda = \frac{1}{2}$, it follows that the sum of a chi-square (n) and an independent chi-square (m) is a chi-square ($n + m$) variable.

So the sum of independent chi-square variables is a chi-square variable.

Question

Suppose that X_1 and X_2 are independent random variables such that $X_1 \sim \chi^2_m$ and $X_2 \sim \chi^2_n$, and let $X = X_1 + X_2$. Use MGFs to prove that $X \sim \chi^2_{m+n}$.

Solution

Since $\chi^2_n \equiv Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$ we have:

$$M_{X_1}(t) = (1 - 2t)^{-\frac{m}{2}} \qquad \text{and} \qquad M_{X_2}(t) = (1 - 2t)^{-\frac{n}{2}}$$

Since X_1 and X_2 are independent:

$$M_X(t) = E(e^{tX}) = E(e^{tX_1 + tX_2}) = E(e^{tX_1} e^{tX_2}) = E(e^{tX_1})E(e^{tX_2}) = M_{X_1}(t)M_{X_2}(t)$$

So $M_X(t) = (1 - 2t)^{-\frac{m}{2}} \times (1 - 2t)^{-\frac{n}{2}} = (1 - 2t)^{-\frac{m+n}{2}}$.

This is the MGF of the χ^2_{m+n} distribution. By the uniqueness property of MGFs, it follows that $X \sim \chi^2_{m+n}$.

This result is useful in many areas of statistics, including generalised linear models (which we will study later in the course).

Normal

Let X be a normal random variable with mean μ_X and standard deviation σ_X, and let Y be a normal random variable with mean μ_Y and standard deviation σ_Y. Let X and Y be independent, and let $Z = X + Y$.

X has MGF $M_X(t) = \exp\left(\mu_X t + \frac{1}{2}\sigma_X^2 t^2\right)$.

Y has MGF $M_Y(t) = \exp\left(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right)$.

So the sum $Z = X + Y$ has MGF:

$$\exp\left(\mu_X\, t + \frac{1}{2}\sigma_X^2\, t^2\right)\exp\left(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right) = \exp\left\{(\mu_X + \mu_Y)t + \frac{1}{2}\left(\sigma_X^2 + \sigma_Y^2\right)t^2\right\}$$

which is the MGF of a normal variable (with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$).

So the sum of independent normal variables is a normal variable.

ie $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

Similarly, it can be shown that:

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

The variance of the *difference* is the *sum* of the variances (as we saw in the general case earlier).

These are important results to remember and are quotable in the Subject CS1 exam.

Question

If X and Y are independent standard normal variables, determine the distribution of $2X - Y$.

Solution

The resulting distribution is normal. We just need to fill in the mean and variance to obtain:

$$2X - Y \sim N(2 \times 0 - 0, 2^2 \times 1 + (-1)^2 \times 1) = N(0, 5)$$

The chapter summary starts on the next page so that you can keep
all the chapter summaries together for revision purposes.

Chapter 3 Summary

Two discrete random variables X and Y have joint probability function (PF), $P(X = x, Y = y)$. This defines how the probability is split between the different combinations of the variables. The joint PF satisfies:

$$\sum_x \sum_y P(X = x, Y = y) = 1 \quad \text{and} \quad P(X = x, Y = y) \geq 0$$

Two continuous random variables X and Y have joint probability density function (PDF), $f_{X,Y}(x, y)$. The joint PDF satisfies:

$$\int_x \int_y f_{X,Y}(x, y)\, dx\, dy = 1 \quad \text{and} \quad f_{X,Y}(x, y) \geq 0$$

We can use the joint PDF to calculate probabilities as follows:

$$P(x_1 < X < x_2, y_1 < Y < y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x, y)\, dx\, dy$$

The joint distribution function, for both discrete and continuous random variables is given by:

$$F(x, y) = P(X \leq x, Y \leq y)$$

For continuous random variables $f(x, y) = \dfrac{\partial^2}{\partial x \partial y} F(x, y)$.

The marginal distribution, *eg* $P(X = x)$ or $f_X(x)$, can be calculated using:

$$P(X = x) = \sum_y P(X = x, Y = y) \qquad f_X(x) = \int_y f_{X,Y}(x, y)\, dy$$

The conditional distribution, *eg* $P(X = x \mid Y = y)$ or $f_{X|Y=y}(x \mid y)$, can be calculated using:

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \qquad f_{X|Y=y}(x, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

The expectation of any function, $E[g(X, Y)]$, can be calculated using:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) P(X = x, Y = y) \quad \text{or} \quad \int_y \int_x g(x, y) f_{X,Y}(x, y)\, dx\, dy$$

The covariance, $\text{cov}(X,Y)$, can be calculated using:

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

The correlation coefficient, $\rho(X,Y) = \text{corr}(X,Y)$, is given by:

$$\text{corr}(X,Y) = \rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

The random variables X and Y are uncorrelated if and only if:

$$\text{corr}(X,Y) = 0 \iff \text{cov}(X,Y) = 0 \iff E(XY) = E(X)E(Y)$$

The random variables X and Y are independent if, and only if:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all values of x and y.

Independent random variables are always uncorrelated. Uncorrelated random variables are not necessarily independent.

Expectations of sums and products can be calculated using:

$$E(X + Y) = E(X) + E(Y)$$

$$E(XY) = E(X)E(Y) + Cov(X,Y)$$

$$= E(X)E(Y) \qquad\qquad\qquad \text{if } X,Y \text{ independent}$$

The above are also true for functions $g(X)$ and $h(Y)$ of the random variables.

Variances of sums can be calculated using:

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

$$= \text{var}(X) + \text{var}(Y) \qquad\qquad \text{if } X,Y \text{ independent}$$

The convolution of the marginal probability (density) functions of X and Y is the probability (density) function of $Z = X + Y$. $P(Z = z)$ or $f_Z(z)$ is given using the formulae:

$$f_Z = f_X * f_Y = \sum_x P(X = x)P(Y = z - x) \qquad \text{or} \qquad \int_x f_X(x)f_Y(z - x)\,dx$$

For independent random variables $X_1, \ldots, X_n$, and for any constants $c_1, c_2, \ldots, c_n$:

$$E\left(c_1 X_1 + \cdots + c_n X_n\right) = c_1 E\left(X_1\right) + \cdots + c_n E\left(X_n\right)$$

$$\text{var}\left(c_1 X_1 + \cdots + c_n X_n\right) = c_1^2 \,\text{var}\left(X_1\right) + \cdots + c_n^2 \,\text{var}\left(X_n\right)$$

For independent random variables $X_1, \ldots, X_n$

$$Y = X_1 + \cdots + X_n \quad \Rightarrow \quad M_Y(t) = M_{X_1}(t) \ldots M_{X_n}(t)$$

$$= \left[M_X(t)\right]^n \qquad \text{if the } X_i\text{'s are also identical}$$

For independent distributions:

‘ $Bernoulli(p) + \cdots + Bernoulli(p)$ is $Bin(n, p)$ ‘

‘ $Bin(n, p) + Bin(m, p)$ is $Bin(n + m, p)$ ‘

‘ $Geo(p) + \cdots + Geo(p)$ is $NBin(k, p)$ ‘

‘ $NBin(k, p) + NBin(m, p)$ is $NBin(k + m, p)$ ‘

‘ $Exp(\lambda) + \cdots + Exp(\lambda)$ is $Gamma(\alpha, \lambda)$ ‘

‘ $Gamma(\alpha, \lambda) + Gamma(\delta, \lambda)$ is $Gamma(\alpha + \delta, \lambda)$ ‘

‘ $\chi_m^2 + \chi_n^2$ is χ_{m+n}^2 ‘

‘ $Poi(\lambda) + Poi(\mu)$ is $Poi(\lambda + \mu)$ ‘

‘ $N(\mu_1, \sigma_1^2) \pm N(\mu_2, \sigma_2^2)$ is $N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$ ‘

Some of the notation used for the linear combinations of random variables is non-standard and is used simply to convey the results in a concise format.

The practice questions start on the next page so that you can keep
all the chapter summaries together for revision purposes.

.

Chapter 3 Practice Questions

3.1   Let X and Y have joint density function given by:

$$f(x,y) = c(x + 3y) \qquad\qquad 0 < x < 2, 0 < y < 2$$

(i)   Calculate the value of c.

(ii)   Hence, calculate $P(X < 1, Y > 0.5)$.

3.2   The continuous random variables X, Y have the bivariate PDF:

Exam style

$$f(x,y) = 2 \qquad x + y < 1, \ x > 0, \ y > 0$$

(i)   Derive the marginal PDF of Y.   [2]

(ii)   Use the result from part (i) to derive the conditional PDF of X given $Y = y$.   [1]
                                                                                                 [Total 3]

3.3   Show that, for the joint random variables M, N, where

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ for } m = 1, \ 2, \ 3, \ 4 \text{ and } n = 1, \ 2, \ 3$$

the conditional probability functions for M given $N = n$ and for N given $M = m$ are equal to the corresponding marginal distributions.

3.4   Let X and Y have joint density function:

Exam style

$$f_{X,Y}(x,y) = \frac{4}{5}\left(3x^2 + xy\right) \ 0 < x < 1, 0 < y < 1$$

Determine:

(i)   the marginal density function of X [2]

(ii)   the conditional density function of Y given $X = x$ [1]

(iii)   the covariance of X and Y.   [5]
                                                                                                 [Total 8]

3.5   Calculate the correlation coefficient of X and Y, where X and Y have the joint distribution:

			X	
		0	1	2
	1	0.1	0.1	0
Y	2	0.1	0.1	0.2
	3	0.2	0.1	0.1

3.6   Claim sizes on a home insurance policy are normally distributed about a mean of £800 and with a standard deviation of £100. Claims sizes on a car insurance policy are normally distributed about a mean of £1,200 and with a standard deviation of £300. All claims sizes are assumed to be independent.

To date, there have already been home claims amounting to £800, but no car claims. Calculate the probability that after the next 4 home claims and 3 car claims the total size of car claims exceeds the total size of the home claims.

3.7   Two discrete random variables, X and Y, have the following joint probability function:

Exam style

			X	
		1	2	3
	1	0.2	0	0.2
Y	2	0	0.2	0
	3	0.2	0	0.2

Determine:

(i) $E(X)$ [1]

(ii)   the probability distribution of $Y \mid X = 1$ [1]

(iii)   whether X and Y are correlated or not   [2]

(iv)   whether X and Y are independent or not.   [1]

[Total 5]

3.8   The random variables X and Y have joint density function given by:

$$kx^{-\alpha}e^{-y/\beta} \qquad 1 < x < \infty \,, 1 < y < \infty$$

where $\alpha > 1, \beta > 0$, and k is a constant.

Derive an expression for k in terms of α and β.

3.9   Show using convolutions that if X and Y are independent random variables and X has a χ_m^2 distribution and Y has a χ_n^2 distribution, then $X+Y$ has a χ_{m+n}^2 distribution.

3.10   Let X be a random variable with mean 3 and standard deviation 2, and let Y be a random variable with mean 4 and standard deviation 1. X and Y have a correlation coefficient of –0.3. Let $Z = X + Y$.

Exam style

Calculate:

(i) $\text{cov}(X, Z)$ [2]

(ii) $\text{var}(Z)$.   [2]

[Total 4]

3.11 X has a Poisson distribution with mean 5 and Y has a Poisson distribution with mean 10. If $\text{cov}(X, Y) = -12$, calculate the variance of Z where $Z = X - 2Y + 3$.   [2]

Exam style

3.12   Show that if X has a negative binomial distribution with parameters k and p, and Y has a negative binomial distribution with parameters m and p, and X and Y are independent, then $X+Y$ also has a negative binomial distribution, and specify its parameters.

3.13   For a certain company, claim sizes on car policies are normally distributed about a mean of £1,800 and with standard deviation £300, whereas claim sizes on home policies are normally distributed about a mean of £1,200 and with standard deviation £500. Assuming independence among all claim sizes, calculate the probability that a car claim is at least twice the size of a home claim.   [4]

Exam style

3.14   (i)   Two discrete random variables, X and Y, have the following joint probability function:

<div align="center">

X

		1	2	3	4
	1	0.2	0	0.05	0.15
Y	2	0	0.3	0.1	0.2

</div>

        Determine $\text{var}(X \mid Y = 2)$.   [3]

       (ii)   Let U and V have joint density function:

$$f_{U,V}(u,v) = \frac{48}{67}\left(2uv - u^2\right) \quad 0 < u < 1, \frac{u}{2} < v < 2$$

        Determine $E(U \mid V = v)$.   [3]

                                                                                [Total 6]

Chapter 3 Solutions

3.1   (i)   Using the result $\iint\limits_{y\,x} f(x,y)\,dx\,dy = 1$ gives:

$$\int_{y=0}^{2}\int_{x=0}^{2} c(x+3y)\,dxdy = \int_{y=0}^{2} c\left[\tfrac{1}{2}x^2 + 3xy\right]_{x=0}^{2} dy$$

$$= \int_{y=0}^{2} c(2+6y)\,dy$$

$$= c\left[2y + 3y^2\right]_{y=0}^{2}$$

$$= 16c = 1$$

$$\Rightarrow c = \tfrac{1}{16}$$

(ii)   The probability is:

$$P(X<1,Y>0.5) = \int_{y=0.5}^{2}\int_{x=0}^{1} \tfrac{1}{16}(x+3y)\,dx\,dy$$

$$= \int_{y=0.5}^{2} \tfrac{1}{16}\left[\tfrac{1}{2}x^2 + 3xy\right]_{x=0}^{1} dy$$

$$= \int_{y=0.5}^{2} \tfrac{1}{16}\left(\tfrac{1}{2} + 3y\right) dy$$

$$= \tfrac{1}{16}\left[\tfrac{1}{2}y + \tfrac{3}{2}y^2\right]_{y=0.5}^{2}$$

$$= \tfrac{51}{128} \simeq 0.398$$

3.2   (i)   The marginal PDF of Y is:

$$f_Y(y) = \int_{0}^{1-y} 2\,dx = \left[2x\right]_{0}^{1-y} = 2(1-y), \quad 0 < y < 1 \qquad [2]$$

(ii)   The conditional PDF of X given $Y = y$ is:

$$f_{X|Y=y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2}{2(1-y)} = \frac{1}{1-y}, \quad 0 < x < 1-y \qquad [1]$$

3.3   In the chapter, we found that the marginal probability functions for M and N were:

$$P_M(m) = \frac{m}{10} \qquad \text{for} \qquad m = 1, 2, 3, 4$$

and:

$$P_N(n) = \frac{1}{7 \times 2^{n-3}} \qquad \text{for} \qquad n = 1, 2, 3$$

So, dividing the joint probability function by the marginal probability function for N, we obtain:

$$P_{M|N=n}(m,n) = \frac{P_{M,N}(m,n)}{P_N(n)} = \left(\frac{m}{35 \times 2^{n-2}}\right) \div \left(\frac{1}{7 \times 2^{n-3}}\right) = \frac{m}{10}, \quad m = 1, 2, 3, 4$$

for the conditional probability function of M given $N = n$.

Similarly:

$$P_{N|M=m}(m,n) = \frac{P(N=n, M=m)}{P(M=m)} = \left(\frac{m}{35 \times 2^{n-2}}\right) \div \frac{m}{10} = \frac{1}{7 \times 2^{n-3}}, \quad n = 1, 2, 3$$

is the conditional probability function of N given $M = m$.

These are identical to the marginal distributions obtained in the chapter text.

3.4   (i)   ***Marginal density***

$$f_X(x) = \int_{y=0}^{1} \frac{4}{5}\left(3x^2 + xy\right) dy = \left[\frac{4}{5}\left(3x^2 y + \frac{1}{2}xy^2\right)\right]_{y=0}^{1} = \frac{4}{5}\left(3x^2 + \frac{1}{2}x\right)$$ [2]

(ii)   ***Conditional density***

$$f_{Y|X=x}(x,y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{\frac{4}{5}\left(3x^2 + xy\right)}{\frac{4}{5}\left(3x^2 + \frac{1}{2}x\right)} = \frac{3x^2 + xy}{3x^2 + \frac{1}{2}x} = \frac{3x + y}{3x + \frac{1}{2}}$$ [1]

(iii)   ***Covarianc*e**

Using the marginal density function of X:

$$E(X) = \int_{x=0}^{1} \frac{4}{5}\left(3x^3 + \frac{1}{2}x^2\right) dx = \frac{4}{5}\left[\frac{3}{4}x^4 + \frac{1}{6}x^3\right]_{x=0}^{1} = \frac{11}{15}$$ [1]

Obtaining the marginal density function of Y:

$$f_Y(y) = \int_{x=0}^{1} \frac{4}{5}\left(3x^2 + xy\right)dx = \frac{4}{5}\left[x^3 + \frac{1}{2}x^2 y\right]_{x=0}^{1} = \frac{4}{5}\left(1 + \frac{1}{2}y\right)$$

$$\Rightarrow \quad E(Y) = \int_{y=0}^{1} \frac{4}{5}\left(y + \frac{1}{2}y^2\right)dy = \frac{4}{5}\left[\frac{1}{2}y^2 + \frac{1}{6}y^3\right]_{y=0}^{1} = \frac{8}{15} \quad\quad [1]$$

Finally:

$$E(XY) = \int_{x=0}^{1}\int_{y=0}^{1} \frac{4}{5}\left(3x^3 y + x^2 y^2\right)dy\,dx$$

$$= \int_{x=0}^{1} \frac{4}{5}\left[\frac{3}{2}x^3 y^2 + \frac{1}{3}x^2 y^3\right]_{y=0}^{1} dx$$

$$= \int_{x=0}^{1} \frac{4}{5}\left(\frac{3}{2}x^3 + \frac{1}{3}x^2\right)dx$$

$$= \frac{4}{5}\left[\frac{3}{8}x^4 + \frac{1}{9}x^3\right]_{x=0}^{1}$$

$$= \frac{7}{18} \quad\quad [2]$$

Hence:

$$\text{cov}(X,Y) = \frac{7}{18} - \frac{11}{15}\times\frac{8}{15} = -\frac{1}{450} \quad\quad [1]$$

3.5   The covariance of X and Y was obtained in Section 2.4 to be $\text{cov}(X,Y) = 0.02$. The variances of the marginal distributions are:

$$\text{var}(X) = E(X^2) - \left[E(X)\right]^2 = 0^2 \times 0.4 + 1^2 \times 0.3 + 2^2 \times 0.3 - (0.9)^2 = 0.69$$

and: $$\text{var}(Y) = E(Y^2) - \left[E(Y)\right]^2 = 1^2 \times 0.2 + 2^2 \times 0.4 + 3^2 \times 0.4 - (2.2)^2 = 0.56$$

So the correlation coefficient is:

$$corr(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{0.02}{\sqrt{0.69 \times 0.56}} = 0.0322$$

3.6   Let X be the amount of a home insurance claim and Y the amount of a car insurance claim. Then:

$$X \sim N(800, 100^2) \quad Y \sim N(1200, 300^2)$$

We require:

$$P\big((Y_1 + Y_2 + Y_3) > (X_1 + X_2 + X_3 + X_4) + 800\big)$$

$$= P\big((Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4) > 800\big)$$

So we need the distribution of $(Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4)$:

$$(Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4) \sim N(3 \times 1200 - 4 \times 800,\ 3 \times 300^2 + 4 \times 100^2)$$

ie $(Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4) \sim N(400,\ 310\,000)$

Therefore:

$$P\big((Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4) > 800\big) = P\left(Z > \frac{800 - 400}{\sqrt{310,000}}\right)$$

$$= P(Z > 0.718)$$

$$= 1 - P(Z < 0.718) = 0.236$$

3.7   (i)   **Mean**

$$E(X) = 1 \times 0.4 + 2 \times 0.2 + 3 \times 0.4 = 2$$ [1]

Alternatively, you could use the fact that the distribution of X is symmetrical about 2.

(ii)   **Probability distribution of $Y \mid X = 1$**

Using $P(Y = y \mid X = 1) = \dfrac{P(X = 1, Y = y)}{P(X = 1)}$ and $P(X = 1) = 0.4$ gives:

$Y = 1 \mid X = 1$	$Y = 2 \mid X = 1$	$Y = 3 \mid X = 1$
0.5	0	0.5

[1]

(iii)   **Correlated?**

To calculate the correlation coefficient, we first require the covariance.

$E(X) = 2$ from part (i)

$E(Y) = 1 \times 0.4 + 2 \times 0.2 + 3 \times 0.4 = 2$

$E(XY) = 1 \times 0.2 + 2 \times 0 + 3 \times 0.2 + \cdots + 3 \times 0.2 + 6 \times 0 + 9 \times 0.2 = 4$

So $\text{cov}(X, Y) + E(XY) - E(X)E(Y) = 4 - 2 \times 2 = 0$.   [1]

Hence $\text{corr}(X, Y) = \dfrac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = 0$.

Therefore X and Y are uncorrelated.   [1]

(iv)   **Independent?**

X and Y are independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all *x* and *y*.

However $P(X = 1, Y = 1) = 0.2 \neq 0.4 \times 0.4 = P(X = 1)P(Y = 1)$.

So X and Y are not independent.   [1]

3.8   Since the pdf must integrate to 1:

$$\int_{y=1}^{\infty} \int_{x=1}^{\infty} kx^{-\alpha} e^{-y/\beta} \, dx \, dy = 1$$

Integrating over the x values gives:

$$\int_{x=1}^{\infty} kx^{-\alpha} e^{-y/\beta} \, dx = ke^{-y/\beta} \left[\frac{x^{-\alpha+1}}{-\alpha+1} \right]_1^{\infty} = \frac{ke^{-y/\beta}}{\alpha-1}$$

Integrating this over the y values gives:

$$\int_{y=1}^{\infty} \frac{ke^{-y/\beta}}{\alpha-1} \, dy = \frac{k}{\alpha-1} \left[-\beta e^{-y/\beta} \right]_1^{\infty} = \frac{k\beta e^{-1/\beta}}{\alpha-1}$$

Equating this to 1:

$$\frac{k\beta e^{-1/\beta}}{\alpha-1} = 1 \quad \Rightarrow \quad k = \frac{(\alpha-1)e^{1/\beta}}{\beta}$$

3.9   The chi-square distribution is a continuous distribution that can take any positive value. The chi-square distribution with parameter m is in fact a gamma distribution with parameters $m/2$ and $1/2$.

So, using the PDF of the gamma distribution, the PDF of the sum $Z = X + Y$ is given by the convolution formula:

$$f_Z(z) = \int f_X(x) f_Y(z-x) dx$$

$$= \int_0^z \frac{(\frac{1}{2})^{\frac{1}{2}m}}{\Gamma(\frac{1}{2}m)} x^{\frac{1}{2}m-1} e^{-\frac{1}{2}x} \frac{(\frac{1}{2})^{\frac{1}{2}n}}{\Gamma(\frac{1}{2}n)} (z-x)^{\frac{1}{2}n-1} e^{-\frac{1}{2}(z-x)} dx$$

$$= \left(\frac{1}{2}\right)^{\frac{1}{2}(m+n)} e^{-\frac{1}{2}z} \int_0^z \frac{1}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} x^{\frac{1}{2}m-1} (z-x)^{\frac{1}{2}n-1} dx$$

Using the substitution $t = x/z$ gives:

$$f_Z(z) = (½)^{½(m+n)} e^{-½z} \int_0^1 \frac{1}{\Gamma(½m)\Gamma(½n)} (zt)^{½m-1} (z-zt)^{½n-1} z\,dt$$

$$= \frac{(½)^{½(m+n)}}{\Gamma(½m+½n)} z^{½(m+n)-1} e^{-½z} \int_0^1 \frac{\Gamma(½m+½n)}{\Gamma(½m)\Gamma(½n)} t^{½m-1} (1-t)^{½n-1}\,dt$$

Since the last integral represents the total probability for a *Beta*(½m,½n) distribution, we get:

$$f_Z(z) = \frac{(½)^{½(m+n)}}{\Gamma(½m+½n)} z^{½(m+n)-1} e^{-½z} \times P[0 < Beta(½m, ½n) < 1]$$

$$= \frac{(½)^{½(m+n)}}{\Gamma(½m+½n)} z^{½(m+n)-1} e^{-½z}$$

Since this matches the PDF of a χ^2_{m+n} distribution (and Z can take any positive value), Z has a χ^2_{m+n} distribution.

It is much easier to prove this result using MGFs.

3.10   (i)   ***Covariance***

We have

$$\text{cov}(X, Z) = \text{cov}(X, X+Y)$$
$$= \text{cov}(X, X) + \text{cov}(X, Y)$$
$$= \text{var}(X) + \text{cov}(X, Y)$$

Using the correlation coefficient between *X* and *Y* gives:

$$\text{corr}(X, Y) = -0.3 = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\text{cov}(X, Y)}{\sqrt{4 \times 1}}$$

$$\Rightarrow \quad \text{cov}(X, Y) = -0.6$$

Hence:

$$\text{cov}(X, Z) = 4 - 0.6 = 3.4 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad [2]$$

(ii)   ***Variance***

Using $\text{var}(Z) = \text{cov}(Z, Z)$:

$$\text{var}(Z) = \text{cov}(X + Y, X + Y)$$

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

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

$$= 4 + 2 \times -0.6 + 1$$

$$= 3.8 \hspace{4cm} [2]$$

Note: $\text{var}(Z) \neq \text{var}(X) + \text{var}(Y)$ as X and Y are not independent.

3.11   The $+3$ term will not affect the variance, so:

$$\text{var}(Z) = \text{var}(X - 2Y + 3) = \text{var}(X - 2Y)$$

Now:

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

and:

$$\text{cov}(aX, bY) = ab\,\text{cov}(X, Y)$$

So:

$$\text{var}(X - 2Y) = \text{var}(X) + 4\,\text{var}(Y) - 2 \times 2\text{cov}(X, Y) \hspace{2cm} [1]$$

$$= 5 + 4 \times 10 - 4 \times (-12) = 93 \hspace{3cm} [1]$$

3.12   The moment generating function of X is:

$$M_X(t) = \left(\frac{pe^t}{1 - qe^t}\right)^k$$

Similarly, the MGF of Y is:

$$M_Y(t) = \left(\frac{pe^t}{1 - qe^t}\right)^m$$

Since X and Y are independent, we have:

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \left(\frac{pe^t}{1 - qe^t}\right)^k \times \left(\frac{pe^t}{1 - qe^t}\right)^m = \left(\frac{pe^t}{1 - qe^t}\right)^{k+m}$$

This is the MGF of another negative binomial distribution with parameters p and $k + m$. Hence, by uniqueness of MGFs, $X + Y$ has this distribution.

3.13   Let X be the claim size on car policies, so that $X \sim N\left(1800, 300^2\right)$.

Let Y be the claim size on home policies, so that $Y \sim N\left(1200, 500^2\right)$.

We want:

$$P(X > 2Y) = P(X - 2Y > 0)$$ [1]

So we need the distribution of $X - 2Y$:

$$X - 2Y \sim N(1800 - 2 \times 1200 , 300^2 + 4 \times 500^2)$$

$$X - 2Y \sim N\left(-600, 1090\,000\right)$$ [2]

Standardising:

$$z = \frac{0 - (-600)}{\sqrt{1,090,000}} = 0.575$$

So:

$$P(X - 2Y > 0) = P(Z > 0.575) = 1 - P(Z < 0.575)$$
$$= 1 - 0.71735 = 0.283$$ [1]

3.14   (i)   ***Conditional variance***

$$\text{var}(X \mid Y = 2) = E(X^2 \mid Y = 2) - E^2(X \mid Y = 2)$$

$$E(X \mid Y = 2) = \sum x P(X = x \mid Y = 2) = \sum x \frac{P(X = x \cap Y = 2)}{P(Y = 2)}$$

$$= 1 \times \frac{0}{0.6} + 2 \times \frac{0.3}{0.6} + 3 \times \frac{0.1}{0.6} + 4 \times \frac{0.2}{0.6}$$

$$= 2\frac{5}{6}$$ [1]

$$E(X^2 \mid Y = 2) = \sum x^2 P(X = x \mid Y = 2) = \sum x^2 \frac{P(X = x \cap Y = 2)}{P(Y = 2)}$$

$$= 1^2 \times \frac{0}{0.6} + 2^2 \times \frac{0.3}{0.6} + 3^2 \times \frac{0.1}{0.6} + 4^2 \times \frac{0.2}{0.6}$$

$$= 8\frac{5}{6}$$ [1]

So $\text{var}(X \mid Y = 2) = 8\frac{5}{6} - \left(2\frac{5}{6}\right)^2 = \frac{29}{36} = 0.80556$.   [1]

(ii) *Conditional expectation*

We require:

$$E(U \mid V = v) = \int_u u \, f(u \mid v) \, du$$

Now:

$$f(v) = \int_{u=0}^{1} \frac{48}{67}(2uv - u^2) \, du = \frac{48}{67}\left[u^2 v - \frac{1}{3}u^3\right]_{u=0}^{1} = \frac{48}{67}\left[v - \frac{1}{3}\right]$$ [1]

$$\Rightarrow \quad f(u \mid v) = \frac{f(u,v)}{f(v)} = \frac{\frac{48}{67}(2uv - u^2)}{\frac{48}{67}(v - \frac{1}{3})} = \frac{2uv - u^2}{v - \frac{1}{3}}$$ [1]

So:

$$E(U \mid V = v) = \int_{u=0}^{1} \frac{2u^2 v - u^3}{v - \frac{1}{3}} \, du = \left[\frac{\frac{2}{3}u^3 v - \frac{1}{4}u^4}{v - \frac{1}{3}}\right]_{u=0}^{1} = \frac{\frac{2}{3}v - \frac{1}{4}}{v - \frac{1}{3}}$$ [1]

4

Conditional expectation

Syllabus objectives

1.3   Expectations, conditional expectations

    1.3.1   Define the conditional expectation of one random variable given the value of another random variable, and calculate such a quantity.

    1.3.2   Show how the mean and variance of a random variable can be obtained from expected values of conditional expected values, and apply this.

0   Introduction

In this short chapter we will return to the conditional distributions, $f_{Y|X=x}(x,y)$, that we met in the last chapter. We will look at finding their expectation, $E(Y\,|\,X=x)$, and their variance, $\text{var}(Y\,|\,X=x)$. We will then see how we can obtain the unconditional values $E(Y)$ and $\text{var}(Y)$ from them.

We will use conditional expectations in a later chapter to define the regression line $E[Y\,|\,x] = \alpha + \beta x$. They will also feature in other actuarial subjects.

In particular, in CS2 we will introduce the idea of a compound random variable (which is the sum of a random number of random variables). Compound random variables are used to model total claim amounts.

1   The conditional expectation $E[Y \mid X = x]$

Definition: The conditional expectation of Y given $X = x$ is the mean of the conditional distribution of Y given $X = x$.

This mean is denoted $E[Y \mid X = x]$, or just $E[Y \mid x]$.

For a discrete distribution, this is:

$$E[Y \mid X = x] = \sum_{y} y \, P[Y = y \mid X = x] = \sum_{y} y \, \frac{P[Y = y, X = x]}{P[X = x]}$$

Question

Write down the equivalent expression for a continuous distribution.

Solution

$$E[Y \mid X = x] = \int_{y} y f(y \mid x) \, dy = \int_{y} y \, \frac{f(x, y)}{f(x)} \, dy$$

We can calculate numerical values for conditional expectations.

Question

Two random variables X and Y have the following discrete joint distribution:

		Y		
		10	20	30
X	1	0.2	0.2	0.1
	2	0.2	0.3	0

Calculate $E(Y \mid X = 1)$.

Solution

$$E(Y \mid X = 1) = \sum y P(Y = y \mid X = 1)$$

$$= 10 P(Y = 10 \mid X = 1) + 20 P(Y = 20 \mid X = 1) + 30 P(Y = 30 \mid X = 1)$$

$$= 10 \times \frac{0.2}{0.5} + 20 \times \frac{0.2}{0.5} + 30 \times \frac{0.1}{0.5}$$

$$= 10 \times 0.4 + 20 \times 0.4 + 30 \times 0.2$$

$$= 18$$

We can also calculate conditional expectations for continuous joint random variables.

Question

Let X and Y have joint density function given by:

$$f(x, y) = \frac{3}{5} x(x + y) \qquad\qquad 0 < x < 1, \ \ 0 < y < 2$$

Determine the conditional expectation $E[Y \mid X = x]$.

Solution

Using $E[Y \mid X = x] = \int\limits_y y \dfrac{f(x, y)}{f(x)} \, dy$ and recalling from Chapter 3 that $f(x) = \int\limits_y f(x, y) \, dy$:

$$f(x) = \int\limits_{y=0}^{2} \frac{3}{5}\left(x^2 + xy\right) dy$$

$$= \frac{3}{5}\left[x^2 y + \frac{1}{2} xy^2 \right]_{y=0}^{2}$$

$$= \frac{3}{5}\left(2x^2 + 2x\right) = \frac{6}{5} x(x + 1)$$

Hence:

$$E[Y \mid X = x] = \int_{y=0}^{2} y \frac{\frac{3}{5}x(x+y)}{\frac{6}{5}x(x+1)} \, dy$$

$$= \int_{y=0}^{2} y \frac{x+y}{2(x+1)} \, dy$$

$$= \int_{y=0}^{2} \frac{xy + y^2}{2(x+1)} \, dy$$

$$= \left[\frac{\frac{1}{2}xy^2 + \frac{1}{3}y^3}{2(x+1)} \right]_{y=0}^{2}$$

$$= \frac{2x + \frac{8}{3}}{2(x+1)}$$

$$= \frac{x + \frac{4}{3}}{x+1}$$

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

2   The random variable $E[Y|X]$

The conditional expectation $E[Y \mid X = x] = g(x)$, say, is, in general, a function of x. It can be thought of as the observed value of a random variable $g(X)$. The random variable $g(X)$ is denoted $E[Y \mid X]$.

We saw in the previous question that $E[Y \mid X = x] = \dfrac{3x+4}{3(x+1)}$ which is a function of x. So $E[Y \mid X]$ is the random variable $\dfrac{3X+4}{3(X+1)}$.

Observe that, although $E[Y \mid X]$ is an expectation, it does not have a single numerical value, as it is a function of the random variable X.

Note: $E[Y \mid X]$ is also referred to as the regression of Y on X.

In a later chapter the regression line will be defined as $E[Y \mid x] = \alpha + \beta x$.

$E[Y \mid X]$, like any other function of X, has its own distribution, whose properties depend on those of the distribution of X itself. Of particular importance is the expected value (the mean) of the distribution of $E[Y \mid X]$. The usefulness of considering this expected value, $E[E[Y \mid X]]$, comes from the following result, proved here in the case of continuous variables, but true in general.

Theorem: $E[E[Y \mid X]] = E[Y]$

Proof:

$$E[E[Y \mid X]] = \int E[Y \mid x] f_X(x)\, dx$$
$$= \int \left(\int y f(y \mid x)\, dy \right) f_X(x)\, dx$$
$$= \int\int y f(x,y)\, dx\, dy = E[Y]$$

We are integrating here over all possible values of x and y.

Here $f(y|x)$ represents the density function of the conditional distribution of $Y \mid X = x$. This was written as $f_{Y|X}(x,y)$ in Chapter 3.

The last two steps follow by noting that $f(y|x) = \dfrac{f(x,y)}{f_X(x)}$ and $\int f(x,y)dx = f_Y(y)$ *ie* the marginal PDF of Y.

This formula is given on page 16 of the *Tables.*

Question

(i) Calculate $E[Y]$ from first principles given that the joint density function of X and Y is:

$$f(x,y) = \frac{3}{5}x(x+y) \qquad\qquad 0 < x < 1, \quad 0 < y < 2$$

(ii) Given that $E[Y|X=x] = \dfrac{3x+4}{3(x+1)}$, calculate $E[E(Y|X)]$.

(iii) Hence, confirm that $E[Y] = E[E(Y|X)]$ for this distribution.

Solution

(i) $E(Y) = \int\limits_{y} y f(y)\,dy$, and $f(y) = \int\limits_{x} f(x,y)\,dx$. So:

$$f(y) = \int\limits_{x=0}^{1} \frac{3}{5}\left(x^2 + xy\right)dx = \frac{3}{5}\left[\frac{1}{3}x^3 + \frac{1}{2}x^2 y\right]_{x=0}^{1} = \frac{1}{5} + \frac{3}{10}y$$

$$E(Y) = \int\limits_{y=0}^{2} \frac{1}{5}y + \frac{3}{10}y^2\,dy = \left[\frac{1}{10}y^2 + \frac{1}{10}y^3\right]_{y=0}^{2} = \frac{6}{5} = 1.2$$

(ii) $E[E(Y|X)] = E\left[\dfrac{3X+4}{3(X+1)}\right] = \int\limits_{x} \dfrac{3x+4}{3(x+1)}f(x)\,dx$

But $f(x) = \dfrac{6}{5}x(x+1)$ as we saw in the previous question, so:

$$E[E(Y|X)] = \int\limits_{x=0}^{1} \frac{3x+4}{3(x+1)} \times \frac{6}{5}x(x+1)\,dx$$

$$= \frac{2}{5}\int\limits_{x=0}^{1} 3x^2 + 4x\,dx$$

$$= \frac{2}{5}\left[x^3 + 2x^2\right]_{x=0}^{1} = \frac{6}{5} = 1.2$$

(iii) Comparing our answers in parts (i) and (ii), we can see that $E[Y] = E[E(Y|X)]$.

We can also deal with situations where a random variable depends on the value of a parameter, which can itself be treated as a random quantity.

For example, consider a portfolio of motor policies. Claim amounts arising in the portfolio might have a gamma distribution with parameters α and λ. However, different policyholders might have different values of α. If this is true, we can represent the variability of α over the portfolio by giving it its own probability distribution. So we might decide that α could be treated as having an exponential distribution over the whole portfolio. We can then deduce the mean and variance of a randomly chosen claim.

Question

The random variable K has an $Exp(\lambda)$ distribution. For a given value of K, the random variable X has a $Poisson(K)$ distribution.

(i)   Obtain an expression for $E[X|K]$.

(ii)   Hence, calculate $E[X]$.

Solution

(i)   If $K = k$, then X has a $Poisson(k)$ distribution, which has mean k. So $E[X|K=k]=k$, and this can be written as $E[X|K]=K$.

(ii) $E[X]=E[E[X|K]]=E[K]=\dfrac{1}{\lambda}$.

3 The random variable $\text{var}[Y \mid X]$ and the '$E[V] + \text{var}[E]$' result

The variance of the conditional distribution of Y given $X = x$ is denoted $\text{var}[Y \mid x]$, where:

$$\text{var}[Y \mid x] = E\left[\{Y - E[Y \mid x]\}^2 \mid x\right] = E[Y^2 \mid x] - \left(E[Y \mid x]\right)^2$$

$\text{var}[Y \mid x]$ is the observed value of a random variable $\text{var}[Y \mid X]$ where:

$$\text{var}[Y \mid X] = E[Y^2 \mid X] - (E[Y \mid X])^2 = E[Y^2 \mid X] - \{g(X)\}^2$$

Hence $E[\text{var}[Y \mid X]] = E[E[Y^2 \mid X]] - E[\{g(X)\}^2] = E[Y^2] - E[\{g(X)\}^2]$ and so:

$$E[Y^2] = E[\text{var}[Y \mid X]] + E[\{g(X)\}^2]$$

So the variance of Y, $\text{var}[Y] = E(Y^2) - [E(Y)]^2$, is given by:

$$E[\text{var}(Y \mid X)] + E[\{g(X)\}^2] - [E\{g(X)\}]^2 = E[\text{var}(Y \mid X)] + \text{var}[g(X)]$$

ie $\text{var}[Y] = E[\text{var}[Y \mid X]] + \text{var}[E[Y \mid X]]$.

This formula is given on page 16 of the *Tables*.

Question

Evaluate $\text{var}[Y \mid X = 1]$ for the joint distribution:

		Y		
		10	20	30
X	1	0.2	0.2	0.1
	2	0.2	0.3	0

Solution

$\text{var}[Y \mid X = 1] = E(Y^2 \mid X = 1) - E^2(Y \mid X = 1)$.

We know that $E(Y \mid X = 1) = 18$ from Section 1.

We now want $E(Y^2 \mid X = 1)$:

$$E(Y^2 \mid X = 1) = \sum y^2 P(Y = y \mid X = 1) = 10^2 P(Y = 10 \mid X = 1)$$

$$+ 20^2 P(Y = 20 \mid X = 1) + 30^2 P(Y = 30 \mid X = 1)$$

$$= 100 \times \frac{0.2}{0.5} + 400 \times \frac{0.2}{0.5} + 900 \times \frac{0.1}{0.5} = 380$$

So $\text{var}[Y \mid X = 1] = 380 - 18^2 = 56$.

Question

The random variable K has an $Exp(\lambda)$ distribution. For a given value of K, the random variable X has a $Poisson(K)$ distribution.

Obtain an expression for $\text{var}[X \mid K]$. Hence derive an expression for $\text{var}(X)$.

Solution

If $K = k$, X has a $Poisson(k)$ distribution, which has variance k.

So $\text{var}[X \mid K = k] = k$ which can be written as $\text{var}[X \mid K] = K$.

Using the result given in this section, we have:

$$\text{var}[X] = E\big[\text{var}(X \mid K)\big] + \text{var}\big[E(X \mid K)\big] = E[K] + \text{var}[K]$$

But K has an exponential distribution. So, using the formulae for the mean and variance of an exponential random variable, we have:

$$\text{var}[X] = E[K] + \text{var}[K] = \frac{1}{\lambda} + \frac{1}{\lambda^2} = \frac{\lambda + 1}{\lambda^2}$$

Chapter 4 Summary

$E(Y|X)$ is the mean of the conditional distribution of Y given X (which was defined in Chapter 3). The formulae for the conditional mean are:

$$E[Y|X=x] = \sum_i y_i P[Y=y_i|X=x] = \sum_i y_i \frac{P[Y=y_i, X=x]}{P[X=x]} \qquad \text{(discrete case)}$$

$$E[Y|X=x] = \int_y y\, f(y|x)\, dy = \int_y y\, \frac{f(x,y)}{f(x)}\, dy \qquad \text{(continuous case)}$$

$\mathrm{var}(Y|X)$ is the variance of the conditional distribution of Y given X. It is given by:

$$\mathrm{var}(Y|X) = E(Y^2|X) - E^2(Y|X)$$

The unconditional mean and variance can be found from the conditional mean and variance using the formulae:

$$E[Y] = E[E(Y|X)]$$

$$\mathrm{var}[Y] = E[\mathrm{Var}(Y|X)] + \mathrm{var}[E(Y|X)]$$

The practice questions start on the next page so that you can
keep the chapter summaries together for revision purposes.

Chapter 4 Practice Questions

4.1   Calculate $E(X | Y = 10)$ for the joint distribution:

		Y	
	10	20	30
X 1	0.2	0.2	0.1
2	0.2	0.3	0

4.2   The random variable V has a Poisson distribution with mean 5. For a given value of V, the
Exam style   random variable U is distributed as follows:

$$U | (V = v) \sim U(0, v)$$

Obtain the mean and variance of the marginal distribution of U.   [4]

4.3   (i)   Given that X and Y are continuous random variables, prove from first principles that:
Exam style
$$E(Y) = E\big[E(Y | X)\big]$$ [3]

(ii)   The random variable X has a gamma distribution with parameters $\alpha = 3$ and $\lambda = 2$. Y is
a related variable with conditional mean and variance of:

$$E(Y | X = x) = 3x + 1 \qquad\qquad \text{var}(Y | X = x) = 2x^2 + 5$$

Calculate the unconditional mean and standard deviation of Y.   [5]
[Total 8]

4.4   Suppose that a random variable X has a standard normal distribution, and the conditional
Exam style   distribution of a Poisson random variable Y, given the value of $X = x$, has expectation
$g(x) = x^2 + 1$.

Determine $E(Y)$ and $\text{var}(Y)$.   [5]

4.5   The table below shows the bivariate probability distribution for two discrete random variables X
        and Y :

	$X = 0$	$X = 1$	$X = 2$
$Y = 1$	0.15	0.20	0.25
$Y = 2$	0.05	0.15	0.20

Calculate the value of $E(X \mid Y = 2)$.

Chapter 4 Solutions

4.1 $E(X \mid Y = 10) = \sum x P(X = x \mid Y = 10)$

$= 1 P(X = 1 \mid Y = 10) + 2 P(X = 2 \mid Y = 10)$

$= 1 \times \dfrac{0.2}{0.4} + 2 \times \dfrac{0.2}{0.4}$

$= 1 \times 0.5 + 2 \times 0.5 = 1.5$

Alternatively, we can see this directly by noting that if we know that $Y = 10$, then X is equally likely to be 1 or 2. Since this is a symmetrical distribution, the conditional mean is just 1.5.

4.2   We are given in the question that:

$U \mid V = v \sim U(0, v) \qquad V \sim Poi(5)$

So:

$E(V) = 5 \qquad var(V) = 5$ [½]

and:

$E(U \mid V) = \frac{1}{2} V \qquad var(U \mid V) = \frac{1}{12} V^2$ [½]

Using the formulae on page 16 of the *Tables*, we have:

$E[U] = E\big[E[U \mid V]\big] \qquad var[U] = var\big[E[U \mid V]\big] + E\big[var[U \mid V]\big]$

Therefore:

$E[U] = E\big[E(U \mid V)\big] = E\left[\frac{1}{2} V\right] = \frac{1}{2} E[V] = 2\frac{1}{2}$ [1]

$var[U] = var\big[E(U \mid V)\big] + E\big[var(U \mid V)\big]$

$= var\left[\frac{1}{2} V\right] + E\left[\frac{1}{12} V^2\right]$

$= \frac{1}{4} var[V] + \frac{1}{12} E[V^2]$ [1]

Since $E[V^2] = var[V] + E^2[V]$, we have:

$var[U] = \frac{1}{4} \times 5 + \frac{1}{12}(5 + 5^2) = 3\frac{3}{4}$ [1]

4.3   (i)   **Proof**

$E(Y \mid X = x)$ is a function of x so using $E[g(x)] = \int_x g(x) f(x) dx$, we have:

$$E\big[E(Y \mid X)\big] = \int_x E(Y \mid x) f(x) dx$$ [1]

Using the definition of $E(Y \mid X = x) = \int_y y f(y \mid x) dy$ gives:

$$E\big[E(Y \mid X)\big] = \int_x \left(\int_y y f(y \mid x) dy \right) f(x) dx$$

Using the definition $f(y \mid x) = \dfrac{f(x,y)}{f(x)}$ gives:

$$E\big[E(Y \mid X)\big] = \int_x \left(\int_y y \frac{f(x,y)}{f(x)} dy \right) f(x) dx$$ [1]

$$= \int_x \int_y y f(x,y) dy dx$$

$$= \int_y y \left(\int_x f(x,y) dx \right) dy$$

Since integrating the joint density function, $f(x,y)$, over all values of x gives the marginal density function, $f(y)$, we have:

$$E\big[E(Y \mid X)\big] = \int_y y f(y) dy = E(Y)$$ [1]

(ii)   **Calculate the unconditional mean and variance**

The mean and variance of X are given by:

$$E(X) = \frac{\alpha}{\lambda} = \frac{3}{2} = 1.5 \qquad \text{var}(X) = \frac{\alpha}{\lambda^2} = \frac{3}{4} = 0.75$$ [1]

Using the result from part (i), *ie* $E(Y) = E[E(Y \mid X)]$:

$$E(Y) = E[3X + 1] = 3E[X] + 1 = 3 \times 1.5 + 1 = 5.5$$ [1]

Using the result $var(Y) = var[E(Y \mid X)] + E[var(Y \mid X)]$ from page 16 of the *Tables*:

$$var(Y) = E[2X^2 + 5] + var[3X + 1]$$

$$= 2E[X^2] + 5 + 9\,var[X] \tag{1}$$

Using the fact that $E(X^2) = var(X) + E^2(X) = 0.75 + 1.5^2 = 3$: $\qquad$ [1]

$$var(Y) = 2 \times 3 + 5 + 9 \times 0.75 = 17.75$$

So the standard deviation is $\sqrt{17.75} = 4.21$. $\qquad$ [1]

4.4 $\quad$ We have $X \sim N(0,1)$. So:

$$E(X) = 0 \quad \text{and} \quad var(X) = 1$$

We also have $(Y \mid X = x) \sim Poi(x^2 + 1)$. Hence:

$$E(Y \mid X = x) = x^2 + 1 \quad \text{and} \quad var(Y \mid X = x) = x^2 + 1 \tag{1}$$

Using the expectation formula gives:

$$E(Y) = E[E(Y \mid X = x)] = E[X^2 + 1] = E(X^2) + 1 = 1 + 1 = 2 \tag{1}$$

Now using the variance formula:

$$var(Y) = E[var(Y \mid X)] + var[E(Y \mid X)] = E(X^2 + 1) + var(X^2 + 1)$$

$$= E(X^2) + 1 + var(X^2) \tag{1}$$

Now $E(X^2) = var(X) + E^2(X) = 1 + 0 = 1$. However, we'll have to do $var(X^2)$ from first principles:

$$var(X^2) = E(X^4) - E^2(X^2)$$

Looking up moments for the $N(0,1)$ on page 10 of the *Tables*, we see that:

$$E(X^4) = \frac{1}{2^{4/2}} \frac{\Gamma(1+4)}{\Gamma(1+\frac{4}{2})} = \frac{1}{2^2} \frac{\Gamma(5)}{\Gamma(3)} = \frac{1}{4} \frac{4!}{2!} = 3$$

Using $E(X^2) = var(X) + E^2(X) = 1 + 0 = 1$ again, gives:

$$var(X^2) = E(X^4) - E^2(X^2) = 3 - 1^2 = 2$$

Hence:

$$var(Y) = E(X^2) + 1 + var(X^2) = 1 + 1 + 2 = 4 \tag{2}$$

4.5 $E(X \mid Y = 2) = \sum_x x P(X \mid Y = 2) = \sum_x x \dfrac{P(X = x, Y = 2)}{P(Y = 2)}$

$= 0 \times \dfrac{0.05}{0.4} + 1 \times \dfrac{0.15}{0.4} + 2 \times \dfrac{0.2}{0.4} = 1.375$

End of Part 1

What next?

1. Briefly **review** the key areas of Part 1 and/or re-read the **summaries** at the end of Chapters 1 to 4.

2. Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 1. If you don't have time to do them all, you could save the remainder for use as part of your revision.

3. Attempt **Assignment X1**.

Time to consider …

… 'learning and revision' products

Marking – Recall that you can buy *Series Marking* or more flexible *Marking Vouchers* to have your assignments marked by ActEd. Results of surveys suggest that attempting the assignments and having them marked improves your chances of passing the exam. One student said:

> 'The insight into my interpretation of the questions compared with that of the model solutions was helpful. Also, the pointers as to how to shorten the amount of work required to reach an answer were appreciated.'

Face-to-face and Live Online Tutorials – If you haven't yet booked a tutorial, then maybe now is the time to do so. Feedback on ActEd tutorials is extremely positive:

> 'I would not pass exams without ActEd's lovely, clever, patient tutors. I don't know how you managed to find so many great teachers. Thank you!'

Online Classroom – Alternatively / additionally, you might consider the Online Classroom to give you access to ActEd's expert tuition and additional support:

> 'Please do an online classroom for everything. It is amazing.'

You can find lots more information, including demos and our *Tuition Bulletin*, on our website at www.ActEd.co.uk.

Buy online at www.ActEd.co.uk/estore

5

The Central Limit Theorem

Syllabus objectives

1.5   Central Limit Theorem - statement and application

   1.5.1   State the Central Limit Theorem for a sequence of independent, identically distributed random variables.

   1.5.2.   Generate simulated values from a given distribution and compare the sampling distribution with the Normal.

0   Introduction

The Central Limit Theorem is perhaps the most important result in statistics. It provides the basis for large-sample inference about a population mean when the population distribution is unknown and more importantly does not need to be known. It also provides the basis for large-sample inference about a population proportion, for example, in initial mortality rates at given age *x*, or in opinion polls and surveys. It is one of the reasons for the importance of the normal distribution in statistics.

We will study statistical inference in (Hypothesis testing).

Basically, the Central Limit Theorem gives us an approximate distribution of the mean, $\overline{X}$, from *any* distribution. The usefulness of this, though not apparent now, will become clear in the next four chapters.

The Central Limit Theorem can also be used to give approximations to other distributions. This is useful if we are calculating probabilities that would take too long otherwise. For example, $P(X < 30)$ where $X \sim Bin(100, 0.3)$ would require us to work out 30 probabilities and then add them all up. If we use a normal approximation, the calculation of the probability is much simpler.

1   The Central Limit Theorem

1.1   Definition

If $X_1, X_2, \ldots, X_n$ is a sequence of independent, identically distributed (iid) random variables with finite mean μ and finite (non-zero) variance σ^2 then the distribution of $\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ approaches the standard normal distribution, $N(0,1)$, as $n \to \infty$.

It is not necessary to be able to prove this result. Remember that $\overline{X}$ is the sample mean, calculated as $\overline{X} = \dfrac{1}{n} \sum\limits_{i=1}^{n} X_i$.

1.2   Practical uses

The way the Central Limit Theorem is used in practice is to provide useful normal approximations to the distributions of particular functions of a set of iid random variables.

Therefore both $\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ and $\dfrac{\sum X_i - n\mu}{\sqrt{n\sigma^2}}$ are approximately distributed as $N(0,1)$ for large *n*.

The second of these expressions can be obtained from the first just by multiplying top and bottom through by the sample size n.

Alternatively the unstandardised forms can be used. Thus $\overline{X}$ is approximately $N(\mu, \sigma^2 / n)$ and $\sum X_i$ is approximately $N(n\mu, n\sigma^2)$.

In fact the expressions for the mean and variance are exact, that is, $E(\overline{X}) = \mu$ and $\mathrm{var}(\overline{X}) = \dfrac{\sigma^2}{n}$. It is the shape of the curve that is approximate.

As a notation the symbol '$\doteqdot$' is used to mean 'is approximately distributed', so we can write the statements in the preceding paragraph as $\overline{X} \doteqdot N(\mu, \sigma^2 / n)$ and $\sum X_i \doteqdot N(n\mu, n\sigma^2)$.

An obvious question is: what is large *n*?

A common answer is simply $n \geq 30$ but this is too simple an answer. A fuller answer is that it depends on the shape of the population, that is, the distribution of X_i, and in particular how skewed it is.

If this population distribution is fairly symmetric even though non-normal, then $n = 10$ may be large enough; whereas if the distribution is very skewed, $n = 50$ or more may be necessary.

In other words, the closer the original distribution is to being symmetrical, the better the approximation given by the Central Limit Theorem.

Question

It is assumed that the number of claims arriving at an insurance company per working day has a mean of 40 and a standard deviation of 12. A survey was conducted over 50 working days. Calculate the probability that the sample mean number of claims arriving per working day was less than 35.

Solution

Using the notation given in the Core Reading, $\mu = 40$, $\sigma = 12$, $n = 50$.

The Central Limit Theorem states that $\overline{X} \overset{.}{\sim} N(40, 12^2/50)$.

We want $P(\overline{X} < 35)$. Standardising in the usual way:

$$P(\overline{X} < 35) \approx P\left(Z < \frac{35 - 40}{\sqrt{12^2/50}}\right)$$

$$= P(Z < -2.946) = 1 - P(Z < 2.946) = 1 - 0.99839 = 0.00161$$

We can also use the Central Limit Theorem to answer questions about the distribution of $\sum X_i$, rather than $\overline{X}$.

Question

The cost of repairing a vehicle following an accident has mean \$6,200 and standard deviation \$650. A study was carried out into 65 vehicles that had been involved in accidents. Calculate the probability that the total repair bill for the vehicles exceeded \$400,000.

Solution

Using the notation given in the Core Reading, we have $\mu = 6,200$, $\sigma = 650$, $n = 65$. Also let $Z \sim N(0,1)$.

We want the probability that the total repair bill, T is greater than 400,000. The Central Limit Theorem states that:

$$T \overset{.}{\sim} N(65 \times 6200, 65 \times 650^2) = N(403000, 5240^2)$$

So the probability is found as follows:

$$P(T > 400,000) \approx P\left(Z > \frac{400,000 - 403,000}{5,240}\right) = P(Z > -0.572) = P(Z < 0.572) = 0.71634$$

2   Normal approximations

We can use Central Limit Theorem to obtain approximations to the binomial, Poisson and gamma distributions. This is useful for calculating probabilities and obtaining confidence intervals and carrying out hypothesis tests on a piece of paper. However, it is easy for a computer to calculate exact probabilities, confidence intervals and hypothesis tests. Hence, these approximations are not as important as they used to be.

2.1   Binomial distribution $Bin(n, p)$

Let X_i be iid Bernoulli random variables, that is, $Bin(1, p)$, so that

$$P(X_i = 1) = p$$

$$P(X_i = 0) = 1 - p$$

In other words X_i is the number of successes in a single Bernoulli trial.

Consider $X_1, X_2, ..., X_n$, a sequence of such variables. This is precisely the binomial situation and $X = \sum X_i$ is the number of successes in the *n* trials.

So $X = \sum X_i \sim Bin(n, p)$. Also note that $\dfrac{X}{n} = \bar{X}$. As a result of the Central Limit Theorem it can be said that, for large *n*:

$$\bar{X} \doteq N\left(\mu, \sigma^2 / n\right) \text{ or } \sum X_i \doteq N(n\mu, n\sigma^2)$$

For the Bernoulli distribution:

$$\mu = E[X_i] = p \quad \text{and} \quad \sigma^2 = \text{var}[X_i] = p(1 - p)$$

Therefore $\sum X_i \doteq N(np, np(1 - p))$ for large *n*, which is of course the normal approximation to the binomial.

Basically, we approximate using a normal distribution, which has the same mean and variance as the binomial distribution.

Question

Given that $X \sim Bin(n, p)$, derive the mean and variance of $\bar{X}$, and hence write down the distribution of $\bar{X}$.

Solution

Since $\bar{X} = \dfrac{\sum X_i}{n}$, then:

$$E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}E\left(\sum X_i\right) = \frac{1}{n}np = p$$

$$\operatorname{var}(\bar{X}) = \operatorname{var}\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2}\operatorname{var}\left(\sum X_i\right) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Therefore $\bar{X} \doteq N\left(p, \dfrac{p(1-p)}{n}\right)$.

What is 'large n'? A commonly quoted rule of thumb is that the approximation can be used only when both np and $n(1-p)$ are greater than 5. The 'only when' is a bit severe. It is more a case of the approximation is less good if either is less than 5. However, this rule of thumb agrees with the answer that it depends on the symmetry/skewness of the population.

Note that when $p = 0.5$ the Bernoulli distribution is symmetrical. In this case both np and $n(1-p)$ equal 5 when $n = 10$, and so the rule of thumb suggests that $n = 10$ is large enough.

As p moves away from 0.5 towards either 0 or 1 the Bernoulli distribution becomes more severely skewed. For example, when $p = 0.2$ or 0.8 the rule of thumb gives $n = 25$ as large enough, but, when $p = 0.05$ or 0.95 the rule of thumb gives $n = 100$ as large enough.

Recall from Chapter 1, that the binomial distribution can also be approximated by the Poisson distribution. This approximation was valid when *n* was large and p was small. This contrasts with the normal approximation, which requires *n* to be large and p to be close to ½ (although, as *n* gets larger the normal approximation works well even if p is not close to ½).

2.2   Poisson distribution

Let X_i, $i = 1, 2, ..., n$ be iid $Poi(\lambda)$ random variables.

So $\mu = E[X_i] = \lambda$ and $\sigma^2 = \operatorname{var}[X_i] = \lambda$.

The Central Limit Theorem implies that

$$\sum X_i \doteq N(n\lambda, n\lambda) \text{ for large } n$$

But $\sum X_i \sim Poi(n\lambda)$ and so, for large n, $Poi(n\lambda) \doteq N(n\lambda, n\lambda)$, or, equivalently,

$Poi(\lambda) \doteq N(\lambda, \lambda)$ for large λ.

Again, we are approximating using a normal distribution, which has the same mean and variance as the Poisson distribution.

Question

Show that $\sum X_i \sim Poi(n\lambda)$, where X_i is $Poi(\lambda)$ for all i.

Solution

Recall that the Poisson distribution is additive, *ie*:

$$X \sim Poi(\lambda) \text{ and } Y \sim Poi(\mu) \;\Rightarrow\; X + Y \sim Poi(\lambda + \mu)$$

Therefore $\sum X_i \sim Poi(n\lambda)$.

A rule of thumb for this one is that the approximation is good if $\lambda > 5$. However since extensive tables for a range of values of λ are available, it is only needed in practice for much larger values of λ.

Remember that the Poisson distribution is the limiting case of the binomial with $\lambda = np$ as $n \to \infty$ and $p \to 0$. So this is consistent with the rule for the binomial.

The normal approximations to the binomial and Poisson distributions (both discrete) are the most commonly used in practice, and they are needed as the direct calculation of probabilities is computationally awkward without them.

This was the point mentioned in the introduction. To calculate $P(X < 30)$ where $X \sim Bin(100, 0.3)$, we'd need to work out 30 probabilities and then add them all up.

2.3 Gamma distribution

Let X_i, $i = 1, 2, \ldots, n$ be a sequence of iid exponential (λ) variables and let Y be their sum.

The exponential distribution has mean $\mu = 1/\lambda$ and variance $\sigma^2 = 1/\lambda^2$.

Therefore for large n, $Y = \sum X_i \doteq N(n/\lambda, n/\lambda^2)$

Therefore Y, which is $Gamma(n, \lambda)$, will have a normal approximation for large values of n.

Recall that if $X_i \sim Exp(\lambda)$ then $\sum X_i \sim Gamma(n, \lambda)$.

Since $\chi_k^2 \equiv Gamma(k/2, 1/2)$, χ_k^2 will have a normal approximation $N(k, 2k)$ for large values of its degrees of freedom k.

These approximations are poorer than those used for the binomial and Poisson distributions due to the skewness of the Gamma distribution. It is therefore preferable to make use of the *exact* result from Chapter 2 that if $X \sim Gamma(\alpha, \lambda)$ then $2\lambda X \sim \chi_{2\alpha}^2$. We can then use the χ^2 tables to obtain the probabilities.

3   The continuity correction

When dealing with the normal approximations to the binomial and Poisson distributions, which are both discrete, a discrete distribution is being approximated by a continuous one. When using such an approximation the change from discrete to continuous must be allowed for.
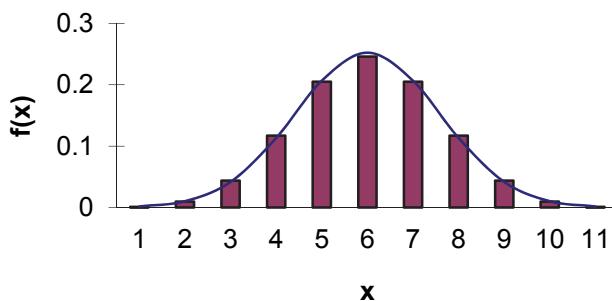
For an integer-valued discrete distribution, such as the binomial or Poisson, it is perfectly reasonable to consider individual probabilities such as $P(X = 4)$. However if X is continuous, such as the normal, $P(X = 4)$ is not meaningful and is taken to be zero. For a continuous variable it is sensible to consider only the probability that X lies in some interval.

For a continuous distribution it is not useful to think about the probability of a random variable being exactly equal to a value: for example, for a continuous distribution:

$$P(X = 4) = P(4 \leq X \leq 4) = \int_4^4 f(x)\,dx = 0$$

To allow for this a continuity correction must be used. Essentially it corresponds to treating the integer values as being rounded to the nearest integer.

The diagram below illustrates the problem. The bars correspond to the probabilities for a $Bin(10, 0.5)$ distribution, whereas the graph corresponds to the probability density function for the normal approximation.

Since the binomial is a discrete distribution there are no probabilities for non-integer values, whereas the normal approximation can take any value. To compensate for the 'gaps' between the bars, we suppose that they are actually rounded to the nearest integer. For example, the $x = 6$ bar is assumed to represent values between $x = 5.5$ and $x = 6.5$.

So to use the continuity correction in practice, for example,
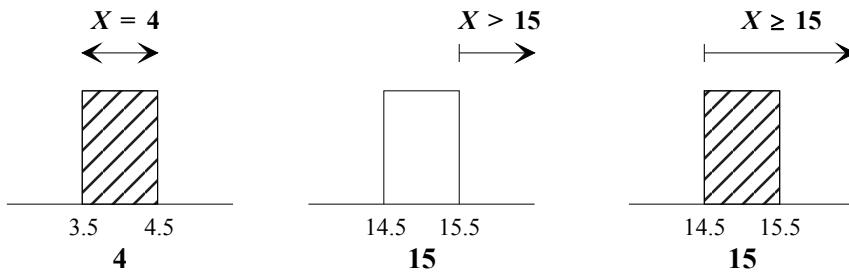
$X = 4$	**is equivalent to**	**'$3.5 < X < 4.5$'**
$X > 15$	**is equivalent to**	**'$X > 15.5$'**
$X \geq 15$	**is equivalent to**	**'$X > 14.5$'**

Take the first example. All values that are contained in the interval $3.5 < X < 4.5$ become 4 when rounded to the nearest whole number. Similarly, values in the interval $X > 15.5$, become values in the interval $X > 15$ when rounded to the nearest whole number.

Alternatively, considering the bars on the graph:

$X = 4$ must, obviously, include all of the $X = 4$ bar which goes from 3.5 to 4.5.

$X > 15$ must not include the $X = 15$ bar (as it is a strict inequality), therefore it should start from 15.5 (the upper end of the 15 bar).

$X \geq 15$ includes the $X = 15$ bar and higher, therefore it should start from 14.5 (the lower end of the 15 bar).
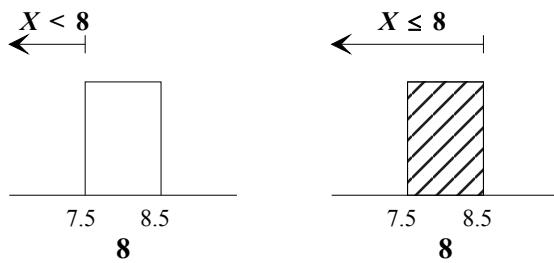
Question

Draw the corresponding diagrams for:

(i) $X < 8$ (ii) $X \leq 8$

Hence give the continuity-corrected inequalities.

Solution

(i) $X < 8$ must not include the $X = 8$ bar (as it is a strict inequality), therefore it should start from 7.5 (the lower end of the 8 bar). This gives $X < 7.5$.

(ii) $X \leq 8$ includes the $X = 8$ bar and lower, therefore it should start from 8.5 (the upper end of the 8 bar). This gives $X < 8.5$.

Let's now see how to calculate a normal approximation to a probability in a discrete distribution, allowing correctly for the continuity correction.

Question

Let *X* be a Poisson variable with parameter 20. Use the normal approximation to obtain a value for $P(X \leq 15)$ and use tables to compare with the exact value.

Solution

We have:

$$X \sim Poi(20) \quad \Rightarrow X \doteq N(20, 20) \Rightarrow \frac{X - 20}{\sqrt{20}} \doteq N(0, 1)$$

$$P(X \leq 15) \equiv P(X < 15.5) \text{ : using continuity correction}$$

$$\approx P\left(Z < \frac{15.5 - 20}{\sqrt{20}}\right) = P(Z < -1.006)$$

$$= 1 - 0.84279 \text{, interpolating in tables to be as accurate as possible}$$

$$= 0.15721.$$

From Poisson tables, $P(X \leq 15) = 0.15651$.

Error $= 0.0007$ or a 0.45% relative error.

It was mentioned earlier that approximations to the binomial and Poisson distributions are used because the direct calculation of probabilities is computationally awkward.

We are now in a position to look at the following example.

Question

The average number of calls received per hour by an insurance company's switchboard is 5. Calculate the probability that in a working day of eight hours, the number of telephone calls received will be:

(i)   exactly 36

(ii)   between 42 and 45 inclusive.

Assuming that the number of calls has a Poisson distribution, calculate the exact probabilities and also the approximate probabilities using a normal approximation.

Solution

If the number of calls per day is X, then $X \sim Poi(40)$. The exact probabilities are:

(i) $P(X = 36) = \dfrac{40^{36} e^{-40}}{36!} = 0.0539$

(ii)   In order to calculate this, we sum the probabilities of getting 42, 43, 44 and 45:

$$P(42 \leq X \leq 45) = \frac{40^{42} e^{-40}}{42!} + \frac{40^{43} e^{-40}}{43!} + \frac{40^{44} e^{-40}}{44!} + \frac{40^{45} e^{-40}}{45!}$$

$$= 0.0585 + 0.0544 + 0.0495 + 0.0440$$

$$= 0.2064$$

The normal approximation to this Poisson distribution would be $N(40, 40)$. Calculating the probabilities again, and using continuity corrections:

(i) $P(X = 36) \approx P(35.5 < X < 36.5)$

$$= P\left(\frac{35.5 - 40}{\sqrt{40}} < Z < \frac{36.5 - 40}{\sqrt{40}} \right)$$

$$= \Phi(-0.553) - \Phi(-0.712)$$

$$= 0.7617 - 0.7099 = 0.0518$$

(ii) $P(42 \leq X \leq 45) \approx P(41.5 < X < 45.5)$

$$= P\left(\frac{41.5 - 40}{\sqrt{40}} < Z < \frac{45.5 - 40}{\sqrt{40}} \right)$$

$$= P(0.237 < Z < 0.870)$$

$$= \Phi(0.870) - \Phi(0.237) = 0.8078 - 0.5937 = 0.2141$$

It is evident that in most cases using an approximation makes the calculations easier, and that the values obtained are fairly close to the exact probabilities.

Question

Use a normal approximation to calculate an approximate value for the probability that an observation from a *Gamma*(25,50) random variable falls between 0.4 and 0.8.

Solution

The mean and variance of a general gamma distribution are $\dfrac{\alpha}{\lambda}$ and $\dfrac{\alpha}{\lambda^2}$, so here the mean and variance are 0.5 and 0.01 respectively. If X is the gamma random variable, then we will use $X \sim N(0.5, 0.01)$:

$$P(0.4 < X < 0.8) \approx P(-1 < Z < 3)$$

$$= \Phi(3) - \Phi(-1)$$

$$= \Phi(3) - \left[1 - \Phi(1)\right]$$

$$= 0.99865 - 0.15866 = 0.840$$

No continuity correction is required, as we started with a continuous distribution.

The exact answer is 0.8387.

We can also use the Central Limit Theorem to calculate approximate probabilities relating to a sample mean obtained from a random sample from a continuous distribution.

Question

Calculate the approximate probability that the mean of a sample of 10 observations from a *Beta*(10,10) random variable falls between 0.48 and 0.52.

Solution

Using the formulae on page 13 of the *Tables, t*he *Beta*(10,10) distribution has mean $\dfrac{10}{10+10} = 0.5$ and variance:

$$\frac{10 \times 10}{(10+10)^2(10+10+1)} = 0.01190$$

We have a sample of 10 values. From the Central Limit Theorem, $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ (approximately),

so here $\overline{X} \sim N\left(0.5, \dfrac{0.01190}{10}\right)$, and:

$$P(0.48 < \overline{X} < 0.52) \approx P(-0.5798 < Z < 0.5798) = \Phi(0.5798) - \Phi(-0.5798)$$

$$= \Phi(0.5798) - (1 - \Phi(0.5798)) = 0.71897 - 0.28103 = 0.43794$$

No continuity correction is required, as the beta distribution is continuous.

4   Comparing simulated samples

This section of the Core Reading refers to the use of R to simulate random samples. This material is not explained in detail here; we cover it in the material for the second paper of Subject CS1.

We saw in a previous unit how to use R to simulate samples from standard distributions. We can then obtain the sum or mean of each of these samples.

The following R code uses a loop to obtain the means of 1,000 samples of size 40 from a Poisson distribution with mean 5. It then stores these sample means in the vector xbar:

```
set.seed(23)
xbar<-rep(0,1000)
for (i in 1:1000)
{x<-rpois(40,5);xbar[i]<-mean(x)}
```

Note that we have used the `set.seed` **function so that you can obtain exactly the same results for your simulation.**
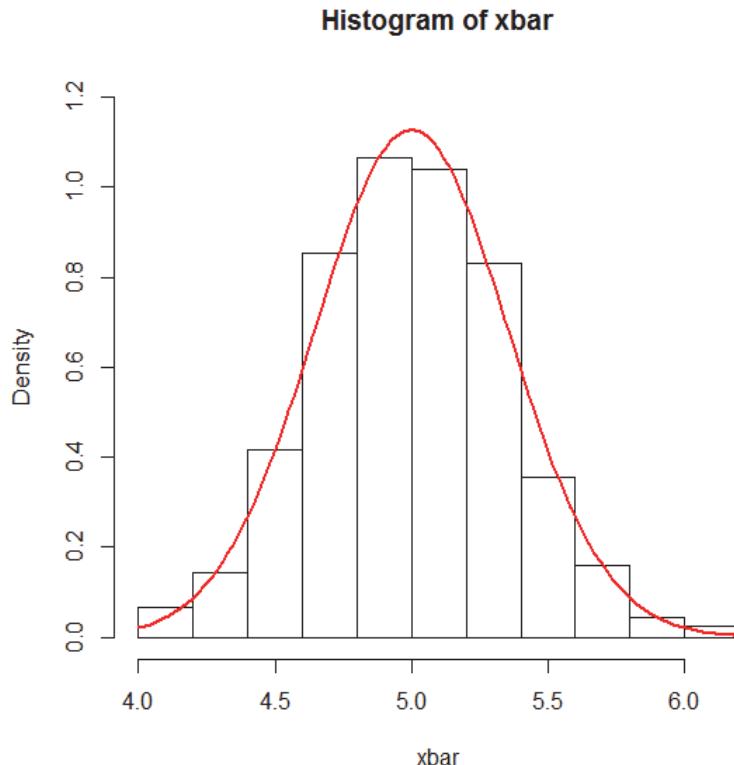
The Central Limit Theorem tells us that the distribution of the sample means will approximately have a *N*(5,0.125) distribution.

The simulated mean and variance of $\bar{x}$ are 5.01135 and 0.1250763 which are very close.

We can compare our observed distribution of the sample means with the Central Limit Theorem by a histogram of the sample means (using the R function `hist`) and superimposing the normal distribution curve (using the R function `curve`):

```
hist(xbar, prob=TRUE, ylim=c(0,1.2))
curve(dnorm(x,mean=5,sd=sqrt(0.125)), add=TRUE, lwd=2,
col="red")
```

Histogram of xbar

Another method of comparing the distribution of our sample means, $\bar{x}$, with the normal distribution is to examine the quantiles.
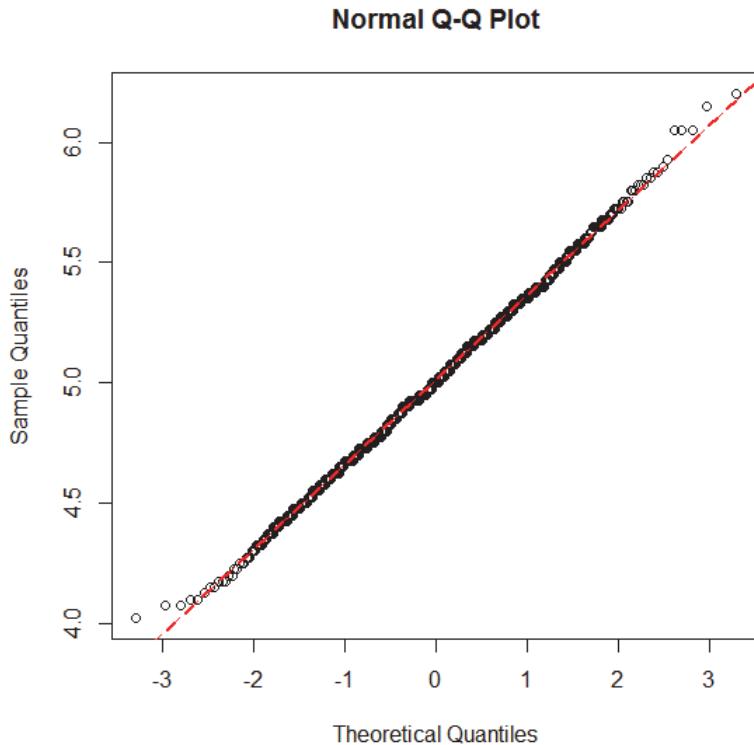
R | In R we can find the quantiles of $\bar{x}$ using the quantile function. Using the default setting (type 7) to obtain the sample lower quartile, median and upper quartile gives 4.775, 5.000 and 5.250, respectively. However in Subject CS1 we prefer to use type 5 or type 6.

In R, we can find the quartiles of the normal distribution using the `qnorm` function. This gives a lower quartile, median and upper quartile of 4.762, 5.000 and 5.238, respectively.

We observe that our distribution of the sample means is slightly more spread out in the tails – which is what we observed in the previous diagram.

R | A quick way to compare all the quantiles in one go is by drawing a QQ-plot using the R function `qqnorm`.

Normal Q-Q Plot

If our sample quantiles coincide with the quantiles of the normal distribution we would observe a perfect diagonal line (which we have added to the diagram for clarity). For our example we can see that $\bar{x}$ and the normal distribution are very similar except in the tails where we see that $\bar{x}$ has a lighter lower tail and a heavier upper tail than the normal distribution.

Chapter 5 Summary

The Central Limit Theorem

If $X_1, \dots, X_n$ are independent and identically distributed random variables with mean μ and variance σ^2, then:

$$\sum X_i \overset{\cdot}{\sim} N\left(n\mu, n\sigma^2\right) \quad \Rightarrow \quad \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \overset{\cdot}{\sim} N(0,1) \quad \text{as } n \to \infty$$

$$\overline{X} \overset{\cdot}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \overset{\cdot}{\sim} N(0,1) \qquad \text{as } n \to \infty$$

Normal approximations

$$\left. \begin{array}{ll} Bin(n,p) \overset{\cdot}{\sim} N(np, npq) & np > 5, nq > 5 \\[2mm] Poi(\lambda) \overset{\cdot}{\sim} N(\lambda, \lambda) & \lambda \text{ large} \end{array} \right\} \text{with continuity correction}$$

$$Gamma(\alpha, \lambda) \overset{\cdot}{\sim} N\left(\frac{\alpha}{\lambda}, \frac{\alpha}{\lambda^2}\right) \qquad\qquad \alpha \text{ large}$$

$$\chi_k^2 \overset{\cdot}{\sim} N(k, 2k) \qquad\qquad\qquad\qquad k \text{ large}$$

The practice questions start on the next page so that you can keep
all the chapter summaries together for revision purposes.

.

Chapter 5 Practice Questions

5.1   The number of claims arising in a month under a home insurance policy follows a Poisson distribution with mean 0.075. Calculate the approximate probability that at least 50 claims in total arise in a month under a group of 500 independent such policies.

5.2   If X has a gamma distribution with parameters $\alpha = 10$ and $\lambda = 0.2$, calculate the probability that X exceeds 80

Exam style

        (a)   using a normal distribution

        (b)   using a chi-squared distribution.

        Explain which of these answers is more accurate.   [5]

5.3   When using the continuity correction with a random variable X that can take any integer value, write down expressions that are equivalent to the following:

        (i) $X < 7$

        (ii) $X = 0$

        (iii) $X \geq -2$

        (iv) $5 < X \leq 10$

        (v) $3 \leq X < 8$

        (vi) $4 \leq 10X < 48$.

5.4   The probability of any given policy in a portfolio of term assurance policies lapsing before it expires is considered to be 0.15. For a group of 100 such policies, calculate the approximate probability that more than 20 will lapse before they expire.

5.5   A company issues questionnaires to clients to obtain feedback on the clarity of their brochure. It is thought that 5% of clients do not find the brochure helpful.

Exam style

        Calculate the approximate probability that in a sample of 1,000 responses, the number, *N*, of clients who do not find the brochure helpful satisfies $40 < N < 70$.   [5]

5.6   In a certain large population 45% of people have blood group A. A random sample of 300 individuals is chosen from this population.

Exam style

        Calculate an approximate value for the probability that more than 115 of the sample have blood group A.   [3]

5.7   Consider a random sample of size 16 taken from a normal distribution with mean $\mu = 25$ and variance $\sigma^2 = 4$. Let the sample mean be denoted $\bar{X}$.

Exam style

        State the distribution of $\bar{X}$ and hence calculate the probability that $\bar{X}$ assumes a value greater than 26.   [3]

5.8   Suppose that the sums assured under policies of a certain type are modelled by a distribution
       with mean £8,000 and standard deviation £3,000. Consider a group of 100 independent policies
       of this type.

       Calculate the approximate probability that the total sum assured under this group of policies
       exceeds £845,000.   [3]

5.9   A computer routine selects one of the integers 1, 2, 3, 4, 5 at random and replicates the process a
       total of 100 times. Let S denote the sum of the 100 numbers selected.

       Calculate the approximate probability that S assumes a value between 280 and 320 inclusive. [5]

5.10   The random variable Y has a gamma distribution with parameters α (>1) and λ.

       (i)   (a)   Show that the mode of Y is given by:

$$\frac{\alpha - 1}{\lambda}$$

                (b)   By considering the relative locations of the mean and mode using sketches of the
                         gamma distribution, state how you would expect the distribution to behave in the

                         limit as $\alpha \to \infty$, but where λ is varied so that the mean $\dfrac{\alpha}{\lambda}$ has a constant

                         value μ.

       (ii)   Given that $\alpha = 50$ and $\lambda = 0.2$, calculate the value of $P(Y > 350)$ using:

                (a)   the chi-squared distribution

                (b)   the Central Limit Theorem.

       (iii)   Explain the reason for the difference between the answers obtained in part (ii).

5.11   (i) $X_1, \ldots, X_n$ are independent and identically distributed $Gamma(\alpha, \lambda)$ random variables.
                Show, using moment generating functions, that $\bar{X}$ has a $Gamma(n\alpha, n\lambda)$ distribution.

       (ii)   If the random variable T, representing the total lifetime of an individual light bulb, has an
                $Exp(\lambda)$ distribution, where $1/\lambda = 2,000$ hours, calculate the probability that the average
                lifetime of 10 bulbs will exceed 4,000 hours.

Chapter 5 Solutions

5.1   The number of claims arising from an individual policy in a month has a *Poi*(0.075) distribution. Hence, the number of claims arising in a month from 500 independent such policies has a *Poi*(37.5) distribution. This is approximated by $N(37.5, 37.5)$.

$$P(X \geq 50) \text{ becomes } P(X > 49.5) \qquad \text{(continuity correction)}$$

$$\approx P\left(Z > \frac{49.5 - 37.5}{\sqrt{37.5}} \right)$$

$$= P(Z > 1.960)$$

$$= 1 - \Phi(1.960)$$

$$= 0.025$$

5.2   (a)   If X is gamma with $\alpha = 10$ and $\lambda = 0.2$, then $E(X) = \dfrac{10}{0.2} = 50$, and $\text{var}(X) = \dfrac{10}{0.2^2} = 250$.

So we will use a $N(50, 250)$ distribution.   [1]

So:

$$P(X > 80) \approx P\left[N(50, 250) > 80 \right] = P\left[Z > \frac{80 - 50}{\sqrt{250}} \right] = 1 - Z\left[N(0,1) \leq 1.89737 \right]$$

Interpolating in the normal distribution tables, we find that:

$$P(X > 80) = 1 - 0.97111 = 0.02889$$ [1]

ie about 2.9%.

        (b)   We now use the result that if X is $Gamma(\alpha, \lambda)$, then $2\lambda X$ has a $\chi^2_{2\alpha}$ distribution. So if X is $Gamma(10, 0.2)$, then $0.4X$ is χ^2_{20}, and the required probability is:

$$P(X > 80) = P(0.4X > 32) = P\left[\chi^2_{20} > 32 \right]$$ [1]

From page 166 of the *Tables*, we see that the probability that χ^2_{20} is less than 32 is 0.9567. So the required probability is $1 - 0.9567 = 0.0433$, or about 4.3%.   [1]

The answer in (b) will be more accurate, since we have not used an approximation. The chi-squared result is exact.   [1]

5.3   (i) $X < 7$ becomes $X < 6.5$

       (ii) $X = 0$ becomes $-0.5 < X < 0.5$

       (iii) $X \geq -2$ becomes $X > -2.5$

       (iv) $5 < X \leq 10$ becomes $5.5 < X < 10.5$

       (v) $3 \leq X < 8$ becomes $2.5 < X < 7.5$

       (vi)   If X can take integer values then $10X$ takes values such as 10, 20, 30, … . So from the
                inequality in the question, $10X$ can actually be 10, 20, 30 or 40, which means that X can
                be 1, 2, 3 or 4. So $1 \leq X < 5$, and using a continuity correction on these values, this
                becomes $0.5 < X < 4.5$.

5.4   Let X be the number of policies lapsing before they expire. $X \sim Bin(100, 0.15)$, which is
       approximately $N(15, 12.75)$.

       Using a continuity correction:

            $P(X > 20)$ becomes $P(X > 20.5)$

$$\approx P\left(Z > \frac{20.5 - 15}{\sqrt{12.75}} \right)$$

$$= 1 - \Phi(1.54)$$

$$= 1 - 0.93822 = 0.06178$$

       So the approximate probability that more than 20 policies will lapse is 0.062.

       The exact answer is 0.0663.

5.5   We have $N \sim Bin(1000, 0.05)$. Using a normal approximation:

            $N \stackrel{.}{\sim} N(50, 47.5)$ [2]

       Using a continuity correction $P(40 < N < 70) \approx P(40.5 < N < 69.5)$. Hence: [1]

            $P(40 < N < 70) \approx P(N < 69.5) - P(N < 40.5)$

$$= P\left(Z < \frac{69.5 - 50}{\sqrt{47.5}} \right) - P\left(Z < \frac{40.5 - 50}{\sqrt{47.5}} \right)$$

$$= P(Z < 2.829) - [1 - P(Z < 1.378)]$$

$$= 0.99766 - [1 - 0.9159]$$

$$= 0.91356 \qquad\qquad [2]$$

5.6   Let X be the number of individuals with blood group A.

$$X \sim Bin(300, 0.45) \stackrel{.}{\approx} N(135, 74.25)$$ [1]

Using a continuity correction $P(X > 115)$ becomes $P(X > 115.5)$: [1]

$$P\left(Z > \frac{115.5 - 135}{\sqrt{74.25}} \right) = P(Z > -2.263) = P(Z < 2.263) = 0.988$$ [1]

5.7   If our population is normal, we do not need the central limit theorem. The distribution of $\bar{X}$ is exactly normal:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} \right)$$ [1]

Hence:

$$P(\bar{X} > 26) = P\left(Z > \frac{26 - 25}{2/\sqrt{16}} \right) = P(Z > 2)$$

$$= 1 - 0.97725 = 0.02275$$ [2]

5.8   Let X_i be the sum assured under the i th policy.

We require:

$$P\left[\sum_{i=1}^{100} X_i > 845,000 \right]$$

Now, according to the Central Limit Theorem:

$$\sum_{i=1}^{100} X_i \sim N\left(100 \times 8000, \ 100 \times 3000^2 \right) \text{ (approximately)}$$ [1]

Therefore:

$$P\left[\sum_{i=1}^{100} X_i > 845,000 \right] \approx P\left(Z > \frac{845,000 - 800,000}{30,000} \right) = P(Z > 1.5)$$

$$= 1 - 0.93319 = 0.06681$$ [2]

5.9   We have the sum of 100 discrete uniform random variables, X_i $(i = 1, 2, \cdots, 100)$. Using the formulae from page 10 of the *Tables*, with $a = 1$, $b = 5$ and $h = 1$, we get:

$$E(X_i) = \frac{1 + 5}{2} = 3$$

$$var(X_i) = \frac{1}{12}(5 - 1)(5 - 1 + 2) = 2$$ [1]

Using the Central Limit Theorem:

$$S = \sum_{i=1}^{100} X_i \overset{\cdot}{\sim} N(300, 200)$$
[1]

Using a continuity correction, the probability is:

$$P(280 \le S \le 320) \approx P(279.5 < S < 320.5)$$
[1]

Standardising this:

$$P(279.5 < S < 320.5) = P(S < 320.5) - P(S < 279.5)$$

$$= P\left(Z < \frac{320.5 - 200}{\sqrt{300}}\right) - P\left(Z < \frac{279.5 - 200}{\sqrt{300}}\right)$$

$$= P(Z < 1.44957) - P(Z < -1.44957)$$

$$= P(Z < 1.44957) - [1 - P(Z < 1.44957)]$$

$$= 2P(Z < 1.44957) - 1$$

$$= 2 \times 0.92641 - 1$$

$$= 0.85282$$
[2]

5.10   (i)(a)   *Mode*

The mode is the maximum of the PDF $f(y)$:

$$f(y) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} \qquad y > 0$$

Differentiating and setting the derivative equal to zero gives:

$$\frac{d}{dy} f(y) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \left[(\alpha-1) y^{\alpha-2} e^{-\lambda y} - \lambda y^{\alpha-1} e^{-\lambda y} \right]$$

$$\Rightarrow \quad y^{\alpha-2} e^{-\lambda y} [(\alpha-1) - \lambda y] = 0$$

Alternatively we could have differentiated the log of the PDF.

This gives:

$$y = 0 \qquad \text{or} \qquad y = \frac{\alpha-1}{\lambda}$$

Since $f(y) \ge 0$ and $f(0) = 0$, the first solution of zero **must** be a minimum and therefore the second solution **must** be a maximum.

Alternatively, the second solution can be shown to be a maximum by considering the second derivative:

$$\frac{d^2}{dy^2}f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-\lambda y}\left[(\alpha-1)(\alpha-2)y^{\alpha-3} - 2\lambda(\alpha-1)y^{\alpha-2} + \lambda^2 y^{\alpha-1}\right]$$

Substituting $y = \dfrac{\alpha-1}{\lambda}$ *gives:*

$$\frac{d^2}{dy^2}f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-(\alpha-1)}\frac{(\alpha-1)^\alpha}{\lambda^\alpha}\left[\frac{\lambda^3(\alpha-2)}{(\alpha-1)^2} - \frac{\lambda^3}{(\alpha-1)}\right]$$

$$= -\frac{1}{\Gamma(\alpha)}e^{-(\alpha-1)}(\alpha-1)^\alpha\frac{\lambda^3}{(\alpha-1)^2}$$
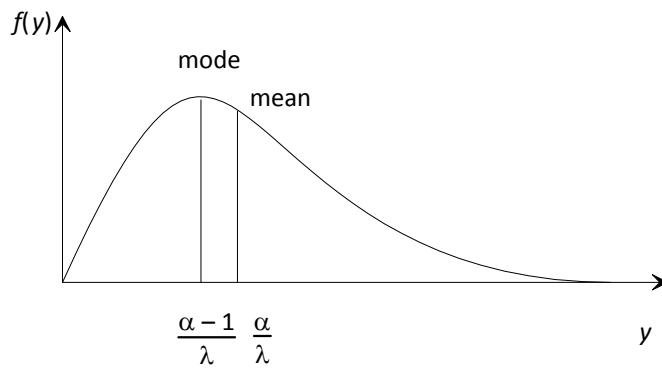
To ensure this is negative, we require $(\alpha-1)^\alpha$ *to be positive, hence we have a maximum if* $\alpha > 1$ *which was given in the question.*

(i)(b)   *Sketch locations of mode and median*

We are letting $\alpha \to \infty$, but keeping μ constant. The mean is $\dfrac{\alpha}{\lambda}$, which will remain constant.

The mode is $\dfrac{\alpha-1}{\lambda} = \dfrac{\alpha}{\lambda} - \dfrac{1}{\lambda} = \mu - \dfrac{\mu}{\alpha}$, which will be less than the mean μ, but will tend to μ as $\alpha \to \infty$.

So, for large α, the distribution will look like this:

The mean and mode are very close together.

In fact, the distribution approaches a normal distribution in the limit.

(ii)(a) *Probability using chi-squared distribution*

$Y \sim Gamma(50, 0.2)$. Using the relationship $Y \sim Gamma(\alpha, \lambda) \implies 2\lambda Y \sim \chi^2_{2\alpha}$:

$$P(Y > 350) = P(2\lambda Y > 2\lambda \times 350)$$

$$= P(0.4Y > 140)$$

$$= P(\chi^2_{100} > 140)$$

Using the χ^2 probabilities on page 169 of the *Tables* gives a value of approximately 0.5%.

(ii)(b) *Probability using normal approximation*

The mean and variance of the gamma distribution are:

$$E(Y) = \frac{\alpha}{\lambda} = \frac{50}{0.2} = 250 \qquad var(Y) = \frac{\alpha}{\lambda^2} = \frac{50}{0.2^2} = 1,250$$

Because the gamma distribution can be represented as the sum of a number of exponential distributions, the CLT tells us that, for large α, the gamma distribution can be approximated by a normal distribution.

Using the normal approximation to the gamma gives:

$$Y \sim Gamma(50, 0.2) \stackrel{.}{\sim} N(250, 1250)$$

Hence:

$$P(Y > 350) = P\left(Z > \frac{350 - 250}{\sqrt{1,250}} \right)$$

$$= P(Z > 2.828)$$

$$= 1 - 0.99766 = 0.234\%$$

(iii) *Explain the differences*

The gamma is only symmetrical when $\alpha \to \infty$. For smaller values it is still positively skewed. As a consequence, the tail will be thicker than a symmetrical distribution and the corresponding tail probabilities will be higher.

5.11 (i) ***Show mean has a gamma distribution***

We have:

$$M_{\overline{X}}(t) = E\left(e^{t\overline{X}}\right) = E\left(e^{\frac{t}{n}(X_1 + \cdots + X_n)}\right) = E\left(e^{\frac{t}{n}X_1} \ldots e^{\frac{t}{n}X_n}\right)$$

$$= M_{X_1}\left(\frac{t}{n}\right) \ldots M_{X_n}\left(\frac{t}{n}\right) \qquad\qquad \text{by independence}$$

$$= \left[M_X\left(\frac{t}{n}\right)\right]^n \qquad\qquad \text{as } X_i\text{'s identical}$$

$$= \left(1 - \frac{t}{n\lambda}\right)^{-n\alpha}$$

This is the MGF of a *Gamma*$(n\alpha, n\lambda)$ distribution. Hence, by the uniqueness property of MGFs, $\overline{X}$ has a *Gamma*$(n\alpha, n\lambda)$ distribution.

(ii) ***Probability that average lifetime of 10 bulbs exceeds 4,000 hours***

The individual lifetimes T have an *Exp*(λ) distribution, which is the same as the *Gamma*$(1, \lambda)$ distribution. So, using the result from part (i) we have:

$$\overline{T} \sim Gamma(10 \times 1\,,\, 10 \times \tfrac{1}{2,000}) \equiv Gamma(10, 0.005)$$

Using the result from page 12 of the *Tables*, the probability that the average lifetime $\overline{T}$ will exceed 4,000 hours is:

$$P(\overline{T} > 4,000) = P(\chi^2_{20} > 2 \times 0.005 \times 4,000) = P(\chi^2_{20} > 40)$$

From page 166 of the *Tables*, this is 0.005. So the probability that the average lifetime will exceed 4,000 hours is 0.5%.

6

Sampling and statistical Inference

Syllabus objectives

2.2   Random sampling and sampling distributions

   2.2.1   Explain what is meant by a sample, a population and statistical inference.

   2.2.2   Define a random sample from a distribution of a random variable.

   2.2.3   Explain what is meant by a statistic and its sampling distribution.

   2.2.4   Determine the mean and variance of a sample mean and the mean of a sample variance in terms of the population mean and variance and the sample size.

   2.2.5   State and use the basic sampling distributions for the sample mean and the sample variance for random samples from a normal distribution.

   2.2.6   State and use the distribution of the t-statistic for random samples from a normal distribution.

   2.2.7   State and use the F distribution for the ratio of two sample variances from independent samples taken from normal distributions.

0   Introduction

When a sample is taken from a population the sample information can be used to infer certain things about the population. For example, to estimate a population quantity or test the validity of a statement made about the population.

A population quantity could be its mean or variance, for example. So we might be testing the mean from a normal distribution, say.

In this chapter we will be looking at taking a sample from a distribution and calculating its mean and variance. If we were to keep taking samples from the same distribution and calculating the mean and variance for each of the samples, we would find that these values also form probability distributions.

The distributions of the sample mean and sample variance are called sampling distributions and will be used extensively in Chapters 8 and 9 to construct confidence intervals and carry out hypothesis tests.

Part of this work will explain mathematically why the sample variance is usually defined to be

$$S^2 = \frac{1}{n-1}\left[\sum X^2 - n\overline{X}^2\right] \text{ rather than } S^2 = \frac{1}{n}\left[\sum X^2 - n\overline{X}^2\right].$$

We will also make use of the Central Limit Theorem from Chapter 5 to obtain the asymptotic distribution of the sample mean.

Finally, this chapter will look at the *t*-distribution and the *F*-distribution in greater detail. You will require a copy of the *Formulae and Tables for the Actuarial Examinations* to be able to progress through this chapter.

1   Basic definitions

The statistical method for testing assertions such as 'smoking reduces life expectancy' involves selecting a sample of individuals from the population and, on the basis of the attributes of the sample, making statistical inferences about the corresponding attributes of the parent population. This is done by assuming that the variation in the attribute in the parent population can be modelled using a statistical distribution. The inference can then be carried out on the basis of the properties of this distribution.

Theoretically this (technique) deals with samples from infinite populations. Actuaries are concerned with sampling from populations of policyholders, policies, claims, buildings, employees, *etc*. Such populations may be looked upon as conceptually infinite but even without doing so, they will be very large populations of many thousands and so the methods for infinite populations will be more than adequate.

1.1   Random samples

A set of items selected from a parent population is a random sample if:

*   the probability that any item in the population is included in the sample is proportional to its frequency in the parent population and

*   the inclusion/exclusion of any item in the sample operates independently of the inclusion/exclusion of any other item.

A random sample is made up of (iid) random variables and so they are denoted by capital *X*'s. We will use the shorthand notation $\underline{X}$ to denote a random sample, that is, $\underline{X} = \left(X_1, X_2, ..., X_n \right)$. An observed sample will be denoted by $\underline{x} = \left(x_1, x_2, ..., x_n \right)$. The population distribution will be specified by a density (or probability function) denoted by $f\left(x; \theta \right)$, where θ denotes the parameter(s) of the distribution.

Due to the Central Limit Theorem, inference concerning a population mean can be considered without specifying the form of the population, provided the sample size is large enough.

Question

Identify the population, the sample and the statistical inference in each of the following examples:

(i)   You are studying 10 cities to establish whether air pollution levels are acceptable in UK cities.

(ii)   You are analysing the burglary claims for last January to get a feel for what the total range of claims might be for the whole year.

Solution

(i)   *Air pollution*

The population consists of all cities in the UK.

The sample consists of the 10 cities selected for study (and the measurements of the pollution levels for these).

The statistical inference required here is to establish whether there are unacceptable pollution levels in UK cities in general.

This is an example of a statistical test.

(ii)   *Burglary claims*

The population consists of all possible claims that could arise during the year.

The sample consists of the amounts paid for each of the January claims.

The statistical inference required here is to find an approximate range for the total claim amount for the year.

This is an example of a confidence interval.

1.2   Definition of a statistic

A statistic is a function of $\underline{X}$ only and does not involve any unknown parameters. Thus

$$\bar{X} = \frac{\sum X_i}{n} \text{ and } S^2 = \frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2 \text{ are statistics whereas } \frac{1}{n}\sum\left(X_i - \mu\right)^2 \text{ is not, unless}$$

of course μ is known.

Note here the difference between μ, which is the population mean (*ie* the mean for all possible observations, which is usually unknown) and $\bar{X}$, which is the sample mean (*ie* the mean of the sample values which we can calculate for any given sample).

We might also be interested in statistics such as $\max X_i$, the highest value in the sample.

A statistic can be generally denoted by $g(\underline{X})$. Since a statistic is a function of random variables, it will be a random variable itself and will have a distribution, its sampling distribution.

2 Moments of the sample mean and variance

In the following section we will look at the statistical properties of the sample mean and sample variance, which are the most important sample statistics.

2.1 The sample mean

Suppose X_i has mean μ and variance σ^2. Recall that the sample mean is $\bar{X} = \dfrac{\sum X_i}{n}$.

Consider first $\sum X_i$:

$$E\left[\sum X_i\right] = \sum E\left[X_i\right] = \sum \mu = n\mu \quad \text{since they are identically distributed}$$

$$\text{var}\left[\sum X_i\right] = \sum \text{var}\left[X_i\right] \quad \text{since they are independent}$$

$$= n\sigma^2 \quad \text{since they are identically distributed}$$

We are using the results from Chapter 3 that $E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_1]$, and if $X_1, \ldots, X_n$ are independent $\text{var}[X_1 + \cdots + X_n] = \text{var}[X_1] + \cdots + \text{var}[X_n]$.

As $\bar{X} = \dfrac{1}{n}\sum X_i$, we can now write down that $E\left[\bar{X}\right] = \mu$ and $\text{var}\left[\bar{X}\right] = \dfrac{1}{n^2}.n\sigma^2 = \dfrac{\sigma^2}{n}$.

Note: $sd\left[\bar{X}\right] = \dfrac{\sigma}{\sqrt{n}}$ is called the standard error of the sample mean.

So we have established that the sample mean $\bar{X}$ has an expected value of μ (*ie* the same as the population mean) and a variance of σ^2/n (*ie* the population variance divided by the sample size). These are very important results and will be used extensively in Chapters 8 and 9.

A consequence of the result for the variance of $\bar{X}$ is that as the sample gets bigger the variance gets smaller. This should be intuitive since a bigger sample produces more accurate results.

2.2 The sample variance

Recall that the sample variance is $S^2 = \dfrac{1}{n-1}\sum (X_i - \bar{X})^2$.

Considering only the mean of S^2, it can be proved that $E\left[S^2\right] = \sigma^2$ as follows:

$$S^2 = \dfrac{1}{n-1}\left[\sum X_i^2 - n\bar{X}^2\right]$$

Taking expectations and noting that for any random variable _Y_, $E[Y^2] = \text{var}[Y] + (E[Y])^2$
(obtained by rearranging $\text{var}(Y) = E(Y^2) - E^2(Y)$) **leads to:**

$$E[S^2] = \frac{1}{n-1}\left(\sum E[X_i^2] - nE[\bar{X}^2]\right)$$

$$= \frac{1}{n-1}\left\{\sum(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\}$$

$$= \frac{1}{n-1}\left\{n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2\right\}$$

$$= \frac{1}{n-1}\left\{(n-1)\sigma^2\right\} = \sigma^2$$

as required.

To work out $E[\bar{X}^2]$, we've used the general formula just mentioned, which tells us that $E[\bar{X}^2] = \text{var}(\bar{X}) + E^2[\bar{X}]$ and then we've used the results we just derived for the sample mean.

So the $n - 1$ denominator is used to make the mean of S^2 equal to the true value of σ^2. This is the motivation behind the definition of the sample variance. Later in Chapter 7, we will discover that this result means that the sample variance is an *unbiased estimator* of the population variance.

There is no general formula for $\text{var}[S^2]$. This depends on the specific distribution of the population. The only one that you will be required to know for Subject CS1 is for a normal population. This is covered in Section 3.2.

Question

The total number of new motor insurance claims reported to a particular branch of an insurance company on successive days during a randomly selected month can be considered to come from a Poisson distribution with $\lambda = 5$. Calculate the mean and variance of a sample mean based on 30 days' figures.

Solution

The Poisson distribution in the question has mean and variance of 5.

If the sample size is 30 then $E[\overline{X}] = 5$ and $\text{var}[\overline{X}] = \dfrac{5}{30} = 0.167$.

We can apply the same theory to situations involving a continuous distribution.

Question

Calculate the mean and variance of the sample mean for samples of size 110 from a parent population which is Pareto with parameters $\alpha = 5$ and $\lambda = 3{,}000$.

Solution

The Pareto distribution has a mean of $\dfrac{\lambda}{\alpha - 1}$, and variance of $\dfrac{\alpha\lambda^2}{(\alpha - 1)^2(\alpha - 2)}$, so the distribution in the question has $\mu = 750$ and $\sigma^2 = 937{,}500$.

Thus $E[\overline{X}] = 750$ and $\text{var}[\overline{X}] = \dfrac{937{,}500}{110} = 8{,}522.7$.

3   Sampling distributions for the normal

3.1   The sample mean

The Central Limit Theorem provides a large-sample approximate sampling distribution for $\bar{X}$ without the need for any distributional assumptions about the population. So for large n:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \stackrel{.}{\sim} N(0, 1) \text{ or } \bar{X} \stackrel{.}{\sim} N\left(\mu, \sigma^2 / n\right)$$

This result is often called the z result.

It transpires that the above result gives the <u>exact</u> sampling distribution of $\bar{X}$ for random samples from a normal population.
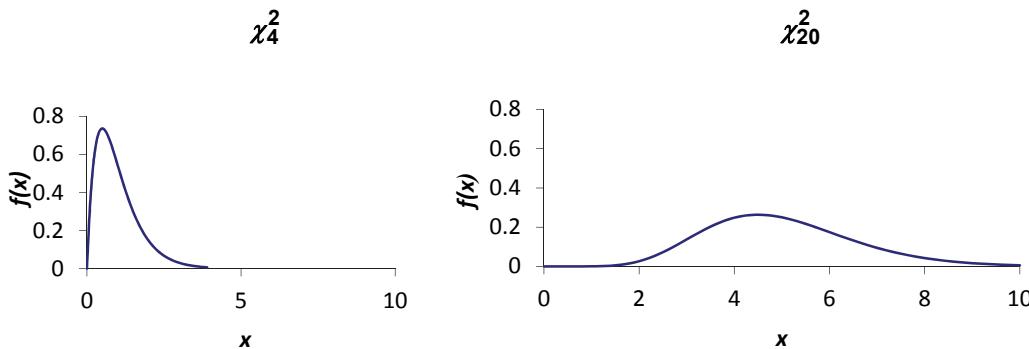
3.2   The sample variance

The sampling distribution of S^2 when sampling from a normal population, with mean μ and variance σ^2, is:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

This is a more advanced result, the proof of which lies outside the Subject CS1 syllabus.

Whereas the distribution of $\bar{X}$ is normal and hence symmetrical, the distribution of S^2 is positively skewed especially so for small n but becoming symmetrical for large n.

χ^2_4 χ^2_{20}

Using the χ^2 result to investigate the first and second order moments of S^2, when sampling from a normal population, and the fact that the mean and variance of χ^2_k are k and $2k$, respectively:

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n - 1 \Rightarrow E[S^2] = \frac{\sigma^2}{n-1} \cdot (n-1) = \sigma^2$$

This is just the result in Section 2.2, in the context of a normal distribution.

We also have:

$$\text{var}\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1) \Rightarrow \text{var}[S^2] = \frac{\sigma^4}{(n-1)^2}.2(n-1) = \frac{2\sigma^4}{n-1}.$$

You need to be able to derive these important results.

For both $\bar{X}$ and S^2 the variances decrease and tend to zero as the sample size n increases. Added to the facts that $E[\bar{X}] = \mu$ and $E[S^2] = \sigma^2$, these imply that $\bar{X}$ gets closer to μ and S^2 gets closer to σ^2 as the sample size increases. These are desirable properties of estimators of μ and σ^2.

Question

Calculate the probability that, for a random sample of 5 values taken from a $N(100, 25^2)$ population

(i) $\bar{X}$ will be between 80 and 120

(ii) S will exceed 41.7.

Solution

(i)   Since $\bar{X} \sim N(100, 25^2 \div 5) = N(100, 125)$:

$$P(80 < \bar{X} < 120) = P\left(\frac{80-100}{\sqrt{125}} < Z < \frac{120-100}{\sqrt{125}}\right)$$

$$= P(-1.789 < Z < 1.789)$$

$$= \Phi(1.789) - \Phi(-1.789)$$

$$= 0.96319 - (1 - 0.96319) = 0.926$$

(ii)   Since $\frac{4S^2}{\sigma^2} \sim \chi_4^2$, we have:

$$P(S > 41.7) = P\left(\frac{4S^2}{\sigma^2} > \frac{4 \times 41.7^2}{25^2}\right) = P(\chi_4^2 > 11.13) = 1 - P(\chi_4^2 < 11.13)$$

Interpolating on page 165 of the *Tables* gives:

$$P(S > 41.7) \simeq 0.0253$$

3.3   Independence of the sample mean and variance

The other important feature when sampling from normal populations is the independence of $\bar{X}$ and S^2. A full proof of this is not trivial but it is a result that is easily appreciated as follows.

Suppose that a sample from some normal distribution has been simulated. The value of $\bar{x}$ does not give any information about the value of s^2.

Remember that changing the mean of a normal distribution shifts the graph to the left or right. Changing the variance squashes the graph up or stretches it out.

However, if the sample is from some exponential distribution, the value of $\bar{x}$ does give information about the value of s^2, as μ and σ^2 are related.

For the exponential distribution these are directly linked since $\mu = \dfrac{1}{\lambda}$ and $\sigma^2 = \dfrac{1}{\lambda^2}$.

Other cases such as Poisson, binomial and gamma can be considered in a similar way, but only the normal has the independence property.

Question

Calculate the probability that, for the sample in the previous question, (i) and (ii) will both occur.

Solution

Since $\bar{X}$ and S^2 are independent, we can factorise the probability:

$$P(80 < \bar{X} < 120 \ \cap \ S > 41.7) = P(80 < \bar{X} < 120) \times P(S > 41.7)$$

Referring back to the previous question, we have already found the probabilities. So:

$$P(80 < \bar{X} < 120 \ \cap \ S > 41.7) = 0.926 \times 0.0253 = 0.023$$

4   The *t* result

The sampling distribution for $\bar{X}$, that is, $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ or $\bar{X} \sim N(\mu, \sigma^2/n)$, will be used in

subsequent units for inference concerning μ when the population variance σ^2 is known. However this is rare in practice, and another result is needed for the realistic situation when σ^2 is unknown. This is the *t* result or the *t* sampling distribution.

The *t* result is similar to the *z* result but with σ replaced by S and $N(0,1)$ replaced by t_{n-1} .

Thus $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

It is not a sampling distribution for $\bar{X}$ alone as it involves a combination of $\bar{X}$ and S .

The t_k variable is defined by:

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi_k^2/k}} \text{ where the } N(0,1) \text{ and } \chi_k^2 \text{ random variables are independent}$$

Then the *t* result above follows from the sampling distributions of the last section, that is, $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is the $N(0,1)$ and $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$ is the χ_k^2 , together with their independence, to obtain $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ when sampling from a normal population.
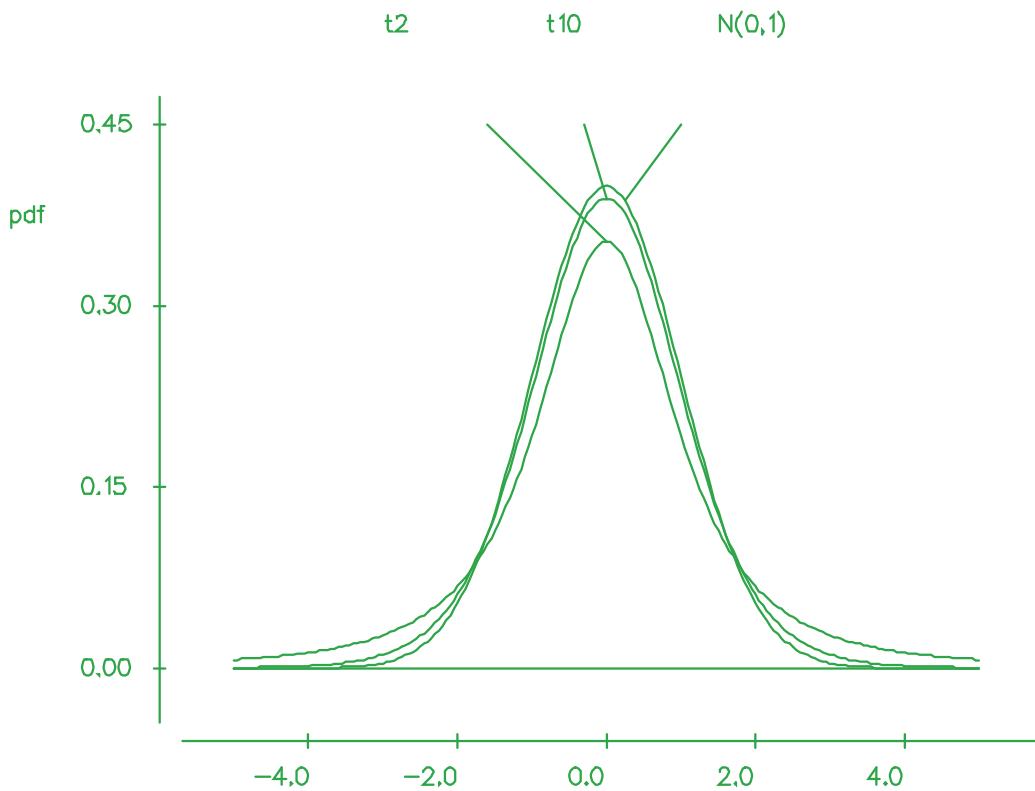
The *t* distribution is symmetrical about zero and its critical points are tabulated.

The tables for the *t* distribution can be found on page 163 of the *Tables*. It has one parameter, which, like the χ^2 distribution, is called the 'number of degrees of freedom'.

When you are using the *t* distribution, you can work out the number of degrees of freedom by remembering that it is the same as the number you divided by when estimating the variance.

So what does the PDF of the *t*-distribution look like?

It looks similar to the standard normal (*ie* symmetrical) **especially for large values of degrees of freedom. The following picture shows a** t_2 **density, a** t_{10} **density and a** $N(0,1)$ **density for comparison.**

In fact, as $k \to \infty$, $t_k \to N(0,1)$.

The t_1 **distribution is also called the Cauchy distribution and is peculiar in that none of its moments exist, not even its mean. However since samples of size 2 are unrealistic, it should not arise as a sampling distribution.**

For $k > 2$, **the** t_k **distribution has mean 0 and variance** $k/(k-2)$.

Question

State the distribution of $\dfrac{\bar{X} - 100}{S/\sqrt{5}}$ for a random sample of 5 values taken from a $N(100, \sigma^2)$

population. Calculate the probability that this quantity will exceed 1.533.

Solution

From previous results $\dfrac{\bar{X} - 100}{S/\sqrt{5}} \sim t_4$.

From the *Tables*, we see that the probability that this quantity will exceed 1.533 is 10%.

We now consider the situation involving two samples from different normal populations.

Question

Independent random samples of size n_1 and n_2 are taken from the normal populations
$N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively.

(i)   Write down the sampling distributions of $\bar{X}_1$ and $\bar{X}_2$ and hence determine the sampling
distribution of $\bar{X}_1 - \bar{X}_2$, the difference between the sample means.

(ii)   Now assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

  (a)   Express the sampling distribution of $\bar{X}_1 - \bar{X}_2$ in standard normal form.

  (b)   State the sampling distribution of $\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}$.

  (c)   Using the $N(0,1)$ distribution from (a) and the χ^2 distribution from (b), apply the
definition of the *t* distribution to find the sampling distribution of $\bar{X}_1 - \bar{X}_2$ when
σ^2 is unknown.

Solution

(i) $\bar{X}_1$ is $N(\mu_1, \sigma_1^2/n_1)$ and $\bar{X}_2$ is $N(\mu_2, \sigma_2^2/n_2)$.

$\bar{X}_1 - \bar{X}_2$ is the difference between two independent normal variables and so is itself

normal, with mean $\mu_1 - \mu_2$ and variance $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$.

(ii)(a)   The variance of $\bar{X}_1 - \bar{X}_2$ is now $\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$ and so standardising gives:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1)$$

(ii)(b)   As $\dfrac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi_{n_1-1}^2$ and $\dfrac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi_{n_2-1}^2$ are independent, (because the samples are

independent), their sum is also χ^2, with $n_1 + n_2 - 2$ degrees of freedom. This is using the

additive property of independent χ^2 distributions (ie $\chi_m^2 + \chi_n^2 \sim \chi_{m+n}^2$), which we

proved, in Chapter 3, Section 4.2.

(ii)(c)   Using the definition of the t distribution:

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi_k^2 / k}}$$

The distribution in part (ii)(a) was $N(0,1)$, and the distribution in part (ii)(b) was $\chi_{n_1+n_2-2}^2$.

So:

$$\frac{\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}}{\sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2(n_1 + n_2 - 2)}}} \sim t_{n_1+n_2-2}$$

The σ^2's cancel to give:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

We will see that $\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}$, which appears in the denominator, is the 'pooled' variance of the two samples. It is just a weighted average of the individual sample variances, using the degrees of freedom as the weightings.

5   The *F* result for variance ratios

The *F* distribution is defined by $F = \dfrac{U / v_1}{V / v_2}$, where *U* and *V* are independent χ^2 random variables with v_1 and v_2 degrees of freedom respectively. Thus if independent random samples of size n_1 and n_2 respectively are taken from normal populations with variances σ_1^2 and σ_2^2, then $\dfrac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}$.

The *F* distribution gives us the distribution of the variance ratio for two normal populations. v_1 and v_2 can be referred to as the number of degrees of freedom in the numerator and denominator respectively.

It should be noted that it is arbitrary which one is the numerator and which is the denominator and so $\dfrac{S_2^2 / \sigma_2^2}{S_1^2 / \sigma_1^2} \sim F_{n_2-1, n_1-1}$.

Since it is arbitrary which value is the numerator and which is the denominator, and since only the upper critical points are tabulated, it is usually easier to put the larger value of the sample variance into the numerator and the smaller sample variance into the denominator.

Alternatively, $F \sim F_{n_1-1, n_2-1} \Leftrightarrow \dfrac{1}{F} \sim F_{n_2-1, n_1-1}$.

This reciprocal form is needed when using tables of critical points, as only upper tail points are tabulated. See 'Formulae and Tables'.

This is an important result and will be used in Chapter 8 in the work on confidence intervals and Chapter 9 in the work on hypothesis tests.

The percentage points for the *F* distribution can be found on pages 170-174 of the *Tables.*

Question

Determine:

(i) $P(F_{9,10} > 3.779)$

(ii) $P(F_{12,14} < 3.8)$

(iii) $P(F_{11,8} < 0.3392)$

(iv)   the value of p such that $P(F_{14,6} < p) = 0.01$.

Solution

By referring to the *Tables* on pages 170 to 174:

(i)   3.779 is greater than 1, so we simply use the upper critical values given:

$$P(F_{9,10} > 3.779) = 0.025$$

since 3.779 is the 2½% point of the $F_{9,10}$ distribution (page 173).

(ii)   Since 3.8 is greater than 1, it is again an upper value and so we use the *Tables* directly. We simply turn the probability around:

$$P(F_{12,14} < 3.8) = 1 - P(F_{12,14} > 3.8) = 1 - 0.01 = 0.99$$

(iii)   Since this is a lower critical point we need to use the $\dfrac{1}{F_{m,n}}$ result:

$$P(F_{11,8} < 0.3392) = P\left(\frac{1}{F_{11,8}} > \frac{1}{0.3392}\right)$$

$$= P\left(F_{8,11} > \frac{1}{0.3392}\right) = P(F_{8,11} > 2.948) = 0.05$$

(iv)   Since only 1% of the distribution is below p, this implies that it must be a lower critical point and so we use the $\dfrac{1}{F_{m,n}}$ result again:

$$P(F_{14,6} < p) = P\left(F_{6,14} > \frac{1}{p}\right) = 0.01 \Rightarrow \frac{1}{p} = 4.456 \Rightarrow p = 0.2244$$

We now apply the *F* result to problems involving sample variances.

Question

For random samples of size 10 and 25 from two normal populations with equal variances, use the F distribution to determine the values of α and β such that $P\left(\dfrac{S_1^2}{S_2^2} > \alpha\right) = 0.05$ and $P\left(\dfrac{S_1^2}{S_2^2} < \beta\right) = 0.05$, where subscript 1 represents the sample of size 10 and subscript 2 represents the sample of size 25.

Solution

Since the population variances are equal, $\dfrac{S_1^2}{S_2^2} \sim F_{9,24}$ and $\dfrac{S_2^2}{S_1^2} \sim F_{24,9}$.

From the table of 5% points for the F distribution on page 172 of the *Tables*, we find that $P(F_{9,24} > 2.300) = 0.05$, and therefore $\alpha = 2.300$.

Now we know that $\dfrac{S_1^2}{S_2^2} < \beta$ is equivalent to $\dfrac{S_2^2}{S_1^2} > 1/\beta$ and $P(F_{24,9} > 2.900) = 0.05$, giving

$$\beta = \frac{1}{2.900} = 0.345.$$

We can use the F distribution to obtain probabilities relating to the ratio of two different sample variances.

Question

Calculate the probability that the sample variance of a sample of 10 values from a normal distribution will be more than 6 times the sample variance of a sample of 5 values from an independent normal distribution with the same variance.

Solution

If X denotes the sample with 10 values and Y denotes the sample with 5 values, we know that as these are from independent normal distributions, $\dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{9,4}$.

Since the population variances are equal, this means that $S_X^2/S_Y^2 \sim F_{9,4}$.

So $P\left(\dfrac{S_X^2}{S_Y^2} > 6\right) = P\left(F_{9,4} > 6\right)$.

From the *Tables* page 172 we see that the tabulated 5% critical value of $F_{9,4}$ is 5.999. So the required probability is just over 5%.

Chapter 6 Summary

The sample mean and sample variance are given by:

$$\bar{X} = \frac{\sum X_i}{n} \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right)$$

We can find their sampling means and variances. For any distribution:

$$E(\bar{X}) = \mu \qquad \text{var}(\bar{X}) = \frac{\sigma^2}{n} \qquad E(S^2) = \sigma^2$$

For a normal distribution only:

$$\text{var}(S^2) = \frac{2\sigma^4}{n-1}$$

The standard deviation of the sample mean is known as the standard error of the sample mean.

To find probabilities involving $\bar{X}$ or S^2 we need their distributions. For a large sample from any distribution (and for any size of sample from a normal distribution):

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ (approximately)}$$

When sampling from a normal population, the sample mean and variance are independent.

For a random sample from a normal distribution, if σ^2 is known:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

If σ^2 is unknown:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Chapter 6 Summary (continued)

For a random sample from a normal distribution:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

If we take random samples from two independent normal distributions:

$$\frac{S_1^2 \big/ \sigma_1^2}{S_2^2 \big/ \sigma_2^2} \sim F_{n_1-1,\,n_2-1}$$

The *t* and *F* distributions are defined as:

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi^2_k/k}} \qquad\qquad F_{m,n} \equiv \frac{\chi^2_m/m}{\chi^2_n/n}$$

To determine probabilities involving the lower tail of the *F* distribution, we use the result:

$$P\big(F_{m,n} < k\big) = P\big(F_{n,m} > 1/k\big)$$

Chapter 6 Practice Questions

6.1   A random sample of n observations is taken from a normal distribution with mean μ and variance σ^2. The sample variance is an observation of a random variable S^2. Using the relationship between the gamma and chi-squared distributions given on page 12 of the *Tables*, derive expressions for $E\left(S^2\right)$ and $\text{var}\left(S^2\right)$.

6.2   (i)   Determine:

        (a) $P(F_{3,9} < 3.863)$ (b) $P(F_{10,10} < 0.269)$

        (ii)   Determine the value of p such that:

        (a) $P(F_{24,30} > p) = 0.10$ (b) $P(F_{18,9} > p) = 99\%$

6.3   A random sample of 10 observations is drawn from a normal distribution with mean μ and

Exam style   standard deviation 15. Independently, a random sample of 25 observations is drawn from a normal distribution with mean μ and standard deviation 12. Let $\bar{X}$ and $\bar{Y}$ denote the respective sample means.

        Evaluate $P(\bar{X} - \bar{Y} > 3)$.   [3]

6.4   Calculate:

        (a) $P(F_{6,8} > 6.371)$

        (b) $P(F_{7,12} > 0.3748)$.

6.5   (i)   (a)   State the definition of a t_k distribution.

                (b)   Hence, using $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$ and $(n-1)S^2\big/\sigma^2 \sim \chi_{n-1}^2$, show that:

                        $$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

        (ii)   (a)   State the definition of an $F_{m,n}$ distribution.

                (b)   Hence, using $(n-1)S^2\big/\sigma^2 \sim \chi_{n-1}^2$, show that for suitably defined samples:

                        $$\frac{S_1^2\big/\sigma_1^2}{S_2^2\big/\sigma_2^2} \sim F_{m-1,n-1}$$

6.6   Evaluate c such that:

        (a) $P(F_{2,15} < c) = 97.5\%$

        (b) $P(F_{8,5} < c) = 5\%$.

6.7   Show that:

$$P(F_{m,n} > a) = b \quad \Leftrightarrow \quad P\left(F_{n,m} < \frac{1}{a}\right) = b$$

6.8   A random sample $x_1, \ldots, x_{10}$ is drawn from a $N(5,4)$ distribution. Evaluate:

        (i) $P\left[\sum X > 60\right]$

        (ii) $P\left[\sum (X - \bar{X})^2 > 34\right]$

        (iii) $P\left[\bar{X} > 4 \text{ and } \sqrt{\frac{1}{9}\sum (X - \bar{X})^2} < 2.6\right]$.

6.9   Let $(X_1, X_2, \ldots, X_9)$ be a random sample from a $N(0, \sigma^2)$ distribution. Let $\bar{X}$ and S^2 denote the
Exam style   sample mean and variance respectively.

        Find the approximate value of $P(\bar{X} > S)$ by referring to an appropriate statistical table.   [3]

6.10   House prices in region X are normally distributed with a mean of £100,000 and a standard
Exam style   deviation of £10,000. House prices in region Y are normally distributed with a mean of £90,000
        and a standard deviation of £5,000. A sample of 10 houses is taken from region X and a sample of
        5 houses from region Y. Calculate the probability that:

        (i)   the region X sample mean is greater than the region Y sample mean   [3]

        (ii)   the difference between the sample means is less than £5,000   [3]

        (iii)   the region X sample variance is less than the region Y sample variance   [3]

        (iv)   the region X sample standard deviation is more than four times greater than the region Y
                   sample standard deviation.   [2]
                                                                                    [Total 11]

6.11   The time taken to process simple home insurance claims has a mean of 20 mins and a standard
        deviation of 5 mins. Stating clearly any assumptions that you make, calculate the probability that:

        (i)   the sample mean of the times to process 5 claims is less than 15 mins   [2]

        (ii)   the sample mean of the times to process 50 claims is greater than 22 mins   [2]

        (iii)   the sample variance of the time to process 5 claims is greater than 6.65 mins   [2]

        (iv)   the sample standard deviation of the time to process 30 claims is less than 7 mins   [2]

        (v)   both (i) and (iii) occur for the same sample of 5 claims.   [1]
                                                                                            [Total 9]

6.12   A statistician suggests that, since a t variable with k degrees of freedom is symmetrical with

        mean 0 and variance $\dfrac{k}{k-2}$ for $k>2$, one can approximate the distribution using the normal

        variable $N\left(0,\dfrac{k}{k-2}\right)$.

        (i)   Use this to obtain an approximation for the upper 5% percentage points for a t variable
                 with:

                 (a)   4 degrees of freedom, and

                 (b)   40 degrees of freedom.   [2]

        (ii)   Compare your answers with the exact values from tables and comment briefly on the
                 result.   [2]
                                                                                            [Total 4]

The solutions start on the next page so that you can
separate the questions and solutions.

Chapter 6 Solutions

6.1   The sampling distribution for S^2 is given by:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Now a χ^2_k is a Gamma with $\alpha = \dfrac{k}{2}$ and $\lambda = \dfrac{1}{2}$. Therefore:

$$E\left[\chi^2_k\right] = \frac{\alpha}{\lambda} = \frac{k/2}{1/2} = k \quad \text{and} \quad \text{var}\left[\chi^2_k\right] = \frac{\alpha}{\lambda^2} = \frac{k/2}{(1/2)^2} = 2k$$

$$E\left(\frac{(n-1)S^2}{\sigma^2}\right) = E\left(\chi^2_{n-1}\right) = n-1$$

$$\Rightarrow \quad \frac{(n-1)}{\sigma^2} E\left(S^2\right) = n-1$$

$$\Rightarrow \quad E\left(S^2\right) = \sigma^2$$

$$\text{var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = \text{var}\left(\chi^2_{n-1}\right) = 2(n-1)$$

$$\Rightarrow \quad \frac{(n-1)^2}{\sigma^4} \text{var}\left(S^2\right) = 2(n-1)$$

$$\Rightarrow \quad \text{var}\left(S^2\right) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}$$

6.2   (i)(a)   3.863 is greater than 1 so we simply use the upper critical values given:

$$P(F_{3,9} < 3.863) = 1 - P(F_{3,9} > 3.863) = 1 - 0.05 = 0.95$$

(i)(b)   Since this is a lower critical point we need to use the $\dfrac{1}{F_{m,n}}$ result:

$$P(F_{10,10} < 0.269) = P\left(F_{10,10} > \frac{1}{0.269}\right) = P(F_{10,10} > 3.717) = 0.025$$

(ii)(a)   Since only 10% of the distribution is above p it must be on the upper tail. So simply reading off from the 10% tables gives:

$$P(F_{24,30} > p) = 0.10 \quad \Rightarrow \quad p = 1.638$$

(ii)(b) Since 99% of the distribution is greater than p it must be on the lower tail. So we need to use the $\dfrac{1}{F_{m,n}}$ result:

$$P(F_{18,9} > p) = P\left(F_{9,18} < \frac{1}{p}\right) = 0.99 \quad \Rightarrow \quad P\left(F_{9,18} > \frac{1}{p}\right) = 0.01$$

Hence reading off the 1% tables gives $\dfrac{1}{p} = 3.597 \Rightarrow p = 0.278$.

6.3 We require $P(\bar{X} - \bar{Y} > 3)$, therefore we need the distribution of $\bar{X} - \bar{Y}$. The distributions of the sample means are:

$$\bar{X} \sim N\left(\mu, \frac{15^2}{10}\right) \quad \bar{Y} \sim N\left(\mu, \frac{12^2}{25}\right)$$

[1]

The mean of the difference is the difference of the means, and the variance of the difference is the sum of the variances:

$$\bar{X} - \bar{Y} \sim N\left(0, \frac{15^2}{10} + \frac{12^2}{25}\right) = N(0, 28.26)$$

[1]

$$P(\bar{X} - \bar{Y} > 3) = P(Z > 0.564)$$

$$= 1 - P(Z < 0.564)$$

$$= 1 - 0.71362$$

$$= 0.28638$$

[1]

6.4 (a) **Probability**

6.371 is greater than 1 so we simply use the upper critical values given:

$$P(F_{6,8} > 6.371) = 0.01$$

(b) **Probability**

Since this is a lower critical point we need to use the $\dfrac{1}{F_{m,n}}$ result:

$$P(F_{7,12} > 0.3748) = P\left(F_{12,7} < \frac{1}{0.3748}\right) = P(F_{12,7} < 2.688) = 1 - P(F_{12,7} > 2.688) = 1 - 0.1 = 0.9$$

6.5   **(i)(a)   *Definition of t distribution***

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi_k^2 / k}}, \text{where } N(0,1) \text{ and } \chi_k^2 \text{ are independent}.$$

(i)(b)   *Show result is t distribution*

Standardising, we get:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

and also we have:

$$\frac{S^2}{\sigma^2} \sim \frac{\chi_{n-1}^2}{n-1}$$

Substituting these into the definition of the t_{n-1} distribution:

$$\frac{\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}}{\sqrt{S^2 / \sigma^2}} = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$$

(ii)(a)   *Definition of F distribution*

$$F_{m,n} \equiv \frac{\chi_m^2 / m}{\chi_n^2 / n}, \text{where } \chi_m^2 \text{ and } \chi_n^2 \text{ are independent}.$$

(ii)(b)   *Show result is F distribution*

Assuming two samples of size *m* and *n*, with sample variances S_1^2 and S_2^2 from normal distributions with variances σ_1^2 and σ_2^2, respectively, then we have:

$$\frac{(m-1)S_1^2}{\sigma_1^2} \sim \chi_{m-1}^2 \quad \Rightarrow \quad \frac{S_1^2}{\sigma_1^2} \sim \frac{\chi_{m-1}^2}{m-1}$$

$$\frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi_{n-1}^2 \quad \Rightarrow \quad \frac{S_2^2}{\sigma_2^2} \sim \frac{\chi_{n-1}^2}{n-1}$$

Hence by the definition of the $F_{m-1,n-1}$ distribution:

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F_{m-1,n-1}$$

6.6   **(a)**   ***Evaluate c***

Since 97.5% of the distribution is below c it must be on the upper tail. So simply reading from the 2½% tables gives:

$$P(F_{2,15} < c) = 0.975 \quad \Rightarrow \quad P(F_{2,15} > c) = 0.025 \quad \Rightarrow \quad c = 4.765$$

(b)   ***Evaluate c***

Since only 5% of the distribution is below c it must be on the lower tail. So we need to use the $\dfrac{1}{F_{m,n}}$ result:

$$P(F_{8,5} < c) = P\left(F_{5,8} > \frac{1}{c} \right) = 0.05 \quad \Rightarrow \quad \frac{1}{c} = 3.688 \Rightarrow c = 0.2711$$

6.7   By taking reciprocals, we obtain:

$$P(F_{m,n} > a) = b \quad \Rightarrow \quad P\left(\frac{1}{F_{m,n}} < \frac{1}{a} \right) = b$$

Now from the definition of the *F* distribution:

$$F_{m,n} \equiv \frac{\chi_m^2 / m}{\chi_n^2 / n} \quad \Rightarrow \quad \frac{1}{F_{m,n}} = \frac{\chi_n^2 / n}{\chi_m^2 / m} = F_{n,m}$$

Hence:

$$P(F_{m,n} > a) = b \quad \Rightarrow \quad P\left(\frac{1}{F_{m,n}} < \frac{1}{a} \right) = b \quad \Rightarrow \quad P\left(F_{n,m} < \frac{1}{a} \right) = b$$

6.8   **(i)**   ***Probability of sum***

Using the result that $\sum X_i \sim N(n\mu, n\sigma^2) = N(50, 40)$ we obtain:

$$P(\sum X_i > 60) = P\left(Z > \frac{60 - 50}{\sqrt{40}} \right) = P(Z > 1.581) = 1 - \Phi(1.581) = 0.0569$$

(ii)   *Probability of central moment*

Since $\sum (X_i - \overline{X})^2 = (n-1)S^2$ and using $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$:

$$P\left[\sum (X_i - \overline{X})^2 > 34\right] = P[9S^2 > 34]$$

$$= P\left[\frac{9S^2}{4} > \frac{34}{4}\right]$$

$$= P\left[\chi^2_9 > 8.5\right]$$

$$= 1 - 0.5154 = 0.485$$

(iii)   *Joint probability*

Since $S = \sqrt{\dfrac{1}{9}\sum (X_i - \overline{X})^2}$ and using the fact that $\overline{X}$ and S^2 are independent when we are sampling from a normal distribution:

$$P[\overline{X} > 4 \text{ and } S < 2.6] = P[\overline{X} > 4]P[S < 2.6]$$

Now:

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n} = \frac{4}{10} = 0.4$$

So $\overline{X} \sim N(5, 0.4)$, and:

$$P[\overline{X} > 4] = P\left(Z > \frac{4-5}{\sqrt{0.4}}\right) = P(Z > -1.581) = \Phi(1.581) = 0.9431$$

Also using $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$:

$$P[S < 2.6] = P\left[\frac{9S^2}{4} < \frac{9 \times 2.6^2}{4}\right] = P[\chi^2_9 < 15.21] = 0.9145$$

Hence $P[\overline{X} > 4 \text{ and } S < 2.6] = 0.9431 \times 0.9145 = 0.862$.

6.9   Using $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$:

$$\frac{\overline{X}}{S/3} \sim t_8 \quad \Rightarrow \quad \frac{3\overline{X}}{S} \sim t_8 \tag{[1]}$$

Considering the probability in the question:

$$P\left(\overline{X} > S\right) = P\left(\frac{\overline{X}}{S} > 1\right) = P\left(\frac{3\overline{X}}{S} > 3\right) = P(t_8 > 3)$$ [1]

From the *Tables*, we can see that this probability lies between 1% and 0.5%. By interpolation we find that the probability is approximately 0.89%.   [1]

6.10   (i)   ***Probability that the mean of X is greater than the mean of Y***

We require $P(\overline{X} > \overline{Y}) = P(\overline{X} - \overline{Y} > 0)$, therefore we need the distribution of $\overline{X} - \overline{Y}$. Working in 1,000's, the distributions of the sample means are:

$$\overline{X} \sim N\left(100, 10\right) \qquad \text{and} \qquad \overline{Y} \sim N\left(90, 5\right)$$

So we obtain:

$$\overline{X} - \overline{Y} \sim N\left(100 - 90, 10 + 5\right) = N(10, 15)$$ [1]

$$P(\overline{X} - \overline{Y} > 0) = P\left(Z > \frac{0 - 10}{\sqrt{15}}\right)$$

$$= P(Z > -2.582)$$

$$= \Phi(2.582)$$

$$= 0.995$$ [2]

(ii)   ***Probability that the difference between means is less than 5,000***

Using the distribution of $\overline{X} - \overline{Y}$ from part (i):

$$P(|\overline{X} - \overline{Y}| < 5) = P(-5 < \overline{X} - \overline{Y} < 5)$$

$$= P(\overline{X} - \overline{Y} < 5) - P(\overline{X} - \overline{Y} < -5)$$ [1]

$$= P\left(Z < \frac{5 - 10}{\sqrt{15}}\right) - P\left(Z < \frac{-5 - 10}{\sqrt{15}}\right)$$

$$= P(Z < -1.291) - P(Z < -3.873)$$ [1]

$$= [1 - \Phi(1.291)] - [1 - \Phi(3.873)]$$

$$= \Phi(3.873) - \Phi(1.291)$$

$$= 0.0983$$ [1]

(iii)   *Probability that the sample variance of X is less than the sample variance of Y*

We require $P(S_X^2 < S_Y^2) = P\left(S_X^2 \big/ S_Y^2 < 1\right)$. Using the definition of the F distribution, we get:

$$\frac{S_X^2 \big/ \sigma_X^2}{S_Y^2 \big/ \sigma_Y^2} = \frac{S_X^2 \big/ S_Y^2}{\sigma_X^2 \big/ \sigma_Y^2} = \frac{S_X^2 \big/ S_Y^2}{10^2 \big/ 5^2} = \frac{S_X^2 \big/ S_Y^2}{4} \sim F_{9,4}$$

Hence:

$$P\left(S_X^2 \big/ S_Y^2 < 1\right) = P\left(\frac{S_X^2 \big/ S_Y^2}{4} < 0.25\right) = P(F_{9,4} < 0.25) \qquad [1]$$

Since this is in the lower tail we need to use the $\dfrac{1}{F_{m,n}}$ result:

$$P(F_{9,4} < 0.25) = P\left(F_{4,9} > \frac{1}{0.25}\right) = P(F_{4,9} > 4) \qquad [1]$$

This value is between 2½% and 5%, and, interpolating, we find that the probability is approximately 4.2%. $\qquad [1]$

(iv)   *Probability that the sample s.d. of X is greater than four times the sample s.d. of Y*

We require $P(S_X > 4S_Y) = P\left(S_X / S_Y > 4\right) = P\left(S_X^2 \big/ S_Y^2 > 16\right)$. Using the result from (iii) we get:

$$P\left(S_X^2 \big/ S_Y^2 > 16\right) = P\left(\frac{S_X^2 \big/ S_Y^2}{4} > 4\right) = P(F_{9,4} > 4) \qquad [1]$$

But from the *Tables*:

$$P\left(F_{9,4} > 3.936\right) = 10\%$$

So our required probability is approximately 10%. $\qquad [1]$

6.11   (i)   *Probability of sample mean ($n = 5$)*

$\bar{X} \sim N(\mu, \sigma^2 / n)$ holds exactly for samples from the normal distribution and approximately for any distribution if *n* is large. Since we only have a sample of size 5, we require that we are sampling from a normal distribution.

$$P(\bar{X} < 15) = P\left(Z < \frac{15 - 20}{\sqrt{5}}\right) \qquad \text{since } \bar{X} \sim N(20, 5)$$

$$= P(Z < -2.236)$$

$$= 1 - \Phi(2.236)$$

$$= 0.0127 \qquad [2]$$

(ii)   *Probability of sample mean* ($n = 50$)

As n is large, we require no assumptions other than it being a random sample, although the answer will be approximate if the sample is not from a normal distribution.

$$P(\overline{X} > 22) = P\left(Z > \frac{22 - 20}{\sqrt{0.5}}\right) \quad \text{since } \overline{X} \sim N(20, 0.5)$$

$$= P(Z > 2.828)$$

$$= 1 - \Phi(2.828)$$

$$= 0.00234 \hspace{4cm} [2]$$

(iii)   *Probability of sample variance*

$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ only holds for samples from a normal distribution. Therefore we require that we are sampling from a normal distribution.

$$P(S^2 > 6.65) = P\left(\frac{4S^2}{\sigma^2} > \frac{4 \times 6.65}{5^2}\right)$$

$$= P(\chi^2_4 > 1.064)$$

$$= 0.9 \hspace{2cm} \text{(from page 168 of the \textit{Tables})} \hspace{1cm} [2]$$

(iv)   *Probability of sample standard deviation*

Again we require that we are sampling from a normal distribution:

$$P(S < 7) = P\left(\frac{29S^2}{\sigma^2} < \frac{29 \times 7^2}{5^2}\right)$$

$$= P(\chi^2_{29} < 56.84) \hspace{4cm} [1]$$

Using the figures from page 169 of the *Tables*, and interpolating, we find that $P(S < 7) \approx 0.998$.

$$\hspace{13cm} [1]$$

(v)   *Probability of (i) and (iii) both occurring*

$\overline{X}$ and S^2 are independent if we are sampling from a normal distribution. So making this assumption, we get:

$$P(\overline{X} < 15 \ \cap \ S^2 > 6.65) = P(\overline{X} < 15) \times P(S^2 > 6.65)$$

$$= 0.0127 \times 0.9$$

$$= 0.0114 \hspace{4cm} [1]$$

6.12   **(i)(a)**   ***Normal approximation for*** t_4

We have:

$$t_4 \sim N(0,2) \text{ (approximately)}$$

We require the value a such that $P(t_4 > a) = 0.05$. Using our approximation, we get:

$$P\left(Z > \frac{a-0}{\sqrt{2}}\right) = 0.05 \quad \Rightarrow \quad \frac{a}{\sqrt{2}} = 1.6449 \quad \Rightarrow \quad a = 2.326 \qquad \qquad [1]$$

(i)(b)   ***Normal approximation for*** t_{40}

We have:

$$t_{40} \sim N\left(0, \frac{40}{38}\right) \text{ (approximately)}$$

We require the value b such that $P(t_{40} > b) = 0.05$. Using our approximation, we get:

$$P\left(Z > \frac{b-0}{\sqrt{40/38}}\right) = 0.05 \quad \Rightarrow \quad \frac{b}{\sqrt{40/38}} = 1.6449 \quad \Rightarrow \quad b = 1.688 \qquad \qquad [1]$$

(ii)   ***Compare approximate results with the exact values***

From the t tables we see that:

$$P(t_4 > 2.132) = 0.05 \qquad \qquad ie \ a = 2.132$$

$$P(t_{40} > 1.684) = 0.05 \qquad \qquad ie \ b = 1.684 \qquad \qquad [1]$$

We can see that the approximation of 2.326 for the upper 5% point of the t_4 distribution is poor, whereas the approximation of 1.688 for the upper 5% point of the t_{40} distribution is quite good. This suggests that the t distribution tends towards the standard normal distribution as the number of degrees of freedom increases. $\qquad \qquad [1]$

7

Point estimation

Syllabus objectives

3.1   Estimation and estimators

    3.1.1   Describe and apply the method of moments for constructing estimators of population parameters.

    3.1.2   Describe and apply the method of maximum likelihood for constructing estimators of population parameters.

    3.1.3   Define the terms: efficiency, bias, consistency and mean squared error.

    3.1.4   Define and apply the property of unbiasedness of an estimator.

    3.1.5   Define the mean square error of an estimator and use it to compare estimators.

    3.1.6   Describe and apply the asymptotic distribution of maximum likelihood estimators.

    3.1.7   Use the bootstrap method to estimate properties of an estimator.

0   Introduction

In many situations we will be interested in the value of an unknown population parameter. For example, we might be interested in the number of claims from a certain portfolio that we receive in a month. Suppose we have the following data relating to 100 one-month periods:

Claims	0	1	2	3	4	5	6
Frequency (number of months)	9	22	26	21	13	6	3

It may be that we know that the Poisson distribution is a good model for the number of claims received, but the natural question is 'what is the value of the Poisson parameter μ?'.

This chapter gives two methods that can be used to estimate the value of the unknown parameter using the information provided by a sample.

The first method is called the method of moments and involves equating the sample moments to the population moments.

The second method is called the method of maximum likelihood and uses differentiation to find the parameter value that would maximise the probability of us getting the particular sample that we observed.

These are not the only methods of obtaining estimates (for example in Subject CS2 we will meet the method of percentiles). The two methods we meet here do not always give the same value for the estimate.

However, later in this chapter we will look at how to decide whether the formulae that we obtain for the parameter estimates give 'good' estimates based upon their 'average' value and their 'spread'.

The expression 'point estimation' refers to the problem of finding a *single number* to estimate the parameter value. This contrasts with 'confidence interval estimation' (covered in the next chapter) where we wish to find a range of possible values.

This is a key topic in most statistics courses.

1   The method of moments

The basic principle is to equate population moments (*ie* the means, variances, *etc* of the theoretical model) **to corresponding sample moments** (*ie* the means, variances, *etc* of the sample data observed) **and solve for the parameter(s).**

1.1   The one-parameter case

This is the simplest case: to equate population mean, $E(X)$, **to sample mean**, $\bar{x}$, **and solve for the parameter, *ie*:**

$$E[X] = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Question

A random sample from an $Exp(\lambda)$ distribution is as follows:

   14.84, 0.19, 11.75, 1.18, 2.44, 0.53

Calculate the method of moments estimate for λ.

Solution

The population mean for an $Exp(\lambda)$ distribution from page 11 of the *Tables* is $E(X) = \frac{1}{\lambda}$.

The sample mean is $\bar{x} = \dfrac{14.84 + 0.19 + 11.75 + 1.18 + 2.44 + 0.53}{6} = 5.155$.

Equating these gives us the method of moments estimate:

$$\frac{1}{\hat{\lambda}} = 5.155 \quad \Rightarrow \quad \hat{\lambda} = 0.1940$$

Because this is an estimate of λ rather than the true value, we distinguish this by putting a 'hat' or similar over the parameter.

We can apply this method to a number of different single parameter distributions. For example, the method works well with a random sample from a Poisson distribution.

Note: For some populations the mean does not involve the parameter, such as the uniform on $(-\theta, \theta)$ or the normal $N(0, \sigma^2)$, in which case a higher-order moment must be used. However such cases are rarely of practical importance.

For example the $U(-\theta, \theta)$ distribution has $E(X) = \frac{1}{2}(-\theta + \theta) = 0$. Clearly setting this equal to the sample mean is not going to be helpful. So what we should do is to use, say, the variance, $\text{var}(X) = \frac{1}{12}[\theta - (-\theta)]^2 = \frac{1}{3}\theta^2$, as this involves the parameter. We could then equate this to the sample variance.

Question

The random sample :

> 2.6, 1.9, 3.8, –4.1, –0.2, –0.7, 1.1, 6.9

is taken from a $U(-\theta, \theta)$ distribution.

By equating the sample and population variances, find an estimate for θ.

Solution

For these sample values, $\sum x_i = 11.3$ and $\sum x_i^2 = 90.97$. So the sample variance is:

$$s^2 = \frac{1}{7}\left(90.97 - 8 \times \left(\frac{11.3}{8}\right)^2\right) = 10.7155$$

So using the formula for the population variance given above, we have:

$$\frac{1}{3}\hat{\theta}^2 = 10.7155$$

Solving this, we find that $\hat{\theta} = 5.67$.

The *estimator* is written in upper case as it is a random variable and will have a sampling distribution. The *estimate* is written in lower case as it comes from an actual sample of numerical values.

Be careful to distinguish between the words 'estimate' and 'estimator'. 'Estimate' refers to a particular numerical value that results from using the formula, *eg* $\hat{\mu} = \overline{x}$ (the lower case denotes actual sample values being used). On the other hand 'estimator' refers to the *random variable* representing any sample, *eg* $\hat{\mu} = \overline{X}$.

1.2   The two-parameter case

With two unknown parameters, we will require two equations.

This involves equating the first and second-order moments of the population and the sample, and solving the resulting pair of equations.

Moments about the origin can be used but the solution is the same (and often more easily obtained) using moments about the mean – apart from the first-order moment being the mean itself.

The first-order equation is the same as in the one-parameter case:

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The second-order equation is:

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

or equivalently:

$$E\left[(X - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

or: $$\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Note that we are not equating sample and population variances here; we are using a denominator of n on the right hand side of the final equation, whereas the sample variance uses a denominator of $n-1$.

Question

Show that these two second-order equations give the same answers for the parameter estimators.

Solution

Starting with the last equation above, our two equations are:

$$E(X) = \bar{x} \quad \text{and} \quad \text{var}(X) = \frac{1}{n}\sum (x_i - \bar{x})^2$$

Expanding the brackets in the second equation gives:

$$\text{var}(X) = \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{1}{n}\left\{\sum x_i^2 - n\bar{x}^2\right\} = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

But our first equation was $E(X) = \bar{x}$ so we have:

$$\text{var}(X) = \frac{1}{n}\sum x_i^2 - E^2(X) \qquad ie \qquad \text{var}(X) + E^2(X) = \frac{1}{n}\sum x_i^2$$

But $E(X^2) = \text{var}(X) + E^2(X)$, so we have:

$$E(X^2) = \frac{1}{n} \sum x_i^2$$

This is the other second-order equation – so they are equivalent.

We can now find method of moments estimators in the two parameter case.

Question

A random sample from a $Bin(n,p)$ distribution yields the following values:

$\qquad$ 4, 2, 7, 4, 1, 4, 5, 4

Calculate method of moments estimates of *n* and *p*.

Solution

There are two unknown parameters so we need two equations. The population mean for a

$Bin(n,p)$ distribution from page 6 of the *Tables* is $E(X) = np$. The sample mean is $\bar{x} = \frac{31}{8} = 3.875$.

Equating these gives $E[X] = \frac{1}{n} \sum_{i=1}^{n} x_i \implies \hat{n}\hat{p} = 3.875$. $\qquad\qquad$ (1)

If we use the first of the second-order equations we see that there is no formula for $E(X^2)$ on

page 6 of the *Tables*. But since $\text{var}(X) = E(X^2) - E^2(X)$ we have:

$$E(X^2) = \text{var}(X) + E^2(X) = np(1-p) + (np)^2$$

We also have $\frac{1}{n} \sum x_i^2 = \frac{143}{8} = 17.875$. Equating this to $E(X^2)$:

$$\hat{n}\hat{p}(1-\hat{p}) + (\hat{n}\hat{p})^2 = 17.875 \qquad\qquad (2)$$

Substituting equation (1) into (2) gives:

$$3.875(1-\hat{p}) + 3.875^2 = 17.875 \implies \hat{p} = 0.2621$$

Hence, $\hat{n} = 14.78$. Since *n* is the number of trials, the true value *cannot* be 14.78. Therefore it is
likely to be 14 or 15.

Alternatively, if we use the second of the second-order equations, we would obtain

$$\text{var}(X) = \hat{n}\hat{p}(1-\hat{p}) \ \text{ and } \ \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 = \frac{143}{8} - 3.875^2 = 2.859375. \ \text{ Equating these gives:}$$

$$\hat{n}\hat{p}(1-\hat{p}) = 2.859375 \qquad\qquad\qquad (3)$$

Substituting equation (1) into (3) gives:

$$3.875(1-\hat{p}) = 2.859375 \ \Rightarrow \ \hat{p} = 0.2621, \text{ a}nd \text{ hence } \hat{n} = 14.78 \text{ as before}$$

We can apply the method of moments to other distributions.

Question

A random sample of size 10 from a Type 2 negative binomial distribution with parameters k and p is as follows:

    1, 1, 0, 1, 1, 1, 3, 2, 0, 5

Calculate method of moments estimates of k and p.

Solution

There are two unknown parameters so we need two equations. The population mean for a Type 2 $NBin(k,p)$ distribution from page 9 of the *Tables* is $E(X) = \dfrac{k(1-p)}{p}$. The sample mean is

$\overline{x} = \dfrac{15}{10} = 1.5$. Equating these gives:

$$E[X] = \frac{1}{n}\sum_{i=1}^{n}x_i \ \Rightarrow \ \frac{\hat{k}(1-\hat{p})}{\hat{p}} = 1.5 \qquad\qquad\qquad (1)$$

If we use the first of the second-order equations we see that there is no formula for $E(X^2)$ on page 9 of the *Tables*. But since $\text{var}(X) = E(X^2) - E^2(X)$ we have:

$$E(X^2) = \text{var}(X) + E^2(X) = \frac{k(1-p)}{p^2} + \left(\frac{k(1-p)}{p}\right)^2$$

We also have $\dfrac{1}{n}\sum x_i^2 = \dfrac{43}{10} = 4.3$. Equating these gives:

$$\frac{\hat{k}(1-\hat{p})}{\hat{p}^2} + \left(\frac{\hat{k}(1-\hat{p})}{\hat{p}}\right)^2 = 4.3 \qquad\qquad\qquad (2)$$

Substituting equation (1) into (2) gives:

$$\frac{1.5}{\hat{p}} + 1.5^2 = 4.3 \quad \Rightarrow \quad \hat{p} = 0.7317$$

Hence, equation (1) gives $\hat{k} = 4.091$.

Alternatively, if we use the second of the second-order equations, we would get $\text{var}(X) = \dfrac{\hat{k}(1-\hat{p})}{\hat{p}^2}$

and $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \dfrac{43}{10} - 1.5^2 = 2.05$. *Equating these gives:*

$$\frac{\hat{k}(1-\hat{p})}{\hat{p}^2} = 2.05 \qquad\qquad (3)$$

Substituting equation (1) into (3) gives:

$$\frac{1.5}{\hat{p}} = 2.05 \quad \Rightarrow \quad \hat{p} = 0.7317$$

and hence $\hat{k} = 4.091$ *as before.*

Note that s^2 with divisor $(n-1)$ is often used in place of the second central sample moment, *ie* we often use the definition of the sample variance quoted on page 22 of the Tables.

So our second-order equation is now:

$$\text{var}(X) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right\}$$

Using this version will *not* give the same estimates as those obtained using the previous second-order equations. However, if n is large there is little difference between the estimates obtained.

The advantage of this method is that s^2 is an unbiased estimator of the population variance. The importance of this property is covered in more detail later.

Question

A random sample from a *Bin*(n,p) distribution yields the following values:

> 4, 2, 7, 4, 1, 4, 5, 4

Find method of moments estimates of *n* and *p* using $\bar{x}$ and s^2 (with a denominator of $n-1$).

Solution

We have sample mean and variance of:

$$\bar{x} = 3.875 \quad \text{and} \quad s^2 = \frac{1}{7}\left\{143 - 8 \times 3.875^2\right\} = 3.26786$$

The population mean and variance are:

$$E[X] = np \qquad \text{and:} \qquad \text{var}[X] = np(1-p)$$

Equating population and sample statistics gives:

$$\hat{n}\hat{p} = 3.875 \quad \text{and} \quad \hat{n}\hat{p}(1-\hat{p}) = 3.26786$$

Solving gives $\hat{p} = 0.1567$ and $\hat{n} = 24.73$ (which are different from the values calculated previously).

We can also apply the method of moments to continuous distributions.

Question

The sample mean and sample variance for a large random sample from a $Gamma(\alpha, \lambda)$ distribution are 10 and 25, respectively. Use the method of moments to estimate α and λ.

Solution

Equating the mean and variance, we get:

$$\frac{\hat{\alpha}}{\hat{\lambda}} = 10 \quad \text{and} \quad \frac{\hat{\alpha}}{\hat{\lambda}^2} = 25$$

Dividing the first equation by the second gives:

$$\hat{\lambda} = \frac{10}{25} = 0.4 \quad \Rightarrow \quad \hat{\alpha} = 10 \times 0.4 = 4$$

For cases with more than two parameters, moments about zero should be used.

For example, if you had 3 parameters to estimate, you would use the set of equations:

$$E[X] = \frac{1}{n}\sum x_i \qquad\qquad E[X^2] = \frac{1}{n}\sum x_i^2 \qquad\qquad E[X^3] = \frac{1}{n}\sum x_i^3$$

This approach can be extended in an obvious way for more than three parameters.

2　The method of maximum likelihood

The method of maximum likelihood is widely regarded as the best general method of finding estimators. In particular maximum likelihood estimators have excellent and usually easily determined asymptotic properties and so are especially good in the large-sample situation.

'Asymptotic' here just means when the samples are very large.

2.1　The one-parameter case

The most important stage in applying the method is that of writing down the likelihood:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

for a random sample $x_1, x_2, \ldots, x_n$ from a population with density or probability function $f(x; \theta)$.

$\prod$ means product, so $\prod_{i=1}^{n} f(x_i)$ would mean $f(x_1) \times f(x_2) \times f(x_3) \times \cdots \times f(x_n)$. The above statement is saying that the likelihood function is the product of the densities (or the probability functions in the case of discrete distributions) calculated for each sample value.

Remember that θ is the parameter whose value we are trying to estimate.

The likelihood is the probability of observing the sample in the discrete case, and is proportional to the probability of observing values in the neighbourhood of the sample in the continuous case.

Notice that the likelihood function is a function of the unknown parameter θ. So different values of θ would give different values for the likelihood. The maximum likelihood approach is to find the value of θ that would have been most likely to give us the particular sample we got. In other words, we need to find the value of θ that maximises the likelihood function.

For a continuous distribution the probability of getting any exact value is zero, but since

$P(X = x) \simeq \int_{x-\varepsilon}^{x+\varepsilon} f(t)\,dt \simeq 2\varepsilon f(x)$, we can see that it is proportional to the PDF.

In most cases taking logs greatly simplifies the determination of the maximum likelihood estimator (MLE) $\hat{\theta}$.

Differentiating the likelihood or log likelihood with respect to the parameter and setting the derivative to zero gives the maximum likelihood estimator for the parameter.

Example

Given a random sample of size n $(ie\ x_1, \ldots, x_n)$ from the exponential population with density $f(x) = \lambda e^{-\lambda x}$, $x > 0$, the MLE, $\hat{\lambda}$, is found as follows:

$$L(\lambda) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

$$\therefore \log L(\lambda) = n \log \lambda - \lambda \sum x_i$$

$$\frac{\partial}{\partial \lambda} \log L(\lambda) = \frac{n}{\lambda} - \sum x_i$$

equating to zero:

$$\frac{n}{\lambda} - \sum x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$$

$$\therefore \text{MLE is } \hat{\lambda} = \frac{1}{\bar{X}}$$

Note that $\frac{1}{\bar{x}}$ is a maximum likelihood estimate, *ie* a numerical value, whereas $\frac{1}{\bar{X}}$ is a maximum likelihood estimator, *ie* a random variable.

It is necessary to check, either formally or through simple logic, that the turning point is a maximum. Generally the likelihood starts at zero, finishes at or tends to zero, and is non-negative. Therefore if there is one turning point it must be a maximum.

The formal approach would be to check that the second derivative is negative. For the above example we get:

$$\frac{d^2}{d\lambda^2} \log L(\lambda) = -\frac{n}{\lambda^2} < 0 \implies \text{max}$$

It is important that we do check, whether formally or through simple logic, *and* state this (together with your working/reasoning) in the exam to receive all the marks.

At the differentiation stage, any terms that do not contain the parameter (λ in this case) will disappear. So when the log-likelihood is written down, any terms that don't contain the parameter can be thought of as 'a constant'.

We can calculate maximum likelihood estimatess for parameters from discrete distributions too.

Question

A random sample of size n (ie $x_1, x_2, \ldots, x_n$) is taken from a $Poi(\mu)$ distribution.

(i)   Derive the maximum likelihood estimator of μ.

(ii)   The sum of a sample of 10 observations from a $Poisson(\mu)$ distribution was 24. Calculate the maximum likelihood estimate $\hat{\mu}$.

Solution

(i)   The likelihood function is:

$$L(\mu) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{x_i}}{x_i!} = \text{constant} \times e^{-n\mu}\mu^{\sum x_i}$$

Taking logs:

$$\ln L(\mu) = \text{constant} - n\mu + \sum x_i \ln \mu$$

Differentiating with respect to μ:

$$\frac{d}{d\mu}\ln L(\mu) = -n + \frac{\sum x_i}{\mu}$$

Setting this equal to zero gives:

$$\hat{\mu} = \frac{\sum x_i}{n} = \overline{x}$$

Differentiating again (to check that it is a maximum):

$$\frac{d^2}{d\mu^2}\ln L(\mu) = -\frac{\sum x_i}{\mu^2} < 0 \implies \text{max}$$

So the estimate (the value obtained for a particular sample) is $\overline{x}$. The estimator (the random variable) is $\overline{X}$.

(ii)   We have $n = 10$ and $\sum x_i = 24$. Hence the estimate is simply:

$$\hat{\mu} = \overline{x} = \frac{24}{10} = 2.4$$

MLEs display the invariance property, which means that if $\hat{\theta}$ is the MLE of θ then the MLE of a function $g(\theta)$ is $g(\hat{\theta})$.

For example, the MLE of $2\theta^2 - 1$ would simply be $2\hat{\theta}^2 - 1$.

Question

The MLEs of the parameters of a lognormal distribution have been found to be $\hat{\mu} = 2$ and $\hat{\sigma}^2 = 0.25$. Derive the maximum likelihood estimate of the mean of the lognormal distribution.

Solution

The formula for the mean θ (say) of a lognormal distribution is (from page 14 of the *Tables*):

$$\theta = e^{\mu + \frac{1}{2}\sigma^2}$$

The invariance property tells us that the MLEs of $\theta, \mu,$ and σ are related by the same equation:

$$\hat{\theta} = e^{\hat{\mu} + \frac{1}{2}\hat{\sigma}^2}$$

So the MLE of the mean is:

$$\hat{\theta} = e^{2 + \frac{1}{2} \times 0.25} = 8.37$$

2.2 The two-parameter case

This is straightforward in principle and the method is the same as the one-parameter case, but the solution of the resulting equations may be more awkward, perhaps requiring an iterative or numerical solution.

The only difference is that a partial derivative is taken with respect to each parameter, before equating each to zero and solving the resulting system of simultaneous equations for the parameters.

So in summary, the steps for finding the maximum likelihood estimator in straightforward cases are:

- Write down the likelihood function, *L*.

- Find $\ln L$ and simplify the resulting expression.

- Partially differentiate $\ln L$ with respect to each parameter to be estimated.

- Set the derivatives equal to zero.

- Solve these equations simultaneously.

In the two-parameter case, the second-order condition that is used to check for maxima is more complicated, and we shall not discuss it here.

Question

Derive the MLEs of μ and σ for a sample of n iid observations from a $N(\mu, \sigma^2)$ distribution.

Solution

The likelihood function is:

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] = \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \times \text{constant}$$

Taking logs:

$$\log L = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 + \text{constant}$$

Differentiating with respect to μ and σ gives:

$$\frac{\partial}{\partial\mu}\log L = \frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i - n\mu\right)$$

$$\frac{\partial}{\partial\sigma}\log L = -\frac{n}{\sigma} - \frac{1}{2}\left(-\frac{2}{\sigma^3}\right)\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{\sigma}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - n\right)$$

Setting these to zero gives:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}$$

Also:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n-1}{n}s^2 \quad \Rightarrow \quad \hat{\sigma} = s\sqrt{\frac{n-1}{n}}$$

2.3 A special case – the uniform distribution

For populations where the range of the random variable involves the parameter, care must be taken to specify when the likelihood is zero and non-zero. Often a plot of the likelihood is helpful.

An example of a random variable where the range involves the parameter is the uniform distribution:

$$f(x) = \frac{1}{b-a} \qquad a < x < b$$

We look at this in the next question – note how we specify when the likelihood is zero (*ie* it does not exist for the specified values of the parameter) and non-zero (*ie* where it does exist for the specified values of the parameter).

The second important feature about this question is that the usual route for finding the maximum using differentiation breaks down.

Question

Derive the maximum likelihood estimate of θ for $U(0,\theta)$ based on a random sample of values $x_1, x_2, ..., x_n$.

Solution

For a sample from the $U(0,\theta)$ distribution we must have $0 \leq x_1, ..., x_n \leq \theta$. Hence $\max x_i \leq \theta$. Thus the likelihood for a sample of size n is:

$$L = \begin{cases} \dfrac{1}{\theta^n} & \text{if } \theta > \max x_i \\ 0 & \text{otherwise} \end{cases}$$

Differentiation doesn't work because $\frac{d}{d\theta} \ln L(\theta) = -\frac{n}{\theta}$ which gives a turning point of $\theta \to \infty$. The second derivative shows the problem $\frac{d^2}{d\theta^2} \ln L(\theta) = \frac{n}{\theta^2} > 0$. We have a minimum as $\theta \to \infty$.

So using common sense, we must find the θ that *maximises* $L(\theta) = \frac{1}{\theta^n}$. We want θ to be as small as possible subject to the constraint that $\theta \geq \max x_i$. Hence $\hat{\theta} = \max x_i$.

2.4   Incomplete samples

The method of maximum likelihood can be applied in situations where the sample is incomplete. For example, truncated data or censored data in which observations are known to be greater than a certain value, or multiple claims where the number of claims is known to be two or more.

Censored data arise when you have information about the full range of possible values but it's not complete (*eg* you only know that there are, say, 6 values greater than 500). Truncated data arise when you actually have no information about part of the range of possible values (*eg* you have no information at all about values greater than 500).

In these situations, as long as the likelihood (the probability of observing the given information) can be written as a function of the parameter(s), then the method can be used. Again in such cases the solution may be more complex, perhaps requiring numerical methods.

For example, suppose a sample yields n observations $(x_1, x_2, ..., x_n)$ and m observations greater than the value y, then the likelihood is given by:

$$L(\theta) = \left[\prod_{i=1}^{n} f(x_i, \theta)\right] \times [P(X > y)]^m$$

Our estimate will be as accurate as possible if we use all the information that we have available. For incomplete samples, we don't know what the values above y are. All we know is that they are greater than y. Since the values above y are unknown we cannot use $L(\theta) = \prod_{i=1}^{n+m} f(x_i, \theta)$. We instead use the formula given.

If the information is more detailed than 'greater than y' we can use a more detailed likelihood function. For example, if we have m observed values between y and z, and p observed values above z, in addition to the n known values, then we would use:

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) \times [P(y < X < z)]^m \times [P(X > z)]^p$$

Question

Claims (in £000s) on a particular policy have a distribution with PDF given by:

$$f(x) = 2cxe^{-cx^2} \qquad x > 0$$

Seven of the last ten claims are given below:

   1.05, 3.38, 3.26, 3.22, 2.71, 2.37, 1.85

The three remaining claims were known to be greater than £6,000. Calculate the maximum likelihood estimate of c.

Solution

We have 7 known claims and 3 claims greater than 6. So the likelihood is:

$$L(c) = \prod_{i=1}^{7} f(x_i) \times [P(X > 6)]^3$$

Since $P(X > 6) = \int_6^\infty 2cxe^{-cx^2}\,dx = \left[-e^{-cx^2}\right]_6^\infty = e^{-c\times 6^2}$ and $\sum_{i=1}^7 x_i^2 = 49.91$, we have a likelihood function of the form:

$$L(c) = \prod_{i=1}^7 2cx_i e^{-cx_i^2} \times \left[e^{-c\times 6^2}\right]^3$$

$$= \text{constant} \times c^7 e^{-c\sum_{i=1}^7 x_i^2} \times e^{-108c}$$

$$= \text{constant} \times c^7 e^{-157.91c}$$

The log-likelihood is:

$$\ln L(c) = \text{constant} + 7\ln c - 157.91c$$

Differentiating the log likelihood gives:

$$\frac{d}{dc}\ln L(c) = \frac{7}{c} - 157.91$$

Setting this equal to zero:

$$\frac{7}{\hat{c}} - 157.91 = 0 \;\;\Rightarrow\;\; \hat{c} = \frac{7}{157.91} = 0.0443$$

Differentiating again to check we get a maximum:

$$\frac{d^2}{dc^2}\ln L(c) = -\frac{7}{c^2} < 0 \;\;\Rightarrow\;\; \text{max}$$

If we have some claims about which nothing is known (*ie* we don't even know whether there are any claims of a particular type), then the data is said to be truncated, rather than censored. We need to take a slightly different approach here.

Question

The number of claims in a year on a pet insurance policy are distributed as follows:

No. of claims, n	0	1	2	≥ 3
$P(N = n)$	5θ	3θ	θ	$1 - 9\theta$

Information from the claims file for a particular year showed that there were 60 policies with 1 claim, 24 policies with 2 claims and 16 policies with 3 or more claims. There was no information about the number of policies with no claims.

Obtain the maximum likelihood estimate of θ.

Solution

Since we have no information at all about zero claims, we need to determine the truncated distribution. All we do is omit the zero claims probability and scale up the remaining probabilities (which only total to $1-5\theta$) so that they now total to 1:

No. of claims, n	1	2	≥ 3
$P(N=n)$	$\dfrac{3\theta}{1-5\theta}$	$\dfrac{\theta}{1-5\theta}$	$\dfrac{1-9\theta}{1-5\theta}$

These probabilities can also be thought of as conditional probabilities, ie the first probability in the table is actually $P(N=1\,|\,N>0)$. Using the definition of conditional probability, we obtain:

$$P(N=1\,|\,N>0)=\frac{P(N=1)}{P(N>0)}=\frac{3\theta}{1-5\theta}$$

and we obtain the same probabilities as before.

The likelihood is:

$$L(\theta\,|\,N>0)=\text{constant}\times\big[P(N=1)\big]^{60}\times\big[P(N=2)\big]^{24}\times\big[P(N\geq3)\big]^{16}$$

So:

$$L(\theta\,|\,N>0)=\text{constant}\times\left(\frac{3\theta}{1-5\theta}\right)^{60}\times\left(\frac{\theta}{1-5\theta}\right)^{24}\times\left(\frac{1-9\theta}{1-5\theta}\right)^{16}$$

$$=\text{constant}\times\frac{\theta^{84}(1-9\theta)^{16}}{(1-5\theta)^{100}}$$

The constant arises from the fact that we don't know which of the 60 policies had 1 claim, *etc* and so there is some combinatorial factor to account for this.

The log-likelihood is:

$$\ln L(\theta\,|\,N>0)=\text{constant}+84\ln\theta+16\ln(1-9\theta)-100\ln(1-5\theta)$$

Differentiating and setting equal to zero gives:

$$\frac{d}{d\theta}\ln L(\theta\,|\,N>0)=\frac{84}{\theta}-\frac{9\times16}{1-9\theta}+\frac{5\times100}{1-5\theta}$$

$$\Rightarrow\;\; 84(1-9\hat\theta)(1-5\hat\theta)-144\hat\theta(1-5\hat\theta)+500\hat\theta(1-9\hat\theta)=0$$

$$\Rightarrow\;\; 84-820\hat\theta=0$$

$$\Rightarrow\;\; \hat\theta=0.102$$

Differentiating again to check we get a maximum:

$$\frac{d^2}{d\theta^2}\ln L(\theta\,|\,N>0) = -\frac{84}{\theta^2} - \frac{9\times9\times16}{(1-9\theta)^2} + \frac{5\times5\times100}{(1-5\theta)^2} < 0 \text{ when } \theta = 0.102 \implies \text{ max}$$

Independent samples

For independent samples from two populations which share a common parameter, the overall likelihood is the product of the two separate likelihoods.

Question

The number of claims, X, per year arising from a low-risk policy has a Poisson distribution with mean μ. The number of claims, Y, per year arising from a high-risk policy has a Poisson distribution with mean 2μ.

A sample of 15 low-risk policies had a total of 48 claims in a year and a sample of 10 high-risk policies had a total of 59 claims in a year. Determine the maximum likelihood estimate of μ based on this information.

Solution

The likelihood for these 15 low-risk and 10 high-risk policies is:

$$L(\mu) = \prod_{i=1}^{15} P(X = x_i) \times \prod_{j=1}^{10} P(Y = y_j) = \prod_{i=1}^{15} \frac{\mu^{x_i}}{x_i!}e^{-\mu} \times \prod_{j=1}^{10} \frac{(2\mu)^{y_j}}{y_j!}e^{-2\mu}$$

$$= \text{constant} \times \mu^{\sum_{i=1}^{15} x_i} e^{-15\mu} \times \mu^{\sum_{j=1}^{10} y_j} e^{-20\mu}$$

$$= \text{constant} \times \mu^{48}e^{-15\mu} \times \mu^{59}e^{-20\mu} = \text{constant} \times \mu^{107}e^{-35\mu}$$

The log-likelihood is:

$$\ln L(\mu) = \text{constant} + 107\ln\mu - 35\mu$$

Differentiating and setting equal to zero gives:

$$\frac{d}{d\mu}\ln L(\mu) = \frac{107}{\mu} - 35 \implies \hat{\mu} = \frac{35}{107} = 3.057$$

Differentiating again to check we get a maximum:

$$\frac{d^2}{d\mu^2}\ln L(\mu) = -\frac{107}{\mu^2} < 0 \implies \text{ max}$$

3   Unbiasedness

Consideration of the sampling distribution of an estimator can give an indication of how good it is as an estimator. Clearly the aim is for the sampling distribution of the estimator to be located near the true value and have a small spread.

If we have a random sample $X = (X_1, X_2, ..., X_n)$ from a distribution with an unknown parameter θ and $g(\underline{X})$ is an estimator of θ, it seems desirable that $E[g(\underline{X})] = \theta$.

This is the property of unbiasedness.

You can think of an unbiased estimator as one whose mean value equals the true parameter value.

Question

Show that the estimator for μ obtained in the question on page 12 is unbiased.

Solution

In the question we had a $Poi(\mu)$ distribution and our estimator was $\hat{\mu} = \overline{X}$. To show that this is unbiased we need to show that $E(\hat{\mu}) = \mu$, ie $E(\overline{X}) = \mu$.

We have:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

Since $X_i \sim Poi(\mu)$ we have $E(X_i) = \mu$. Hence:

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \times n\mu = \mu$$

So the estimator $\hat{\mu} = \overline{X}$ is unbiased.

If an estimator is biased, its bias is given by $E[g(\underline{X})] - \theta$, ie it is a measure of the difference between the expected value of the estimator and the parameter being estimated.

If the bias is greater than zero, the estimator is said to be positively biased *ie* it tends to overestimate the true value. Alternatively, the bias could be less than zero, leading to a negatively biased estimator that would tend to underestimate the true value.

Question

The following are estimators for the variance of a distribution having mean μ and variance σ^2. Obtain the bias for each estimator:

(i) $\qquad S^2 = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$

(ii) $\qquad \hat{\sigma}^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$.

Solution

(i)   The formula for the bias of S^2 is:

$$bias(S^2) = E(S^2) - \sigma^2$$

Consider $E(S^2)$:

$$E(S^2) = E\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] = E\left[\frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) \right]$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2) \right)$$

But since:

$$E(X_i^2) = \text{var}(X_i) + E^2(X_i) = \sigma^2 + \mu^2$$

$$E(\bar{X}^2) = \text{var}(\bar{X}) + E^2(\bar{X}) = \frac{\sigma^2}{n} + \mu^2$$

So we get:

$$E(S^2) = \frac{1}{n-1} \left(\sum_{i=1}^{n} (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2 \right) \right)$$

$$= \frac{1}{n-1} \left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \right)$$

$$= \frac{1}{n-1} (n-1)\sigma^2$$

$$= \sigma^2$$

So the bias is:

$$bias(S^2) = E(S^2) - \sigma^2 = \sigma^2 - \sigma^2 = 0$$

This means that S^2 is an unbiased estimator of σ^2.

(ii)   Since $\hat{\sigma}^2 = \dfrac{n-1}{n} S^2$ we can use the result from part (i) to get:

$$E(\hat{\sigma}^2) = E\left[\frac{n-1}{n} S^2\right] = \frac{n-1}{n} E(S^2) = \frac{n-1}{n} \sigma^2$$

So the bias is:

$$bias(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

The property of unbiasedness is not preserved under non-linear transformations of the estimator/parameter.

So, for example, the fact that S^2 is an unbiased estimator of the population variance does not mean that *S* is an unbiased estimator of the population standard deviation.

As indicated earlier unbiasedness seems to be a desirable property. However it is not necessarily an essential property for an estimator. There are many common situations in which a biased estimator is better than an unbiased one, and, in fact, better than the best unbiased estimator.

The importance of unbiasedness is secondary to that of having a small mean square error.

An unbiased estimator is simply one that for different samples will give the true value on average. However, it could be that some of the estimates are too large and some are too small – but on *average* they give the true value. So we need some way of measuring the 'spread' of the estimates obtained for different samples. That measure is the mean square error and is covered in the next section.

Therefore a biased estimator whose value does not deviate very far from the true value (*ie* has a small spread) would be preferable to an unbiased one whose values are 'all over the place' – as the biased estimator would be more reliable (*ie* no matter what sample we had, the estimate is still likely to be closer to the true value).

4   Mean square error

As biased estimators can be better than unbiased ones a measure of efficiency is needed to compare estimators generally. That measure is the mean square error.

The mean square error (MSE) of an estimator $g(\underline{X})$ for θ is defined by:

$$\mathsf{MSE}(g(\underline{X})) = E\left[\left(g(\underline{X}) - \theta\right)^2\right]$$

Note that this is a function of θ.

Thus the mean square error is the second moment of $g(\underline{X})$ about θ and an estimator with a lower MSE is said to be more efficient.

The MSE of a particular estimator can be worked out directly as an integral using the density of the sampling distribution of $g(\underline{X})$, or using the density of $\underline{X}$ itself.

However it is usually much easier to use the alternative expression:

$$\mathsf{MSE} = \mathsf{Variance} + \mathsf{bias}^2$$

as this makes use of quantities that are already known or can easily be obtained.

This expression can be proved as follows:

(Simplifying things by dropping the $(\underline{X})$ and writing simply *g*.)

$$\mathsf{MSE}(g) = E\left[\left(g - \theta\right)^2\right]$$

$$= E\left[\left\{\left(g - E[g]\right) + \left(E[g] - \theta\right)\right\}^2\right]$$

$$= E\left[\left(g - E[g]\right)^2\right] + 2\left(E[g] - \theta\right)E\left[g - E[g]\right] + \left[E[g] - \theta\right]^2$$

$$= \mathsf{var}[g] + 0 + \mathsf{bias}^2[g] \quad \text{as required}$$

Note: If the estimator $g(\underline{X})$ is unbiased, then MSE = variance.

Question

Obtain the MSE of the estimator for μ obtained in the question on page 12.

Solution

In the question, we had a $Poi(\mu)$ distribution and our estimator was $\hat{\mu} = \overline{X}$. The MSE is given by:

$$MSE(\hat{\mu}) = \text{var}(\hat{\mu}) + bias^2(\hat{\mu})$$

Earlier on page 21 we showed that the estimator was unbiased, *ie* $bias(\hat{\mu}) = 0$. So:

$$MSE(\hat{\mu}) = \text{var}(\hat{\mu}) + 0^2 = \text{var}(\hat{\mu}) = \text{var}(\overline{X})$$

But:
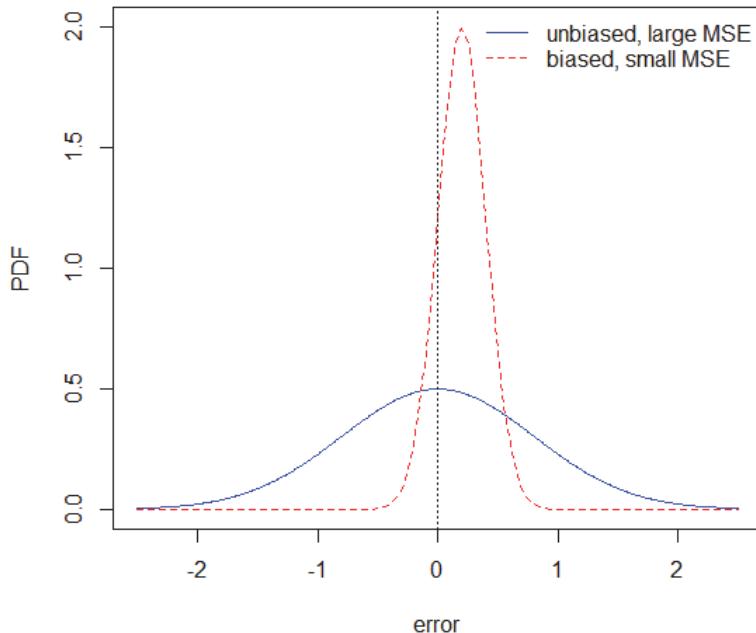
$$\text{var}(\overline{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) \quad \text{since } X_i \text{ are independent}$$

As $X_i \sim Poi(\mu)$ we have $\text{var}(X_i) = \mu$. Hence:

$$\text{var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mu = \frac{1}{n^2} \times n\mu = \frac{\mu}{n}$$

So the MSE is $\dfrac{\mu}{n}$.

The following diagram gives the sampling distributions of two estimators: one is unbiased but has a large variance, the other is biased with a much smaller variance. This illustrates a situation in which a biased estimator is better than an unbiased one.

It is clear that an estimator with a 'small' MSE is a good estimator. It is also desirable that an estimator gets better as the sample size increases. Putting these together suggests that it is desirable that $\text{MSE} \to 0$ as $n \to \infty$. This property is known as consistency.

Question

The estimator, $\hat{\sigma}^2$, is used to estimate the variance of a $N(\mu, \sigma^2)$ distribution based on a random sample of n observations:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

(i)   Determine the mean square error of $\hat{\sigma}^2$.

(ii)   Determine whether $\hat{\sigma}^2$ is consistent.

Solution

(i)   Now:

$$\hat{\sigma}^2 = \frac{(n-1)}{n} S^2 \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\Rightarrow \quad \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Hence the mean of $\hat{\sigma}^2$ is obtained from:

$$E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = n-1 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$$

This gives a bias of:

$$bias(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

The variance of $\hat{\sigma}^2$ is determined as follows:

$$var\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = 2(n-1) \quad \Rightarrow \quad var(\hat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4$$

Therefore the MSE of $\hat{\sigma}^2$ is given by:

$$MSE(\hat{\sigma}^2) = var(\hat{\sigma}^2) + \left[bias(\hat{\sigma}^2)\right]^2$$

$$= \frac{2(n-1)}{n^2}\sigma^4 + \left(-\frac{\sigma^2}{n}\right)^2$$

$$= \frac{2n-1}{n^2}\sigma^4$$

(ii)   Since the MSE, $\dfrac{2n-1}{n^2}\sigma^4$, tends to zero as $n \to \infty$, it *is* consistent.

5   Asymptotic distribution of MLEs

Given a random sample of size _n_ from a distribution with density (or probability function in the discrete case) $f(x; \theta)$, the maximum likelihood estimator $\hat{\theta}$ is such that, for large _n_, $\hat{\theta}$ is approximately normal, and is unbiased with variance given by the Cramér-Rao lower bound, that is:

$$\hat{\theta} \dot{\sim} N(\theta, \text{CRLB})$$

where $\text{CRLB} = \dfrac{1}{nE\left\{\left[\dfrac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right\}}$.

The MLE can therefore be called asymptotically efficient in that, for large _n_, it is unbiased with a variance equal to the lowest possible value of unbiased estimators.

The Core Reading is saying that the CRLB gives a lower bound for the variance of an *unbiased* estimator of a parameter (which is the same as its mean square error). So no *unbiased* estimator can have a smaller variance than the CRLB.

This is potentially a very useful result as it provides an approximate distribution for the MLE when the true sampling distribution may be unknown or impossible to determine easily, and hence may be used to obtain approximate confidence intervals.

Confidence intervals will be covered in a later chapter.

The result holds under very general conditions with only one major exclusion: it does not apply in cases where the range of the distribution involves the parameter, such as the uniform distribution.

This is due to a discontinuity, so the derivative in the formula doesn't make sense.

There are two useful alternative expressions for the CRLB based on the likelihood itself. Noting that $L(\theta)$ is really $L(\theta, \underline{X})$, these are:

$$\text{CRLB} = \frac{1}{E\left\{\left[\dfrac{\partial}{\partial\theta}\log L(\theta, \underline{X})\right]^2\right\}} \quad \text{and} \quad \text{CRLB} = \frac{1}{-E\left[\dfrac{\partial^2}{\partial\theta^2}\log L(\theta, \underline{X})\right]}$$

The second formula is normally easier to work with (as we would have calculated the second derivative of the log-likelihood when checking that we get a maximum). This formula is given on page 23 of the *Tables*.

Question

Derive the CRLB for estimators of μ, for a sample $X_1, \ldots, X_n$ from a *Poi(μ)* distribution.

Solution

The likelihood is:

$$L(\mu) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{X_i}}{X_i!} = \text{constant} \times e^{-n\mu}\mu^{\sum X_i}$$

So:

$$\ln L(\mu) = \text{constant} - n\mu + \sum X_i \ln\mu$$

Differentiating with respect to μ gives:

$$\frac{d}{d\mu}\ln L(\mu) = -n + \frac{\sum X_i}{\mu}$$

Setting this equal to zero would give the MLE of $\hat{\mu} = \overline{X}$.

Differentiating again (which we would have done to check we get a maximum):

$$\frac{d^2}{d\mu^2}\ln L(\mu) = -\frac{\sum X_i}{\mu^2}$$

Finding the expectation of this (noting that only the X_i's are random variables):

$$E\left[\frac{d^2}{d\mu^2}\ln L(\mu)\right] = -\frac{1}{\mu^2}\sum E[X_i] = -\frac{1}{\mu^2}\sum\mu = -\frac{1}{\mu^2}n\mu = -\frac{n}{\mu}$$

So, from the second formula for the CRLB:

$$CRLB = -1 \bigg/ E\left[\frac{d^2}{d\mu^2}\ln L(\mu)\right] = \frac{\mu}{n}$$

In fact, in this case, the maximum likelihood estimator $\hat{\mu} = \overline{X}$ is unbiased and has variance μ/n. So, the CRLB *can* be attained by the variance.

We can find the CRLB for estimators of parameters from continuous distributions.

Question

(i)   Show that the CRLB for unbiased estimators of μ, based on a random sample of n

        observations from a $N(\mu, \sigma^2)$ distribution with known variance σ^2, is given by $\dfrac{\sigma^2}{n}$.

(ii)   Show that the variance of the maximum likelihood estimator $\hat{\mu} = \overline{X}$ attains the CRLB.

Solution

(i)   From the question on page 14, we see that:

$$\frac{\partial}{\partial \mu} \ln L(\mu) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(X_i - \mu) = \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} X_i - n\mu \right)$$

Setting this equal to zero and rearranging gave the MLE $\hat{\mu} = \bar{X}$.

Note: We have changed x_i to X_i as we are working with the estimator.

Differentiating again gives:

$$\frac{\partial^2}{\partial \mu^2} \ln L(\mu) = -\frac{n}{\sigma^2}$$

Since there are no X_i's, all the values are constants, and hence:

$$E\left[\frac{\partial^2}{\partial \mu^2} \ln L(\mu) \right] = E\left[-\frac{n}{\sigma^2} \right] = -\frac{n}{\sigma^2}$$

So, from the second formula for the CRLB:

$$CRLB = -1 \bigg/ E\left[\frac{\partial^2}{\partial \mu^2} \ln L(\mu) \right] = \frac{\sigma^2}{n}$$

(ii)   From a previous chapter we saw that if $X \sim N(\mu, \sigma^2)$ then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} \right)$ so $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$.

Hence the variance of the MLE attains the CRLB.

What follows now is an example to illustrate the fact that if we want to obtain the CRLB for the variance, σ^2, we can't just take the CRLB for the standard deviation, σ, and square it. The reason for this is that the formula for the CRLB of σ is:

$$CRLB(\sigma) = -\frac{1}{E\left[\dfrac{d^2}{d\sigma^2} \ln L(\sigma) \right]}$$

whereas the formula for the CRLB of $v = \sigma^2$ is:

$$CRLB(v) = -\frac{1}{E\left[\dfrac{d^2}{dv^2} \ln L(v) \right]}$$

There will be no simple connection between the derivatives.

Question

Derive the CRLB for estimators of the variance of a $N(\mu, \sigma^2)$ distribution, where μ is known, based on a random sample of n observations.

Solution

We need to work in terms of the population variance σ^2, which we will write as v. The likelihood function is:

$$L(v) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{1}{2v}(X_i - \mu)^2\right] = v^{-\frac{n}{2}} \exp\left[-\frac{1}{2v}\sum_{i=1}^{n}(X_i - \mu)^2\right] \times \text{constant}$$

Taking logs:

$$\ln L(v) = -\frac{n}{2}\ln v - \frac{1}{2v}\sum_{i=1}^{n}(X_i - \mu)^2 + \text{constant}$$

Differentiating with respect to *v* gives:

$$\frac{\partial}{\partial v}\log L(v) = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Differentiating again:

$$\frac{\partial^2}{\partial v^2}\ln L(v) = \frac{n}{2v^2} - \frac{1}{v^3}\sum_{i=1}^{n}(X_i - \mu)^2 \;\; ie \;\; \frac{n}{2v^2} - \frac{1}{v^2}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

We need to determine the expectation of this. We will use the fact that $X_i \sim N(\mu, \sigma^2)$, so $Z_i = \left(\dfrac{X_i - \mu}{\sigma}\right) \sim N(0,1)$ and hence:

$$E(Z_i^2) = \text{var}(Z_i) + E^2(Z_i) = 1 + 0^2 = 1$$

So we have:

$$E\left[\frac{\partial^2}{\partial v^2}\ln L(v)\right] = \frac{n}{2v^2} - \frac{1}{v^2}\sum_{i=1}^{n}E\left[\left(\frac{X_i - \mu}{\sigma}\right)^2\right]$$

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

$$= \frac{n}{2v^2} - \frac{1}{v^2}\sum_{i=1}^{n}1$$

$$= \frac{n}{2v^2} - \frac{n}{v^2} = -\frac{n}{2v^2}$$

Hence:

$$CRLB = -1\Big/ E\left[\frac{\partial^2}{\partial v^2}\log L\right] = \frac{2v^2}{n} = \frac{2\sigma^4}{n}$$

We now consider the CRLB for a random sample of observations from an exponential distribution.

Question

Given a random sample of n observations from an $Exp(\lambda)$ distribution, determine the CRLB for unbiased estimators of:

(i) λ

(ii)   the population mean, $\mu = \frac{1}{\lambda}$.

Comment on the results.

Solution

(i)   Using the Core Reading example from page 11, we have:

$$L(\lambda) = \prod_{i=1}^{n}\lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda\sum_{i=1}^{n}X_i}$$

$$\Rightarrow \quad \ln L(\lambda) = n\ln\lambda - \lambda\sum_{i=1}^{n}X_i$$

Differentiating this with respect to λ gives:

$$\frac{d}{d\lambda}\ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i$$

Setting this equal to zero gives the estimator $\hat{\lambda} = \dfrac{1}{\bar{X}}$.

Differentiating again with respect to λ gives:

$$\frac{d^2}{d\lambda^2}\ln L(\lambda) = -\frac{n}{\lambda^2}$$

Since there are no X_i's, all the values are constants and hence:

$$E\left[\frac{d^2}{d\lambda^2}\ln L(\lambda)\right] = E\left[-\frac{n}{\lambda^2}\right] = -\frac{n}{\lambda^2}$$

So, from the second formula for the CRLB:

$$CRLB = -1 \Big/ E\left[\frac{d^2}{d\lambda^2}\ln L(\lambda)\right] = \frac{\lambda^2}{n}$$

(ii)   We are estimating the mean of an *Exp*(λ) distribution, ie $\mu = \dfrac{1}{\lambda}$, therefore we need to work in terms of μ and differentiate with respect to μ.

The likelihood function for the sample is:

$$L(\mu) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \frac{1}{\mu^n} e^{-\sum X_i/\mu}$$

$$\Rightarrow \quad \ln L(\mu) = -n\ln\mu - \frac{1}{\mu}\sum X_i$$

Differentiating with respect to μ:

$$\frac{d}{d\mu}\ln L(\mu) = -\frac{n}{\mu} + \frac{\sum X_i}{\mu^2}$$

Differentiating again with respect to μ:

$$\frac{d^2}{d\mu^2}\ln L(\mu) = \frac{n}{\mu^2} - \frac{2\sum X_i}{\mu^3}$$

Finding the expectation of this:

$$E\left[\frac{d^2}{d\mu^2}\ln L(\mu)\right] = \frac{n}{\mu^2} - \frac{2}{\mu^3}\sum E[X_i]$$

Since for $X_i \sim Exp(\lambda)$ we have $E(X_i) = \mu$, we get:

$$E\left[\frac{d^2}{d\mu^2}\ln L(\mu)\right] = \frac{n}{\mu^2} - \frac{2}{\mu^3}\sum \mu = \frac{n}{\mu^2} - \frac{2}{\mu^3}n\mu = -\frac{n}{\mu^2}$$

So, from the second formula for the CRLB:

$$CRLB = -1\bigg/ E\left[\frac{d^2}{d\mu^2}\log L\right] = \frac{\mu^2}{n}$$

Comment

Although $\mu = \frac{1}{\lambda}$ we see that $CRLB(\mu) \neq \frac{1}{CRLB(\lambda)}$.

In fact we actually have $CRLB(\mu) = \frac{\mu^2}{n} = \frac{1}{n\lambda^2}$.

6   Comparing the method of moments with MLE

We now compare the method of moments and the method of maximum likelihood.

Essentially maximum likelihood is regarded as the better method.

In the usual one-parameter case the method of moments estimator is always a function of the sample mean $\overline{X}$ and this must limit its usefulness in some situations. For example in the case of the uniform distribution on $(0, \theta)$ the method of moments estimator is $2\overline{X}$ and this can result in inadmissible estimates which are greater than θ.

For example, supposing we had the following data from $U(0, \theta)$:

> 4.5, 1.8, 2.7, 0.9, 1.3

This gives $\overline{x} = 2.24$. Since the method of moments estimator is $\hat{\theta} = 2\overline{X}$, we have $\hat{\theta} = 4.48$. But this estimate for the upper limit is inadmissible as one of the data values is greater than this.

Nevertheless in many common applications such as the binomial, Poisson, exponential and normal cases both methods yield the same estimator.

In some situations such as the gamma with two unknown parameters the simplicity of the method of moments gives it a possible advantage over maximum likelihood which may require a complicated numerical solution.

To obtain the MLE of α from a gamma distribution requires the differentiation of $\Gamma(\alpha)$, which will require numerical methods.

7   The bootstrap method

This section of the Core Reading refers to the use of R in bootstrapping. This material is not explained in detail here; we cover it in the material for the second paper of Subject CS1.

7.1   Introduction to bootstrap

The bootstrap method is a computer intensive estimation method and can be used to estimate the properties of an estimator. It is mainly distinguished in two types: parametric and non-parametric bootstrap.

Suppose that we want to make inferences about parameter θ using observed data $(y_1, y_2, \ldots, y_n)$ which follow a distribution with cumulative distribution function $F(y; \theta)$.

Usually inference is based on the *sampling distribution* of an estimator $\hat{\theta}$. A sampling distribution is obtained either by theoretical results, or is based on a large number of samples from $F(y; \theta)$.

For example, suppose we have a sample $(y_1, y_2, \ldots, y_n)$ from an exponential distribution with parameter λ and we wish to make inferences about λ. The CLT tells us that asymptotically $\bar{Y} \sim N(1/\lambda, 1/n\lambda^2)$ and we can use this sampling distribution to estimate quantities of interest (*eg* for confidence intervals or tests about λ). However, there will be cases where assumptions or asymptotic results may not hold (or we may not want to use them – *eg* when samples are small).

Then one alternative option is to use the bootstrap method. Bootstrap allows us to avoid making assumptions about the *sampling* distribution of a statistic of interest, by instead forming an *empirical sampling* distribution of the statistic. This is generally achieved by resampling based on the available sample.

7.2   Non-parametric (full) bootstrap

The main idea behind non-parametric bootstrap, when estimating a parameter θ, can be described as follows.

Construct the empirical distribution, $\hat{F}_n$, of the data:

$$\hat{F}_n(y) = \frac{1}{n}\{\text{Number of } y_i \le y\}$$

Then perform the following steps:

Draw a sample of size n from $\hat{F}_n$.

This is the bootstrap sample $(y_1^*, y_2^*, \ldots, y_n^*)$ with y^* selected *with replacement* from $(y_1, y_2, \ldots, y_n)$.

Obtain an estimate $\hat{\theta}^*$ from the bootstrap sample.

This is done in the same way as $\hat{\theta}$ is obtained from the original sample.

Repeat steps 1 and 2, say, *B* times.

Provided that *B* is sufficiently large, the output set of estimates $(\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*)$ will provide the empirical distribution of $\hat{\theta}^*$, which serves as an estimate of the sampling distribution of $\hat{\theta}$, and is referred to as the *bootstrap empirical distribution* of $\hat{\theta}$.

Schematically, this can be thought as

$$y_1, y_2, \dots, y_n \rightarrow \left. \begin{array}{l} \text{sample 1:} \quad \left(y_1^*, y_2^*, \dots, y_n^*\right) \rightarrow \hat{\theta}_1^* \\ \text{sample 2:} \quad \left(y_1^*, y_2^*, \dots, y_n^*\right) \rightarrow \hat{\theta}_2^* \\ \quad\vdots \\ \text{sample } B: \quad \left(y_1^*, y_2^*, \dots, y_n^*\right) \rightarrow \hat{\theta}_B^* \end{array} \right\} \rightarrow \text{Bootstrap empirical distribution of } \hat{\theta}.$$

The bootstrap distribution of $\hat{\theta}$ can then be used for any desired inference regarding the estimator $\hat{\theta}$, and particularly to estimate its properties. For example we can:

estimate the mean of estimator $\hat{\theta}$ by using the sample mean of the bootstrap estimates $(\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*)$:

$$\hat{E}(\hat{\theta}) = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j^* \; ;$$

estimate its median, using the 0.5 empirical quantile of the bootstrap estimates $\hat{\theta}_j^*$;

estimate the *variance* of estimator $\hat{\theta}$ by using the sample variance of the bootstrap estimates $(\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*)$:

$$\widehat{var}(\hat{\theta}) = \frac{1}{B-1} \left\{ \sum_{j=1}^{B} \left(\hat{\theta}_j^*\right)^2 - \frac{1}{B} \left(\sum_{j=1}^{B} \hat{\theta}_j^* \right)^2 \right\} \; ;$$

estimate a $(1-\alpha)\%$ *confidence interval* for $\hat{\theta}$ by:

$$\left(k_{\alpha/2}, \; k_{1-\alpha/2} \right)$$

where k_α denotes the α *th* empirical quantile of the bootstrap values $\hat{\theta}^*$. Confidence intervals are described in Chapter 8.

Example

Suppose we have the following sample of 10 values (to 2 DP) from an *Exp(λ)* distribution with unknown parameter λ:

        0.61   6.47   2.56   5.44   2.72   0.87   2.77   6.00   0.14   0.75

We can use the following R code to obtain a single resample with replacement from this original sample.

```
sample.data <- c(0.61, 6.47, 2.56, 5.44, 2.72, 0.87, 2.77, 6.00,
0.14, 0.75)

sample(sample.data, replace=TRUE)
```

Note that this is non-parametric as we are ignoring the *Exp(λ)* assumption to obtain a new sample.

The following R code obtains $B = 1,000$ estimates $(\hat{\lambda}_1^*, \hat{\lambda}_2^*, ..., \hat{\lambda}_{1,000}^*)$ using $\hat{\lambda}_j^* = 1/\bar{y}_j^*$ and stores them in the vector `estimate`:

```
set.seed(47)

estimate<-rep(0,1000)

for (i in 1:1000)

{x<-sample(sample.data, replace=TRUE);

estimate[i]<-1/mean(x)}
```
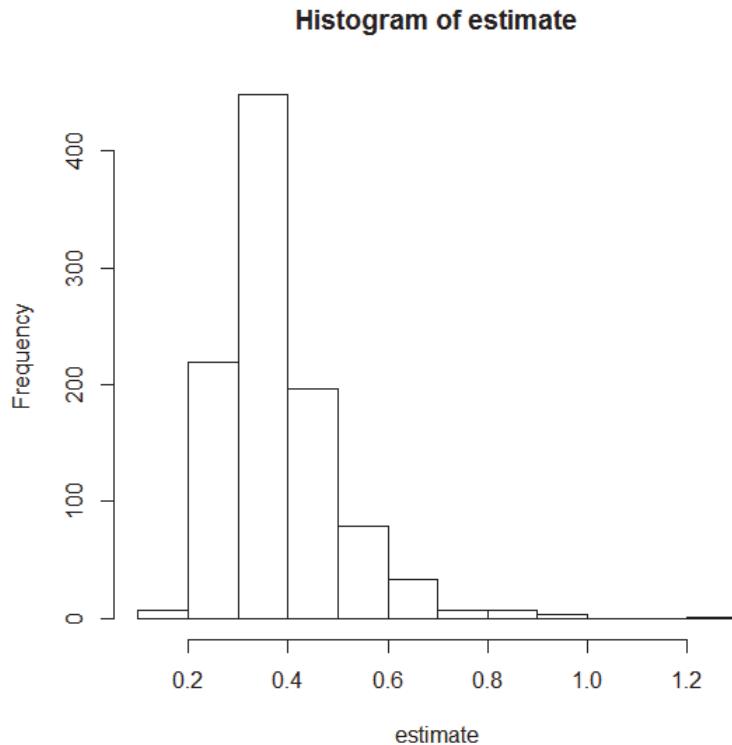
An alternative would be to use:

```
set.seed(47)

estimate <- replicate(1000, 1/mean(sample(sample.data,
replace=TRUE)))
```

This gives us the following empirical sampling distribution of $\hat{\lambda}$:

Histogram of estimate

We can obtain estimates for the mean, standard error and 95% confidence interval of the estimator $\hat{\lambda}$ using the following R code:

```
mean(estimate)

sd(estimate)

quantile(estimate, c(0.025,0.975))
```

7.3   Parametric bootstrap

If we are *prepared to assume* that the sample is considered to come from a given distribution, we first obtain an estimate of the parameter of interest $\hat{\theta}$ (*eg* using maximum likelihood, or method of moments). Then we use the assumed distribution, with parameter equal to $\hat{\theta}$, to draw the bootstrap samples. Once the bootstrap samples are available, we proceed as with the non-parametric method before.

Example

Using our sample of 10 values (to 2 DP) from an *Exp*(λ) distribution with unknown parameter λ:

        0.61   6.47   2.56   5.44   2.72   0.87   2.77   6.00   0.14   0.75

Our estimate would for λ would be $\hat{\lambda} = 1/\bar{y} = 1/2.833 = 0.3530$. We now use the *Exp*(0.3530) distribution to generate the bootstrap samples.

Note that this is parametric as we are using the exponential distribution to obtain new samples.

We can use the following R code to obtain $B = 1,000$ estimates $(\hat{\lambda}_1^*, \hat{\lambda}_2^*,..., \hat{\lambda}_{1,000}^*)$ using $\hat{\lambda}_j^* = 1/\overline{y}_j^*$ and stores them in the vector `param.estimate`:

```
set.seed(47)

param.estimate<-rep(0,1000)

for (i in 1:1000)

{x<-rexp(10,rate=1/mean(sample.data));

param.estimate[i]<-1/mean(x)}
```
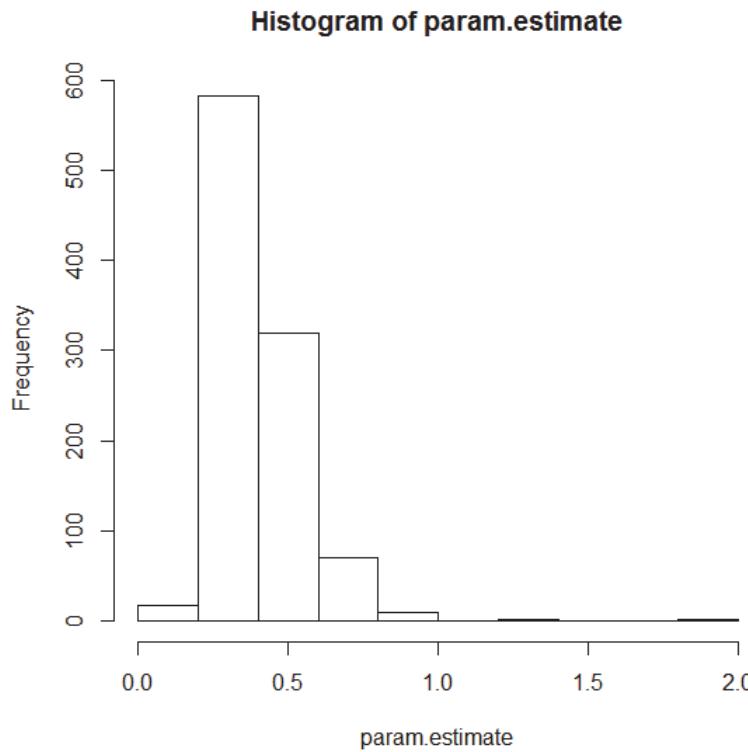
An alternative would be to use:

```
param.estimate <- replicate(1000,
1/mean(rexp(10,rate=1/mean(sample.data)))
```

This gives us the following empirical sampling distribution of $\hat{\lambda}$:

Histogram of param.estimate

Various inferences can then be made using the bootstrap estimates $(\hat{\lambda}_1^*, \hat{\lambda}_2^*,..., \hat{\lambda}_B^*)$ as before.

Bootstrap methodology can also be used in other, more complicated, scenarios – for example in regression analysis or generalised linear model settings.

The chapter summary starts on the next page so that you can keep
all the chapter summaries together for revision purposes.

Chapter 7 Summary

We have covered two methods in this chapter for estimating parameters.

The method of moments technique equates the population moments to the sample moments using the formulae:

1 parameter $E(X) = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

2 parameters $E(X) = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ $E(X^2) = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2$ or $\text{var}(X) = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

alternatively $E(X) = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ $\text{var}(X) = S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

The method of maximum likelihood has the following stages:

* find the likelihood $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$

* find $\ln L$

* find θ that solves $\dfrac{\partial}{\partial \theta}\ln L(\theta) = 0$

* check for maximum $\dfrac{\partial^2}{\partial \theta^2}\ln L(\theta) < 0$.

If the range of the distribution is a function of the parameter the maximum must be found from first principles.

Three properties of estimators are bias, mean square error (MSE) and consistency.

The bias of an estimator is given by $E[g(\underline{X})] - \theta$ where $g(\underline{X})$ is the estimator.

$g(\underline{X})$ is an unbiased estimator of θ if $E[g(\underline{X})] = \theta$.

Chapter 7 Summary (continued)

The mean square error of an estimator is given by $E[(g(\underline{X}) - \theta)^2]$ where $g(\underline{X})$ is the estimator.

An easier formula is $\text{var}[g(\underline{X})] + bias^2[g(\underline{X})]$.

An estimator is consistent if the mean square error tends to zero as n tends to infinity, where n is the size of the sample.

A good estimator has a small MSE, is unbiased and consistent.

The Cramér-Rao lower bound gives a lower bound for the variance of an unbiased estimator. It can be used to obtain confidence intervals. Its formula is:

$$CRLB(\theta) = -\frac{1}{E\left[\dfrac{\partial^2}{\partial \theta^2} \ln L(\theta, \underline{X})\right]}$$

The value of the CRLB depends on the parameter you are estimating. To use this formula, the likelihood must be expressed in terms of the correct parameter.

The asymptotic distribution of an MLE is:

$$\hat{\theta} \stackrel{.}{\sim} N(\theta, CRLB)$$

Chapter 7 Practice Questions

7.1   A random sample from a *Poisson(μ)* distribution is as follows:

        4, 2, 7, 3, 1, 2, 5, 4, 0, 2

        Calculate the method of moments estimate for μ.

7.2   Using the method of moments, estimate the mean and variance of the heights of 10-year old
        children, assuming these conform to a normal distribution, based on a random sample of 5 such
        children whose heights are:

        124cm, 122cm, 130cm, 125cm and 132cm.

7.3   Waiting times in a post office queue have an *Exp(λ)* distribution. Ten people had waiting times
        (in minutes) of:

        1.6   0.9   1.1   2.1   0.7   1.5   2.3   1.7   3.0   3.4

        A further six people had waiting times of more than 4 minutes. Based on these data calculate the
        maximum likelihood estimate of λ.

7.4   The number of claims arising in a year on a certain type of insurance policy has a Poisson
        distribution with parameter λ.

        The insurer's claim file shows that claims were made on 238 policies during the last year with the
        following frequency distribution for the number of claims:

Number of claims	Frequency
1	174
2	50
3	10
4	4
≥ 5	0

        No information is available from the *policy* file, that is, only data concerning those policies on
        which claims were made can be used in the estimation of the claim rate λ. (This is why there is
        no entry for the number of claims being 0 in the table.)

        (i)   Show that the truncated probability function is given by:

$$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!(1-e^{-\lambda})} \qquad x=1,2,3,\ldots \qquad\qquad\qquad [3]$$

        (ii)   Hence show that both the method of moments estimate and the MLE of λ satisfy the
                equation $\lambda = \bar{x}(1-e^{-\lambda})$, where $\bar{x}$ is the mean number of claims for policies that have at
                least one claim.   [7]

(iii)   Solve this equation, by any means, for the given data and calculate the resulting estimate of λ to two decimal places.   [3]

(iv)   Hence, estimate the percentage of all policies with no claims during the year.   [1]

[Total 14]

7.5   Determine the mean square error of $\hat{\mu} = \overline{X}$ which is used to estimate the mean of a $N(\mu, \sigma^2)$ distribution based on a random sample of *n* observations.

7.6   Suppose that unbiased estimators X_1 and X_2 of a parameter θ have been determined by two

Exam style   independent methods, and suppose that $\text{var}(X_1) = \sigma^2$ and that $\text{var}(X_2) = \phi\sigma^2$, where $\phi > 0$.

Let Y be the combination given by $Y = \alpha X_1 + \beta X_2$, where α and β denote non-negative weights.

(i)   Derive the relationship satisfied by α and β so that Y is also an unbiased estimator of θ.

[2]

(ii)   Determine the variance of Y in terms of ϕ and σ^2 if, additionally, the weights are chosen such that the variance of Y is a minimum.   [4]

7.7   A random sample $x_1, x_2, \ldots, x_n$ is taken from a population, which has the probability distribution

Exam style   function $F(x)$ and the density function $f(x)$. The values in the sample are arranged in order and the minimum and maximum values x_{MIN} and x_{MAX} are recorded.

(i)   Show that the distribution function of X_{MAX} is $[F(x)]^n$, and find a corresponding formula for the distribution function of X_{MIN}.   [3]

The original distribution is now believed to be a *Pareto*$(\alpha, 1)$ distribution, *ie* the probability density function is:

$$f(x) = \frac{\alpha}{(1+x)^{\alpha+1}}, \quad x \geq 0$$

(ii)   Find the distribution function of X, and hence find the distribution function of X_{MAX}.   [2]

(iii) Show that the probability density function for the distribution of X_{MIN}, is:

$$f_{X_{MIN}}(x) = \frac{n\alpha}{(1+x)^{n\alpha+1}} \quad x \geq 0$$ [2]

(iv)   A random sample of 25 values gives a sample value for x_{MIN} of 23. Use the distribution of X_{MIN} to obtain a maximum likelihood estimate of α.   [3]

(v)   The same random sample gives a value of x_{MAX} of 770. Obtain an equation for the maximum likelihood estimator of α using x_{MAX}. Comment on the difficulty of solving this equation.   [3]

(vi)   Outline what further information you would need here in order to obtain a method of moments estimate of α .   [1]
[Total 14]

7.8   A random sample of eight observations from a distribution are given below:

         4.8   7.6   1.2   3.5   2.9   0.8   0.5   2.3

(i)   Derive the method of moments estimates for:

         (a) λ from an *Exp*(λ) distribution

         (b) v from a χ_v^2 distribution.

(ii)   Derive the method of moments estimators for:

         (a)   *k* and *p* from a Type 2 negative binomial distribution

         (b) μ and σ^2 from a lognormal distribution.

7.9   Show that the likelihood that an observation from a *Poisson*(λ) distribution takes an odd value (*ie* 1, 3, 5,...) is $\frac{1}{2}(1 - e^{-2\lambda})$.

7.10   A discrete random variable has a probability function given by:

x	2	4	5
$P(X = x)$	$\frac{1}{8} + 2\alpha$	$\frac{1}{2} - 3\alpha$	$\frac{3}{8} + \alpha$

(i)   Give the range of possible values for the unknown parameter α .

A random sample of 30 observations gave respective frequencies of 7, 6 and 17.

(ii)   Calculate the method of moments estimate of α .

(iii)   Write down an expression for the likelihood of these data and hence show that the maximum likelihood estimate $\hat{\alpha}$ satisfies the quadratic equation:

$$180\hat{\alpha}^2 + \frac{111}{8}\hat{\alpha} - \frac{91}{32} = 0$$

(iv)   Hence determine the maximum likelihood estimate and explain why one root is rejected as a possible estimate of α .

7.11 A motor insurance portfolio produces claim incidence data for 100,000 policies over one year. The table below shows the observed number of policyholders making 0, 1, 2, 3, 4, 5, and 6 or more claims in a year.

No. of claims	No. of policies
0	87,889
1	11,000
2	1,000
3	100
4	10
5	1
≥ 6	–
Total	100,000

(i) Using the method of moments, estimate the parameter of a Poisson distribution to fit the above data and hence calculate the expected number of policies giving rise to the different numbers of claims assuming a Poisson model. [3]

(ii) Show that the estimate of the Poisson parameter calculated from the above data using the method of moments is also the maximum likelihood estimate of this parameter. [4]

(iii) Using the method of moments, estimate the two parameters of a Type 2 negative binomial distribution to fit the above data and hence calculate the expected number of policies giving rise to the different numbers of claims assuming a negative binomial model. [6]

You may use the relationship:

$$P(X = x) = \frac{k + x - 1}{x} \times q \times P(X = x - 1)$$

for the negative binomial distribution.

(iv) Explain briefly why you would expect a negative binomial distribution to fit the above data better than a Poisson distribution. [2]

[Total 15]

7.12   A random sample $X_1, \ldots, X_n$ is taken from a normal distribution with mean μ and variance σ^2.

Exam style

(i)   State the distribution of $\dfrac{\sum (X_i - \overline{X})^2}{\sigma^2}$.   [1]

It is decided to estimate the variance, σ^2, using the following estimator:

$$\hat{\sigma}^2 = \frac{1}{n+b} \sum (X_i - \overline{X})^2$$

where b is a constant.

(ii)   (a)   Use part (i) to obtain the bias of $\hat{\sigma}^2$.

         (b)   Hence, show that $\hat{\sigma}^2$ is unbiased when $b = -1$.   [3]

(iii)   (a)   Show, using parts (i) and (ii)(a), that the mean square error of $\hat{\sigma}^2$ is given by:

$$MSE(\hat{\sigma}^2) = \frac{2(n-1) + (1+b)^2}{(n+b)^2} \sigma^4$$

         (b)   Determine whether the estimator, $\hat{\sigma}^2$, is consistent.

         (c)   Show that the mean square error of $\hat{\sigma}^2$ is minimised when $b = 1$.   [7]

                  You may assume that the turning point is a minimum.

(iv)   Comment on the best choice for the value of b.   [2]
                                                                                    [Total 13]

The solutions start on the next page so that you can
separate the questions and solutions.

Chapter 7 Solutions

7.1   The population mean for a *Poisson*(μ) from page 7 of the *Tables* is just μ .

The sample mean is $\bar{x} = \dfrac{30}{10} = 3$.

Equating population mean to sample mean gives:

$$\mu = 3$$

Since this is an estimate of the true value of μ we write $\hat{\mu} = 3$.

7.2   The sample moments are:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{633}{5} = 126.6 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{80,209}{5} = 16,041.8$$

The population moments are $E(X) = \mu$ and $E(X^2) = \text{var}(X) + E^2(X) = \sigma^2 + \mu^2$. Equating the sample and population moments gives:

$$\hat{\mu} = 126.6$$

$$\hat{\mu}^2 + \hat{\sigma}^2 = 16,041.8 \quad \Rightarrow \quad \hat{\sigma}^2 = 14.24$$

Alternatively, using $\bar{x} = 126.6$ *and* $s^2 = \frac{1}{4}\left\{80,209 - 5 \times 126.6^2\right\} = 17.8$ *and equating these to the*

population moments of $E(X) = \mu$ *and* $\text{var}(X) = \sigma^2$ *gives:*

$$\hat{\mu} = \bar{x} = 126.6 \quad \text{and} \quad \hat{\sigma}^2 = s^2 = 17.8$$

7.3   Using the likelihood formula given for censored data in Section 2.4:

$$L(\lambda) = \left[\prod_{i=1}^{10} f(x_i)\right] \times \left[P(X > 4)\right]^6 = \lambda^{10} e^{-\lambda \sum x_i} \times \left(e^{-4\lambda}\right)^6$$

since $f(x_i) = \lambda e^{-\lambda x_i}$ and $P(X > 4) = \int_{4}^{\infty} \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_{4}^{\infty} = e^{-4\lambda}$.

Taking logs:

$$\ln L(\lambda) = 10\ln \lambda - \lambda \sum x_i - 24\lambda$$

Since $\sum x_i = 18.3$ we get:

$$\ln L(\lambda) = 10\ln \lambda - 42.3\lambda$$

Differentiating and equating to zero gives:

$$\frac{d}{d\lambda}\ln L(\lambda) = \frac{10}{\hat{\lambda}} - 42.3 = 0 \quad \Rightarrow \quad \hat{\lambda} = 0.2364$$

Checking that it's a maximum:

$$\frac{d^2}{d\lambda^2}\ln L(\lambda) = -\frac{10}{\lambda^2} < 0 \quad \Rightarrow \quad \text{max}$$

7.4   (i)   Since only policies with claims are included, we must use a truncated Poisson distribution:

$$P(X=x) = k\frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 1,2,3,\dots \qquad\qquad [1]$$

where k is the constant of proportionality to ensure that the sum of the probabilities is 1.

For the ordinary Poisson distribution:

$$\sum_x P(X=x) = 1 \qquad \Rightarrow \quad P(X \geq 1) = 1 - P(X=0) = 1 - e^{-\lambda} \qquad\qquad [1]$$

So our probability function can be written as:

$$k\sum_{x=1}^{\infty} P(X=x) = 1 \quad \Rightarrow \quad k(1 - e^{-\lambda}) = 1 \quad \Rightarrow \quad k = \frac{1}{(1 - e^{-\lambda})} \qquad\qquad [1]$$

(ii)   We will first use the method of moments technique, so we need the mean of the truncated Poisson distribution:

$$E[X] = \sum_{x=1}^{\infty} x\frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})} = \sum_{x=0}^{\infty} x\frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})} = \frac{1}{(1 - e^{-\lambda})}\sum_{x=0}^{\infty} x\frac{\lambda^x e^{-\lambda}}{x!} \qquad\qquad [2]$$

since the $x = 0$ term is zero.

The sum is the mean of the Poisson distribution (found by summing $x \times \text{PF}$), so we get:

$$E[X] = \frac{1}{(1 - e^{-\lambda})}\sum_{x=0}^{\infty} x\frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{(1 - e^{-\lambda})} \times \lambda = \frac{\lambda}{(1 - e^{-\lambda})} \qquad\qquad [1]$$

So the method of moments equation is $\bar{x} = \dfrac{\lambda}{1 - e^{-\lambda}}$ or $\lambda = \bar{x}(1 - e^{-\lambda})$, as required.

Now using maximum likelihood, the likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!(1 - e^{-\lambda})} = \text{constant} \times \frac{\lambda^{\sum x_i} e^{-n\lambda}}{(1 - e^{-\lambda})^n} \qquad\qquad [1]$$

where the constant incorporates the factorial factor.

Taking logs:

$$\log L = \text{constant} + (\sum x_i)\log \lambda - n\lambda - n\log(1 - e^{-\lambda})$$ [1]

Differentiating with respect to λ:

$$\frac{d}{d\lambda}\log L = \frac{\sum x_i}{\lambda} - n - \frac{ne^{-\lambda}}{1 - e^{-\lambda}} = \frac{n\overline{x}(1 - e^{-\lambda}) - n\lambda(1 - e^{-\lambda}) - n\lambda e^{-\lambda}}{\lambda(1 - e^{-\lambda})}$$

$$= \frac{n\overline{x}(1 - e^{-\lambda}) - n\lambda}{\lambda(1 - e^{-\lambda})}$$

Equating to zero gives $\lambda = \overline{x}(1 - e^{-\lambda})$ as required.   [2]

(iii)   From the data:

$$\overline{x} = \frac{174 \times 1 + 50 \times 2 + 10 \times 3 + 4 \times 4}{238} = \frac{320}{238}$$ [1]

So $\frac{320}{238}(1 - e^{-\lambda}) = \lambda$ or $\frac{\lambda}{(1 - e^{-\lambda})} - \frac{320}{238} = 0$.

Using trial and error on the second equation we get:

$$\lambda = 0.6 \implies LHS = -0.0147$$

$$\lambda = 0.7 \implies LHS = 0.0460$$

Using linear interpolation:

$$\lambda = 0.6 + \frac{0 - (-0.0147)}{0.0460 - (-0.0147)}(0.7 - 0.6) = 0.624$$ [2]

Alternatively we could use a systematic method such as Newton-Raphson.

(iv)   Now $P(X = 0) = e^{-\lambda}$. By the invariance property, the maximum likelihood estimate of this probability is:

$$e^{-\hat{\lambda}} = e^{-0.624} = 0.536$$

So we estimate that 54% of policies have no claims.   [1]

7.5   The MSE is given by:

$$MSE(\hat{\mu}) = \text{var}(\hat{\mu}) + bias^2(\hat{\mu})$$

$$= \text{var}(\overline{X}) + bias^2(\overline{X})$$

where:

$$bias(\overline{X}) = E(\overline{X}) - \mu$$

When $X \sim N(\mu, \sigma^2)$ we have $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ so $E(\overline{X}) = \mu$ and $\text{var}(\overline{X}) = \dfrac{\sigma^2}{n}$. Hence:

$$bias(\overline{X}) = \mu - \mu = 0$$

Therefore:

$$MSE(\hat{\mu}) = \text{var}(\overline{X}) + 0^2 = \frac{\sigma^2}{n}$$

7.6   (i)   Since X_1 and X_2 are unbiased estimators of θ this means that:

$$E[X_1] = E[X_2] = \theta \qquad\qquad\qquad [1]$$

$$\Rightarrow \quad E(Y) = E(\alpha X_1 + \beta X_2) = \alpha E(X_1) + \beta E(X_2) = (\alpha + \beta)\theta$$

Hence, if Y is unbiased for θ, then $\alpha + \beta = 1$.   [1]

(ii)   Now we have $\text{var}(X_1) = \sigma^2$ and $\text{var}(X_2) = \phi\sigma^2$. Also X_1 and X_2 are independent, so:

$$\text{var}(Y) = \text{var}(\alpha X_1 + \beta X_2) = \alpha^2 \text{var}(X_1) + \beta^2 \text{var}(X_2)$$

$$= \sigma^2[\alpha^2 + (1-\alpha)^2 \phi] \qquad\qquad\qquad [1]$$

To obtain the minimum, we set the derivative equal to zero:

$$\frac{d}{d\alpha}\text{var}(Y) = \sigma^2[2\alpha - 2(1-\alpha)\phi]$$

$$\Rightarrow \quad \alpha = (1-\alpha)\phi \quad \Rightarrow \quad \alpha = \frac{\phi}{1+\phi} \qquad\qquad [1]$$

Checking it's a minimum:

$$\frac{d^2}{d\alpha^2}\text{var}(Y) = \sigma^2[2 + 2\phi] > 0 \quad \Rightarrow \quad \text{min} \qquad\qquad [1]$$

So:

$$\text{var}(Y) = \sigma^2 \left[\left(\frac{\phi}{1+\phi} \right)^2 + \left(1 - \frac{\phi}{1+\phi} \right)^2 \phi \right]$$

$$= \sigma^2 \left[\frac{\phi^2}{(1+\phi)^2} + \frac{\phi}{(1+\phi)^2} \right]$$

$$= \frac{\phi}{1+\phi} \sigma^2 \qquad\qquad [1]$$

7.7   (i)   Consider the value of X_{MAX}. This will be less than some value x, say, if and only if all the sample values are less than x. The probability of this happening is just $[F(x)]^n$. So:

$$F_{X_{MAX}}(x) = [F(x)]^n \qquad\qquad [1]$$

Using similar logic, X_{MIN} will be greater than some number x if and only if all the sample values are greater than x. So:

$$P(X_{MIN} \geq x) = P(\text{all } X_i \geq x) = [1 - F(x)]^n$$

$$\Rightarrow \quad F_{X_{MIN}}(x) = 1 - [1 - F(x)]^n \qquad\qquad [2]$$

(ii)   The distribution function of X is given by:

$$F(x) = \int_0^x f(t)dt = \int_0^x \alpha(1+t)^{-\alpha-1}dt = \left[-(1+t)^{-\alpha} \right]_0^x = 1 - (1+x)^{-\alpha} \qquad\qquad [1]$$

where $x \geq 0$.

Hence $F_{X_{MAX}}(x) = [F(x)]^n = \left(1 - (1+x)^{-\alpha} \right)^n$. $\qquad\qquad [1]$

(iii)   Similarly $F_{X_{MIN}}(x) = 1 - [1 - F(x)]^n = 1 - \left[(1+x)^{-\alpha} \right]^n = 1 - (1+x)^{-n\alpha} \quad x \geq 0$ $\qquad [1]$

This has the same form as the original distribution function, so X_{MIN} has a Pareto distribution with parameters $n\alpha$ and 1. So the density function of X_{MIN} is:

$$f_{X_{MIN}}(x) = \frac{n\alpha}{(1+x)^{n\alpha+1}} \quad x \geq 0 \qquad\qquad [1]$$

(iv)   The likelihood function for α, based on a *single value* of X_{MIN}, is:

$$L(\alpha) = \frac{n\alpha}{(1+x)^{n\alpha+1}}$$

$$\Rightarrow \quad \log L(\alpha) = \log n + \log \alpha - (n\alpha + 1)\log(1+x)$$

$$\Rightarrow \quad \frac{\partial}{\partial \alpha} \log L(\alpha) = \frac{1}{\alpha} - n\log(1+x)$$

$$\Rightarrow \quad \hat{\alpha} = \frac{1}{n\log(1+x)} \qquad\qquad [2]$$

Substituting in $n = 25$ and $x = 23$, we get $\hat{\alpha} = 0.01259$. $\qquad$ [1]

(v)   Applying the same approach to X_{MAX}, we have (using the derivative of $F_{X_{MAX}}(x)$ from earlier) a likelihood function of:

$$L(\alpha) = f_{X_{MAX}}(x) = n\left(1 - \left(1+x\right)^{-\alpha}\right)^{n-1} \alpha(1+x)^{-\alpha-1}$$

$$\Rightarrow \log L(\alpha) = \log n + (n-1)\log\left[1 - (1+x)^{-\alpha}\right] + \log \alpha - (\alpha+1)\log(1+x)$$

$$\Rightarrow \frac{\partial}{\partial \alpha} \log L(\alpha) = \frac{1}{\alpha} + (n-1)\times\frac{(1+x)^{-\alpha}\log(1+x)}{1-(1+x)^{-\alpha}} - \log(1+x) = 0 \qquad [2]$$

Substituting in $n = 25$ and $x = 770$ we get:

$$\frac{1}{\alpha} - \log 771 + 24 \times \frac{771^{-\alpha}\log 771}{1 - 771^{-\alpha}} = 0 \qquad\qquad [1]$$

This equation cannot be solved algebraically. A numerical method will be needed to solve it.

(vi)   We cannot use the usual method of moments approach unless we know all the individual sample values (or at least the mean of the sample). So we do not have sufficient information to use the method of moments approach here. $\qquad$ [1]

7.8   (i)   ***Method of moments (one unknown)***

We have one unknown and so require only one equation:

$$E(X) = \frac{1}{n}\sum x_i = \overline{x}$$

For our data we have $\overline{x} = 2.95$.

(i)(a) *Exponential*

Using the formula for the mean of an exponential distribution:

$$\frac{1}{\hat{\lambda}} = 2.95 \quad \Rightarrow \quad \hat{\lambda} = 0.33898$$

(i)(b) *Chi-square*

Since $\chi_v^2 = Gamma(\frac{v}{2}, \frac{1}{2})$, we get $E(X) = \frac{\alpha}{\lambda} = \frac{v/2}{1/2} = v$. Hence $\hat{v} = 2.95$.

(ii) *Method of moments (two unknowns)*

We have two unknowns and so require two equations.

Either: $\quad E(X) = \frac{1}{n} \sum x_i = \bar{x} \quad$ and $\quad E(X^2) = \frac{1}{n} \sum x_i^2$

or: $\quad E(X) = \frac{1}{n} \sum x_i = \bar{x} \quad$ and $\quad var(X) = s^2$

For our data we have $\bar{x} = 2.95$, $\frac{1}{8} \sum x_i^2 = 13.635$ and $s^2 = 5.6371$.

(ii)(a) *Negative binomial*

Using the first method gives:

$$E(X) = \frac{\hat{k}(1-\hat{p})}{\hat{p}} = 2.95$$

$$E(X^2) = var(X) + E^2(X) = \frac{\hat{k}(1-\hat{p})}{\hat{p}^2} + \left(\frac{\hat{k}(1-\hat{p})}{\hat{p}} \right)^2 = 13.635$$

Substituting the first equation into the second gives:

$$\frac{2.95}{\hat{p}} + 2.95^2 = 13.635 \quad \Rightarrow \quad \frac{2.95}{\hat{p}} = 4.9325 \quad \Rightarrow \quad \hat{p} = 0.59807$$

Hence, substituting this back into the first equation gives $\hat{k} = 4.3896$.

Using the second method gives:

$$E(X) = \frac{\hat{k}(1-\hat{p})}{\hat{p}} = 2.95 \quad \text{and} \quad var(X) = \frac{\hat{k}(1-\hat{p})}{\hat{p}^2} = 5.6371$$

Substituting the first equation into the second gives:

$$\frac{2.95}{\hat{p}} = 5.6371 \quad \Rightarrow \quad \hat{p} = 0.52331$$

Hence, substituting this back into the first equation gives $\hat{k} = 3.2386$.

(ii)(b)   **Lognormal**

Using the first method gives:

$$E(X) = e^{\hat{\mu} + \frac{1}{2}\hat{\sigma}^2} = 2.95 \quad \text{and} \quad E(X^2) = e^{2\hat{\mu} + 2\hat{\sigma}^2} = 13.635$$

Rewriting the second equation gives:

$$e^{2(\hat{\mu} + \frac{1}{2}\hat{\sigma}^2)} e^{\hat{\sigma}^2} = 2.95^2 e^{\hat{\sigma}^2} = 13.635 \quad \Rightarrow \quad \hat{\sigma}^2 = 0.44903$$

Substituting this into the first equation gives $\hat{\mu} = 0.85729$.

Using the second method gives:

$$E(X) = e^{\hat{\mu} + \frac{1}{2}\hat{\sigma}^2} = 2.95 \quad \text{and} \quad var(X) = e^{2\hat{\mu} + \hat{\sigma}^2}(e^{\sigma^2} - 1) = 5.6371$$

Substituting the first equation into the second gives:

$$var(X) = 2.95^2(e^{\hat{\sigma}^2} - 1) = 5.6371 \quad \Rightarrow \quad \hat{\sigma}^2 = 0.49942$$

Hence, substituting this into the first equation gives $\hat{\mu} = 0.83210$.

7.9   We have:

$$P[Poisson(\lambda) \text{ is odd}] = P(X = 1) + P(X = 3) + P(X = 5) + \cdots$$

$$= e^{-\lambda}\left[\lambda + \frac{\lambda^3}{3!} + \frac{\lambda^5}{5!} + \cdots\right]$$

To sum the series in the square bracket, note that:

$$e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots$$

$$e^{-\lambda} = 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \cdots$$

So $\frac{1}{2}(e^{\lambda} - e^{-\lambda}) = \lambda + \frac{\lambda^3}{3!} + \cdots$, which is the required series.

So the required probability is:

$$e^{-\lambda} \times \frac{1}{2}(e^{\lambda} - e^{-\lambda}) = \frac{1}{2}(1 - e^{-2\lambda})$$

7.10   (i)   **Range of values**

Since $0 \le P(X = x) \le 1$, using this for each of the probabilities gives lower bounds for α of $-\frac{1}{16}, -\frac{1}{6}$ and $-\frac{3}{8}$. Hence, $\alpha \ge -\frac{1}{16}$. We also obtain upper bounds for α of $\frac{7}{16}, \frac{1}{6}$ and $\frac{5}{8}$.

Hence, $\alpha \le \frac{1}{6}$.

(ii)   **Method of moments estimator**

We have one unknown, so we will use $E(X) = \bar{x}$.

$$E(X) = 2\left(\frac{1}{8} + 2\alpha\right) + 4\left(\frac{1}{2} - 3\alpha\right) + 5\left(\frac{3}{8} + \alpha\right) = \frac{33}{8} - 3\alpha$$

From the data, we have:

$$\bar{x} = \frac{7 \times 2 + 6 \times 4 + 17 \times 5}{30} = \frac{123}{30} = 4.1$$

Therefore:

$$\frac{33}{8} - 3\hat{\alpha} = 4.1 \implies \hat{\alpha} = 0.008\dot{3}$$

Note that this value lies between the limits derived in part (i).

(iii)   **Maximum likelihood**

The likelihood of obtaining the observed results is:

$$L(\alpha) = constant \times \left(\frac{1}{8} + 2\alpha\right)^7 \times \left(\frac{1}{2} - 3\alpha\right)^6 \times \left(\frac{3}{8} + \alpha\right)^{17}$$

Taking logs and differentiating gives:

$$\implies \ln L(\alpha) = constant + 7\ln\left(\frac{1}{8} + 2\alpha\right) + 6\ln\left(\frac{1}{2} - 3\alpha\right) + 17\ln\left(\frac{3}{8} + \alpha\right)$$

$$\implies \frac{d}{d\theta}\ln L(\alpha) = \frac{14}{\frac{1}{8} + 2\alpha} - \frac{18}{\frac{1}{2} - 3\alpha} + \frac{17}{\frac{3}{8} + \alpha}$$

Equating this to zero to find the maximum value of θ gives:

$$\frac{14}{\frac{1}{8} + 2\hat{\alpha}} - \frac{18}{\frac{1}{2} - 3\hat{\alpha}} + \frac{17}{\frac{3}{8} + \hat{\alpha}} = 0$$

$$\implies 14\left(\frac{1}{2} - 3\hat{\alpha}\right)\left(\frac{3}{8} + \hat{\alpha}\right) - 18\left(\frac{1}{8} + 2\hat{\alpha}\right)\left(\frac{3}{8} + \hat{\alpha}\right) + 17\left(\frac{1}{8} + 2\hat{\alpha}\right)\left(\frac{1}{2} - 3\hat{\alpha}\right) = 0$$

$$\implies 14\left(\frac{3}{16} - \frac{5}{8}\hat{\alpha} - 3\hat{\alpha}^2\right) - 18\left(\frac{3}{64} + \frac{7}{8}\hat{\alpha} + 2\hat{\alpha}^2\right) + 17\left(\frac{1}{16} + \frac{5}{8}\hat{\alpha} - 6\hat{\alpha}^2\right) = 0$$

$$\implies 180\hat{\alpha}^2 + \frac{111}{8}\hat{\alpha} - \frac{91}{32} = 0$$

(iv)   *MLE*

Solving the quadratic equation gives:

$$\hat{\alpha} = \frac{-\frac{111}{8} \pm \sqrt{\left(\frac{111}{8}\right)^2 - 4 \times 180 \times -\frac{91}{32}}}{360} = -0.170, 0.0929$$

The maximum likelihood estimate is 0.0929.

The other solution of -0.170 does not lie between the bounds calculated in (i). It is not feasible as it is less than the smallest possible value for α of -0.0625 .

7.11   (i)   ***Expected results using method of moments (Poisson)***

The sample mean for the data is:

$$\frac{1}{100,000} \times (87,889 \times 0 + 11,000 \times 1 + 1,000 \times 2 + \cdots) = 0.13345$$

The mean of a Poisson distribution with parameter λ is λ .

So the method of moments estimate of λ is 0.13345.   [1]

The expected numbers, based on this estimate, are (using the iterative formula):

$x = 0$: $100,000 \, e^{-0.13345} = 87,507$

$x = 1$: $0.13345 \times 87,507 = 11,678$

$x = 2$: $\dfrac{0.13345}{2} \times 11,678 = 779$

$x = 3$: $\dfrac{0.13345}{3} \times 779 = 35$

$x = 4$: $\dfrac{0.13345}{4} \times 35 = 1$

$x = 5$: $\dfrac{0.13345}{5} \times 1 = 0$

$x \geq 6$: $100,000 - 87,507 - 11,678 - 779 - 35 - 1 - 0 = 0$ [2]

(ii)   **MLE (Poisson)**

The likelihood of obtaining n_0 0's, n_1 1's *etc* (making a total of n), assuming the numbers conform to a Poisson distribution, is the multinomial probability:

$$L(\lambda) = \frac{n!}{n_0! \, n_1! \, n_2! \ldots} (e^{-\lambda})^{n_0} (\lambda e^{-\lambda})^{n_1} \left(\frac{\lambda^2 e^{-\lambda}}{2!} \right)^{n_2} \ldots$$

$$= constant \times \lambda^{n_1 + 2n_2 + 3n_3 + \cdots} e^{-\lambda(n_0 + n_1 + n_2 + \cdots)}$$

$$= constant \times \lambda^{13,345} e^{-100,000\lambda} \qquad\qquad\qquad [1]$$

So the log likelihood is:

$$\ln L(\lambda) = 13,345 \ln \lambda - 100,000\lambda + constant \qquad\qquad\qquad [1]$$

Differentiating with respect to λ to maximise this:

$$\frac{d}{d\lambda} \ln L(\lambda) = \frac{13,345}{\lambda} - 100,000 \qquad\qquad\qquad [1]$$

This is zero when:

$$\lambda = 13,345 / 100,000 = 0.13345 \qquad\qquad\qquad [1]$$

Since the second derivative is negative, this is the MLE of λ. It is the same as the method of moments estimate.

(iii)   **Expected results using method of moments (negative binomial)**

The second (non-central) sample moment for the data is:

$$\frac{1}{100,000} \times (87,889 \times 0^2 + 11,000 \times 1^2 + 1,000 \times 2^2 + \cdots) = 0.16085$$

The mean and second non-central moment of the negative binomial distribution with parameters k and p are $\dfrac{kq}{p}$ and $\dfrac{kq}{p^2} + \left(\dfrac{kq}{p} \right)^2$.

So the method of moments estimators of k and p satisfy the equations:

$$\frac{kq}{p} = 0.13345 \quad \text{and} \quad \frac{kq}{p^2} + \left(\frac{kq}{p} \right)^2 = 0.16085 \qquad\qquad [2]$$

From the second equation:

$$\frac{kq}{p^2} = 0.16085 - \left(\frac{kq}{p} \right)^2 = 0.16085 - (0.13345)^2 = 0.14304 \qquad\qquad [\tfrac{1}{2}]$$

Using the first equation gives:

$$p = \frac{0.13345}{0.14304} = 0.93295$$ [½]

$$q = 1 - p = 1 - 0.93295 = 0.06705$$ [½]

and $$k = \frac{0.13345 \times 0.93295}{0.06705} = 1.8569$$ [½]

The expected numbers, based on these estimates, are:

$x = 0:$ $\qquad 100,000(0.93295)^{1.8569} = 87,909$

$x = 1:$ $\qquad \dfrac{1.8569}{1} \times 0.06705 \times 87,909 = 10,945$

$x = 2:$ $\qquad \dfrac{2.8569}{2} \times 0.06705 \times 10,945 = 1,048$

$x = 3:$ $\qquad \dfrac{3.8569}{3} \times 0.06705 \times 1,048 = 90$

$x = 4:$ $\qquad \dfrac{4.8569}{4} \times 0.06705 \times 90 = 7$

$x = 5:$ $\qquad \dfrac{5.8569}{5} \times 0.06705 \times 7 = 1$

$x \geq 6:$ $\qquad 100,000 - 87,909 - 10,945 - 1,048 - 90 - 7 - 1 = 0$ [2]

Note: we have made use of the negative binomial recursive relationship given in the question.

(iv)   *Why negative binomial is a better fit*

For a Poisson distribution, the mean and variance are the same. Since the sample mean and variance (which, for a sample as large as this, should be very close to the true values) are 0.13345 and 0.14304, which differ significantly, this suggests that the Poisson distribution may not be a suitable model here. [1]

The negative binomial distribution has more flexibility and can accommodate different values for the mean and variance (provided the variance exceeds the mean). [1]

7.12   (i)   *Distribution*

Using the result given on page 22 of the Tables:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$ [1]

(ii)(a) *Bias*

The bias of $\hat{\sigma}^2$ is given by $bias(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2$. From part (i) we have:

$$E\left[\frac{\sum(X_i - \overline{X})^2}{\sigma^2}\right] = (n-1) \qquad\qquad [\frac{1}{2}]$$

Since $(n+b)\hat{\sigma}^2 = \sum(X_i - \overline{X})^2$, we have:

$$E\left[\frac{(n+b)\hat{\sigma}^2}{\sigma^2}\right] = (n-1)$$

$$\frac{(n+b)}{\sigma^2}E[\hat{\sigma}^2] = (n-1)$$

$$E[\hat{\sigma}^2] = \frac{(n-1)}{(n+b)}\sigma^2 \qquad\qquad [1]$$

Therefore the bias is given by:

$$bias(\hat{\sigma}^2) = \frac{(n-1)}{(n+b)}\sigma^2 - \sigma^2 = -\frac{(1+b)}{(n+b)}\sigma^2 \qquad\qquad [\frac{1}{2}]$$

(ii)(b) *Unbiased*

Substituting $b = -1$ into the bias gives:

$$bias(\hat{\sigma}^2) = -\frac{(1-1)}{(n-1)}\sigma^2 = 0 \qquad\qquad [1]$$

Hence, $\hat{\sigma}^2$ is an unbiased estimator of σ^2 when $b = -1$.

(iii)(a) *Mean square error*

The mean square error of $\hat{\sigma}^2$ is given by $MSE(\hat{\sigma}^2) = var(\hat{\sigma}^2) + bias^2(\hat{\sigma}^2)$. From part (i) we have:

$$var\left[\frac{\sum(X_i - \overline{X})^2}{\sigma^2}\right] = 2(n-1) \qquad\qquad [1]$$

Since $(n+b)\hat{\sigma}^2 = \sum(X_i - \bar{X})^2$, we have:

$$\text{var}\left[\frac{(n+b)\hat{\sigma}^2}{\sigma^2}\right] = 2(n-1)$$

$$\frac{(n+b)^2}{\sigma^4}\text{var}[\hat{\sigma}^2] = 2(n-1)$$

$$\text{var}[\hat{\sigma}^2] = \frac{2(n-1)}{(n+b)^2}\sigma^4 \qquad [1]$$

Using this and the bias from (ii)(a), the mean square error is given by:

$$MSE(\hat{\sigma}^2) = \frac{2(n-1)}{(n+b)^2}\sigma^4 + \frac{(1+b)^2}{(n+b)^2}\sigma^4$$

$$= \frac{2(n-1)+(1+b)^2}{(n+b)^2}\sigma^4 \qquad [1]$$

(iii)(b)   *Consistent*

As $n \to \infty$, the mean square error becomes:

$$MSE(\hat{\sigma}^2) \to \frac{2}{n}\sigma^4 \to 0$$

So $\hat{\sigma}^2$ is consistent. $\qquad [1]$

(iii)(c)   *Minimum mean square error*

Differentiating with respect to b using the quotient rule gives:

$$\frac{d}{db}MSE(\hat{\sigma}^2) = \frac{2(1+b)(n+b)^2 - [2(n-1)+(1+b)^2]\times 2(n+b)}{(n+b)^4}\sigma^4 \qquad [2]$$

Substituting $b=1$ into this expression gives:

$$\frac{d}{db}MSE(\hat{\sigma}^2)\bigg|_{b=1} = \frac{2\times 2(n+1)^2 - [2(n-1)+4]\times 2(n+1)}{(n+1)^4}\sigma^4$$

$$= \frac{4(n+1)^2 - 4(n+1)^2}{(n+1)^4}\sigma^4$$

$$= 0 \qquad [1]$$

So the MSE is minimised when $b=1$.

Alternatively, students may attempt to find the value of b that makes this zero as follows:

$$2(1+b)(n+b)^2 = [2(n-1)+(1+b)^2] \times 2(n+b)$$

$$(1+b)(n+b) = [2(n-1)+(1+b)^2]$$

$$n+b+bn+b^2 = 2n-1+2b+b^2$$

$$b(n-1) = n-1$$

$$b = 1$$

(iv)   *Best estimator*

All values of b give consistent estimators. When $b = -1$, the estimator $\hat{\sigma}^2 = \dfrac{1}{n-1}\sum(X_i - \bar{X})^2$ is

unbiased, whereas when $b = 1$, the estimator $\hat{\sigma}^2 = \dfrac{1}{n+1}\sum(X_i - \bar{X})^2$ has the smallest MSE, but it

is biased.

Since a smaller MSE is more important than being unbiased, we should choose $b = 1$.   [1]

However, there will be little difference between the estimators when n is large as the mean square errors and biases both tend to zero.   [1]

8

Confidence intervals

Syllabus objectives

3.2   Confidence intervals

3.2.1   Define in general terms a confidence interval for an unknown parameter of a distribution based on a random sample.

3.2.2   Derive a confidence interval for an unknown parameter using a given sampling distribution.

3.2.3   Calculate confidence intervals for the mean and the variance of a normal distribution.

3.2.4   Calculate confidence intervals for a binomial probability and a Poisson mean, including the use of the normal approximation in both cases.

3.2.5   Calculate confidence intervals for two-sample situations involving the normal distribution, and the binomial and Poisson distributions using the normal approximation.

3.2.6   Calculate confidence intervals for a difference between two means from paired data.

3.2.7   Use the bootstrap method to obtain confidence intervals.

0   Introduction

In the previous chapter we used the method of moments and the method of maximum likelihood to obtain estimates for the population parameter(s). For example, we might have the following numbers of claims from a certain portfolio that we receive in 100 different monthly periods:
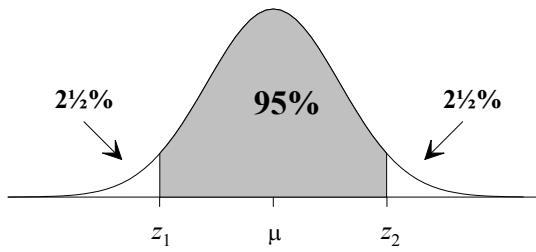
Claims	0	1	2	3	4	5	6
Frequency	9	22	26	21	13	6	3

Assuming a Poisson distribution with parameter μ for the number of claims in a month, our estimate of μ using the methods given in the previous chapter would be $\mu = \overline{x} = 2.37$.

The problem is that this might not be the correct value of μ. In this chapter we look at constructing confidence intervals that have a high probability of containing the correct value. For example, a 95% confidence interval for μ means that there is a 95% probability that it contains the true value of μ.

Confidence intervals will be constructed using the sampling distributions given in Chapter 6. For example, when sampling from a $N(\mu, \sigma^2)$ distribution where σ^2 is known:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \;\; \Rightarrow \;\; Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

If we require a 95% confidence interval, then we can read off the 2½% and 97½% z-values of ± 1.96 from the normal tables and substitute these z-values into the equation, along with our values of $\overline{X}$, σ^2 and n. Rearranging this equation gives our confidence interval for μ of

$$\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} \; .$$

It is important to realise that the formula for the endpoints of this interval contains $\overline{X}$, and so the endpoints are random variables. We can obtain numerical values for these endpoints by collecting some sample data and replacing $\overline{X}$ by the observed sample mean $\overline{x}$. Naturally, different samples may lead to different endpoints. If we sample repeatedly, 95% of the intervals we obtain should contain the true value of μ.

1   Confidence intervals in general

A confidence interval provides an 'interval estimate' of an unknown parameter (as opposed to a 'point estimate'). It is designed to contain the parameter's value with some stated probability. The width of the interval provides a measure of the precision accuracy of the estimator involved.

A $100(1-\alpha)\%$ confidence interval for θ is defined by specifying random variables $\hat{\theta}_1(\underline{X})$, $\hat{\theta}_2(\underline{X})$ such that $P\left(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})\right) = 1 - \alpha$.

Rightly or wrongly, $\alpha = 0.05$ leading to a 95% confidence interval, is by far the most common case used in practice and we will tend to use this in most of our illustrations.

Thus $P\left(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})\right) = 0.95$ specifies $\left(\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})\right)$ as a 95% confidence interval for θ. This emphasises the fact that it is the interval and not θ that is random. In the long run, 95% of the realisations of such intervals will include θ and 5% of the realisations will not include θ.

Confidence intervals are not unique. In general they should be obtained via the sampling distribution of a good estimator, in particular the maximum likelihood estimator. Even then there is a choice between one-sided and two-sided intervals and between equal-tailed and shortest-length intervals although these are often the same, *eg* for sampling distributions that are symmetrical about the unknown value of the parameter.

We will see some examples of these shortly.

2   Derivation of confidence intervals

2.1   The pivotal method

There is a general method of constructing confidence intervals called the pivotal method.

This method requires the finding of a pivotal quantity of the form $g(\underline{X}, \theta)$ with the following properties:

(1)   it is a function of the sample values and the unknown parameter θ

(2)   its distribution is completely known

(3)   it is monotonic in θ.

The distribution in condition (2) must not depend on θ. 'Monotonic' means that the function either consistently increases or decreases with θ.

The equation

$$\int_{g_1}^{g_2} f(t)\, dt = 0.95, \quad \text{(where } f(t) \text{ is the known probability (density) of } g(\underline{X}, \theta))$$

defines two values, g_1 and g_2, such that

$$P\big(g_1 < g(\underline{X}, \theta) < g_2\big) = 0.95$$

g_1 and g_2 are usually constants.

We are assuming here that X has a continuous distribution. We will look shortly at examples based on discrete distributions.

If $g(\underline{X}, \theta)$ is monotonic increasing in θ, then:

$$g(\underline{X}, \theta) < g_2 \Leftrightarrow \theta < \theta_2 \text{ for some number } \theta_2$$

$$g_1 < g(\underline{X}, \theta) \Leftrightarrow \theta_1 < \theta \text{ for some number } \theta_1$$

and if $g(\underline{X}, \theta)$ is monotonic decreasing in θ, then:

$$g(\underline{X}, \theta) < g_2 \Leftrightarrow \theta_1 < \theta$$

$$g_1 < g(\underline{X}, \theta) \Leftrightarrow \theta < \theta_2$$

resulting in (θ_1, θ_2) being a 95% confidence interval for θ.

Fortunately in most practical situations such quantities $g(\underline{X}, \theta)$ do exist, although an approximation to the method is needed for the binomial and Poisson cases.

In sampling from a $N(\mu, \sigma^2)$ distribution with known value of σ^2, a pivotal quantity is:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

which is $N(0,1)$.

For example, given a random sample of size 20 from the normal population $N(\mu, 10^2)$ which yields a sample mean of 62.75, an equal-tailed 95% confidence interval for μ is:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 62.75 \pm 1.96 \frac{10}{\sqrt{20}} = 62.75 \pm 4.38$$

This is a symmetrical confidence interval since it is of the form $\theta \pm \beta$. For symmetrical confidence intervals, we can write down the interval using the '$\pm$' notation, where the two values indicate the upper and lower limits. Alternatively, we can write this confidence interval in the form $(58.37, 67.13)$. Here we are using the pivotal quantity $\frac{\bar{X} - \mu}{10 / \sqrt{20}}$, which has a $N(0,1)$ distribution, irrespective of the value of μ.

The normal mean illustration shows that confidence intervals are not unique.

Another 95% interval, with unequal tails, is $\left(\bar{X} - 1.8808 \frac{\sigma}{\sqrt{n}}, \bar{X} + 2.0537 \frac{\sigma}{\sqrt{n}} \right)$.

However, there would not be much reason to use this one in practice.

Question

Show that both this and the first interval given above are 95% confidence intervals. Calculate the length of each of these intervals.

Solution

For the second confidence interval:

$$P(-1.8808 < Z < 2.0537) = P(Z < 2.0537) - P(Z < -1.8808)$$
$$= P(Z < 2.0537) - (1 - P(Z < 1.8808))$$
$$= 0.98000 - (1 - 0.97000) = 0.95$$

This interval has length $3.9345 \frac{\sigma}{\sqrt{n}}$.

For the first confidence interval:

$$P(-1.96 < Z < 1.96) = P(Z < 1.96) - P(Z < -1.96)$$

$$= 2 \times P(Z < 1.96) - 1$$

$$= 2 \times 0.975 - 1 = 0.95$$

This interval has length $3.92 \dfrac{\sigma}{\sqrt{n}}$.

Other intervals which are of some use in practice are the one-sided 95% intervals:

$$\left(-\infty, \ \bar{X} + 1.6449 \frac{\sigma}{\sqrt{n}} \right) \text{ and } \left(\bar{X} - 1.6449 \frac{\sigma}{\sqrt{n}}, \ \infty \right)$$

Since the normal distribution is symmetrical about the value of the unknown parameter, it is quite easy to see that the equal-tailed interval is the shortest-length interval for that level of confidence.

Question

The average IQ of a random sample of 50 university students was found to be 132. Calculate a symmetrical 95% confidence interval for the average IQ of university students, assuming that IQs are normally distributed. It is known from previous studies that the standard deviation of IQs among students is approximately 20.

Solution

Since the distribution is normal, we know that $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, when σ is known.

From the *Tables* we know that $0.95 = P(-1.96 < Z < 1.96)$, so:

$$0.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

Rearranging to obtain limits for μ:

$$0.95 = P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

Using $n = 50$, $\sigma = 20$ and $\bar{X} = 132$ from the question, we obtain:

$$126.5 < \mu < 137.5$$

So a symmetrical 95% confidence interval for the average IQ is $(126.5, \ 137.5)$.

2.2   Confidence limits

The 95% confidence interval $\left(\bar{X} - 1.96\dfrac{\sigma}{\sqrt{n}}, \ \bar{X} + 1.96\dfrac{\sigma}{\sqrt{n}} \right)$ for μ is often expressed as:

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

This is quite informative as it gives the point estimator $\bar{X}$ together with the indication of its accuracy. However, this cannot always be done so simply using a confidence interval.

Also one-sided confidence intervals correspond to specifying an upper or lower confidence limit only.

If an exam question asks for a 'confidence interval', it means a two-sided symmetrical confidence interval. If the examiners require any other type of confidence interval, they will explicitly ask for it.

2.3   Sample size

A very common question asked of a statistician is:

'How large a sample is needed?'

This question cannot be answered without further information, namely:

(1)   the accuracy of estimation required

(2)   an indication of the size of the population standard deviation σ.

The latter information may not readily be available, in which case a small pilot sample may be needed or a rough guess based on previous studies in similar populations.

As a consequence of the Central Limit Theorem, a confidence interval that is derived from a large sample will tend to be narrower than the corresponding interval derived from a small sample, since the variation in the observed values will tend to 'average out' as the sample size is increased. Market research companies often need to be confident that their results are accurate to within a given margin (*eg* $\pm3\%$). In order to do this, they will need to estimate how big a sample is required in order to obtain a narrow enough confidence interval.

Example

A company wishes to estimate the mean claim amount for claims under a certain class of policy during the past year. Extensive past records from previous years suggest that the standard deviation of claim amounts is likely to be about £45.

If the company wishes to estimate the mean claim amount such that a 95% confidence interval is of width '$\pm£5$', determine the sample size needed to achieve this accuracy of estimation.

Solution

The resulting confidence interval will be $\bar{x} \pm 1.96 \dfrac{\sigma}{\sqrt{n}}$.

The standard deviation σ can be taken to be 45 and so we require n such that:

$$1.96 \times \frac{45}{\sqrt{n}} = 5 \;\Rightarrow\; \sqrt{n} = 1.96 \times \frac{45}{5} = 17.64 \;\Rightarrow\; n = 311.2$$

So a sample of size 312, or perhaps 320 to err on the safe side (since the variance is only a rough guess) would be required.

Question

Calculate how big a sample would be needed to have a 99% confidence interval of width $\pm£1$.

Solution

The answer can be calculated from the equation:

$$2.5758 \times \frac{45}{\sqrt{n}} = 1 \;\Rightarrow\; n = 13,436$$

The figure of 2.5758 can be found on page 162 of the *Tables*.

In this case we need a substantially bigger sample size.

3   Confidence intervals for the normal distribution

3.1   The mean

The previous section dealt with confidence intervals for a normal mean μ in the case where the standard deviation σ was known. In practice this is unlikely to be the case and so we need a different pivotal quantity for the realistic case when σ is unknown.

Fortunately there is a similar pivotal quantity readily available and that is the *t* result:

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$$

where S is the sample standard deviation.

The resulting confidence interval, in the form of symmetrical 95% confidence limits, is:

$$\bar{X} \pm t_{0.025, n-1} \frac{S}{\sqrt{n}}$$

$t_{0.025, n-1}$ is used to denote the upper 2.5% point of the t distribution with $n-1$ degrees of freedom, and is defined by:

$$P\left(t_{n-1} > t_{0.025, n-1}\right) = 0.025$$

For example, from page 163 of the *Tables*, $t_{0.025, 10}$ is equal to 2.228.

This is a small sample confidence interval for μ. For large samples t_{n-1} becomes like $N(0,1)$ and the Central Limit Theorem justifies the resulting interval without the requirement that the population is normal.

The normality of the population is an important assumption for the validity of the t interval especially when the sample size is very small, for example, in single figures. However the t interval is quite robust against departures from normality especially as the sample size increases. Normality can be checked by inspecting a diagram, such as a dotplot, of the data. This can also be used to identify substantial skewness or outliers which may invalidate the analysis.

Question

Calculate a 95% confidence interval for the average height of 10-year-old children, assuming that heights have a $N(\mu, \sigma^2)$ distribution (where μ and σ are unknown), based on a random sample of 5 children whose heights are: 124cm, 122cm, 130cm, 125cm and 132cm.

Solution

Since the sample comes from a normal distribution, we know that:

$\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ has a t_{n-1} distribution, where S^2 is the sample variance

From the *Tables*, we find that $t_{0.025,4} = 2.776$, *ie* $0.95 = P(-2.776 < t_4 < 2.776)$.

So:

$$0.95 = P\left(-2.776 < \dfrac{\overline{X} - \mu}{S/\sqrt{n}} < 2.776\right)$$

Rearranging the inequality to isolate μ gives:

$$0.95 = P\left(\overline{X} - 2.776\, S/\sqrt{n} < \mu < \overline{X} + 2.776\, S/\sqrt{n}\right)$$

Using the calculated values for the sample ($n = 5$, $\overline{x} = 126.6$, and $s^2 = 17.8$) gives:

$$(121.4,\ 131.8)$$

When calculating a numerical confidence interval, we must drop the probability notation. This is required since mu is not a random variable and hence expressions such as $P\left(121.4 < \mu < 131.8\right) = 0.95$) do not make sense.

The R function for a symmetrical 95% confidence interval for the mean with unknown variance is:

```
t.test(<sample data>, conf=0.95)
```

For small samples from a non-normal distribution, confidence intervals can be constructed empirically in R using the bootstrap method used in Chapter 7 Section 7. For example, a non-parametric 95% confidence interval for the mean could be obtained by:

```
quantile(replicate(1000,mean(sample(<sample data>,replace=TRUE))),
probs=c(0.025,0.975))
```

3.2 The variance

For the estimation of a normal variance σ^2, there is again a pivotal quantity readily available:

$$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

The resulting 95% confidence interval for the variance σ^2 is:

$$\left(\frac{(n-1)S^2}{\chi^2_{0.025,n-1}}, \frac{(n-1)S^2}{\chi^2_{0.975,n-1}} \right)$$

or for the standard deviation σ :

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{0.025,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{0.975,n-1}}} \right)$$

Note: Due to the skewness of the χ^2 distribution, these confidence intervals are not symmetrical about the point estimator S^2, and are also not the shortest-length intervals.

So we can't write these using the '$\pm$' notation.

The above intervals require the normality assumption for the population but are considered fairly robust against departures from normality for reasonable sample sizes.

There is no built-in function for calculating confidence intervals for the variance in R. We can use R to calculate the results of the formula from scratch or use a bootstrap method if the assumptions are not met.

Question

Calculate:

(i)   an equal-tailed 95% confidence interval and

(ii)   a 95% confidence interval of the form $(0, L)$

for the standard deviation of the heights of the children in the population based on the information given in the last question.

Solution

Since the sample comes from a normal distribution, we know that the quantity $\dfrac{4S^2}{\sigma^2}$ has a χ^2_4 distribution.

(i)   From the *Tables*, we find that:

$$0.95 = P(0.4844 < \chi^2_4 < 11.14)$$

So:

$$0.95 = P\left(0.4844 < \frac{4S^2}{\sigma^2} < 11.14 \right)$$

Replacing S^2 by 17.8, the sample variance calculated in the solution to the previous question, and dropping the probability notation (since σ^2 is not a random variable), we have:

$$0.4844 < \frac{4 \times 17.8}{\sigma^2} < 11.14 \;\;\Rightarrow\;\; \frac{71.2}{11.14} < \sigma^2 < \frac{71.2}{0.4844}$$

$$\Rightarrow\;\; 6.39 < \sigma^2 < 147.0 \;\;\Rightarrow\;\; 2.53 < \sigma < 12.1$$

So, an equal-tailed 95% confidence interval for the standard deviation is (2.53, 12.1).

(ii)   From the *Tables* we find that:

$$0.95 = P(0.7107 < \chi_4^2)$$

So:

$$0.95 = P\left(0.7107 < \frac{4S^2}{\sigma^2} \right)$$

Replacing S^2 by 17.8, the sample variance calculated in the solution to the previous question, and dropping the probability notation (since σ^2 is not a random variable), we have:

$$\sigma^2 < \frac{4 \times 17.8}{0.7107} \;\;\Rightarrow\;\; \sigma^2 < 100.2 \;\;\Rightarrow\;\; \sigma < 10.0$$

So, a one-sided 95% confidence interval for the standard deviation is $(0, 10.0)$.

To find a confidence interval with an upper limit, *ie* of the form $(0, L)$, we need to start with the lower 5% point of the χ_4^2 distribution, *ie* the point which is exceeded by 95% of the distribution.

If we had wanted to find a confidence interval with a lower limit, *ie* of the form (L, ∞), we would need to start by finding the upper 5% point of the χ_4^2 distribution, *ie* the point which is exceeded by only 5% of the distribution.

4   Confidence intervals for binomial & Poisson parameters

Both these situations involve a discrete distribution which introduces the difficulty of probabilities not being exactly 0.95, and so 'at least 0.95' is used instead. Also when not using the large-sample normal approximations, the pivotal quantity method must be adjusted.

One approach is to use a quantity $h(\underline{X})$ whose distribution involves θ such that:

$$P\big(h_1(\theta) < h(\underline{X}) < h_2(\theta)\big) \geq 0.95$$

Then if both $h_1(\theta)$ and $h_2(\theta)$ are monotonic increasing (or both decreasing), the inequalities can be inverted to obtain a confidence interval as before.

4.1   The binomial distribution

If X is a single observation from $Bin(n,\theta)$, the maximum likelihood estimator is:
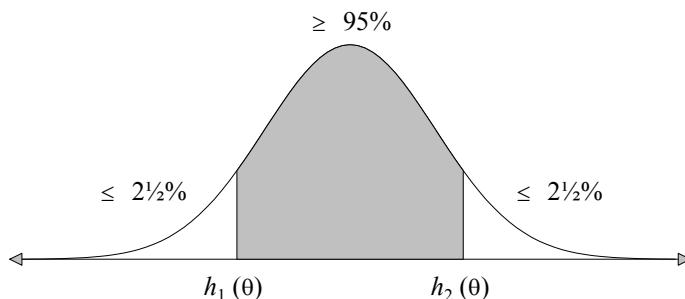
$$\hat{\theta} = \frac{X}{n}$$

What follows is a slight diversion from our aim of obtaining a confidence interval for θ. It is just demonstrating that the method is sound.

Using X as the quantity $h(\underline{X})$, it is necessary to find if $h_1(\theta)$ and $h_2(\theta)$ exist such that $P\big(h_1(\theta) < X < h_2(\theta)\big) \geq 0.95$, where with equal tails $P\big(X \leq h_1(\theta)\big) \leq 0.025$ and $P\big(X \geq h_2(\theta)\big) \leq 0.025$.

We can have at most 2.5% in the lower (or upper) tail, so we need to be very careful about finding the values of h_1 and h_2.

There is no explicit expression for the pivotal quantity $h(\underline{X})$.

For the $Bin(20,0.3)$ case:

$$P\big(X \leq 1\big) = 0.0076 \text{ and } P\big(X \leq 2\big) = 0.0355 \quad \therefore h_1(\theta) = 1$$

Also:

$$P(X \geq 11) = 0.0171, \; P(X \geq 10) = 0.0480 \quad \therefore h_2(\theta) = 11$$

Question

For the binomial distribution with parameters $n = 20$, and $\theta = 0.4$, calculate the values of h_1 and h_2.

Solution

If X is $Bin(20, 0.4)$, then using page 188 of the *Tables*, $P(X \leq 3) = 0.0160$ and $P(X \leq 4) = 0.0510$, so $h_1 = 3$.

Also $P(X \geq 13) = 0.0210$ and $P(X \geq 12) = 0.0565$, so $h_2 = 13$.

h_1 and h_2 have higher values than for the $Bin(20, 0.3)$ case.

So $h_1(\theta)$ and $h_2(\theta)$ do exist and increase with θ.

We're back on track. We can move on to obtain our confidence interval for θ.

Therefore the inequalities can be inverted as follows:

$$X \leq h_1(\theta) \Rightarrow \theta \geq \theta_1(X)$$

$$X \geq h_2(\theta) \Rightarrow \theta \leq \theta_2(X)$$

These are the *tail* probabilities. So the inequalities involving θ_1 and θ_2 are defining the tails. Our confidence interval is the region *not* covered by these tail inequalities:

This gives a 95% confidence interval of the form $\theta_2(X) < \theta < \theta_1(X)$.

Note: The lower limit $\theta_2(X)$ comes from the upper tail probabilities and the upper limit $\theta_1(X)$ from the lower tail probabilities.

We'll see this is the case in the question on the next page.

However since there are no explicit expressions for $h_1(\theta)$ and $h_2(\theta)$, there are no expressions for $\theta_1(X)$ and $\theta_2(X)$ and they will have to be calculated numerically.

So, adopting the convention of including the observed x in the tails, θ_1 and θ_2 can be found by solving:

$$\sum_{r=x}^{n} b(r; n, \theta_1) = 0.025 \quad \text{and} \quad \sum_{r=0}^{x} b(r; n, \theta_2) = 0.025$$

Here $b(r; n, \theta)$ denotes $P(X = r)$ when $X \sim Bin(n, \theta)$.

These can be expressed in terms of the distribution function $F(x; \theta)$:

$$1 - F(x - 1; \theta_1) = 0.025 \text{ and } F(x; \theta_2) = 0.025$$

Note: Equality can be attained as θ has a continuous range $(0, 1)$ and the 'discrete' problem does not arise.

The R function for an exact 95% confidence interval for the proportion is:

```
binom.test(x,n, conf=0.95)
```

Question

We have obtained a value of 1 from the binomial distribution with parameters $n = 20$ and θ. Construct a 95% symmetrical confidence interval for θ.

Solution

We need θ_1 such that $P(X \leq 1) = 0.025$ under $Bin(20, \theta_1)$, and θ_2 such that $P(X \geq 1) = 0.025$ under $Bin(20, \theta_2)$.

For the first equation, we have $(1 - \theta_1)^{20} + 20(1 - \theta_1)^{19} \theta_1 = 0.025$.

Solving this we obtain $\theta_1 = 0.249$.

A numerical method will be needed here, or trial and improvement. One approach would be to write the equation in the form $(1 - \theta_1)^{19}(1 + 19\theta_1) = 0.025$, then iterate using

$$\theta_{n+1} = 1 - \left(\frac{0.025}{1 + 19\theta_n} \right)^{\frac{1}{19}} \text{ starting with } \theta_1 = 0.5.$$

For the second equation we have $(1 - \theta_2)^{20} = 0.975$.

Solving this we obtain $\theta_2 = 0.00127$.

Our confidence interval is then $(0.00127, 0.249)$.

The normal approximation

It is no bother for a computer to calculate an exact confidence interval for the binomial parameter p even if n is 'large'. However, on a piece of paper we use the normal approximation to the binomial distribution.

$$\frac{X - n\theta}{\sqrt{n\theta(1 - \theta)}} \text{ can be used as a pivotal quantity.}$$

Solving the resulting equations for θ would not be easy.

However $\dfrac{X - n\theta}{\sqrt{n\hat{\theta}\left(1-\hat{\theta}\right)}}$, with $\hat{\theta}$ in place of θ (in the denominator only), can be used in a simpler

way and yields the standard 95% confidence interval used in practice, namely:

$$\frac{X \pm 1.96\sqrt{n\hat{\theta}(1-\hat{\theta})}}{n}$$

or $\hat{\theta} \pm 1.96\sqrt{\dfrac{\hat{\theta}(1-\hat{\theta})}{n}}$, where $\hat{\theta} = \dfrac{X}{n}$.

Question

In a one-year mortality investigation, 45 of the 250 ninety-year-olds present at the start of the investigation died before the end of the year. Assuming that the number of deaths has a binomial distribution with parameters $n = 250$ and q, calculate a symmetrical 90% confidence interval for the unknown mortality rate q.

Solution

Since 250 is a large sample, we know that $\dfrac{X - nq}{\sqrt{nq(1-q)}} \sim N(0,1)$ approximately.

Since $P(-1.6449 < Z < 1.6449) = 0.90$, we can say that:

$$P\left(-1.6449 < \frac{X - 250q}{\sqrt{250q(1-q)}} < 1.6449 \right) = 0.90$$

Rearranging this:

$$P\left(\frac{X}{250} - 1.6449\sqrt{\frac{q(1-q)}{250}} < q < \frac{X}{250} + 1.6449\sqrt{\frac{q(1-q)}{250}} \right) = 0.90$$

Replacing X by the observed value of 45 gives $q = \dfrac{45}{250}$.

Therefore a symmetrical 90% confidence interval for q is $(0.140, 0.220)$.

Question

Repeat the question on page 15 using the normal approximation. Comment on the answer obtained.

Solution

A 95% symmetrical confidence interval is given by:

$$\frac{X}{n} \pm 1.96 \sqrt{\frac{\frac{X}{n}\left(1 - \frac{X}{n}\right)}{n}}$$

From the question, we know that $x = 1$ and $n = 20$. Substituting these into the formula, we get the confidence interval to be $(-0.046, 0.146)$.

Since the value of n is so small, the normal approximation is not really appropriate. This is highlighted by the lower limit which is not sensible, as p must be between 0 and 1. The upper limit is not even close to the accurate value either.

The reason why the accuracy is so poor in this case is that the distribution is skew. Since we got 1 out of 20, the value of p can be estimated as 0.05. So the value of $np \approx 20 \times 0.05 = 1$ is nowhere near big enough to justify a normal approximation, where we usually require $np \geq 5$.

4.2 The Poisson distribution

The Poisson situation can be tackled in a very similar way to the binomial for both large and small sample sizes.

If $X_i, i = 1, 2, ..., n$ are independent $Poi(\lambda)$ variables, that is, a random sample of size n from $Poi(\lambda)$, then $\sum X_i \sim Poi(n\lambda)$.

Using $\sum X_i$ as a single observation from $Poi(n\lambda)$ is equivalent to the random sample of size n from $Poi(\lambda)$. This is similar to the single binomial situation.

Recall that a $Bin(n, p)$ distribution arises from the sum of n Bernoulli trials with probability of success p.

Given a single observation X from a $Poi(\lambda)$ distribution, then $P\big(h_1(\lambda) < X < h_2(\lambda)\big) \geq 0.95$, where $h_1(\lambda)$ and $h_2(\lambda)$ are increasing functions of λ.

Inverting this gives $P(\lambda_1(X) < \lambda < \lambda_2(X)) = 0.95$.

The resulting 95% confidence interval for λ is given by (λ_1, λ_2) where:

$$\sum_{r=x}^{\infty} p(r; \lambda_1) = 0.025 \quad \text{and} \quad \sum_{r=0}^{x} p(r; \lambda_2) = 0.025$$

Here $p(r : \lambda_1)$ denotes $P(X = r)$ where $X \sim Poi(\lambda_1)$.

or:

$$1 - F(x - 1; \lambda_1) = 0.025 \quad \text{and} \quad F(x; \lambda_2) = 0.025$$

The R function for an exact 95% confidence interval for the Poisson parameter λ is:

```
poisson.test(x,n, conf=0.95)
```

Question

We have obtained a value of 1 from $Poi(\lambda)$. Calculate a symmetrical 90% confidence interval for λ.

Solution

We need $P(X \geq 1) = 0.05$ under $Poi(\lambda_1)$, and $P(X \leq 1) = 0.05$ under $Poi(\lambda_2)$.

The first equation is $1 - e^{-\lambda_1} = 0.05 \implies e^{-\lambda_1} = 0.95$, which gives $\lambda_1 = 0.0513$.

The second equation is $e^{-\lambda_2} + \lambda_2 e^{-\lambda_2} = 0.05$. Solving this numerically, for example by using the iterative equation $\lambda = \log\left(\dfrac{1 + \lambda}{0.05}\right)$, we obtain $\lambda_2 = 4.74$.

Therefore a symmetrical 90% confidence interval for λ is $(0.0513, 4.74)$. Not surprisingly this is very wide, since we only have 1 sample value.

The normal approximation

Again, it is easy for a computer to calculate an exact confidence interval for λ even for a large sample from Poisson(λ), or a single observation from Poisson(λ) where λ is large. However, on a piece of paper a normal approximation can be used either from

$$\sum X_i \sim \text{Poi}(n\lambda) \rightarrow N(n\lambda, n\lambda) \text{ or from the Central Limit Theorem as } \bar{X} \rightarrow N\left(\lambda, \frac{\lambda}{n}\right).$$

$\dfrac{\overline{X} - \lambda}{\sqrt{\lambda / n}}$ **can then be used as a pivotal quantity yielding a confidence interval. However, as in the binomial case, the standard confidence interval in practical use comes from** $\dfrac{\overline{X} - \lambda}{\sqrt{\hat{\lambda} / n}}$ **where** $\hat{\lambda} = \overline{X}$.

This clearly gives $\overline{X} \pm 1.96 \sqrt{\dfrac{\overline{X}}{n}}$ **as an approximate 95% confidence interval for** λ .

Question

In a one-year investigation of claim frequencies for a particular category of motorists, the total number of claims made under 5,000 policies was 800. Assuming that the number of claims made by individual motorists has a $Poi(\lambda)$ distribution, calculate a symmetrical 90% confidence interval for the unknown average claim frequency λ .

Solution

Since the sample comes from a Poisson distribution, we know that $\dfrac{\overline{X} - \lambda}{\sqrt{\lambda / n}} \sim N(0,1)$ (approximately). Here $n = 5,000$.

From the *Tables* we find that $P(-1.6449 < Z < 1.6449) = 0.90$.

So:

$$P(-1.6449 < \frac{\overline{X} - \lambda}{\sqrt{\lambda / n}} < 1.6449) = 0.90$$

which we can rearrange to give:

$$P\left(\overline{X} - 1.6449 \sqrt{\frac{\hat{\lambda}}{n}} < \lambda < \overline{X} + 1.6449 \sqrt{\frac{\hat{\lambda}}{n}} \right) = 0.90$$

Replacing n by 5,000, $\overline{X}$ by 0.16 and $\hat{\lambda}$ by 0.16, the confidence interval is $(0.151, 0.169)$.

5   Confidence intervals for two-sample problems

A comparison of the parameters of two populations can be considered by taking independent random samples from each population.

The importance of the independence is illustrated by noting that:

$$\text{var}\left[\bar{X}_1 - \bar{X}_2\right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

when the samples are independent.

If the samples are not independent, then a covariance term will be included:

$$\text{var}\left[\bar{X}_1 - \bar{X}_2\right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\text{cov}\left[\bar{X}_1, \bar{X}_2\right]$$

This covariance term can clearly have a substantial effect in the non-independent case.

The most common form of non-independence is due to paired data.

5.1   Two normal means

Case 1 (known population variance)

If $\bar{X}_1$ and $\bar{X}_2$ are the means from independent random samples of size n_1 and n_2 respectively taken from normal populations which have known variances σ_1^2 and σ_2^2 respectively, then the equal-tailed $100(1-\alpha)\%$ confidence interval for the difference in the population means is given by:

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

So for example, when $\alpha = 5\%$, we have $z_{\alpha/2} = z_{2.5\%} = 1.9600$.

> **There is no built-in function for calculating the above confidence interval in R. We can use R to calculate the results of the formula from scratch or use a bootstrap method if the assumptions are not met.**

Case 2 (unknown population variance)

If $\bar{X}_1, \bar{X}_2, S_1$ and S_2, are the means and standard deviations from independent random samples of size n_1 and n_2 respectively taken from normal populations which have equal variances, then the equal-tailed $100(1-\alpha)\%$ confidence interval for the difference in the population means is given by:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

This formula is given on page 23 of the *Tables*.

In any practical situation consideration must be made as to whether n_1 and n_2 are large or small and whether σ_1^2 and σ_2^2 are known or unknown. In the case of the t result it should be noted that there is the additional assumption of equality of variances and this should be checked by plotting the data in a suitable way and/or using the formal approach in Section 5.2.

Note: The pooled estimator S_p^2 is based on the maximum likelihood estimator but adjusted to give an unbiased estimator.

Remember that the number of degrees of freedom for the t distribution is the same as the number used in the denominator of the pooled sample variance formula. S_1^2 and S_2^2 are the sample variances calculated in the usual way.

Question

A motor company runs tests to investigate the fuel consumption of cars using a newly developed fuel additive. Sixteen cars of the same make and age are used, eight with the new additive and eight as controls. The results, in miles per gallon over a test track under regulated conditions, are as follows:

Control	27.0	32.2	30.4	28.0	26.5	25.5	29.6	27.2
Additive	31.4	29.9	33.2	34.4	32.0	28.7	26.1	30.3

Calculate a 95% confidence interval for the increase in miles per gallon achieved by cars with the additive. State clearly any assumptions required for this analysis.

Solution

Assuming that the samples come from normal distributions with the same variance and that the samples are independent, we know that $\dfrac{(\bar{A} - \bar{C}) - (\mu_A - \mu_C)}{S_p\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_C}}} \sim t_{n_A + n_C - 2}$, where $\bar{A}$ and $\bar{C}$ are the

sample means, μ_A and μ_C are the underlying population means, n_A and n_C are the sample sizes and s_p is the pooled sample standard deviation.

We now replace $\bar{A}$ by 30.75, $\bar{C}$ by 28.3, S_p^2 by 5.96, and n_A and n_C by 8 to obtain the

confidence interval. (The individual sample variances are $s_A^2 = \dfrac{48.06}{7}$ and $s_C^2 = \dfrac{35.38}{7}$.)

For a symmetric confidence interval, we need the upper 2.5% point of t_{14}, which is 2.145.

Substituting these values in we get the symmetrical 95% confidence interval to be:

$$2.45 \pm 2.145 \sqrt{\frac{5.96}{4}} = (-0.168, 5.068)$$

In R we can use the function `t.test` **with the argument** `var.equal = TRUE` **to obtain a confidence interval for the difference between the means with unknown but equal variances.**

The `t.test` **function can also obtain confidence intervals for the difference between the means with unknown but non-equal variances.**

Again, we could use the bootstrap method to construct empirical confidence intervals if the assumptions of the above formulae are not met.

5.2   Two population variances

For the comparison of two population variances it is more natural to consider the ratio σ_1^2 / σ_2^2 than the difference $\sigma_1^2 - \sigma_2^2$. This follows logically from the concept of variance, but also from a technical point of view there is a pivotal quantity readily available for the ratio of normal variances but not for their difference.

It is $\dfrac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}$.

The resulting confidence interval is given by:

$$\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{n_1-1, n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \cdot F_{n_2-1, n_1-1}$$

where F_{n_1-1, n_2-1} is the relevant percentage point from the *F* distribution. Notice that the order of the degrees of freedom is different in the two *F* distributions here.

It should be said that in practice the estimation of σ_1^2 / σ_2^2 is not a common objective. However the same *F* result is used for the more common objective of 'testing' whether σ_1^2 and σ_2^2 may be equal, which is relevant for the *t* result for comparing population means. The acceptability of the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ can be determined simply by confirming that the value 1 is included in the confidence interval for σ_1^2 / σ_2^2.

What this is saying is that if the number 1 lies in the confidence interval, then 1 is one of the many reasonable values that the variance ratio can take. So we are not unhappy about the assumption that $\sigma_1^2 / \sigma_2^2 = 1$, *ie* $\sigma_1^2 = \sigma_2^2$. The alternative way of checking equality is to use the hypothesis test detailed in Chapter 9.

Question

For the fuel additive data in the previous question, calculate a 90% confidence interval for the

ratio $\dfrac{\sigma_C^2}{\sigma_A^2}$ of the variances of the fuel consumption distributions both without and with the

additive, and comment on the equality of variances assumption needed for the analysis in that question.

Solution

For two independent random samples from $N(\mu_A, \sigma_A^2)$ and $N(\mu_C, \sigma_C^2)$, $\dfrac{S_A^2/\sigma_A^2}{S_C^2/\sigma_C^2} \sim F_{n_A-1, n_C-1}$,

where n_A and n_C are the sample sizes, and S_A^2 and S_C^2 are the sample variances.

From the previous question, $s_A^2 = 6.8657$ and $s_C^2 = 5.0543$.

From the *Tables*, we know that $0.90 = P\left(\dfrac{1}{3.787} < F_{7,7} < 3.787\right)$, which gives us:

$$0.90 = P\left(\dfrac{1}{3.787} < \dfrac{S_A^2/\sigma_A^2}{S_C^2/\sigma_C^2} < 3.787\right)$$

Rearranging this to give $\dfrac{\sigma_C^2}{\sigma_A^2}$ (and dropping the probability notation since $\dfrac{\sigma_C^2}{\sigma_A^2}$ is not a random

variable), we get $0.1944 < \dfrac{\sigma_C^2}{\sigma_A^2} < 2.788$.

So the confidence interval is therefore $(0.1944, 2.788)$.

Since the value of 1 lies well within this interval, the assumption of equality of variances needed in the previous question appears to be justified.

The R code for this confidence interval is `var.test` **or we could use a bootstrap method if the assumptions are not met.**

5.3 Two population proportions

The comparison of population proportions corresponds to comparing two binomial probabilities on the basis of single observations X_1, X_2 from $Bin(n_1, \theta_1)$ and $Bin(n_2, \theta_2)$ respectively.

Considering only the case where n_1 and n_2 are large, so that the normal approximation can be used, the pivotal quantity used in practice is:

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\dfrac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \dfrac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}} \doteq N(0,1)$$

where $\hat{\theta}_1, \hat{\theta}_2$ are the MLEs $\dfrac{X_1}{n_1}, \dfrac{X_2}{n_2}$, respectively.

The R code for this confidence interval is `prop.test` with the argument `correct=FALSE`.

Question

In a one-year mortality investigation, 25 of the 100 ninety-year-old males and 20 of the 150 ninety-year-old females present at the start of the investigation died before the end of the year. Assuming that the numbers of deaths follow independent binomial distributions, calculate a symmetrical 95% confidence interval for the difference between male and female mortality rates at this age.

Solution

Since the samples come from independent binomial distributions, we know that, approximately:

$$\frac{\left(\dfrac{X_1}{n_1} - \dfrac{X_2}{n_2}\right) - (p_1 - p_2)}{\sqrt{\dfrac{\dfrac{X_1}{n_1}\left(1 - \dfrac{X_1}{n_1}\right)}{n_1} + \dfrac{\dfrac{X_2}{n_2}\left(1 - \dfrac{X_2}{n_2}\right)}{n_2}}} \sim N(0,1)$$

Calling $\dfrac{X_1}{n_1} = \hat{p}_1$, and $\dfrac{X_2}{n_2} = \hat{p}_2$, and using the *Tables*, we know that:

$$0.95 = P\left(-1.96 < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}} < 1.96\right)$$

Replacing $\hat{p}_1$ by 0.25 and $\hat{p}_2$ by 0.133, the inequality becomes:

$$0.016 < p_1 - p_2 < 0.218$$

So a symmetrical 95% confidence interval for the difference in mortality rates is $(0.016, 0.218)$.

5.4   Two Poisson parameters

Considering the comparison of two Poisson parameters (λ_1 and λ_2) when the normal approximation can be used:

$$\bar{X}_i \text{ is an estimator of } \lambda_i \text{ such that } \bar{X}_i \rightarrow N\left(\lambda_i, \frac{\hat{\lambda}_i}{n_i}\right)$$

Therefore $\bar{X}_1 - \bar{X}_2$ is an estimator of $\lambda_1 - \lambda_2$ such that:

$$\bar{X}_1 - \bar{X}_2 \rightarrow N\left(\lambda_1 - \lambda_2, \frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)$$

Using $\hat{\lambda}_i = \bar{X}_i$, an approximate 95% confidence interval for $\lambda_1 - \lambda_2$ is given by:

$$\bar{X}_1 - \bar{X}_2 \pm 1.96\sqrt{\left(\frac{\bar{X}_1}{n_1} + \frac{\bar{X}_2}{n_2}\right)}$$

We are assuming that the two samples are independent.

There is no built in function for calculating the above confidence interval in R. We can use R to calculate the results of the formula from scratch. However, the function `poisson.test` can be used to obtain a confidence interval for the ratio of the two Poisson parameters.

Question

In a one-year investigation of claim frequencies for a particular category of motorists, there were 150 claims from the 500 policyholders aged under 25 and 650 claims from the 4,500 remaining policyholders. Assuming that the numbers of claims made by the individual motorists in each category have independent Poisson distributions, calculate a 99% confidence interval for the difference between the two Poisson parameters.

Solution

Since the samples come from independent Poisson distributions, we know that

$\dfrac{(\bar{X}_1 - \bar{X}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\dfrac{\bar{X}_1}{n_1} + \dfrac{\bar{X}_2}{n_2}}} \sim N(0,1)$, where subscripts 1 and 2 refer to young and old drivers

respectively.

From the *Tables*, we know that $0.99 = P(-2.5758 < Z < 2.5758)$. This gives us:

$$0.99 = P\left(-2.5758 < \frac{\left(\bar{X}_1 - \bar{X}_2 \right) - (\lambda_1 - \lambda_2)}{\sqrt{\dfrac{\bar{X}_1}{n_1} + \dfrac{\bar{X}_2}{n_2}}} < 2.5758 \right)$$

Replacing $\bar{X}_1$ by 0.3, $\bar{X}_2$ by 0.1444, n_1 by 500 and n_2 by 4,500 and rearranging, the inequality becomes:

$$0.0908 < \lambda_1 - \lambda_2 < 0.2203$$

So the confidence interval for $\lambda_1 - \lambda_2$ is $(0.0908, 0.2203)$.

6   Paired data

Paired data is a common example of comparison using non-independent samples.

Essentially having paired or matched data means that there is one sample:

$$(X_{11}, X_{21}), (X_{12}, X_{22}), (X_{13}, X_{23}), ..., (X_{1n}, X_{2n})$$

rather than two separate samples:

$$(X_{11}, X_{12}, X_{13}, ..., X_{1n}) \text{ and } (X_{21}, X_{22}, X_{23}, ..., X_{2n})$$

The paired situation is really a single sample problem, that is, a problem based on a sample of n pairs of observations. (In the independent two-sample situation the sample sizes need not, of course, be equal.)

Paired data can arise in the form of 'before and after' comparisons.

We will see one of these in the next question.

Investigations using paired data are usually better than two-sample investigations in the sense that the estimation is more accurate.

When finding confidence intervals, this means the confidence interval derived from the paired data will usually be narrower.

Paired data are analysed using the differences $D_i = X_{1i} - X_{2i}$ and estimation of $\mu_D = \mu_1 - \mu_2$ is considered. A z result or a t result can be used, but the latter will be more common as it is unlikely that the variances of the differences will be known. Assuming normality of the population of such differences (but not necessarily the normality of the X_1 and X_2 populations), the pivotal quantity for the t result is:

$$\frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \sim t_{n-1}$$

Note that S_D is calculated from the values of D.

The resulting 95% confidence interval for μ_D will be $\bar{D} \pm t_{0.025, n-1} \dfrac{S_D}{\sqrt{n}}$.

Question

The average blood pressure $\bar{b}$ for a group of 10 patients was 77.0 mmHg. The average blood pressure $\bar{a}$ after they were put on a special diet was 75.0 mmHg. Assuming that variation in blood pressure follows a normal distribution, calculate a 95% symmetrical confidence interval for the reduction in blood pressure attributable to the special diet. Assess the effectiveness of the diet in reducing the patients' blood pressure. It is known that $\sum (b_i - a_i)^2 = 68$.

Solution

Since this is a paired sample from a normal distribution, we know that $\dfrac{\bar{D} - (\mu_A - \mu_B)}{S_D / \sqrt{n}} \sim t_{n-1}$

where $D = A - B$.

From the *Tables*, we know that $0.95 = (-2.262 < t_9 < 2.262)$, so:

$$0.95 = \left(-2.262 < \frac{\bar{D} - (\mu_A - \mu_B)}{S_D / \sqrt{n}} < 2.262\right)$$

We can now replace n by 10, $\bar{D}$ by -2.0 and S_D by:

$$s_D = \sqrt{\frac{\sum (b_i - a_i)^2 - n(\bar{b} - \bar{a})^2}{n - 1}} = \sqrt{\frac{68 - 10 \times 4}{9}} = 1.764$$

We now obtain:

$$-3.26 < \mu_A - \mu_B < -0.74$$

The required confidence interval is $(-3.26, -0.74)$.

Since this interval does not include the value 0 (which would be the value if there was no difference in the average blood pressure before and after), the diet seems to be effective.

A plot of the sample differences can be used to check on normality but recall that the *t* result is robust as *n* increases. Also the Central Limit Theorem means that it can be safely used for large *n*.

From a practical viewpoint:

(i) **When confronted with 'two-sample' data, consideration should be made of whether the data may in fact be paired. One way is to draw a scatterplot and calculate the correlation coefficient to see whether there is any relationship in the 'pairs' of data points. If there is a strong relationship, the data source should be checked to see if the data were paired by design.**

(ii) **If a paired problem is analysed as though it involved independent samples, then the results would be invalid because the assumption of independence is violated. On the other hand, if independent samples are analysed as though they were paired, then the results would be valid although they would be making inefficient use of the data due to the discarding of possible information about the means and variances of the two separate populations.**

Obviously the ideal approach is to ask the person who collected the data whether any pairing was used.

The R code for this confidence interval is `t.test` **with the argument** `paired=TRUE`.

The fuel consumption data given earlier in the chapter were not paired data. There is no way to link a specific item of data in the control group to the corresponding item of data in the additive group. So we analyse the data using the two-sample t situation.

On the other hand, suppose that we had measured the fuel consumption of 8 cars without a fuel additive, and then re-measured the fuel consumption *of the same 8 cars* with the fuel additive. This is now a paired situation. A data item from the first sample is linked to a specific item in the second sample. In this situation we would treat the data as being paired, and would subtract the figure for control consumption for each car from the figure for the same car when using the additive.

The chapter summary starts on the next page so that you can keep
all the chapter summaries together for revision purposes.

Chapter 8 Summary

A confidence interval gives us a range of valuse in which we believe the true parameter value lies, together with an associated probability. There are a number of different situations for which we can find confidence intervals.

For a single sample from a normal distribution:

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \sigma^2 \text{ known} \qquad\qquad \frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1} \quad \sigma^2 \text{ unknown}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

For samples from two independent normal distributions:

$$\frac{\left(\overline{X}_1-\overline{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1) \qquad \frac{\left(\overline{X}_1-\overline{X}_2\right)-\left(\mu_1-\mu_2\right)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

$$\sigma^2 \text{ known} \qquad\qquad\qquad\qquad \sigma^2 \text{ unknown}$$

where:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

To compare the variances of two independent normal populations:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

For a sample from a binomial distribution:

$$\frac{\hat{p}-p}{\sqrt{\hat{p}\hat{q}/n}} \sim N(0,1) \quad \text{or} \quad \frac{X-np}{\sqrt{n\hat{p}\hat{q}}} \sim N(0,1) \quad \text{(approximately)}$$

For samples from two independent binomial distributions:

$$\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}} \sim N(0,1) \quad \text{(approximately) where} \quad \hat{p}_1 = \frac{X_1}{n_1}, \hat{p}_2 = \frac{X_2}{n_2}$$

Chapter 8 Summary (continued)

For a sample from a Poisson distribution:

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/n}} \sim N(0,1) \quad \text{or} \quad \frac{\sum X - n\lambda}{\sqrt{n\hat{\lambda}}} \sim N(0,1) \quad \text{(approximately)}$$

For samples from two independent Poisson distributions:

$$\frac{(\hat{\lambda}_1 - \hat{\lambda}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\dfrac{\hat{\lambda}_1}{n_1} + \dfrac{\hat{\lambda}_2}{n_2}}} \sim N(0,1) \quad \text{(approximately) where} \quad \hat{\lambda}_1 = \bar{X}_1, \hat{\lambda}_2 = \bar{X}_2$$

General confidence intervals for parameters can be found, using the pivotal method, and the formulae given above.

The confidence interval for two normal means (unknown variances) requires that the variances are the same.

For paired data we subtract the paired values to come up with a new variable *D* and then follow one of the other standard confidence interval calculations.

Chapter 8 Practice Questions

8.1   An experiment was carried out to find out the number of hours that actuarial students spend watching television per week. It was discovered that for a sample of 10 students, the following times were spent watching television:

        8, 4, 7, 5, 9, 7, 6, 9, 5, 7

        Calculate a symmetrical 95% confidence interval for the mean time an actuarial student spends watching television per week. Write down the assumptions needed to find this confidence interval.

8.2   A researcher investigating attitudes to Sunday shopping reports that, in a sample of 8 interviewees, 7 were in favour of more opportunities to shop on Sunday. Use the binomial distribution to calculate an *exact* 95% confidence interval for the underlying proportion in favour of this idea.

8.3   An opinion poll of 1,000 voters found that 450 favoured Party P. Calculate an approximate 99% confidence interval for the proportion of voters who favour Party P. Comment on the likelihood of more than 50% of the voters voting for Party P in an election.

8.4   Two inspectors carry out property valuations for an estate agency. Over a particular week they each go out to similar properties. The table below shows their valuations (in £000s):

Exam style

A	102	98	93	86	92	94	89	97	
B	86	88	92	95	98	97	94	92	91

(i)   (a)   Make an informative plot of these figures.

        (b)   Use the diagrams from part (i)(a) to comment on the possible assumption of equal variance for the two underlying populations.

        (c)   With the equal variance assumption of (b), calculate a 95% confidence interval for this common variance.

        (d)   Calculate a 95% confidence interval for the mean difference between the valuations by A and B, and comment briefly on the result.   [11]

(ii)   The estate agency employing the inspectors decides to test their valuations by sending them each to the same set of eight houses, independently and without knowledge that the other is going. The resulting valuations (in £000s) follow:

Property								
	1	2	3	4	5	6	7	8
A	94	98	102	132	118	121	106	123
B	92	96	111	129	111	122	101	118

(a)   Make an informative plot of these figures.

(b)   Calculate a 90% confidence interval for the mean difference between valuations by A and B, and comment briefly on the result.   [6]

[Total 17]

8.5   The ordered remission times (in weeks) of 20 leukaemia patients are given in the table:

1	1	2	2	3
4	4	5	5	8
8	8	11	11	12
12	15	17	22	23

Suppose the remission times can be regarded as a random sample from an exponential distribution with density:

$$f(X;\lambda) = \lambda e^{-\lambda x}, x > 0$$

(i)   (a)   Determine the maximum likelihood estimator $\hat{\lambda}$ of λ.

(b)   Calculate the large-sample approximate variance of $\hat{\lambda}$.

(c)   Hence calculate an approximate 95% confidence interval for λ.   [7]

(ii)   Using the fact that $2\lambda n\bar{X}$ has a χ^2_{2n} distribution, calculate an exact 95% confidence interval for λ, and comment briefly on how it compares with your interval in (i)(c).   [3]

[Total 10]

8.6   Heights of males with classic congenital adrenal hyperplasia (CAH) are assumed to be normally distributed.

        Determine the minimum sample size to ensure that a 95% confidence interval for the mean height has a maximum width of 10cm, if:

        (i)   a previous sample had a standard deviation of 8.4 cm

        (ii)   the population standard deviation *is* 8.4 cm.

8.7   (i)   A sample value of 2 is obtained from a Poisson distribution with mean μ. Calculate an *exact* two-sided 90% confidence interval for μ.

        (ii)   A sample of 30 values from the same Poisson distribution has a mean of 2. Use these data values to construct an approximate 90% confidence interval for μ.

8.8   An office manager wants to analyse the variability in the time taken for her typists to complete a given task. She has given seven typists the task and the results are as follows (in minutes):

                15, 17.2, 13.7, 11.2, 18, 15.1, 14

        The manager wants a 95% confidence interval for the true standard deviation of time taken of the form (k, ∞). Calculate the value of k.

8.9   The amounts of individual claims arising under a certain type of general insurance policy are
Exam style   known from past experience to conform to a lognormal distribution in which the standard deviation equals 1.8 times the mean. An actuary has found that the lower and upper limits of a 95% confidence interval for the mean claim amount are £4,250 and £4,750. Evaluate the lower and upper limits of a 95% confidence interval for the lognormal parameter μ.   [3]

8.10   A general insurance company is debating introducing a new screening programme to reduce the claim amounts that it needs to pay out. The programme consists of a much more detailed application form that takes longer for the new client department to process. The screening is applied to a test group of clients as a trial whilst other clients continue to fill in the old application form. It can be assumed that claim payments follow a normal distribution.

Exam style

The claim payments data for samples of the two groups of clients are (in £100 per year):

Without screening	24.5	21.7	35.2	15.9	23.7	34.2	29.3	21.1	23.5	28.3
With screening	22.4	21.2	36.3	15.7	21.5	7.3	12.8	21.2	23.9	18.4

(i)   (a)   Calculate a 95% confidence interval for the difference between the mean claim amounts.

       (b)   Comment on your answer.   [6]

(ii)   (a)   Calculate a 95% confidence interval for the ratio of the population variances.

       (b)   Hence, comment on the assumption of equal variances required in part (i).   [4]

(iii)   Assume that the sample sizes taken from the clients with and without screening are always equal to keep processing easy. Calculate the minimum sample size so that the width of a 95% confidence interval for the difference between mean claim amounts is less than 10, assuming that the samples have the same variances as in part (i).   [3]

[Total 13]

Chapter 8 Solutions

8.1   The sample mean and variance are:

$$\overline{x} = \frac{67}{10} = 6.7$$

$$s^2 = \frac{1}{9}\left\{475 - 10 \times 6.7^2\right\} = 2.9$$

So the confidence interval is given by:

$$6.7 \pm t_{0.025, 9}\sqrt{\frac{2.9}{10}}$$

From the *Tables* with $\alpha = 0.025$, $t_{0.025, 9} = 2.262$, so our confidence interval is $(5.48, 7.92)$.

We have assumed that the numbers of hours that actuarial students spend watching television has a normal distribution.

8.2   The number in a sample of 8 who are in favour has a *Bin*$(8, p)$ distribution, where p is the true underlying proportion in favour. We want the value of p for which the probability of getting 7 or more in favour in a sample of 8 is 0.025. This will give the lower end of the confidence interval for p. We also want the value of p for which the probability of getting 7 or fewer in favour is 0.025. This will give us the upper end of the interval.

The probability of getting 7 or more in favour is:

$$\binom{8}{7}p^7(1-p) + p^8 = 0.025$$

Rearranging the equation:

$$p^7(8 - 7p) = 0.025$$

Using trial and error, or goalseek in Excel to solve this equation we obtain:

$$p = 0.4735$$

For the upper end of the interval, we have:

$$1 - p^8 = 0.025$$

which we can solve directly to give $p = 0.9968$. So a 95% confidence interval for p is $(0.4735, 0.9968)$.

8.3 Assuming that the sample comes from a binomial distribution, we know that the quantity

$$\frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1) \text{ or } \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1). \text{ Here } n = 1,000 \text{ and } X \text{ is the number who favour}$$

Party P.

From the *Tables* we find that $0.99 = P(-2.5758 < Z < 2.5758)$, so:

$$0.99 = P\left(-2.5758 < \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} < 2.5758 \right)$$
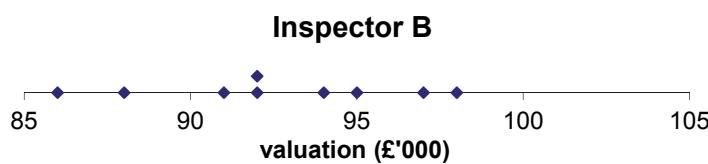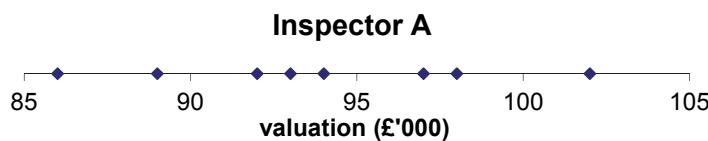
Rearranging this to give us p, and replacing p by $\hat{p}$ under the square root:

$$0.99 = P\left(\frac{X}{n} - 2.5758\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \frac{X}{n} + 2.5758\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Replacing X by 450 and $\hat{p}$ by $\frac{450}{1,000}$, we get the confidence interval to be $0.409 < p < 0.491$.

Since this 99% confidence interval doesn't contain the value $p = 0.5$ (or higher values of p), it is unlikely that Party P will gain more than 50% of the votes.

8.4 (i)(a) The dotplots are as follows:

Inspector A

valuation (£'000)

Inspector B

valuation (£'000)

[2]

(i)(b) B appears to have a slightly smaller spread (but it is hard to tell with so few data points). The difference in the spread doesn't appear to be significant, so the assumption of equal variances could be allowed to stand. [1]

(i)(c)   For Inspector A, we have $n_A = 8$, $\sum x_A = 751$, $\sum x_A^2 = 70,683$, giving:

$$s_A^2 = \frac{1}{7}\left[70,683 - \frac{751^2}{8}\right] = 26.125$$ [1]

For Inspector B, we have $n_B = 9$, $\sum x_B = 833$, $\sum x_B^2 = 77,223$, giving:

$$s_B^2 = \frac{1}{8}\left[77,223 - \frac{833^2}{9}\right] = 15.528$$ [1]

The common (or pooled) variance is given by:

$$s_P^2 = \frac{7 \times 26.125 + 8 \times 15.528}{7 + 8} = 20.473$$ [1]

The pivotal quantity is $\dfrac{15 S_P^2}{\sigma_P^2} \sim \chi_{15}^2$. This gives a 95% confidence interval for σ_P^2 of:

$$\left(\frac{15 \times 20.473}{27.49}, \frac{15 \times 20.473}{6.262}\right) = (11.2, 49.0)$$ [1]

(i)(d)   The confidence interval is calculated using:

$$(\overline{x}_A - \overline{x}_B) \pm t_{0.025,15}\sqrt{s_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \left(\frac{751}{8} - \frac{833}{9}\right) \pm 2.131\sqrt{20.473\left(\frac{1}{8} + \frac{1}{9}\right)}$$ [2]

This gives a confidence interval of $(-3.37, 6.00)$.   [1]

Since this interval contains zero, there is insufficient evidence at the 5% level to suggest that there is a difference in the valuations given by each of the two inspectors.   [1]

(ii)(a)   We are looking at paired data, so we need to examine the differences. The dotplot for the differences is as follows:

[2]

(ii)(b)   For the differences we have $n_D = 8$, $\sum x_D = 14$, $\sum x_D^2 = 198$, giving:

$$\bar{x}_D = 1.75 \qquad s_D^2 = \frac{1}{7}\left[198 - \frac{14^2}{8}\right] = 24.786 \qquad [1]$$

The confidence interval is calculated using:

$$\bar{x}_D \pm t_{0.05,7}\sqrt{\frac{s_D^2}{n_D}} = \frac{14}{8} \pm 1.895\sqrt{\frac{24.786}{8}} \qquad [1]$$

This gives a confidence interval of $(-1.59, 5.09)$. Since this interval contains zero there is insufficient evidence to suggest that A and B give different valuations.   [2]

8.5   (i)(a)   The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

$$\Rightarrow \ \ln L(\lambda) = n\ln\lambda - \lambda\sum x_i \ \Rightarrow \ \frac{d}{d\lambda}\ln L(\lambda) = \frac{n}{\lambda} - \sum x_i \qquad [2]$$

Setting the derivative equal to zero to obtain the MLE:

$$\Rightarrow \ \frac{n}{\hat{\lambda}} - \sum x_i = 0 \ \Rightarrow \ \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}} \qquad [1]$$

Checking it's a maximum:

$$\frac{d^2}{d\lambda^2}\ln L(\lambda) = -\frac{n}{\lambda^2} < 0 \ \Rightarrow \ \max \qquad [½]$$

For these data, $\hat{\lambda} = \dfrac{1}{8.7} = 0.11494$.   [½]

(i)(b) $CRLB = -\dfrac{1}{E\left[\dfrac{d^2}{d\lambda^2}\ln L(\lambda)\right]} = \dfrac{1}{E\left[\dfrac{n}{\lambda^2}\right]} = \dfrac{1}{\dfrac{n}{\lambda^2}} = \dfrac{\lambda^2}{n}$ [1]

For these data values, our estimate of the CRLB is $\hat{\lambda}^2/n = 0.000661$.   [1]

(i)(c)   Since $\hat{\lambda} \sim N(\lambda, CRLB)$ approximately, the confidence interval is given by $\hat{\lambda} \pm 1.96\sqrt{CRLB}$ which, using our CRLB estimate, gives $(0.06457, 0.1653)$.   [1]

(ii)   Since $2\lambda n\overline{X} \sim \chi^2_{2n}$, we have $40\lambda\overline{X} \sim \chi^2_{40}$. The lower and upper 2.5% points of χ^2_{40} are 24.43 and 59.34. So:

$$P\left(24.43 < 2n\lambda\overline{X} < 59.34\right) = 0.95$$

Hence a 95% confidence interval for λ is:

$$\left(\frac{24.43}{40\overline{x}}, \frac{59.34}{40\overline{x}}\right) = \left(\frac{24.43}{348}, \frac{59.34}{348}\right) = \left(0.07020, 0.1705\right) \quad\quad [1]$$

This confidence interval is narrower as it is based upon the exact result, whereas in part (i)(c) it was based on a relatively small sample of 20. A larger sample would have given a narrower interval.   [2]

8.6   (i)   ***Sample size needed (unknown variance)***

Using the pivotal quantity $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$, gives a 95% confidence interval of:

$$\overline{x} \pm t_{0.025;n-1}\frac{s}{\sqrt{n}}$$

The width of this confidence interval is $2 \times t_{0.025;n-1}\dfrac{s}{\sqrt{n}}$, so we require:

$$2 \times t_{0.025;n-1}\frac{8.4}{\sqrt{n}} < 10 \quad\Rightarrow\quad \frac{t_{0.025;n-1}}{\sqrt{n}} < 0.5952$$

Using the values from page 163 of the *Tables*, we find that:

$$\frac{t_{0.025;12}}{\sqrt{13}} = \frac{2.179}{\sqrt{13}} = 0.6043$$

and:

$$\frac{t_{0.025;13}}{\sqrt{14}} = \frac{2.160}{\sqrt{14}} = 0.5773$$

Therefore we need a sample size of at least 14 individuals.

(ii)   ***Sample size needed (known variance)***

Using the pivotal quantity of $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, gives a 95% confidence interval of:

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

The width of this confidence interval is $2 \times 1.96 \dfrac{\sigma}{\sqrt{n}}$, so we require:

$$2 \times 1.96 \frac{8.4}{\sqrt{n}} < 10 \quad \Rightarrow \quad 3.29 < \sqrt{n} \quad \Rightarrow \quad n > 10.8$$

Therefore we need a sample size of at least 11 individuals.

8.7   (i)   ***Exact confidence interval***

We require:

$$P(X \geq 2) = 0.05 \quad \text{under} \quad Poi(\mu_1)$$

$$P(X \leq 2) = 0.05 \quad \text{under} \quad Poi(\mu_2)$$

From the first equation:

$$0.95 = P(X = 0) + P(X = 1) = e^{-\mu_1} + \mu_1 e^{-\mu_1}$$

Solving this numerically we obtain $\mu_1 = 0.36$.

From the second equation:

$$0.05 = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= e^{-\mu_2} + \mu_2 e^{-\mu_2} + \frac{\mu_2^2}{2} e^{-\mu_2}$$

Solving this numerically we obtain $\mu_2 = 6.3$.

So the confidence interval is $(0.36, 6.3)$.

(ii)   ***Approximate confidence interval***

Since n is large enough to use a normal approximation, the pivotal quantity is:

$$\frac{\sum X - n\lambda}{\sqrt{n\lambda}} \sim N(0,1) \quad \text{or} \quad \frac{\hat{\lambda} - \lambda}{\sqrt{\dfrac{\hat{\lambda}}{n}}} \sim N(0,1) \quad \text{(approximately)}$$

where $\hat{\lambda} = \bar{X}$. Hence, a 90% confidence interval can be obtained for λ from:

$$\frac{\sum X \pm 1.6449\sqrt{n\hat{\lambda}}}{n} \quad \text{or} \quad \hat{\lambda} \pm 1.6449\sqrt{\frac{\hat{\lambda}}{n}}$$

Replacing n by 30, $\sum X$ by 60 and $\hat{\lambda}$ by 2 gives:

$$\frac{60 \pm 1.6449\sqrt{30 \times 2}}{30} \quad \text{or} \quad 2 \pm 1.6449\sqrt{\frac{2}{30}}$$

So the confidence interval is:

$$(1.58, 2.42)$$

8.8 The confidence interval is based on the distributional result:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

We have:

$$\bar{x} = \frac{104.2}{7} = 14.88571$$

$$s^2 = \frac{1}{6}\left\{1,581.98 - 7 \times 14.88571^2\right\} = 5.148$$

So a 95% one-sided confidence interval for the variance is given by:

$$\left(\frac{6 \times 5.148}{\chi_{0.05;6}^2}, \infty\right) = \left(\frac{30.888}{12.59}, \infty\right) = (2.45, \infty)$$

So a 95% one-sided confidence interval for the standard deviation is $(1.57, \infty)$.

8.9 The formulae for the mean and variance of a lognormal distribution are:

$$E[X] = e^{\mu + \frac{1}{2}\sigma^2} \quad \text{and} \quad var(X) = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$$

Since the standard deviation equals 1.8 times the mean, we know that:

$$e^{\mu + \frac{1}{2}\sigma^2}(e^{\sigma^2} - 1)^{\frac{1}{2}} = 1.8e^{\mu + \frac{1}{2}\sigma^2} \qquad [1]$$

So:

$$(e^{\sigma^2} - 1)^{\frac{1}{2}} = 1.8 \quad \Rightarrow \quad \sigma^2 = 1.4446 \qquad [1]$$

The 95% confidence interval for the mean corresponds to the inequality:

$$4,250 < e^{\mu + \frac{1}{2}\sigma^2} < 4,750$$

Solving for μ gives:

$$\log 4,250 - \frac{1}{2}\sigma^2 < \mu < \log 4,750 - \frac{1}{2}\sigma^2$$

Using the value found for σ^2, this is:

$$7.632 < \mu < 7.744 \qquad [1]$$

So the lower limit of the confidence interval for μ is 7.632 and the upper limit is 7.744.

8.10   (i)(a)   ***Mean difference confidence interval***

Using the subscript 1 to refer to 'without screening', and 2 to refer to 'with screening', the pivotal quantity is:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_P\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$ [1]

Calculating the required values:

$$\overline{x}_1 = \frac{257.4}{10} = 25.74 \qquad\qquad \overline{x}_2 = \frac{200.7}{10} = 20.07$$ [1]

$$s_1^2 = \frac{1}{9}\left\{6,951.16 - 10 \times 25.74^2\right\} = 36.1871$$ [½]

$$s_2^2 = \frac{1}{9}\left\{4,553.97 - 10 \times 20.07^2\right\} = 58.4357$$ [½]

The pooled sample variance is given by:

$$s_P^2 = \frac{1}{18}(9 \times 36.1871 + 9 \times 58.4357) = 47.3114$$ [1]

Hence, a 95% confidence interval is given by:

$$(25.74 - 20.07) \pm 2.101\sqrt{47.3114}\sqrt{\frac{2}{10}} = (-0.793, 12.1)$$ [1]

Alternatively, the confidence interval for $\mu_2 - \mu_1$ is $(-12.1, 0.793)$.

(i)(b)   ***Comment***

Since the confidence interval contains the value 0, there is insufficient evidence to conclude that the new screening programme significantly reduces the mean claim amount.   [1]

(ii)(a)   ***Ratio of variances confidence interval***

The pivotal quantity is:

$$\frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F_{n_1 - 1, n_2 - 1}$$ [1]

Hence, a 95% confidence interval is given by:

$$\frac{S_1^2 / S_2^2}{F_{0.025; n_1 - 1, n_2 - 1}} < \sigma_1^2 / \sigma_2^2 < \frac{S_1^2 / S_2^2}{F_{0.975; n_1 - 1, n_2 - 1}}$$

Replacing S_1^2 by 36.1871 and S_2^2 by 58.4357, we obtain:

$$\frac{0.6193}{4.026} < \sigma_1^2 \big/ \sigma_2^2 < \frac{0.6193}{1/4.026}$$ [2]

So the confidence interval is:

$$(0.154, 2.49)$$

Alternatively, the confidence interval for $\sigma_2^2 \big/ \sigma_1^2$ *is* $(0.401, 6.50)$.

(ii)(b)   *Comment*

Since the confidence interval contains 1, this means that we are reasonably confident that the population variances are the same.   [1]

(iii)   *Sample size*

The width of the confidence interval is:

$$2 \times t_{2.5\%; 2n-2} \sqrt{\frac{2}{n}} \sqrt{\frac{36.1871(n-1) + 58.4357(n-1)}{2n-2}} = \frac{19.455\, t_{2.5\%; 2n-2}}{\sqrt{n}}$$ [1]

This must be less than 10, so using the percentage points of the t distribution from page 163 of the *Tables*, we see that:

$$n = 15 \quad \Rightarrow \quad \frac{19.455 t_{0.025, 2n-2}}{\sqrt{15}} = \frac{19.455 \times 2.048}{\sqrt{15}} = 10.3 > 10$$

and: $$n = 16 \quad \Rightarrow \quad \frac{19.455 \times 2.042}{\sqrt{16}} = 9.93 < 10$$

The minimum sample size is 16.   [2]

End of Part 2

What next?

1.   Briefly **review** the key areas of Part 2 and/or re-read the **summaries** at the end of Chapters 5 to 8.

2.   Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 2. If you don't have time to do them all, you could save the remainder for use as part of your revision.

3.   Attempt **Assignment X2**.

Time to consider …

… 'revision' products

Flashcards – These are available in both paper and eBook format. One student said:

  'The paper-based Flashcards are brilliant.'

You can find lots more information, including samples, on our website at www.ActEd.co.uk.

Buy online at www.ActEd.co.uk/estore

9

Hypothesis testing

Syllabus objectives

3.3   Hypothesis testing and goodness of fit

    3.3.1   Explain what is meant by the terms null and alternative hypotheses, simple and composite hypotheses, type I and type II errors, test statistic, likelihood ratio, critical region, level of significance, probability-value and power of a test.

    3.3.2   Apply basic tests for the one-sample and two-sample situations involving the normal, binomial and Poisson distributions, and apply basic tests for paired data.

    3.3.3   Apply the permutation approach to non-parametric hypothesis tests.

    3.3.4   Use a chi-square test to test the hypothesis that a random sample is from a particular distribution, including cases where parameters are unknown.

    3.3.5   Explain what is meant by a contingency (or two-way) table, and use a chi-square test to test the independence of two classification criteria.

0   Introduction

In many research areas, such as medicine, education, advertising and insurance, it is necessary to carry out statistical tests. These tests enable researchers to use the results of their experiments to answer questions such as:
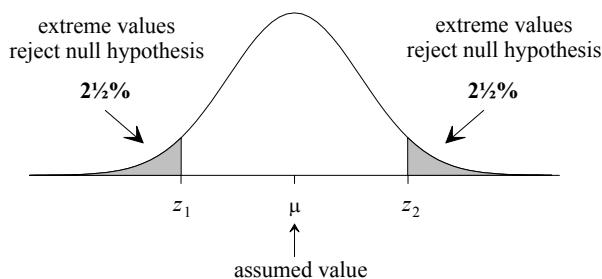
- Is drug A a more effective treatment for AIDS than drug B?

- Does training programme T lead to improved staff efficiency?

- Are the severities of large individual private motor insurance claims consistent with a lognormal distribution?

A hypothesis is where we make a statement about something; for example the mean lifetime of smokers is less than that of non-smokers. A hypothesis test is where we collect a representative sample and examine it to see if our hypothesis holds true.

Hypothesis tests are closely linked to the confidence intervals we developed in Chapter 8. For example, when we were sampling from a $N(\mu, \sigma^2)$ distribution (σ^2 known) we used:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

By substituting in $\bar{X}, \sigma^2$ and n, we found the values of μ that corresponded to 95% of the data being in the 'centre'. For hypothesis tests, we now assume a value of μ based on our hypothesis and can calculate a probability value for the test assuming our initial value of μ is correct. If we find that our sample mean is unlikely to occur given our hypothesised value of μ, we naturally conclude that it is likely that our sample does *not* come from this distribution with the assumed value of μ. In this case we would *reject* the 'null' hypothesis. If, however our sample mean is not very extreme, it would be fair to say that it probably does have the assumed value of μ. In this case we would not reject the 'null' hypothesis.

Because of the similarity with Chapter 8, most of the formulae used are identical. The only exceptions are for the binomial and Poisson distributions.

Finally, we can develop our estimation work from Chapter 7. For example, given the number of claims from a certain portfolio that we receive in 100 monthly periods:

Claims	0	1	2	3	4	5	6
Frequency	9	22	26	21	13	6	3

Assuming a Poisson distribution with parameter μ, our estimate using the methods given in Chapter 7 would be $\mu = \bar{X} = 2.37$. We then obtained a confidence interval for the mean in Chapter 8. But all of this work is appropriate *only* if the distribution is Poisson. Hence, we will see in this chapter how to carry out a test of whether our sample does or does not conform to this distribution.

This chapter has traditionally formed one of the longer questions of the Statistics exam. Spend your time wisely.

1   Hypotheses, test statistics, decisions and errors

1.1   The testing procedure

The standard approach to carrying out a statistical test involves the following steps:

- specify the hypothesis to be tested

- select a suitable statistical model

- design and carry out an experiment/study

- calculate a test statistic

- calculate the probability value

- determine the conclusion of the test.

We will not be concerned here with the design of the experiment. We will assume that an experiment, based on an appropriate statistical model, has already been conducted and the results are available.

1.2   Hypotheses

In Sections 1-6 of this chapter a hypothesis is a statement about the value of an unknown parameter in the model.

The basic hypothesis being tested is the null hypothesis, denoted H_0 – it can sometimes be regarded as representing the current state of knowledge or belief about the value of the parameter being tested (the 'status quo' hypothesis). In many situations a difference between two populations is being tested and the null hypothesis is that there is <u>no</u> difference.

In a test, the null hypothesis is contrasted with the alternative hypothesis, denoted H_1.

Where a hypothesis completely specifies the distribution, it is called a simple hypothesis. Otherwise it is called a composite hypothesis.

For example, when testing the null hypothesis $H_0 : \mu = 0.8$ against the alternative hypothesis $H_1 : \mu = 0.6$, both of the hypotheses are simple. However when testing $H_0 : \mu = 0.8$ against $H_1 : \mu < 0.8$, H_1 is a composite hypothesis.

A test is a rule which divides the sample space (the set of possible values of the data) into two subsets, a region in which the data are consistent with H_0, and its complement, in which the data are inconsistent with H_0. The tests discussed here are designed to answer the question 'Do the data provide sufficient evidence to justify our rejecting H_0?'.

1.3   One-sided and two-sided tests

In a test of whether smoking reduces life expectancies, the hypotheses would be:

H_0 : smoking makes no difference to life expectancy

H_1 : smoking reduces life expectancy

This is an example of a one-sided test, since we are only considering the possibility of a reduction in life expectancy *ie* a change in one direction. However we could have specified the hypotheses:

H_0 : smoking makes no difference to life expectancy

H_1 : smoking affects life expectancy

This is a two-sided test, since the alternative hypothesis considers the possibility of a change in either direction, *ie* an increase or a decrease.

1.4   Test statistics

The actual decision is based on the value of a suitable function of the data, the test statistic. The set of possible values of the test statistic itself divides into two subsets, a region in which the value of the test statistic is consistent with H_0 , and its complement, the critical region (or rejection region), in which the value of the test statistic is inconsistent with H_0 . If the test statistic has a value in the critical region, H_0 is rejected. The test statistic (like any statistic) must be such that its distribution is completely specified when the value of the parameter itself is specified (and in particular 'under H_0' *ie* when H_0 is true).

In exam questions the test statistic is generally calculated from data given in the question. For details of how to reach a conclusion in practice, see Section 3.1.

1.5   Errors

The level of significance of the test, denoted α, is the probability of committing a Type I error, *ie* it is the probability of rejecting H_0 when it is in fact true. The probability of committing a Type II error, denoted β, is the probability of accepting H_0 when it is false. An ideal test would be one which simultaneously minimises α and β – this ideal however is not attainable in practice.

Question

A random variable *X* is believed to follow an *Exp*(λ) distribution. In order to test the null hypothesis $\mu = 20$ against the alternative hypothesis $\mu = 30$, where $\mu = 1/\lambda$, a single value is observed from the distribution. If this value is less than 28, H_0 is accepted, otherwise H_0 is rejected. Calculate the probabilities of:

(i)   a Type I error

(ii)   a Type II error.

Solution

(i)   The probability of a Type I error is given by:

$$P(\text{reject } H_0 \text{ when } H_0 \text{ true}) = P\left(X > 28 \text{ when } X \sim Exp\left(1/20\right)\right)$$

$$= 1 - F_X(28) = e^{-28/20} = 0.2466$$

The CDF of the exponential distribution is given on Page 11 of the Tables.

(ii)   The probability of a Type II error is given by:

$$P(\text{do not reject } H_0 \text{ when } H_0 \text{ false}) = P\left(X < 28 \text{ when } X \sim Exp\left(1/30\right)\right)$$

$$= F_X(28) = 1 - e^{-28/30} = 0.6068$$

In this case we were forced to choose between $H_0 : \mu = 20$ and $H_1 : \mu = 30$. So saying that H_0 is false is the same as saying that $\mu = 30$.

Since we've only got one value in our sample here, not surprisingly, the probabilities of Type I and Type II errors are quite big.

The probability of a Type I error is also referred to as the 'size' of the test, which will normally be a small number such as 0.05 (say).

The power of a test is the probability of rejecting H_0 when it is false, so that the power equals $1 - \beta$. In general, this will be a function of the unknown parameter value. **For simple hypotheses the power is a single value, but for composite hypotheses it is a function being defined at all points in the alternative hypothesis.**

A test with a high power is said to be 'powerful' as it is very effective at demonstrating a positive result.

Question

Give an expression in terms of μ for the power of the test in the question on the previous page. Comment on how the power is affected by the value of μ.

Solution

The power is the probability of rejecting H_0 when the true value of the parameter μ is some value other than $\mu = 20$. In terms of μ this is:

$$P\big(X > 28 \mid X \sim Exp(1/\mu)\big) = 1 - F_X(28) = e^{-28/\mu}$$

If μ is large (1,000, say), then the power will be close to 1, since the test will reject $H_0 : \mu = 20$ very easily. Conversely if μ is small (10, say), then the power will be close to 0, since the test will not reject $H_0 : \mu = 20$ very easily.

R can calculate the power of a one-sample t-test (covered in Section 3.1) using the function:

```
power.t.test.
```

2   Classical testing, significance and *p*-values

2.1   'Best' tests

The classical approach to finding a 'good' test (called the Neyman-Pearson theory) fixes the value of α , *ie* the level of significance required and then tries to find such a test for which the other error probability, β , is as small as possible for every value of the parameter specified by the alternative hypothesis. This can also be described as finding the 'most powerful' test.

The key result in the search for such a test is the Neyman-Pearson lemma, which provides the 'best' test (smallest β) in the case of two simple hypotheses. For a given level, the critical region (and in fact the test statistic) for the best test is determined by setting an upper bound on the likelihood ratio L_0/L_1 , where L_0 and L_1 are the likelihood functions of the data under H_0 and H_1 respectively.

The Neyman-Pearson lemma

Formally, if *C* is a critical region of size α and there exists a constant *k* such that $L_0/L_1 \leq k$ inside *C* and $L_0/L_1 \geq k$ outside *C*, then *C* is a most powerful critical region of size α for testing the simple hypothesis $\theta = \theta_0$ against the simple alternative hypothesis $\theta = \theta_1$.

So a Neyman-Pearson test rejects H_0 if:

$$\frac{\text{Likelihood under } H_0}{\text{Likelihood under } H_1} < \text{critical value}$$

Question

A random variable *X* is believed to follow an *Exp*(λ) distribution. In order to test the null hypothesis $\mu = 20$ against the alternative hypothesis $\mu = 30$, where $\mu = 1/\lambda$, a single value is observed from the distribution. If this value is less than 28, H_0 is accepted, otherwise H_0 is rejected.

Show that this is a Neyman-Pearson test.

Solution

Given a single value from an exponential distribution, the Neyman-Pearson criterion is 'reject H_0 if $L_0/L_1 < \text{critical value}$.' Using the null and alternative hypotheses, the test becomes:

$$\frac{\dfrac{1}{20}e^{-\frac{x}{20}}}{\dfrac{1}{30}e^{-\frac{x}{30}}} < \text{constant}$$

This reduces to $e^{-\frac{x}{60}} < \text{constant}$, or $x > \text{constant}$. This was exactly the form of the test that we used (we rejected H_0 when $x > 28$). So this is a Neyman-Pearson test.

Common tests are often such that the null hypothesis is simple, *eg* $H_1 : \theta = \theta_0$, against a composite alternative, *eg* $H_1 : \theta \neq \theta_0$, which is two-sided, and $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$, which are one-sided.

Here it is only in certain special cases (usually one-sided cases) that a single test is available which is best (*ie* uniformly most powerful) for all parameter values. In cases where a single best test in the sense of the Neyman–Pearson Lemma is unavailable, another approach is used to derive sensible tests. This approach, which is a generalisation of the Lemma, produces tests which are referred to as likelihood ratio tests.

Likelihood ratio tests

The critical region (and test statistic) for the test are determined by setting an upper bound on the ratio ($\max L_0 / \max L$), where $\max L_0$ is the maximum value of the likelihood L under the restrictions imposed by the null hypothesis, and $\max L$ is the overall maximum value of L for all allowable values of all parameters involved. Likelihood ratio tests are used, for example, in survival models with covariates (see Subject CS2).

In the most common case when H_0 and H_1 together cover all possible values for the parameters, this generalised test rejects H_0 if:

$$\frac{\max(\text{Likelihood under } H_0)}{\max(\text{Likelihood under } H_0 + H_1)} < \text{critical value}$$

Important results include the case of sampling from a $N(\mu, \sigma^2)$ distribution. The method leads to the test statistic:

$$\frac{\overline{X} - \mu_0}{S / \sqrt{n}} \sim t_{n-1} \text{ under } H_0 : \mu = \mu_0$$

for tests on the value of the mean μ.

We're assuming here that σ^2 is unknown. If it is known, then the *z*-test is the 'best' test.

The method also leads to the test statistic:

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2 \text{ under } H_0 : \sigma^2 = \sigma_0^2$$

for tests on the value of the variance σ^2.

2.2   p-values

Under the 'classical' Neyman-Pearson approach, with a fixed predetermined value of α, a test will produce a decision as to whether to reject H_0. But merely comparing the observed test statistic with some critical value and concluding *eg* 'using a 5% test, reject H_0' or 'reject H_0 with significance level 5%' or 'result significant at 5%' (all equivalent statements) does not provide the recipient of the results with clear detailed information on the strength of the evidence against H_0.

A more informative approach is to calculate and quote the probability value (*p*-value) of the observed test statistic. This is the observed significance level of the test statistic – the probability, assuming H_0 is true, of observing a test statistic at least as 'extreme' (inconsistent with H_0) as the value observed.

The *p*-value is the lowest level at which H_0 can be rejected.

The smaller the *p*-value, the stronger is the evidence against the null hypothesis.

For example, when testing $H_0 : \theta = 0.5$ *vs* $H_1 : \theta = 0.4$, where θ is the probability of a coin coming up heads, and 82 heads have been observed in 200 tosses, the *p*-value of the result is:

$$P(X \leq 82) \text{ where } X \sim \text{Bin}(200, 0.5)$$

$$P\left(Z < \frac{82.5 - 100}{\sqrt{50}} \right) = P(Z < -2.475) = 0.0067$$

H_0 is therefore extremely unlikely – probability < 0.01 – and there is very strong evidence against H_0 and in favour of H_1. A good way of expressing the result is: 'we have very strong evidence against the hypothesis that the coin is fair (*p*-value 0.007) and conclude that it is biased against heads'.

Testing does not prove that any hypothesis is true or untrue. Failure to detect a departure from H_0 means that there is not enough evidence to justify rejecting H_0, so H_0 is accepted in this sense only, whilst realising that it may not be true. This attitude to the acceptance of H_0 is a feature of the fact that H_0 is usually a precise statement, which is almost certainly not exactly true.

Question

A random variable *X* is believed to follow an $Exp(\lambda)$ distribution. In order to test the null hypothesis $\mu = 20$ against the alternative hypothesis $\mu = 30$, where $\mu = 1/\lambda$, a single value is observed from the distribution. If this value of *X* is less than *k*, H_0 is accepted, otherwise H_0 is rejected.

(i)   Calculate the value of *k* that gives a test of size 5%.

(ii)   Determine is the probability of a Type II error in this case.

Solution

(i)   We want:

$$0.05 = \int_k^\infty \frac{1}{20} e^{-\frac{x}{20}} dx = \left[-e^{-\frac{x}{20}} \right]_k^\infty = e^{-\frac{k}{20}}$$

So:

$$k = -20\ln 0.05 = 59.9$$

(ii)   The probability of a Type II error is:

$$\int_0^k \frac{1}{30} e^{-\frac{x}{30}} dx = \left[-e^{-\frac{x}{30}} \right]_0^k = 1 - e^{-1.997} = 0.864$$

Note that a *p*-value of less than 5% is considered 'significant' – so that the null hypothesis is rejected. If an exam question does not state the level of the test, you should assume that it is 5%.

3 Basic tests – single samples

3.1 Testing the value of a population mean

Situation: random sample, size n, from $N(\mu, \sigma^2)$ – sample mean $\bar{X}$

Testing: H_0 : $\mu = \mu_0$

(a) σ known: test statistic is $\bar{X}$, and $\dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$ under H_0

(b) σ unknown: test statistic is $\dfrac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t_{n-1}$ under H_0

For large samples, $N(0,1)$ can be used in place of t_{n-1}. Further, the Central Limit Theorem justifies the use of a normal approximation for the distribution of $\bar{X}$ in sampling from any reasonable population, and s^2 is a good estimate of σ^2, so the requirement that we are sampling from a normal distribution is not necessary in either case (a) or (b) when we have a large sample.

Question

The average IQ of a sample of 50 university students was found to be 105. Carry out a statistical test to determine whether the average IQ of university students is greater than 100, assuming that IQs are normally distributed. It is known from previous studies that the standard deviation of IQs among students is approximately 20.

Solution

We are testing:

$$H_0 : \mu = 100 \quad vs \quad H_1 : \mu > 100 \ (\sigma \text{ known})$$

Under H_0, $\dfrac{\bar{X} - 100}{\sigma / \sqrt{n}} \sim N(0,1)$.

The test statistic is $\dfrac{105 - 100}{20 / \sqrt{50}} = 1.768$.

We need to draw a conclusion and there are two ways of doing this.

Method 1:

Calculate the probability of getting a result as extreme as the test statistic (*ie* the *p*-value). If $Z \sim N(0,1)$:

$$P(Z > 1.768) = 1 - 0.96147 = 0.03853$$

We are carrying out a 5% one-tailed test. The probability we have obtained is less than 5%, so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the average IQ of university students is greater than 100.

Method 2:

From the *Tables*, $P(Z > 1.6449) = 0.05$, so 1.6449 is the critical value for a one-tailed 5% test. The test statistic of 1.768 exceeds this critical value, so we reach the same conclusion as we did for Method 1.

Question

Test using a 5% significance level whether the average IQ of university students is greater than 103, based on the sample in the previous question.

Solution

We are testing:

$$H_0 : \mu = 103 \quad vs \quad H_1 : \mu > 103$$

Under H_0:

$$\frac{\overline{X} - 103}{\sigma / \sqrt{n}} \sim N(0,1)$$

The observed value of the test statistic is:

$$\frac{105 - 103}{20 / \sqrt{50}} = 0.707$$

This is less than 1.6449 (the upper 5% point of a $N(0,1)$ distribution) so we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the average IQ of university students is not more than 103.

Alternatively, using probability values, we have $P(Z > 0.707) \simeq 0.24$. This is greater than 0.05, so we have insufficient evidence to reject H_0 at the 5% level.

Question

The annual rainfall in centimetres at a certain weather station over the last ten years has been as follows:

17.2   28.1   25.3   26.2   30.7   19.2   23.4   27.5   29.5   31.6

Scientists at the weather station wish to test whether the average annual rainfall has increased from its former long-term value of 22 cm. Test this hypothesis at the 5% level, stating any assumptions that you make.

Solution

We are testing:

$$H_0 : \mu = 22 \quad vs \quad H_1 : \mu > 22$$

Assuming that annual rainfall measurements are independent and normally distributed, then under H_0 :

$$\frac{\overline{X} - 22}{S/\sqrt{n}} \sim t_{n-1}$$

We have:

$$s^2 = \frac{1}{9}\left(6{,}895.73 - 10 \times 25.87^2\right) = 22.57$$

So the observed value of the test statistic is:

$$\frac{25.87 - 22}{\sqrt{22.57}/\sqrt{10}} = 2.576$$

Since this is greater than 1.833 (the upper 5% point of the t_9 distribution), we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the long-term average annual rainfall has increased from its former level.

Alternatively, using probability values, we have $P(t_9 > 2.576) \simeq 0.0166$. This is less than 0.05, so we have sufficient evidence to reject H_0 at the 5% level.

R **R can carry out a hypothesis test for the mean with unknown variance at the 5% level using:**

```
t.test(<sample data>, conf=0.95)
```

For small samples from a non-normal distribution then an empirical distribution of the statistic can be constructed in R using the bootstrap method (see Chapter 7, Section 7), from which we can calculate the critical value(s) and obtain an estimate of the *p*-value.

3.2   Testing the value of a population variance

Situation: random sample, size *n*, from $N(\mu, \sigma^2)$ – sample variance S^2.

Testing: $H_0 : \sigma^2 = \sigma_0^2$

Test statistic is $\dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$ under H_0

For large samples, the test works well even if the population is not normally distributed.

Question

Carry out a statistical test to assess whether the standard deviation of the heights of 10-year-old children is equal to 3cm, based on the random sample of 5 heights in cm given below. Assume that heights are normally distributed.

124,   122,   130,   125,   132

Solution

We are testing:

$$H_0 : \sigma = 3 \quad vs \quad H_1 : \sigma \neq 3$$

Under H_0 :

$$\frac{4S^2}{3^2} \sim \chi_4^2$$

We have:

$$s^2 = \frac{1}{4}\left(80,209 - 5 \times 126.6^2\right) = 17.8$$

So the observed value of the test statistic is:

$$\frac{4 \times 17.8}{3^2} = 7.91$$

Our statistic of 7.91 lies between 0.4844 and 11.14 (the lower and upper 2½% points of the χ_4^2 distribution). So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the standard deviation of the heights of 10-year-old children is 3cm.

Alternatively, using probability values, we have $P(\chi_4^2 > 7.91) \simeq 0.0952$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.0952 = 0.190$, which is greater than 0.05 so we have insufficient evidence to reject H_0 at the 5% level.

Question

The annual rainfall in centimetres at a certain weather station over the last ten years has been as follows:

      17.2 28.1 25.3 26.2 30.7 19.2 23.4 27.5 29.5 31.6

Assuming this data is taken from a normal distribution test at the 5% level whether the standard deviation of the annual rainfall at the weather station is equal to 4 cm.

Solution

We are testing:

$$H_0 : \sigma = 4 \quad vs \quad H_1 : \sigma \neq 4$$

The test is two-sided. Assuming independence and normality, then under H_0 :

$$\frac{9S^2}{4^2} \sim \chi_9^2$$

Using the sample variance calculated earlier, the observed value of the test statistic is:

$$\frac{9 \times 22.57}{16} = 12.69$$

This is between the upper and lower 2½% points of χ_9^2 (2.700 and 19.02), so we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the standard deviation of the rainfall is 4 cm.

Alternatively, using probability values, we have $P(\chi_9^2 > 12.69) \simeq 0.1775$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.1775 = 0.355$. Since this is greater than 0.05 we have insufficient evidence to reject H_0 at the 5% level.

There is no built-in function to carry out a hypothesis test for the variance in R. We can use R to calculate the value of the statistic from scratch or use a bootstrap method if the assumptions are not met.

For example, if we are unsure whether our sample comes from a normal distribution, a bootstrap method would be more appropriate here.

3.3   Testing the value of a population proportion

<u>Situation</u>: *n* binomial trials with *P*(success) = *p* ; we observe *x* successes.

<u>Testing:</u> $H_0 : p = p_0$.

<u>Test statistic</u> is $X \sim Bin(n, p_0)$ under H_0 .

For large *n*, use the normal approximation to the binomial (with continuity correction), *ie* use:

$$\frac{\dfrac{X \pm \frac{1}{2}}{n} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \doteqdot N(0,1)$$

or:

$$\frac{X \pm \frac{1}{2} - np}{\sqrt{np(1-p)}} \doteqdot N(0,1)$$

When carrying out tests of this type you can work out whether you need to add or subtract the $\frac{1}{2}$ in the continuity correction if you remember that you always adjust the value of X *towards* the mean of the distribution under H_0 . For large values of n , this will make little difference unless the test statistic is close to the critical value.

Question

In a one-year mortality investigation, 45 of the 250 ninety-year-olds present at the start of the investigation died before the end of the year. Assuming that the number of deaths has a $Bin(250, q)$ distribution, test whether this result is consistent with a mortality rate of $q = 0.2$ for this age.

Solution

We are testing:

$$H_0 : q = 0.2 \quad vs \quad H_1 : q \neq 0.2$$

Under H_0 :

$$\frac{X/n - 0.2}{\sqrt{\dfrac{0.2 \times 0.8}{n}}} \sim N(0,1) \text{ approximately}$$

Using the observed values, $n = 250$ and $x = 45$, the test statistic with continuity correction is:

$$\frac{45.5/250 - 0.2}{\sqrt{\dfrac{0.2 \times 0.8}{250}}} = -0.712$$

Note that since the mean is $np = 250 \times 0.2 = 50$, our continuity correction involves adjusting 45 towards the mean. So we have to *add* 0.5.

Our statistic of -0.712 lies between ± 1.960 (the lower and upper 2½% points of the $N(0,1)$ distribution). So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the true mortality rate for this age is 0.2.

Alternatively, using probability values, we have $P(Z < -0.712) = 0.238$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.238 \simeq 0.48$, which is greater than 0.05 so we have insufficient evidence to reject H_0 at the 5% level.

Question

A new gene has been identified that makes carriers of it particularly susceptible to a particular degenerative disease. In a random sample of 250 adult males born in the UK, 8 were found to be carriers of the disease. Test whether the proportion of adult males born in the UK carrying the gene is less than 10%.

Solution

We are testing:

$$H_0 : p = 0.1 \quad vs \quad H_1 : p < 0.1$$

Under H_0:

$$\frac{X/n - 0.1}{\sqrt{\dfrac{0.1 \times 0.9}{n}}} \doteqdot N(0,1)$$

The observed value of the test statistic, with continuity correction adjusted towards the mean, is:

$$\frac{8.5/250 - 0.1}{\sqrt{\dfrac{0.1 \times 0.9}{250}}} = -3.479$$

We are carrying out a one-sided test. The value of the test statistic is less than -1.6449 (the lower 5% point of the $N(0,1)$ distribution) so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the proportion of male carriers in the population is less than 10%.

Alternatively, using probability values, we have $P(Z < -3.479) \simeq 0.00025$. This is less than 0.05, so we have sufficient evidence to reject H_0 at the 5% level. In fact, we have sufficient evidence to reject H_0 at even the 0.05% level.

R **R can carry out an *exact* hypothesis test for *p* at the 5% level using:**

```
binom.test(x,n, conf=0.95)
```

3.4 Testing the value of the mean of a Poisson distribution

Situation: random sample, size *n*, from Poi(λ) distribution.

Testing: $H_0 : \lambda = \lambda_0$

Test statistic is sample sum $\sum X_i \sim$ Poi $(n\lambda_0)$ under H_0. In the case where *n* is small and $n\lambda_0$ is of moderate size, probabilities can be evaluated directly (or found from tables, if available).

For large samples (or indeed whenever the Poisson mean is large) a normal approximation can be used for the distribution of the sample sum or sample mean. Recall that $\sum X_i \sim$ Poi $(n\lambda) \rightarrow N(n\lambda, n\lambda)$.

Test statistic is $\bar{X}$, and $\dfrac{\bar{X} - \lambda_0}{\sqrt{\lambda_0/n}} \sim N(0,1)$ under H_0

or we can use $\sum X_i$, and $\dfrac{\sum X_i - n\lambda_0}{\sqrt{n\lambda_0}} \sim N(0,1)$ under H_0.

Using the second version it is easier to incorporate a continuity correction.

The first version has continuity correction $0.5/n$, whereas the second version has continuity correction 0.5.

Question

In a one-year investigation of claim frequencies for a particular category of motorists, the total number of claims made under 5,000 policies was 800. Assuming that the number of claims made by individual motorists has a *Poi*(λ) distribution, test at the 1% level whether the unknown average claim frequency λ is less than 0.175.

Solution

We are testing:

$$H_0 : \lambda = 0.175 \quad vs \quad H_1 : \lambda < 0.175$$

Under H_0:

$$\frac{\overline{X} - 0.175}{\sqrt{0.175/n}} \sim N(0,1)$$

Using the observed values, $n = 5,000$ and $\overline{x} = 0.16$, the test statistic, with continuity correction, is:

$$\frac{\frac{800.5}{5,000} - 0.175}{\sqrt{0.175/5,000}} = -2.519$$

This is less than -2.3263, the lower 1% point of the $N(0,1)$ distribution. So we have sufficient evidence at the 1% level to reject H_0. Therefore it is reasonable to conclude that the true claim frequency is less than 0.175.

Alternatively, using probability values, we have $P(Z < -2.519) = 0.0059$. Since this is less than 0.01, we have sufficient evidence to reject H_0 at the 1% level.

Question

A random sample of 500 policies of a particular kind revealed a total of 116 claims during the last year. Test the null hypothesis $H_0 : \lambda = 0.18$ against the alternative $H_1 : \lambda > 0.18$, where λ is the annual claim frequency, *ie* the average number of claims per policy.

Solution

We are testing:

$$H_0 : \lambda = 0.18 \quad vs \quad H_1 : \lambda > 0.18$$

Assuming that the underlying claim frequency has a Poisson distribution, then under H_0:

$$\frac{\overline{X} - 0.18}{\sqrt{0.18/n}} \doteq N(0,1)$$

The observed value of the test statistic, with continuity correction, is:

$$\frac{\frac{115.5}{500} - 0.18}{\sqrt{0.18/500}} = 2.688$$

We are carrying out a one-sided test. The value of the test statistic is greater than 1.6449 (the upper 5% point of the $N(0,1)$ distribution) so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the true claim frequency is more than 0.18.

Alternatively, using probability values, we have $P(Z > 2.688) \simeq 0.0036$, ie 0.36%. This is less than 0.05, so we have sufficient evidence to reject H_0 at the 5% level. In fact, we have sufficient evidence to reject H_0 even at the 0.5% level.

R can carry out an *exact* hypothesis test for λ at the 5% level using:

```
poisson.test(x,n, conf=0.95)
```

4   Basic tests – two independent samples

4.1   Testing the value of the difference between two population means

Situation: independent random samples, sizes n_1 and n_2 from $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ respectively.

Testing: $H_0 : \mu_1 - \mu_2 = \delta$

(a) σ_1^2, σ_2^2 **known**

test statistic: $z = \dfrac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

> **R** There is no built-in function for calculating the above hypothesis test in R. We can use R to calculate the results of the statistic from scratch or use a bootstrap method if the assumptions are not met.

(b) σ_1^2, σ_2^2 **unknown – much the more usual situation**

Large samples: use S_i^2 **to estimate** σ_i^2. We will now use a t distribution.

Further, the Central Limit Theorem justifies the use of a normal approximation for the distribution of the test statistic in sampling from any reasonable populations, so the requirement that we are sampling from *normal* distributions is not necessary when we have large samples.

Small samples: under the assumption $\sigma_1^2 = \sigma_2^2 \left(= \sigma^2 \text{ say} \right)$, **this common variance is estimated by** S_p^2, **and the test statistic is** $t = \dfrac{\bar{x}_1 - \bar{x}_2 - \delta}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ **which is distributed as *t* with**

$n_1 + n_2 - 2$ **degrees of freedom under** H_0.

Remember that $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$.

> **R** R can carry out a hypothesis test for the difference between the means with unknown variance using the function `t.test`. We set the argument `var.equal = TRUE` for small samples.

Again, we could use the bootstrap method to construct an empirical distribution of the statistic if the assumptions are not met.

Question

The average blood pressure for a control group C of 10 patients was 77.0 mmHg. The average blood pressure in a similar group T of 10 patients on a special diet was 75.0 mmHg. Carry out a statistical test to assess whether patients on the special diet have lower blood pressure.

You are given that $\displaystyle\sum_{i=1}^{10} c_i^2 = 59,420$ and $\displaystyle\sum_{i=1}^{10} t_i^2 = 56,390$.

Solution

We are testing:

$$H_0 : \mu_C = \mu_T \quad vs \quad H_1 : \mu_C > \mu_T$$

If we assume that blood pressures are normally distributed and that the variance of the underlying distribution for each group is the same, then under H_0:

$$\frac{(\overline{C}-\overline{T})-(0)}{S_P\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}} \sim t_{m+n-2}$$

We have:

$$
\begin{aligned}
s_P^2 &= \frac{1}{m+n-2}\left[(m-1)s_C^2 + (n-1)s_T^2\right] \\[2mm]
&= \frac{1}{m+n-2}\left[\sum_{i=1}^{m}(c_i-\overline{c})^2 + \sum_{i=1}^{n}(t_i-\overline{t})^2\right] \\[2mm]
&= \frac{1}{m+n-2}\left[\sum_{i=1}^{m}c_i^2 - m\overline{c}^2 + \sum_{i=1}^{n}t_i^2 - n\overline{t}^2\right] \\[2mm]
&= \frac{1}{10+10-2}\left[59,420 - 10\times 77.0^2 + 56,390 - 10\times 75.0^2\right] = 15.00 = 3.873^2
\end{aligned}
$$

Note that, as mentioned previously, the number of degrees of freedom to use with a t-test is the same as the denominator used when calculating the estimate of the variance ie 18 in this case.

Using the observed values of $m=10$, $n=10$, $\overline{t}=75.0$, $\overline{c}=77.0$, and $s_P^2 = 3.873^2$, the test statistic is:

$$\frac{(77.0-75.0)}{3.873\sqrt{\dfrac{1}{10}+\dfrac{1}{10}}} = 1.15$$

This is less than 1.734, the upper 5% point of the t_{18} distribution. So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that patients on the special diet have the same blood pressure as patients on the normal diet.

Alternatively, using probability values, we have $P(t_{18} > 1.15) \simeq 0.134$. Since this is greater than 0.05, we have insufficient evidence to reject H_0 at the 5% level.

Question

A car manufacturer runs tests to investigate the fuel consumption of cars using a newly developed fuel additive. Sixteen cars of the same make and age are used, eight with the new additive and eight as controls. The results, in miles per gallon over a test track under regulated conditions, are as follows:

Control   27.0 32.2 30.4 28.0 26.5 25.5 29.6 27.2

Additive   31.4 29.9 33.2 34.4 32.0 28.7 26.1 30.3

If μ_C is the mean number of miles per gallon achieved by cars in the control group, and μ_A is the mean number of miles per gallon achieved by cars in the group with fuel additive, test:

(i) $H_0 : \mu_A - \mu_C = 0$ *vs* $H_1 : \mu_A - \mu_C > 0$

(ii) $H_0 : \mu_A - \mu_C = 6$ *vs* $H_1 : \mu_A - \mu_C \neq 6$

Solution

Using C_i for the number of miles per gallon of the cars in the control group and A_i for the number of miles per gallon of the cars with additive, we have:

$$\sum c_i = 226.4, \ \sum c_i^2 = 6,442.5, \ \sum a_i = 246, \ \sum a_i^2 = 7,612.56$$

Our estimate of the pooled sample variance is:

$$s^2 = \frac{1}{m+n-2}\left[\sum c_i^2 - n\bar{c}^2 + \sum a_i^2 - m\bar{a}^2\right]$$

$$= \frac{1}{14}\left(6,442.5 - 8 \times 28.3^2 + 7,612.56 - 8 \times 30.75^2\right) = 5.96$$

(i)   We are testing:

$$H_0 : \mu_A - \mu_C = 0   \ \text{vs}   \ H_1 : \mu_A - \mu_C > 0$$

Assuming that the underlying distributions are normal, then under H_0:

$$\frac{(\bar{A}-\bar{C})-0}{S\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}} \sim t_{m+n-2}$$

The observed value of the test statistic is:

$$\frac{30.75-28.3}{\sqrt{5.96}\sqrt{\dfrac{1}{8}+\dfrac{1}{8}}} = 2.007$$

This is greater than 1.761 (the upper 5% point of the t_{14} distribution) so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the mean performance is greater with the additive than without.

Alternatively, using probability values, we have $P(t_{14} > 2.007) \simeq 0.0340$. This is less than 0.05, so we have sufficient evidence to reject H_0 at the 5% level.

(ii)   We are now testing:

$$H_0 : \mu_A - \mu_C = 6 \quad vs \quad H_1 : \mu_A - \mu_C \neq 6$$

Making the same assumptions as before, under H_0:

$$\frac{(\bar{A}-\bar{C})-6}{S\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}} \sim t_{m+n-2}$$

The observed value of the test statistic is now:

$$\frac{(30.75-28.3)-6}{\sqrt{5.96}\sqrt{\dfrac{1}{8}+\dfrac{1}{8}}} = -2.908$$

This is a two-sided test and our statistic is less than -2.145 (the lower 2.5% point of the t_{14} distribution) so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the difference in the means is not equal to 6.

Alternatively, using probability values, we have $P(t_{14} < -2.908) \simeq 0.00598$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.00598 = 0.0120$. Since this is less than 0.05, we have sufficient evidence to reject H_0 at the 5% level. In fact, we have sufficient evidence to reject H_0 even at the 2.5% level.

4.2   Testing the value of the ratio of two population variances

__Situation__: independent random samples, sizes n_1 and n_2 from $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ respectively. Sample variances S_1^2 and S_2^2.

__Testing__: $H_0 : \sigma_1^2 = \sigma_2^2$ *vs* $H_1 : \sigma_1^2 \neq \sigma_2^2$

This test is a formal prerequisite for the two-sample *t* test, for which the assumption $\sigma_1^2 = \sigma_2^2$ is required. In practice, however, a simple plot of the data is often sufficient to justify the assumption – only if the population variances are very different in size is there any problem with the *t* test.

__Test statistic__: $S_1^2 / S_2^2 \sim F_{n_1-1, n_2-1}$ under H_0

We saw in Chapter 8 that $\dfrac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}$, so it follows that if we are testing the hypothesis $\sigma_1^2 = \sigma_2^2$, we can use the test statistic S_1^2 / S_2^2 and compare it with the critical points in the appropriate *F* table.

Question

The average blood pressure for a control group C of 10 patients was 77.0 mmHg. The average blood pressure in a similar group T of 10 patients on a special diet was 75.0 mmHg. Test whether the variances in the two populations can be considered to be equal.

You are given that $\sum\limits_{i=1}^{10} c_i^2 = 59{,}420$ and $\sum\limits_{i=1}^{10} t_i^2 = 56{,}390$.

Solution

We are testing:

$$H_0 : \sigma_T^2 = \sigma_C^2 \quad vs \quad H_1 : \sigma_T^2 \neq \sigma_C^2$$

Assuming that blood pressures are normally distributed, then under H_0, both populations have the same variance, so that:

$$\frac{S_T^2 / \sigma^2}{S_C^2 / \sigma^2} = \frac{S_T^2}{S_C^2} \sim F_{m-1, n-1}$$

where:

$$s_T^2 = \frac{1}{9}\left(56{,}390 - 10 \times 75^2\right) = 15.56$$

$$s_C^2 = \frac{1}{9}\left(59{,}420 - 10 \times 77^2\right) = 14.44$$

The observed value of the test statistic is:

$$\frac{15.56}{14.44} = 1.077$$

This is a two-sided test and our statistic is between 4.026 and $\frac{1}{4.026} = 0.2484$ (the upper and lower 2½% values from the $F_{9,9}$ distribution). So there is insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that there is no difference in the variances of the two populations.

Alternatively, we can see from page 171 of the Tables that the p-value, $P(F_{9,9} > 1.077)$, is greater than 0.1. But since the test is two-sided the p-value is greater than $2 \times 0.1 = 0.2$. Since this is greater than 0.05, we have insufficient evidence to reject H_0 at the 5% level.

This means that we were justified in carrying out the two-sample *t*-test previously which presupposes equal variances.

Note that had we used $\dfrac{s_C^2}{s_T^2} = \dfrac{14.44}{15.56} = 0.9280$, we would have reached the same conclusion.

R can carry out a hypothesis test for the ratio of the variances using `var.test` **or we could use a bootstrap method if the assumptions are not met.**

4.3   Testing the value of the difference between two population proportions

Both one-sided and two-sided tests can easily be performed on the difference between two binomial probabilities – at least for large samples.

<u>Situation</u>:

> n_1 (large) trials with P(success) $= p_1$; observe x_1 successes.
>
> n_2 (large) trials with P(success) $= p_2$; observe x_2 successes.

<u>Testing:</u> $H_0 : p_1 = p_2$.

<u>Test statistic</u> is $\dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n_1} + \dfrac{\hat{p}(1-\hat{p})}{n_2}}} \sim N(0,1)$ **under** H_0

where $\hat{p}_1, \hat{p}_2$ are the maximum likelihood estimates (MLEs) of p_1 and p_2 respectively, (the sample proportions $\dfrac{X_1}{n_1}, \dfrac{X_2}{n_2}$), and $\hat{p}$ is the MLE of the common p under the null hypothesis, which is the overall sample proportion, namely $\dfrac{X_1 + X_2}{n_1 + n_2}$.

In some textbooks an alternative test statistic is used, namely: $\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$.

The denominator in the Core Reading expression is found by pooling the sample proportions, whereas in the alternative version, the values of $\hat{p}_1$ and $\hat{p}_2$ are used separately.

Since the test is approximate, both approximations are valid. In the exam we would advise you to use the version shown in the Core Reading.

Question

In a one-year mortality investigation, 25 of the 100 ninety-year-old males and 20 of the 150 ninety-year-old females present at the start of the investigation died before the end of the year. Assuming that the numbers of deaths follow binomial distributions, test whether there is a difference between male and female mortality rates at this age.

Solution

We are testing:

$$H_0 : q_M = q_F \quad vs \quad H_1 : q_M \neq q_F$$

If X_M and X_F denote the number of deaths among the males and females, m and f are the sample sizes, and $\hat{q}$ the pooled sample proportion, then, under H_0:

$$\frac{\left(\dfrac{X_M}{m} - \dfrac{X_F}{f} \right) - 0}{\sqrt{\dfrac{\hat{q}(1-\hat{q})}{m} + \dfrac{\hat{q}(1-\hat{q})}{f}}} \sim N(0,1)$$

Using the observed values of $m = 100$, $f = 150$, $x_M = 25$, $x_F = 20$, and $\hat{q} = \dfrac{45}{250}$, the value of the test statistic is:

$$\frac{(0.25 - 0.1333)}{\sqrt{(0.18 \times 0.82)/100 + (0.18 \times 0.82)/150}} = 2.35$$

This is greater than 1.960 (the upper 2½% point of the $N(0,1)$ distribution). So we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that male and female mortality rates are different at this age.

Alternatively, using probability values, we have $P(Z > 2.35) = 0.0093$. Since this test is two-sided, the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.0093 = 0.019$. Since this is less than 0.05 we have sufficient evidence to reject H_0 at the 5% level.

Question

A sample of 100 claims on household policies made during the year just ended showed that 62 were due to burglary. A sample of 200 claims made during the previous year had 115 due to burglary.

Test the hypothesis that the underlying proportion of claims that are due to burglary is higher in the second year than in the first.

Solution

We are testing:

$$H_0 : p_1 = p_2 \quad vs \quad H_1 : p_2 > p_1 \quad (ie \; p_1 - p_2 < 0)$$

where p_1 and p_2 are the proportions of claims due to burglaries in the previous and current years respectively.

If N_1 and N_2 denote the numbers of claims due to burglaries in each year, then, under H_0:

$$\frac{(N_1/200 - N_2/100) - 0}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{200} + \dfrac{\hat{p}(1-\hat{p})}{100}}} \doteqdot N(0,1)$$

The observed value of the test statistic is:

$$\frac{(115/200 - 62/100) - 0}{\sqrt{\dfrac{0.59(1-0.59)}{200} + \dfrac{0.59(1-0.59)}{100}}} = -0.747$$

We are carrying out a one-sided test and the value of our statistic is greater than -1.6449 (the lower 5% point of the $N(0,1)$ distribution). So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the proportion of claims due to burglaries in the year just ended is not greater than the proportion in the previous year.

Alternatively, using probability values, we have $P(Z < -0.747) \simeq 0.228$. Since this is greater than 0.05, we have insufficient evidence to reject H_0 at the 5% level.

R can carry out a hypothesis test for the difference in proportions using `prop.test` **with the argument** `correct=FALSE`.

4.4 Testing the value of the difference between two Poisson means

Situation: independent random samples, sizes n_1 and n_2, from **Poi**(λ_1) and **Poi**(λ_2) distributions. Considering the case in which normal approximations can be used – which is so whenever the sample sizes are large and/or the parameter values are large:

Testing: $H_0 : \lambda_1 = \lambda_2$.

Test statistic is $\dfrac{(\hat{\lambda}_1 - \hat{\lambda}_2)}{\sqrt{\dfrac{\hat{\lambda}}{n_1} + \dfrac{\hat{\lambda}}{n_2}}}$ ~ **N(0,1)**

under H_0 where $\hat{\lambda}_1$, $\hat{\lambda}_2$ are the MLEs (the sample means $\bar{X}_1$, $\bar{X}_2$, respectively) and $\hat{\lambda}$ is the MLE of the common λ under the null hypothesis, which is the overall sample mean.

Again, in some textbooks you may see an alternative test statistic, namely: $\dfrac{(\hat{\lambda}_1 - \hat{\lambda}_2)}{\sqrt{\dfrac{\hat{\lambda}_1}{n_1} + \dfrac{\hat{\lambda}_2}{n_2}}} \sim N(0,1)$.

Similarly to the last section, the Core Reading version has a pooled value for the parameter, whereas the alternative version doesn't. Both are valid approximations.

Question

In a one-year investigation of claim frequencies for a particular category of motorists, there were 150 claims from the 500 policyholders aged under 25 and 650 claims from the 4,500 remaining policyholders. Assuming that the number of claims made by individual motorists in each category has a Poisson distribution, test at the 1% level whether the claim frequency is the same for drivers under age 25 and over age 25.

Solution

We are testing:

$$H_0 : \lambda_Y = \lambda_O \quad vs \quad H_1 : \lambda_Y \neq \lambda_O$$

where we are using Y to represent 'young' and O to represent 'old'.

Under H_0 :

$$\frac{(\overline{Y} - \overline{O}) - 0}{\sqrt{\dfrac{\hat{\lambda}}{m} + \dfrac{\hat{\lambda}}{n}}} \doteq N(0,1) \text{ where } m \text{ and } n \text{ are the sample sizes}$$

The observed value of the test statistic is:

$$\frac{0.300 - 0.144}{\sqrt{\dfrac{0.16}{500} + \dfrac{0.16}{4,500}}} = 8.25$$

We are carrying out a two-sided test and our statistic is much greater than $+2.5758$ (the upper ½% point of the $N(0,1)$ distribution). So we easily have sufficient evidence to reject H_0 at the 1% level. Therefore it is reasonable to conclude that the claim frequencies are different for younger and older drivers.

Alternatively, using probability values, we have $P(Z > 8.25) \ll 0.0005\%$. Doubling this (as this test is two-sided) gives a p-value that is still less than 0.001%. So we have sufficient evidence to reject H_0, even at the 0.001% level.

In fact, although the hypotheses weren't posed in this way in the question, we can conclude that the claim frequency is higher for the younger drivers.

R | **There is no built-in function for calculating the above statistic in R. We can use R to calculate the result from scratch or use a bootstrap method. However, R can carry out a hypothesis test for the *ratio* of the two Poisson parameters using** `poisson.test.`

5   Basic test – paired data

In testing for a difference between two population means, the use of independent samples can have a major drawback. Even if a real difference does exist, the variability among the responses within each sample can be large enough to mask it. The random variation within the samples will mask the real difference between the populations from which they come. One way to control this variability external to the issue in question is to use a pair of responses from each subject, and then work with the differences within the pairs. The aim is to remove as far as possible the subject-to-subject variation from the analysis, and thus to 'home in' on any real difference between the populations.

<u>Assumption</u>: differences constitute a random sample from a normal distribution.

<u>Testing</u>: $H_0 : \mu_D \left(= \mu_1 - \mu_2\right) = \delta$

<u>Test statistic</u> is $\dfrac{\bar{D} - \delta}{S_D / \sqrt{n}} \sim t_{n-1}$ under H_0.

We can use $N(0,1)$ for t, and do not require the 'normal' assumption, if n is large.

Question

The average blood pressure $\bar{B}$ for a group of 10 patients was 77.0 mmHg. The average blood pressure $\bar{A}$ after they were put on a special diet was 75.0 mmHg. Carry out a statistical test to assess whether the special diet reduces blood pressure.

You are given that $\displaystyle\sum_{i=1}^{10} (b_i - a_i)^2 = 68.0$.

Solution

We are testing:

$$H_0 : \mu_A = \mu_B \quad vs \quad H_1 : \mu_A < \mu_B \text{ where } A \text{ is after and } B \text{ is before}$$

We can calculate the difference in blood pressure within each pair, *ie* $D_i = A_i - B_i$. If we assume that blood pressures are normally distributed, then under H_0, the D_i's also have a normal distribution. So we can apply a one-sample t test to the D_i's, based on the sample variance s_D^2:

$$\frac{\bar{D} - (\mu_A - \mu_B)}{S_D / \sqrt{n}} \sim t_{n-1}$$

For our samples:

$$\bar{d} = \bar{a} - \bar{b} = 75.0 - 77.0 = -2$$

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}d_i^2 - n\bar{d}^2\right]$$

$$= \frac{1}{9}\left[68.0 - 10(-2.0)^2\right] = 3.111 = 1.764^2$$

So, the observed value of the test statistic is:

$$\frac{75.0 - 77.0}{1.764/\sqrt{10}} = -3.59$$

This is less than -1.833, the lower 5% point of the t_9 distribution. So we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the special diet does reduce blood pressure.

Alternatively, using probability values, we have $P(t_9 < -3.59) \simeq 0.0037$, which is less than 0.05. So we have sufficient evidence to reject H_0 at the 5% level. In fact, we have sufficient evidence to reject it at even the 0.5% level.

Note that when we performed the two-sample t-test earlier, we were unable to reach this conclusion because the reduction was masked by other factors.

Question

In order to increase the efficiency with which employees in a certain organisation can carry out a task, 5 employees are sent on a training course. The time in seconds to carry out the task both before and after the training course is given below for the 5 employees:

	A	B	C	D	E
Before	42	51	37	43	45
After	38	37	32	40	48

Test whether the training course has had the desired effect.

Solution

We are testing:

$$H_0 : \mu_A = \mu_B \quad vs \quad H_1 : \mu_A < \mu_B \ (ie \ \mu_A - \mu_B < 0)$$

where A is 'After' and B is 'Before'.

Taking the differences $d = b - a$ (so that a positive value of d represents an improvement in performance), we have:

4	14	5	3	−3

Applying a one-sample *t*-test to the D values (and assuming that the underlying distributions are normal):

$$\frac{\bar{D} - (\mu_B - \mu_A)}{S_D / \sqrt{n}} \sim t_{n-1}$$

For our sample values:

$$\bar{d} = \frac{23}{5} = 4.6 \text{ and } s_D^2 = \frac{1}{4}\sum(d_i - \bar{d})^2 = \frac{1}{4}(255 - 5 \times 4.6^2) = 6.107^2$$

So the observed value of our test statistic is:

$$\frac{4.6 - 0}{6.107 / \sqrt{5}} = 1.684$$

This is a one-sided test and our statistic is greater than 2.132 (the upper 5% critical value of the t_4 distribution). So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the training course does not increase employees' efficiency.

Alternatively, using probability values, we have $P(t_4 > 1.684) \simeq 0.0874$, which is greater than 0.05. So we have insufficient evidence to reject H_0 at the 5% level.

R can carry out this hypothesis test using `t.test` **with the argument** `paired=TRUE`**.**

6 Tests and confidence intervals

You may have noticed that we've been using some of the same examples in this chapter as in Chapter 8. This is because statistical tests and confidence intervals are very closely related. The methods are basically the same in each case, except that they work opposite ways round. Confidence intervals start from a probability and find a range of parameters associated with this. Statistical tests start with a possible value (or values) for the parameter and associate a probability value with this.

There are very close parallels between the inferential methods for tests and confidence intervals. In many situations there is a direct link between a confidence interval for a parameter and tests of hypothesised values for it.

A confidence interval for θ can be regarded as a set of acceptable hypothetical values for θ, so a value θ_0 contained in the confidence interval should be such that the hypothesis $H_0 : \theta = \theta_0$ will be accepted in a corresponding test. This generally proves to be the case.

In some situations there is a difference between the manner of construction of the confidence interval and that of the construction of the test statistic which is actually used. For example the confidence interval for the difference between two proportions (based on normal approximations) is constructed in a different way from that used for the test statistic in the corresponding test, where an estimate of a common proportion (under H_0) is used. As a result, in this and similar cases there is only an approximate match (albeit a good one) between the confidence interval and the corresponding test.

One useful consequence of this relationship between tests and confidence intervals is that if you have a 95% confidence interval for a parameter, you can immediately apply a 5% test on the value of that parameter simply by observing whether or not the interval contains the proposed value.

Question

A researcher has found 95% confidence intervals for the average daily vitamin C consumption (in milligrams) in three countries. For country A it is $(75, 95)$, for country B it is $(40, 50)$ and for country C it is $(55, 65)$. Comment on whether you think that people are getting sufficient vitamin C in each country. (The recommended daily allowance is 60mg.)

Solution

Country A

The 95% confidence interval is $(75, 95)$, which contains only values above 60. So in a 5% test of $H_0 : \mu = 60$ *vs* $H_1 : \mu > 60$ we reject H_0 and conclude that people are getting more than enough vitamin C.

Country B

The 95% confidence interval is $(40, 50)$, which contains only values below 60. So in a 5% test of $H_0 : \mu = 60$ *vs* $H_1 : \mu < 60$ we reject H_0 and conclude that people are not getting enough vitamin C.

Country C

The 95% confidence interval is $(55,65)$, which contains the value 60. So in a 5% test we cannot reject H_0 and we conclude that people are getting the recommended daily allowance.

Country C

7 Non-parametric tests

The tests we have been considering so far all make assumptions about the distribution of the variables of interest within the population. If these assumptions are not correct, then the level of statistical significance can be affected.

It is possible to devise tests which make no distributional assumptions. Such tests are termed *non-parametric*. They have the advantages of being applicable under conditions in which the tests in the previous sections should not be used.

For example, whilst the two sample *t*-test is robust for departures from normality and equal variances for large samples, it is not appropriate for small samples with a non-normal distribution.

Hence, we need to use a test which doesn't make any distributional assumptions about the data or the test statistic. These tests are called non-parametric tests.

However, some non-parametric tests do not use all the information available. For example, the Signs Test in Subject CS2 uses the signs of the differences between two samples while ignoring their magnitude.

By using only some of the information, the test won't be as accurate.

7.1 Permutation approach

One way of constructing a non-parametric test is to consider all possible permutations of the data subject to some criterion. For example, consider a test of the difference between the means of two independent samples of sizes n_A and n_B. The null hypothesis is that there is no difference in the mean of the two samples.

Label the two samples as *A* and *B*, and consider all possible ways of selecting the $n_A + n_B$ elements on the combined sample such that n_A of them are in category *A* and n_B are in category *B*. Each of these permutations will produce a test statistic (the mean difference), and the mean differences from all possible permutations will provide a distribution of mean differences. Assuming that each permutation is equally likely, we can calculate the *p*-value of the mean difference in the data we have (the permutation actually observed).

The null hypothesis is that the distributions of both categories are the same and hence the means (or any other statistic such as the medians) are the same. In which case, a data point is equally likely to have been assigned to either group.

We can then calculate the *p*-value for our observed statistic of the sampling distribution. This will be the proportion of permutations that lead to test statistics at least as extreme (relative to an alternative hypothesis) as the actual labelling of the data.

If the two samples are stored in vectors `xA` **and** `xB` **then sample R code for obtaining the permutation sampling distribution for the difference in the means is as follows:**

```
results <- c(xA,xB)

index <- 1:length(results)

p<-combn(index,nA)

n<-ncol(p)

dif<-rep(0,n)

for (i in 1:n) {

dif[i]<-mean(results[p[,i]])-mean(results[-p[,i]])

}
```

If our observed statistic is `T` **and our alternative hypothesis is** $H_1 : \mu_1 > \mu_2$ **then the *p*-value is calculated as follows:**

```
length(dif[dif>=T])/length(dif)
```

Alternatively, we can use the `permTS` **function in the** `perm` **package or the** `oneway_test` **function in the** `coin` **package or the** `perm.test` **function in the** `exactRankTests` **package (though this only works if the observed values are integers).**

Similar approaches can be used for tests for paired data where the pairs are kept together. This is equivalent to calculating the permutations of the signs of the differences of the pairs.

The permutation approach is not new, but it has become much more feasible in recent years with the advent of powerful computers, which can undertake the calculation of the many permutations involved in all but the smallest problems.

However, for larger samples the number of permutations grows rapidly and this becomes computationally expensive. Hence, we usually resort to resampling methods.

For example, two groups of size 20 result in 137,846,528,820 combinations. Resampling methods reduce the number of combinations and thus the computation time.

These techniques will be described more fully in the CS1B course.

8   Chi-square tests

These tests are relevant to category or count data. Each sample value falls into one or other of several categories or cells. The test is then based on comparing the frequencies actually observed in the categories/cells with the frequencies expected under some hypothesis, using the test statistic

$$\sum \frac{(f_i - e_i)^2}{e_i}$$

where f_i and e_i are the observed and expected frequencies respectively in the ith category/cell, and the summation is taken over all categories/cells involved. This statistic has, approximately, a chi-square (χ^2) distribution under the hypothesis on the basis of which the expected frequencies were calculated.

The statistic is often written as $\sum \frac{(O_i - E_i)^2}{E_i}$, to show which is the observed value. Note that the values of O_i and E_i should be numbers rather than proportions or percentages.

8.1   Goodness of fit

This is investigating whether it is reasonable to regard a random sample as coming from a specified distribution, *ie* whether a particular model provides a 'good fit' to the data.

Degrees of freedom

Suppose there are *k* cells, so *k* terms in the summation which produces the statistic, and that the sample size is $n = \sum f_i$. The expected frequencies also sum to *n*, so knowing any $k - 1$ of them automatically gives you the last one. There is a dependence built in to the *k* terms which are added up to produce the statistic – and this is the reason why the degrees of freedom of the basic statistic is $k - 1$ and not *k*.

Further, for each parameter of the distribution specified by the null hypothesis which must be estimated from the observed data, another degree of dependence is introduced in the expected frequencies – for each parameter estimated another degree of freedom is lost. The theory behind this assumes that the maximum likelihood estimators are used. So the number of degrees of freedom is reduced by the number of parameters estimated from the observed data.

The 'accuracy' of the chi-square approximation

The test statistic is only approximately, not exactly, distributed as χ^2. The presence of the expected frequencies e_i in the denominators of the terms to be added up is important — dividing by very small e_i values causes the resulting terms to be somewhat large and 'erratic', and the tail of the distribution of the statistic may not match that of the χ^2 distribution very well. So, in practice, it is best not to have too many small e_i values, which can be done by combining cells and suffering the consequent loss of information/degrees of freedom. The most common recommendation is not to use any e_i which is less than 5. (However, the statistic is more robust than that and in practice a less conservative

approach, such as ensuring that all e_i are greater than 1 and that not more than 20% of them are less than 5, may be taken.)

Question

In testing whether a die is fair, a suitable model is:

$$P(X = i) = \frac{1}{6}, \quad i = 1,2,3,4,5,6 \text{ where } X \text{ is the number thrown}$$

and the hypotheses may be:

H_0 : Number thrown has the distribution specified in the model

H_1 : Number thrown does not have the distribution specified in the model

If the die is thrown 300 times, with the following results,

x:	1	2	3	4	5	6
f_i :	43	56	54	47	41	59

Carry out a χ^2 test to determine whether the data comes from a fair die.

Solution

Under H_0, $300 \times \frac{1}{6} = 50$ occurrences of each face of the die would be expected, so $e_i = 50$, $i = 1,2,3,4,5,6$. The values of $(f_i - e_i)$, the differences between observed and expected frequencies, are then:

$$-7, 6, 4, -3, -9, 9$$

which of course sum to zero.

The value of the test statistic is then:

$$\frac{49}{50} + \frac{36}{50} + \frac{16}{50} + \frac{9}{50} + \frac{81}{50} + \frac{81}{50} = \frac{272}{50} = 5.44$$

In this illustration, with 6 cells and a fully specified model (no parameters to estimate), the distribution of the test statistic under H_0 is χ_5^2.

This is a one-sided test. We reject H_0 for large values of the statistic (*ie* when the observed and expected values and very different). Since 5.44 is less than 11.07 (the upper 5% point of the χ_5^2 distribution) we have insufficient evidence to reject H_0 at the 5% level.

Alternatively, the **P value is** $P(\chi_5^2 > 5.44)$. **The** probability **tables** (on page 165) **show that** $P(\chi_5^2 > 5.5) = 0.358$, **so** $P(\chi_5^2 > 5.44)$ **is about 0.36. Note also that a** χ_5^2 **variable has mean 5, so we have observed a value much in line with what is expected under the model.**

We have no evidence that the die is not fair. H_0 can stand.

Question

The table below shows the causes of death in elderly men derived from a study in the 1970s. Carry out a chi-square test to determine whether these percentages can still be considered to provide an accurate description of causes of death in 2000.

Cause of death	Proportion of deaths in 1975	Number of deaths in 2000
Cancer	8%	286
Heart disease	22%	805
Other circulatory disease	40%	1,548
Respiratory diseases	19%	755
Other causes	11%	464

Solution

We are testing:

H_0 : the causes of death in 2000 conform to the percentages shown

vs H_1 : the causes of death in 2000 do not conform to the percentages shown

Under H_0 :

$$\sum \frac{(O_i - E_i)^2}{E_i} \sim \chi_f^2$$

where f is the number of degrees of freedom.

The expected values for each category are calculated by multiplying the total number of deaths by the percentage for that category. For example the expected number of deaths from heart disease is $0.22 \times 3,858 = 848.8$.

The table below shows the observed and expected figures is (where $C_i = (O_i - E_i)^2 / E_i$):

Cause of death	Actual, O_i	Expected, E_i	C_i
Cancer	286	308.6	1.66
Heart disease	805	848.8	2.26
Other circulatory diseases	1,548	1,543.2	0.01
Respiratory disease	755	733.0	0.66
Other causes	464	424.4	3.7
Total	3,858	3,858	8.29

There are no small groups. The value of the chi-square statistic is 8.29.

There are 5 categories. The E_i's were calculated from the total number of observations. We haven't estimated any parameters. So the number of degrees of freedom is $5 - 1 = 4$.

Chi-square goodness of fit tests are one-sided tests. Our observed value of the test statistic is less than 9.488, the upper 5% point of the χ_4^2 distribution. So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that there has been no change in the pattern of causes of death.

Alternatively, using probability values, we have $P(\chi_4^2 > 8.29) \simeq 0.0819$, which is greater than 0.05. So we have insufficient evidence to reject H_0 at the 5% level.

Question

The numbers of claims made last year by individual motor insurance policyholders were:

Number of claims	0	1	2	3	4+
Number of policyholders	2,962	382	47	25	4

Carry out a chi-square test to determine whether these frequencies can be considered to conform to a Poisson distribution.

Solution

We are testing:

H_0 : the number of claims conform to a Poisson distribution

vs H_1 : the number of claims don't conform to a Poisson distribution

Under H_0:

$$\sum \frac{(O_i - E_i)^2}{E_i} \sim \chi_f^2$$

where f is the number of degrees of freedom.

To find the expected numbers, we must estimate the unknown mean of the Poisson distribution. The MLE of the mean of a Poisson distribution is the mean number of claims. If we assume that no policyholders made more than 4 claims, this is:

$$\hat{\lambda} = \frac{2,962 \times 0 + 382 \times 1 + 47 \times 2 + 25 \times 3 + 4 \times 4}{3,420} = 0.1658$$

The expected values are found by applying the Poisson probabilities calculated using this value for the parameter to the total observed number of claims *ie* 3,420.

The table showing the observed and expected figures is:

Number of claims	Actual	Expected
0	2962	2,897.5
1	382	480.4
2	47	39.8
3	25	2.2
4 or more	4	0.1
Total	3,420	3,420

We calculate the last expected figure by subtraction.

The expected numbers in the last two groups are very small, so we need to combine the last three groups to form a '2 or more' group.

The value of the chi-square statistic is:

$$\chi^2 = \frac{(2,962 - 2,897.5)^2}{2,897.5} + \frac{(382 - 480.4)^2}{480.4} + \frac{(76 - 42.1)^2}{42.1} = 48.89$$

There are now 3 groups. The E_i's were calculated from the total number of observations. We have estimated one parameter. So the number of degrees of freedom is $3 - 1 - 1 = 1$.

We are carrying out a one-sided test. Our observed value of the test statistic far exceeds 7.879, the upper 0.5% point of the χ_1^2 distribution. So we have sufficient evidence to reject H_0 at the 0.5% level. Therefore it is reasonable to conclude that a Poisson model does *not* provide a good model for the number of claims.

Question

On a particular run of a process which bottles a drink, it is thought that the cleansing process of the bottles has partially failed. The bottles have been boxed into crates, each containing six bottles. It is thought that each bottle, independently of all others, has the same chance of containing impurities.

A survey has been conducted, and each bottle in a random sample of 200 crates has been tested for impurities. The table below gives the numbers of crates in the sample which had the respective number of bottles which contained impurities:

Number of impure bottles:	0	1	2	3	4	5	6
Number of crates:	38	70	58	25	6	2	1

Test the goodness of fit of a binomial distribution to these observations.

Solution

We first need an estimate of θ, the proportion of bottles containing impurities. We get this by finding the MLE for θ based on the random sample.

Perhaps the simplest way to calculate the MLE, $\hat{\theta}$, is:

$$\frac{\text{total number of successes (impure bottles)}}{\text{total number of bottles}} = \frac{301}{1,200} = 0.25083333$$

Alternatively, you might see that $\hat{\theta} = \dfrac{\overline{x}}{6}$, where $\overline{x}$ is the mean number of impure bottles per crate. From the data, $\overline{x} = \dfrac{301}{200} = 1.505$, so, given that there are six bottles in each crate,

$\hat{\theta} = \dfrac{\overline{x}}{6} = 0.25083333$.

If you can't spot either of these immediately then you can derive the MLE as follows:

Let the number of bottles with impurities in each crate in a random sample of 200 crates be x_1, $x_2, ..., x_{200}$. Each x_i is an observation from a $Bin(6, \theta)$ distribution, and so the likelihood function for θ is:

$$L(\theta) = \binom{6}{x_1} \theta^{x_1} (1-\theta)^{6-x_1} \dots \binom{6}{x_{200}} \theta^{x_{200}} (1-\theta)^{6-x_{200}}$$

$$= \text{constant} \times \theta^{\sum x_i} (1-\theta)^{1,200-\sum x_i}$$

Taking logs:

$$\log L = \sum x_i \log \theta + (1,200 - \sum x_i) \log(1-\theta)$$

Differentiating with respect to θ and setting the result equal to zero:

$$\frac{\partial}{\partial \theta} \log L = \frac{\sum x_i}{\theta} - \frac{1,200 - \sum x_i}{1-\theta} = 0$$

Solving this we get $\hat{\theta} = \dfrac{\sum x_i}{1,200} = \dfrac{301}{1,200} = 0.25083$.

We can now calculate the expected frequencies. We calculate the probabilities from a $Bin(6, 0.25083)$ distribution, and multiply each probability by 200:

Number of bottles with impurities	Observed	Expected
0	38	35.36
1	70	71.03
2	58	59.46
3	25	26.54
4 or more	9	7.61
Total	200	200

Note that we have combined the last three groups since the expected frequencies are small. In fact we anticipated that the last two groups were going to have small expected numbers and calculated the expected number for the '4 or more' group by subtraction from 200.

The observed value of the chi square statistic is:

$$\chi^2 = \frac{(38-35.36)^2}{35.36} + \frac{(70-71.03)^2}{71.03} + \frac{(58-59.46)^2}{59.46} + \frac{(25-26.54)^2}{26.54} + \frac{(9-7.61)^2}{7.61}$$

$$= 0.59$$

There are now 5 groups. The E_i's were calculated from the total number of observations. We have estimated one parameter. So the number of degrees of freedom is $5-1-1=3$.

We are carrying out a one-sided test. Our observed value of the test statistic has a *p*-value of about 90%. So we have insufficient evidence to reject H_0 at the 90% level. Therefore it is reasonable to conclude that the underlying distribution is binomial.

Indeed the fit is almost 'too good' – the resulting value of the test statistic is suspiciously small.

R can carry out a χ^2 goodness-of-fit test using:

```
chisq.test(<observed freq>, p=<expected probabilities>)
```

8.2   Contingency tables

A contingency table is a two-way table of counts obtained when sample items (people, companies, policies, claims *etc*) are classified according to two category variables. The question of interest is whether the two classification criteria are independent.

H_0: the two classification criteria are independent.

The simple rule for calculating the expected frequency for any cell is then:

$$\frac{\text{row total} \times \text{column total}}{\text{table total}}$$

(*ie* the proportion of data in row *i* is $f_{i.}/f$ so if the criteria are independent, the number expected in cell (i, j) is $(f_{i.}/f) \times f_{.j}$.)

The degrees of freedom associated with a table with *r* rows and *c* columns is:

$$(rc - 1) - (r - 1) - (c - 1) = (r - 1)(c - 1)$$

since the column totals and row totals reduce the number of degrees of freedom.

An important use of this method is with a table of dimension $2 \times c$ (or $r \times 2$) which gives a test for differences among 2 or more population proportions.

Question

For each of three insurance companies, A, B, and C, a random sample of non-life policies of a particular kind is examined. It turns out that a claim (or claims) have arisen in the past year in 23% of the sampled policies for A, in 28% of those for B, and in 20% of those for C.

Test for differences in the underlying proportions of policies of this kind which have given rise to claims in the past year among the three companies in the two situations:

(a)   the sample sizes were 100, 100, and 200 respectively

(b)   the sample sizes were 300, 300, and 600 respectively.

Comment briefly on your results.

Solution

H_0 : population proportions are all equal

H_1 : population proportions are not all equal

(a)   Observed frequencies:

	A	B	C	
✓	23	28	40	91
✗	77	72	160	309
	100	100	200	400

Expected frequencies under H_0 :

	A	B	C	
✓	22.75	22.75	45.50	91
✗	77.25	77.25	154.50	309
	100	100	200	400

Values of $f_i - e_i$:   0.25   5.25   −5.5

                                    −0.25   −5.25   5.5

So:

$$\chi^2 = \frac{0.25^2}{22.75} + \frac{5.25^2}{22.75} + \frac{5.5^2}{45.50} + \frac{0.25^2}{77.25} + \frac{5.25^2}{77.25} + \frac{5.5^2}{154.50}$$

$$= 0.003 + 1.212 + 0.665 + 0.001 + 0.357 + 0.196$$

$$= 2.43$$

on 2df.

where df stands for 'degrees of freedom'.

This is an unremarkable value for χ^2_2 – we have no evidence against H_0, which can stand. No differences among the population proportions have been detected.

(b)   The sample sizes are increased by a factor of 3, but the same percentages with claims as in (a) are assumed. f_i, e_i and $(f_i - e_i)$ all increase by a factor of 3 – so each component of χ^2, and the resulting value, also increase by a factor of 3. So now $\chi^2 = 7.3$.

p-value = $P(\chi_2^2 > 7.3)$, which is just a bit bigger than 0.025.

There is quite strong evidence against H_0 – we conclude that the population proportions are not all equal (p-value about 0.03).

Comments: The observed sample proportions 23%, 28%, and 20% are not 'significantly different' when based on sample sizes of 100, 100, and 200, but *are* 'significantly different' when based on sample sizes which are considerably bigger (300, 300, and 600).

Question

In an investigation into the effectiveness of car seat belts, 292 accident victims were classified according to the severity of their injuries and whether they were wearing a seat belt at the time of the accident. The results were as follows:

	Wearing a seat belt	Not wearing a seat belt
Death	3	47
Severe injury	78	32
Minor injury	103	29

Determine whether the severity of injuries sustained is dependent on whether the victims are wearing a seat belt.

Solution

The hypotheses are:

H_0:   severity of injuries is independent of wearing a seatbelt

H_1:   severity of injuries is not independent of wearing a seatbelt

We can calculate the expected frequencies in each category by multiplying the row and column totals, and dividing by the overall total:

Expected	Wearing a seatbelt	Not wearing a seatbelt
Death	31.5	18.5
Severe injury	69.3	40.7
Minor injury	83.2	48.8

For example, $\frac{184 \times 50}{292} = 31.507$. So we can now calculate the value of the chi square statistic:

$$\chi^2 = \frac{(3-31.5)^2}{31.5} + \cdots + \frac{(29-48.8)^2}{48.8} = 85.39$$

The number of degrees of freedom is $(3-1)(2-1) = 2$.

We are carrying out a one-sided test. Our observed value of the test statistic is far in excess of 10.60, the upper 0.5% point of the χ_2^2 distribution. In fact we could have stopped after working out the first term in the χ^2 value which is already 25.79. So we have sufficient evidence to reject H_0 at the 0.5% level. Therefore it is reasonable to conclude that the level of injury is almost certainly dependent on whether the victim is wearing a seatbelt.

Question

The table below shows the numbers of births during one month at a particular hospital classified according to whether a particular medical characteristic was or wasn't present during childbirth. Determine whether the presence of this characteristic is dependent on the age of the mother.

Age of mother	<20	21-25	26-30	31-35	36+	Total
Characteristic present	10	12	9	4	3	38
Characteristic absent	5	51	38	25	5	124
Total	15	63	47	29	8	162

Solution

The hypotheses are:

H_0 :   the characteristic is independent of the mother's age

H_1 :   the characteristic is not independent of the mother's age

The observed frequencies were:

Age of mother	<20	21-25	26-30	31-35	36+	Total
Characteristic present	10	12	9	4	3	38
Characteristic absent	5	51	38	25	5	124
Total	15	63	47	29	8	162

We can calculate the expected frequencies in each category by multiplying the row and column totals, and dividing by 162:

Age of mother	<20	21-25	26-30	31-35	36+	Total
Characteristic present	3.52	14.78	11.02	6.80	1.88	38
Characteristic absent	11.48	48.22	35.98	22.20	6.12	124
Total	15	63	47	29	8	162

Note that in contingency tables the totals are always the same in the observed and expected tables. This means that in a table with only 2 rows, if you calculate the entries in one of the rows first, you can work out the entries in the other row by subtraction.

Two cells out of 10 cells have expected frequencies less than 5. Since this is not *more* than 20% we can use the table as it is.

So we can now calculate the value of the chi square statistic.

$$\chi^2 = \frac{(10-3.5)^2}{3.5} + \cdots + \frac{(5-6.1)^2}{6.1} = 19.2$$

The number of degrees of freedom is $(5-1)(2-1) = 4$.

We are carrying out a one-sided test. Our observed value of the test statistic exceeds 18.47, the upper 0.1% point of the χ_4^2 distribution. So we have sufficient evidence to reject H_0 at the 0.1% level. Therefore it is reasonable to conclude that the characteristic *is* dependent of the mother's age.

If we decided to combine cells because of the expected values being less than 5, we could have done this by combining adjacent groups as follows:

Age of mother	≤25	26-30	31+	Total
Characteristic present	22	9	7	38
Characteristic absent	56	38	30	124
Total	78	47	37	162

and the expected values are:

Age of mother	≤25	26-30	31+	Total
Characteristic present	18.30	11.02	8.68	38
Characteristic absent	59.70	35.98	28.32	124
Total	78	47	37	162

So we can now calculate the value of the chi square statistic.

$$\chi^2 = \frac{(22-18.30)^2}{18.30} + \cdots + \frac{(30-28.32)^2}{28.32} = 1.89$$

The number of degrees of freedom is $(3-1)(2-1) = 2$.

*We are carrying out a one-sided test. Our observed value of the test statistic does not exceed 5.991, the upper 5% point of the χ_2^2 distribution. So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the characteristic is **not** dependent of the mother's age.*

The results are so different because of the effect of the small expected values.

R can carry out a χ^2 contingency table test using `chisq.test(<table>)`. **Since this automatically applies a continuity correction for 2×2 tables we would need to set the argument** `correct=FALSE` **if we wished to prevent this.**

8.3   Fisher's exact test

A non-parametric permutation approach to contingency tables was devised more than 80 years ago by the great statistician R.A. Fisher. Consider two categorical variables X and Y, each with two categories, X_1, X_2, Y_1 and Y_2. Suppose we have data for n observations, and that of these

n_{X_1} **are in category X_1 on variable X,**

n_{X_2} **are in category X_2 on variable X,**

n_{Y_1} **are in category Y_1 on variable Y, and**

n_{Y_2} **are in category Y_2 on variable Y.**

These data can be represented in a 2×2 contingency table as shown below.

Variable	X_1	X_2	
Y_1			n_{Y_1}
Y_2			n_{Y_2}
	n_{X_1}	n_{X_2}	n

Fisher proposed testing the association between the two categorical variables by working out the probability of each possible permutation of values in the shaded cells consistent with the *marginal totals* n_{X_1}, n_{X_2}, n_{Y_1} and n_{Y_2}. Then, under the null hypothesis of no association, the distribution of ways of allocating the data to the four shaded cells is hypergeometric. This means that, if the number of individuals which have the value X_1 on variable X and the value Y_1 on variable Y is $n_{X_1Y_1}$, then the probability of obtaining this number is given by:

$$P(n_{X_1Y_1}) = \frac{\binom{n_{X_1}}{n_{X_1Y_1}}\binom{n_{X_2}}{n_{Y_1} - n_{X_1Y_1}}}{\binom{n}{n_{Y_1}}} \qquad \text{for } n_{X_1Y_1} \le n_{X_1}, n_{Y_1}$$

The stronger the association between X and Y the more heavily the observations should be concentrated in either cells $\{Y_1, X_1\}$ and $\{Y_2, X_2\}$ or $\{Y_1, X_2\}$ and $\{Y_2, X_1\}$ (*ie* in two opposite corners of the contingency table).

Consider a sample of 10 people, 6 men and 4 women. Of these 3 are colour blind:

	Colour blind	Not	
Male	2	4	6
Female	1	3	4
	3	7	10

Using the formula above the probability of observing 2 colour-blind men from this sample would be:

$$\frac{\binom{3}{2}\binom{7}{4}}{\binom{10}{6}} = \frac{3 \times 35}{210} = \frac{1}{2}$$

$\binom{10}{6}$ gives us the total number of ways of choosing the 6 men from the 10 people.

$\dbinom{3}{2}\dbinom{7}{4}$ gives us the number of ways of choosing 2 men from the 3 colour blind people and the 4 men from the 7 non-colour blind people.

Hence this expression gives us the probability of observing 2 colour blind men from this group of 10 people.

A test can then be constructed by considering the probability of getting a distribution with the observed or a more extreme concentration of observations in two opposite corners.

The only four outcomes which produce a 2×2 table with the same row and column totals are:

3	3
0	4

2	4
1	3

1	5
2	2

0	6
3	1

Using the formula we can calculate the probabilities of each of these outcomes:

$$\frac{\dbinom{3}{3}\dbinom{7}{3}}{\dbinom{10}{6}} \qquad \frac{\dbinom{3}{2}\dbinom{7}{4}}{\dbinom{10}{6}} \qquad \frac{\dbinom{3}{1}\dbinom{7}{5}}{\dbinom{10}{6}} \qquad \frac{\dbinom{3}{0}\dbinom{7}{6}}{\dbinom{10}{6}}$$

These are $\frac{1}{6}$, $\frac{1}{2}$, $\frac{3}{10}$ and $\frac{1}{30}$, respectively.

For a one-tailed test it suffices to consider only distributions which are extreme in the same direction as the observed table, whereas for a two-tailed test distributions should be considered which are extreme in the opposite direction (this can cause complications with small tables as the sampling distribution is not symmetrical).

At the 5% level of significance we should reject the null hypothesis of no association if the probability of getting a distribution which is the same or more extreme than that observed is less than 0.05.

In our example, the probability of observing this result or more extreme (*ie* 2 or more men with colour blindness) would be $\frac{1}{6}+\frac{1}{2}=\frac{2}{3}$. This is not less than 5% - it is actually very likely and so based on these results we would conclude that gender and colour blindness are independent.

On the other hand, if we were to find that our result was rare, we would conclude that the result is not just due to chance, there is some connection between the variables.

Fisher's test was extended to a general $R \times C$ table by Freeman and Halton.

We chose an example with a very small sample as otherwise there would be many combinations which will be time consuming on a piece of paper. However, this test is no bother for a computer.

R can carry out Fisher's Exact Test using the command `fisher.test(<table>)`.

Chapter 9 Summary

Statistical tests can be used to test assertions about populations.

The process of statistical testing involves setting up a null hypothesis and an alternative hypothesis, calculating a test statistic and using this to determine a *p*-value.

The probability of a Type I error is the probability of rejecting H_0 when it is true. This is also called the size (or level) of the test. The probability of a Type II error is the probability of accepting H_0 when it is false. The power of a test is the probability of rejecting H_0 when it is false.

The 'best' test can be found using the likelihood ratio criterion. This leads to the tests detailed overleaf.

The test for two normal means (unknown variances) requires that the variances are the same and uses the pooled sample variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

χ^2 tests can be carried out to test for goodness of fit or to test whether two factors are independent (using contingency tables).

The statistic is $\sum \frac{(O_i - E_i)^2}{E_i}$.

To find the number of degrees of freedom for the goodness of fit test, take the number of cells, subtract 1 if the total of the observed figures has been used in the calculation of the expected numbers (which is usually the case), and then subtract the number of parameters estimated.

To find the number of degrees of freedom for a contingency table calculate $(r - 1)(c - 1)$. If the expected numbers in some cells are small, these should be grouped. One degree of freedom is lost for each cell that is 'lost'.

One-sample normal distribution

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \sigma^2 \text{ known} \qquad\qquad \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \sigma^2 \text{ unknown}$$

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Two-sample normal distribution

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1) \qquad \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1 + n_2 - 2}$$

$$\qquad\qquad \sigma^2 \text{ known} \qquad\qquad\qquad\qquad \sigma^2 \text{ unknown}$$

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1 - 1, n_2 - 1}$$

One-sample binomial

$$\frac{\hat{p} - p_0}{\sqrt{p_0 q_0/n}} \stackrel{.}{\sim} N(0,1) \quad \text{or} \quad \frac{X - np_0}{\sqrt{np_0 q_0}} \stackrel{.}{\sim} N(0,1) \qquad \text{with continuity correction}$$

Two-sample binomial

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}\hat{q}}{n_1} + \dfrac{\hat{p}\hat{q}}{n_2}}} \stackrel{.}{\sim} N(0,1) \qquad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \text{ is the overall sample proportion}$$

One-sample Poisson

$$\frac{\overline{x} - \lambda_0}{\sqrt{\lambda_0/n}} \stackrel{.}{\sim} N(0,1) \quad \text{or} \quad \frac{\sum X - n\lambda_0}{\sqrt{n\lambda_0}} \stackrel{.}{\sim} N(0,1) \qquad \text{with continuity correction}$$

Two-sample Poisson

$$\frac{(\hat{\lambda}_1 - \hat{\lambda}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\dfrac{\hat{\lambda}}{} + \dfrac{\hat{\lambda}}{}}} \stackrel{.}{\sim} N(0,1) \qquad \hat{\lambda} = \frac{n_1 \hat{\lambda}_1 + n_2 \hat{\lambda}_2}{n_1 + n_2} \text{ is the overall sample mean}$$

Chapter 9 Practice Questions

9.1   A statistical test is used to determine whether or not an anti-smoking campaign carried out 5 years ago has led to a significant reduction in the mean number of smoking related illnesses. If the probability value of the test statistic is 7%, determine the conclusion for a test of size:

     (i)   10%

     (ii)   5%.

9.2   A random sample, $x_1, \ldots, x_{10}$, from a normal population gives the following values:

$$9.5 \quad 18.2 \quad 4.69 \quad 3.76 \quad 14.2 \quad 17.13 \quad 15.69 \quad 13.9 \quad 15.7 \quad 7.42$$

$$\sum x_i = 120.19 \qquad \sum x_i^2 = 1{,}693.6331$$

     (i)   Test at the 5% level whether the mean of the whole population is 15 if the variance is:

          (a)   unknown

          (b)   20.

     (ii)   Test at the 5% level whether the population variance is 20.

9.3   A professional gambler has said: '*Flipping* a coin into the air is fair, since the coin rotates about a horizontal axis, and it is equally likely to be either way up when it first clips the ground. So a flicked coin is equally likely to land showing heads or tails. However, *spinning* a coin on a table is not fair, since the coin rotates about a vertical axis, and there is a systematic bias causing it to tilt towards the side where the embossed pattern is heavier. In fact, when a new coin is spun, it is more than twice as likely to land showing tails as it is to land showing heads.'

     After hearing this, you carried out an experiment, spinning a new coin 25 times on a polished table, and found that it showed tails 18 times. Comment on whether the results of your experiment support the gambler's claims about the probabilities when a coin is spun.

9.4   The sample variances of two independent samples from normal populations A and B, which have the same population variance, are $s_A^2 = 12.4$ and $s_B^2 = 25.8$. If the sample sizes are $n_A = 10$ and $n_B = 5$ and the sample means are found to differ by 4.5, test whether the population means are equal.

9.5   Two populations X and Y are known to have the same variance, but the precise distributions are not known. A sample of 5 values from population X and 10 values from population Y had sample variances of $s_X^2 = 47.0$ and $s_Y^2 = 12.6$. Apply a statistical test based on the F distribution to assess whether both populations can be considered to be normally distributed.

9.6   Determine the form of the best test of $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu = \mu_1$, where $\mu_1 > \mu_0$, assuming the distribution of the underlying population is $N(\mu, \sigma^2)$, based on a sample of size n.

9.7   The lengths of a random sample of 12 worms of a particular species have a mean of 8.54 cm and a standard deviation of 2.97 cm. Let μ denote the mean length of a worm of this species. It is required to test:

Exam style

$$H_0 : \mu = 7\text{cm} \quad vs \quad H_1 : \mu \neq 7\text{cm}$$

Assuming that the lengths of worms are normally distributed, calculate the probability-value of these sample results.   [3]

9.8   A general insurance company is debating introducing a new screening programme to reduce the claim amounts that it needs to pay out. The programme consists of a much more detailed application form that takes longer for the new client department to process. The screening is applied to a test group of clients as a trial whilst other clients continue to fill in the old application form. It can be assumed that claim payments follow a normal distribution.

Exam style

The claim payments data for samples of the two groups of clients are (in £100 per year):

Without screening	24.5	21.7	45.2	15.9	23.7	34.2	29.3	21.1	23.5	28.3
With screening	22.4	21.2	36.3	15.7	21.5	7.3	12.8	21.2	23.9	18.4

(i)   Test the hypothesis that the new screening programme reduces the mean claim amount. [5]

(ii)   Formally test the assumption of equal variances required in part (i).   [3]
                                                                                  [Total 8]

9.9   An environmentalist is investigating the possibility that oestrogenic chemicals are leading to a particular type of deformity in a species of amphibians living in a lake. The usual proportion of deformed animals living in unpolluted water is 0.5%. In a sample of 1,000 animals examined, 15 were found to have deformities.

Exam style

(i)   Test whether this provides evidence of the presence of harmful chemicals in the lake.   [3]

Following an extensive campaign to reduce these chemicals in the lake a further sample of 800 animals was examined and 10 were found to have deformities.

(ii)   Test whether there has been a significant reduction in the proportion of deformed animals in the lake.   [3]
                                                                                  [Total 6]

9.10   The total claim amounts (in £m) for home and car insurance over a year for similar sized companies are collected by an independent advisor:

Exam style

Home	13.3	19.2	12.9	15.8	17.6

Car	14.3	21.0	12.8	17.4	22.8

(i)   Test whether the mean home and car claims are equal. State clearly your probability value.   [5]

It was subsequently discovered that the results were actually 5 consecutive years from the same company.

(ii)   Carry out an appropriate test of whether the mean home and car claims are equal.   [3]

[Total 8]

9.11   A random variable X is believed to have probability density function, $f(x)$, where:

Exam style

$$f(x) = 3\lambda^3 (\lambda + x)^{-4} \qquad x > 0$$

In order to test the null hypothesis $\lambda = 50$ against the alternative hypothesis $\lambda = 60$, a single value is observed. If this value is greater than 93.5, H_0 is rejected.

(i)   Calculate the size of the test.   [2]

(ii)   Calculate the power of the test.   [2]

[Total 4]

9.12   In an extrasensory perception experiment carried out in a live television interview, the interviewee who claimed to have extrasensory powers was required to identify the pattern on each of 10 cards, which had been randomly assigned with one of five different patterns. The cards were visible only to the audience who were asked to 'transmit' the patterns to the interviewee. When the interviewee failed to identify any of the cards correctly, she claimed that this was clear proof of the existence of ESP, since there was a strong mind in the audience who was willing her to get the answers wrong.

Exam style

(i)   State the hypotheses implied by the interviewee's conclusion and carry out a 5% test on this basis. Comment on your answer.   [3]

(ii)   State precisely the hypotheses that the interviewer could have specified before the experiment to prevent the interviewee from 'cheating' in this way, and determine the number of cards that would have to be identified correctly to demonstrate the existence of ESP at the 5% level.   [2]

[Total 5]

9.13   An insurer believes that the distribution of the number of claims on a particular type of policy is
        binomial with parameters $n = 3$ and p. A random sample of the number of claims on 153 policies
        revealed the following results:

Number of claims	0	1	2	3
Number of policies	60	75	16	2

(i)   Derive the maximum likelihood estimate of p.   [4]

(ii)   Carry out a goodness of fit test for the binomial model specified in part (i) for the number of
        claims on each policy.   [5]

[Total 9]

9.14   In an investigation into a patient's red corpuscle count, the number of such corpuscles appearing in
        each of 400 cells of a haemocytometer was counted. The results were as follows:

No. of red blood corpuscles	0	1	2	3	4	5	6	7	8
No. of cells	40	66	93	94	62	25	14	5	1

It is thought that a Poisson distribution with mean μ provides an appropriate model for this
situation.

(i)   (a)   Estimate μ, the Poisson parameter.

        (b)   Test the fit of the Poisson model.   [8]

(ii)   For a healthy person, the mean count per cell is known to be equal to 3. For a patient with
        certain types of anaemia, the number of red blood corpuscles is known to be lower than
        this.

        Test whether this patient has one of these types of anaemia.   [3]

[Total 11]

9.15   In a recent study investigating a possible genetic link between individuals' susceptibility to
        developing symptoms of AIDS, 549 men who had been diagnosed HIV positive were classified
        according to whether they carried two particular alleles (DRB1*0702 and DQA1*0201). The
        results were as follows:

Condition of individual	Free of symptoms	Early symptoms	Suffering from AIDS	Total
Alleles present	24	7	17	48
Alleles absent	98	93	310	501
Total	122	100	327	549

Use these results to test whether there is an association between the presence of the alleles and the classification into the three AIDS statuses.   [5]

9.16   Insurance claims (in £) arriving at an office over the last month have been analysed. The results are as follows:

Claim size, c	$0 \leq c < 500$	$500 \leq c < 1,000$	$1,000 \leq c < 2,500$	over 2,500
No. of claims	75	51	22	5

(i)   Assuming that the maximum claim amount is £10,000:

(a)   calculate the sample mean of the data

(b)   test at the 5% level whether an exponential distribution with parameter λ is an appropriate distribution for the claim sizes. You should estimate the value of λ using the method of moments.   [6]

(ii)   An actuary decides to investigate whether claim sizes vary according to the postcode of residence of the claimant. She splits the data into the three different postcodes observed. The results for the first two postcodes are given below:

Postcode 1:

Claim size, c	$0 \leq c < 500$	$500 \leq c < 1,000$	$1,000 \leq c < 2,500$	over 2,500
No. of claims	23	14	7	3

Postcode 2:

Claim size, c	$0 \leq c < 500$	$500 \leq c < 1,000$	$1,000 \leq c < 2,500$	over 2,500
No. of claims	30	16	11	1

Test at the 5% level whether claim sizes are independent of the postcodes.   [8]

[Total 14]

9.17   A politician has said: 'A recent study in a particular area showed that 25% of the 400 teenagers who were living in single-parent families had been in trouble with the police, compared with only 20% of the 1,200 teenagers who were living in two-parent families. Our aim is to reduce the number of single-parent families in order to reduce the crime rates during the next decade.'

(i)   Carry out a contingency table test to assess whether there is a significant association between living in a single-parent family and getting into trouble with the police. Use a 5% level of significance.   [5]

(ii)   Hence, comment on the politician's statement.   [1]

[Total 6]

9.18   A certain species of plant produces flowers which are either red, white or pink. It also produces
        leaves which may be either plain or variegated. For a sample of 500 plants, the distribution of
        flower colour and leaf type was:

	Red	White	Pink
Plain	97	42	77
Variegated	105	148	31

(i)   Test whether these results indicate any association between flower colour and leaf
        type.   [6]

(ii)   A genetic model suggests that the proportions of each combination should be as follows:

	Red	White	Pink
Plain	q	$q/2$	$(1-3q)/2$
Variegated	q	$3q/2$	$(1-5q)/2$

        where q $(0 < q < 1/5)$ is an unknown parameter.

        (a)   Show that the maximum likelihood estimate for q is 0.181.

        (b)   Test whether this genetic model fits the data well.   [12]

(iii)   Comment briefly on your conclusions.   [3]
                                                                                          [Total 21]

9.19   A particular area in a town suffers a high burglary rate. A sample of 100 streets is taken, and in
        each of the sampled streets, a sample of six similar houses is taken. The table below shows the
        number of sampled houses, which have had burglaries during the last six months.

No. of houses burgled	x	0	1	2	3	4	5	6
No. of streets	f	39	38	18	4	0	1	0

(i)   (a)   State any assumptions needed to justify the use of a binomial model for the
                number of houses per street which have been burgled during the last six months.

        (b)   Derive the maximum likelihood estimator of p, the probability that a house of the
                type sampled has been burgled during the last six months.

        (c)   Determine the probabilities for the binomial model using your estimate of p, and,
                without doing a formal test, comment on the fit.   [10]

(ii)   An insurance company works on the basis that the probability of a house being burgled
        over a six-month period is 0.18. Carry out a test to investigate whether the binomial
        model with this value of p provides a good fit for the data.   [7]
                                                                                          [Total 17]

9.20   It is desired to investigate the level of premium charged by two companies for contents policies
       for houses in a certain area. Random samples of 10 houses insured by Company A are compared
       with 10 similar houses insured by Company B. The premiums charged in each case are as follows:

Exam style

| Company A | 117 | 154 | 166 | 189 | 190 | 202 | 233 | 263 | 289 | 331 |
| Company B | 142 | 160 | 166 | 188 | 221 | 241 | 276 | 279 | 284 | 302 |

For these data: $\sum A = 2,134$, $\sum A^2 = 494,126$, $\sum B = 2,259$, $\sum B^2 = 541,463$.

(i)   Draw the data given above on a suitable diagram and hence comment briefly on the
       validity of the assumptions required for a two-sample t test for the premiums of these
       two companies.   [3]

(ii)   Assuming that the premiums are normally distributed, carry out a formal test to check
       that it is appropriate to apply a two-sample t test to these data.   [4]

(iii) Test whether the level of premiums charged by Company B was higher than that charged
       by Company A. State your p-value and conclusion clearly.   [3]

(iv)   Calculate a 95% confidence interval for the difference between the proportions of
       premiums of each company that are in excess of £200. Comment briefly on your result.   [3]

(v)   The average premium charged by Company A in the previous year was £170. Formally
       test whether Company A appears to have increased its premiums since the previous year.
                                                                                                [3]
                                                                                          [Total 16]

The solutions start on the next page so that you can
separate the questions and solutions.

Chapter 9 Solutions

9.1   The hypotheses are:

H_0 :   The campaign has not led to a reduction in smoking related illnesses.

H_1 :   The campaign has led to a reduction in smoking related illnesses.

(i)   *Conclusion for test of size 10%*

Since the calculated probability value (7%) is less than the size of the test (10%), we have sufficient evidence at the 10% level to reject H_0 . Therefore the campaign has led to a reduction in the mean number of smoking related illnesses at the 10% level.

(ii)   *Conclusion for a test of size 5%*

Since the calculated probability value (7%) is greater than the size of the test (5%), we have insufficient evidence at the 5% level to reject H_0 . Therefore the campaign has not led to a reduction in the mean number of smoking related illnesses at the 5% level.

9.2   **(i)(a)   *Test mean when population variance unknown***

We are testing:

$$H_0 : \mu = 15 \quad vs \quad H_1 : \mu \neq 15$$

Since the variance is unknown, the test statistic is $\dfrac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$. From the data, we have:

$$\overline{x} = \frac{120.19}{10} = 12.019$$

$$s^2 = \frac{1}{9}(1,693.6331 - 10 \times 12.019^2) = 27.674$$

This gives a statistic of:

$$t = \frac{12.019 - 15}{\sqrt{27.674}/\sqrt{10}} = -1.792$$

This is greater than the t_9 critical value of -2.262 so there is insufficient evidence at the 5% level to reject H_0 . Therefore it is reasonable to conclude that $\mu = 15$.

Alternatively, using probability values, we have $P(t_9 < -1.792) \approx 0.055$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.055 = 0.11$ which is greater than 0.05 so we have insufficient evidence to reject H_0 at the 5% level.

(i)(b) *Test mean when population variance known*

We are testing:

$$H_0 : \mu = 15 \quad vs \quad H_1 : \mu \neq 15$$

Since the variance is known we can use $\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$. This gives:

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{12.019 - 15}{\sqrt{20} / \sqrt{10}} = -2.108$$

This is less than the critical value of -1.96 so there is sufficient evidence at the 5% level to reject H_0 . Therefore it is reasonable to conclude that $\mu \neq 15$.

Alternatively, using probability values, we have $P(Z < -2.108) = 0.0175$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.0175 = 0.035$ which is less than 0.05 so we have sufficient evidence to reject H_0 at the 5% level.

(ii) *Test variance*

We are testing:

$$H_0 : \sigma^2 = 20 \quad vs \quad H_1 : \sigma^2 \neq 20$$

We know that $\dfrac{(n-1)S^2}{\sigma^2}$ has a χ^2_{n-1} distribution.

The observed value of the test statistic is:

$$\frac{9 \times 27.674}{20} = 12.45$$

The critical values of χ^2_9 are 2.700 and 19.02 for a two-sided test. So we have insufficient evidence at the 5% level to reject H_0 . Therefore it is reasonable to conclude that $\sigma^2 = 20$.

9.3 To test whether tails is more than twice as likely, we use the hypotheses:

$$H_0 : p = \tfrac{2}{3} \quad vs \quad H_1 : p > \tfrac{2}{3}$$

Let X be the number of tails obtained in the experiment, then:

$$X \sim Bin(25, p) \doteqdot N(25p, 25pq) \quad \Rightarrow \quad \frac{X - 25p}{\sqrt{25pq}} \doteqdot N(0,1)$$

Under H_0, the statistic with continuity correction is:

$$z = \frac{17\frac{1}{2} - 16\frac{2}{3}}{\sqrt{5\frac{5}{9}}} = 0.354$$

This is less than the critical value of 1.645, so there is insufficient evidence at the 5% level to reject H_0. Therefore it is reasonable to conclude that $p = \frac{2}{3}$, *ie* the experiment does not provide enough evidence to show that tails is more than twice as likely as heads.

Alternatively, using probability values, we have $P(Z > 0.354) = 0.362$ which is greater than 0.05 so we have insufficient evidence to reject H_0 at the 5% level.

9.4   We are testing:

$$H_0 : \mu_A = \mu_B \quad vs \quad H_1 : \mu_A \neq \mu_B$$

The test statistic is:

$$\frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{S_P \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_A + n_B - 2} \quad \text{where} \quad S_P^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$$

Now the observed value of the pooled variance is:

$$s_P^2 = \frac{9 \times 12.4 + 4 \times 25.8}{13} = 16.52$$

So the value of the test statistic is:

$$\frac{4.5 - 0}{\sqrt{16.52}\sqrt{\frac{1}{10} + \frac{1}{5}}} = 2.021$$

This lies between the t_{13} critical values of ± 2.160, so we have insufficient evidence at the 5% level to reject H_0. Therefore we conclude that $\mu_A = \mu_B$.

Alternatively, using probability values, we have $P(t_{13} > 2.021) \simeq 0.034$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.034 = 0.068$ which is greater than 0.05, so we have insufficient evidence to reject H_0 at the 5% level.

9.5   We are testing:

H_0 :   The populations both have normal distributions

vs H_1 :   At least one of the populations does not have a normal distribution.

If H_0 is true, we know that the statistic $\dfrac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2}$ has an $F_{4,9}$ distribution.

Since we know that $\sigma_X^2 = \sigma_Y^2$, this test statistic is just S_X^2 / S_Y^2, which has an observed value of $47.0 / 12.6 = 3.730$.

The 5% critical values for an $F_{4,9}$ distribution are 0.1123 and 4.718. Since 3.730 lies between these, we have insufficient evidence at the 5% level to reject H_0. Therefore we conclude that the populations are both normal.

This is a slightly unusual application of the F test, which is usually used to test variances for populations that are assumed to have a normal distribution.

9.6   The hypotheses are:

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu = \mu_1 \text{ (where } \mu_1 > \mu_0)$$

Here, we can use the likelihood ratio criterion, which says that we should reject H_0 if:

$$\frac{\text{Likelihood under } H_0}{\text{Likelihood under } H_1} < \text{critical value}$$

Since the populations are normal, this is:

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_0}{\sigma}\right)^2} \Bigg/ \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_1}{\sigma}\right)^2} < constant$$

Cancelling the constants reduces this to:

$$e^{-\frac{1}{2\sigma^2}\sum(x_i - \mu_0)^2} \Bigg/ e^{-\frac{1}{2\sigma^2}\sum(x_i - \mu_1)^2} < constant$$

Taking logs:

$$-\frac{1}{2\sigma^2}\sum(x_i - \mu_0)^2 + \frac{1}{2\sigma^2}\sum(x_i - \mu_1)^2 < constant$$

Multiplying through by $2\sigma^2$ and expanding the squares:

$$-\sum(x_i^2 - 2\mu_0 x_i + \mu_0^2) + \sum(x_i^2 - 2\mu_1 x_i + \mu_1^2) < constant$$

Simplifying this gives $(\mu_0 - \mu_1)\sum x_i < constant$.

Since $\mu_1 > \mu_0$, we have to reverse the inequality when we divide through by the negative constant $\mu_0 - \mu_1$, and the test criterion reduces to:

$$\bar{x} > constant$$

So the best test requires us to reject H_0 if the sample mean exceeds a specified critical value.

9.7   We are testing:

$$H_0 : \mu = 7\text{cm} \quad vs \quad H_1 : \mu \neq 7\text{cm} \qquad (\sigma^2 \text{ unknown})$$

Under H_0, the statistic $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ has a t_{11} distribution.

So the value of our test statistic is:

$$\frac{8.54 - 7}{2.97/\sqrt{12}} = 1.796 \tag{[1]}$$

Comparing this with the tables of the t_{11} distribution, we find that $P(t_{11} > 1.796) = 5\%$.   [1]

Hence, we have a probability value of $5\% \times 2 = 10\%$, as the test is two sided.   [1]

9.8   (i)   ***Test whether new screening programme reduces mean claim amount***

We are testing:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_2 < \mu_1 \tag{[½]}$$

where subscript 1 refers to without screening.

The test statistic is:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2} \text{ where } S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{[½]}$$

Calculating the observed values:

$$\overline{x}_1 = \frac{267.4}{10} = 26.74, \quad \overline{x}_2 = \frac{200.7}{10} = 20.07 \tag{[½]}$$

$$s_1^2 = \frac{1}{9}\left(7,755.16 - 10 \times 26.74^2\right) = 67.2093 \tag{[½]}$$

$$s_2^2 = \frac{1}{9}\left(4,553.97 - 10 \times 20.07^2\right) = 58.4357 \tag{[½]}$$

$$s_P^2 = \frac{9 \times 67.2093 + 9 \times 58.4357}{18} = 62.8225 \tag{[½]}$$

So the value of the test statistic is:

$$\frac{(26.74 - 20.07) - 0}{\sqrt{62.8225}\sqrt{\frac{2}{10}}} = 1.882 \tag{[1]}$$

This is greater than the t_{18} critical value of 1.734, so we have sufficient evidence at the 5% level to reject H_0. Therefore we conclude that $\mu_2 < \mu_1$.   [1]

Alternatively, using probability values, we have $P(t_{18} > 1.882) \simeq 0.04$ which is less than 0.05 so we have sufficient evidence to reject H_0 at the 5% level.

(ii)   **Test equality of variances**

We are testing:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \qquad\qquad [\tfrac{1}{2}]$$

The test statistic is:

$$\frac{S_1^2 \big/ \sigma_1^2}{S_2^2 \big/ \sigma_2^2} \sim F_{n_1 - 1, n_2 - 1} \qquad\qquad [\tfrac{1}{2}]$$

Under H_0, the value of the test statistic is:

$$\frac{67.2093}{58.4357} = 1.150 \qquad\qquad [1]$$

The 5% critical values for an $F_{9,9}$ distribution are 0.2484 and 4.026. Since 1.150 lies between these, we have insufficient evidence at the 5% level to reject H_0. Therefore we conclude that $\sigma_1^2 = \sigma_2^2$ (hence the assumption required for part (i) seems valid).   [1]

9.9   (i)   **Test if chemicals are present**

We are testing the proportion p of defective animals using the hypotheses:

$$H_0 : p = 0.005 \quad vs \quad H_1 : p > 0.005 \qquad\qquad [\tfrac{1}{2}]$$

Let X be the number of deformed animals obtained, then:

$$X \sim Bin(1000, p) \doteq N(1000p, 1000pq) \quad \Rightarrow \quad \frac{X - 1{,}000p}{\sqrt{1{,}000pq}} \doteq N(0,1) \qquad\qquad [\tfrac{1}{2}]$$

Under H_0, the statistic with continuity correction is:

$$\frac{14.5 - 5}{\sqrt{4.975}} = 4.26 \qquad\qquad [1]$$

This is greater than the 1% critical value of 2.3263, so there is sufficient evidence at the 1% level to reject H_0. Therefore we conclude that $p > 0.005$, *ie* there are harmful chemicals present in the lake.   [1]

Alternatively, using probability values, we have $P(Z > 4.26) = 0.00001$, which is very significant.

(ii) Test if there has been a significant reduction in deformed animals

We are testing:

$$H_0 : p_1 = p_2 \quad vs \quad H_1 : p_2 < p_1 \qquad [\frac{1}{2}]$$

where the subscript 1 refers to 'before' and 2 refers to 'after'.

The test statistic is:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n_1} + \dfrac{\hat{p}(1-\hat{p})}{n_2}}} \stackrel{.}{\sim} N(0,1) \qquad [\frac{1}{2}]$$

Here we have:

$$\hat{p}_1 = \frac{15}{1{,}000} = 0.015 \qquad \hat{p}_2 = \frac{10}{800} = 0.0125 \qquad \hat{p} = \frac{25}{1{,}800} = 0.013\dot{8} \qquad [\frac{1}{2}]$$

which gives us a value of 0.450 for our test statistic. $\qquad [\frac{1}{2}]$

This is less than the critical value of 1.6449, so there is insufficient evidence at the 5% level to reject H_0. Therefore it is reasonable to conclude that $p_1 = p_2$ (*ie* there has not been a significant reduction in the proportion of deformed animals in the lake). $\qquad [1]$

Alternatively, using probability values, we have $P(Z > 0.450) = 0.326$, which is greater than 0.05 so we have insufficient evidence to reject H_0 at the 5% level.

9.10 (i) Test whether mean home and car claims are equal

We are testing:

$$H_0 : \mu_H = \mu_C \quad vs \quad H_1 : \mu_H \neq \mu_C \qquad [\frac{1}{2}]$$

The test statistic is:

$$\frac{(\bar{X}_H - \bar{X}_C) - (\mu_H - \mu_C)}{S_P \sqrt{\dfrac{1}{n_H} + \dfrac{1}{n_C}}} \sim t_{n_H + n_C - 2}$$

where $S_P^2 = \dfrac{(n_H - 1)S_H^2 + (n_C - 1)S_C^2}{n_H + n_C - 2}$ $\qquad [\frac{1}{2}]$

Calculating the observed values:

$$\bar{x}_H = \frac{78.8}{5} = 15.76, \quad \bar{x}_2 = \frac{88.3}{5} = 17.66 \qquad [\frac{1}{2}]$$

$$s_1^2 = \frac{1}{4}\left(1{,}271.34 - 5 \times 15.76^2\right) = 7.363 \qquad [\frac{1}{2}]$$

$$s_2^2 = \frac{1}{4}\left(1{,}631.93 - 5 \times 17.66^2\right) = 18.138 \qquad\qquad [½]$$

$$s_P^2 = \frac{4 \times 7.363 + 4 \times 18.138}{8} = 12.7505 \qquad\qquad [½]$$

The value of the test statistic is:

$$\frac{(15.76 - 17.66) - 0}{\sqrt{12.7505}\sqrt{\frac{2}{5}}} = -0.841 \qquad\qquad [1]$$

This lies between the t_8 critical values of ± 2.306, so we have insufficient evidence at the 5% level to reject H_0. Therefore we conclude that $\mu_H = \mu_C$. [1]

Alternatively, using probability values, we have $P(t_8 < -0.841) \simeq 0.21$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.21 = 0.42$ which is much greater than 0.05, so we have insufficient evidence to reject H_0 at the 5% level.

(ii)   **Paired t-test**

Since the data are paired, we are now testing:

$$H_0 : \mu_D = 0 \quad vs \quad H_1 : \mu_D \neq 0$$

The differences D for each pair are:

Sample 2 – Sample 1:

1.0   1.8   −0.1   1.6   5.2 [½]

Now:

$$\bar{x}_D = \frac{9.5}{5} = 1.9 \qquad s_D^2 = \frac{1}{4}\left(33.85 - 5 \times 1.9^2\right) = 3.95 \qquad [½]$$

So our test statistic is:

$$\frac{\bar{x}_D - \mu_D}{s_D / \sqrt{n}} = \frac{1.9 - 0}{\sqrt{3.95/5}} = 2.138 \qquad\qquad [1]$$

This lies between the t_4 critical values of ± 2.776, so we have insufficient evidence at the 5% level to reject H_0. Therefore we conclude that $\mu_D = 0$. [1]

Alternatively, using probability values, we have $P(t_4 > 2.138) \simeq 0.05$. But this test is two-sided, so the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.05 = 0.1$ which is greater than 0.05 so we have insufficient evidence to reject H_0 at the 5% level.

9.11   (i)   ***Size of the test***

The size of a test, α, is the probability of a Type I error *ie* the probability of rejecting H_0 when it is true:

$$\alpha = P(X > 93.5 \text{ when } \lambda = 50)$$ [1]

$$= \int_{93.5}^{\infty} 3 \times 50^3 (50 + x)^{-4} \, dx$$

$$= \left[-50^3 (50 + x)^{-3} \right]_{93.5}^{\infty}$$

$$= 0.0423$$

The size of the test is 4.23%.   [1]

(ii)   ***Power of the test***

The power of a test, $1 - \beta$, is the probability of rejecting H_0 when it is false.

$$1 - \beta = P(X > 93.5 \text{ when } \lambda = 60)$$ [1]

$$= \int_{93.5}^{\infty} 3 \times 60^3 (60 + x)^{-4} \, dx$$

$$= \left[-60^3 (60 + x)^{-3} \right]_{93.5}^{\infty} = 0.0597$$

The power of the test is 5.97%.   [1]

9.12   (i)   ***State the interviewee's hypotheses and test***

The interviewee appears to be assuming (with the benefit of hindsight) a two-sided alternative hypothesis that includes both very good results and very bad results, *ie* the hypotheses (expressed in terms of the probability of a correct identification p) would be:

$$H_0 : p = 0.2 \quad vs \quad H_1 : p \neq 0.2$$ [1]

Under H_0, the number of correctly identified patterns has a *Bin*(10, 0.2) distribution.

The probability of getting as few as 0 correct is:

$$\binom{10}{0} (0.2)^0 (0.8)^{10} = 0.107$$

The additional probability for the other tail can only increase this value. So the result is not significant even at the 10% level.   [1]

So, even after bending the rules, the interviewee has failed to demonstrate her powers.   [1]

(ii)   **Correct hypotheses and number of cards required to be correct**

The hypotheses to use in a one-sided test designed to convince non-believers should be:

$$H_0 : p = 0.2 \quad vs \quad H_1 : p > 0.2$$

Calculating the probabilities for the $Bin(10, 0.2)$ distribution (iteratively) shows that:

$$P[Bin(10, 0.2) \leq 4] = 0.1074 + 0.2684 + 0.3020 + 0.2013 + 0.0881$$

$$= 0.9672 \tag{1}$$

So the interviewee would have to identify at least 5 cards correctly to demonstrate the existence of ESP at the 5% level. (The actual size of the test is 3.28%.) [1]

9.13   (i)   **Maximum likelihood estimate of p**

The likelihood of observing the given sample is:

$$L = C \left[(1-p)^3 \right]^{60} \left[3p(1-p)^2 \right]^{75} \left[3p^2(1-p) \right]^{16} \left[p^3 \right]^2$$

$$= K(1-p)^{180} p^{75} (1-p)^{150} p^{32} (1-p)^{16} p^6$$

$$= K(1-p)^{346} p^{113} \tag{1}$$

where C is a constant arising from the fact that the sample can occur in different orders.

Taking logs:

$$\ln L = \ln K + 346 \ln(1-p) + 113 \ln p$$

Differentiating with respect to p:

$$\frac{d \ln L}{dp} = -\frac{346}{1-p} + \frac{113}{p} \tag{1}$$

Setting this equal to zero gives:

$$346 \hat{p} = 113(1 - \hat{p})$$

$$\Rightarrow \quad 459 \hat{p} = 113$$

$$\Rightarrow \quad \hat{p} = \frac{113}{459} = 0.246 \tag{1}$$

Checking that we do have a maximum:

$$\frac{d^2 \ln L}{dp^2} = -\frac{346}{(1-p)^2} - \frac{113}{p^2} < 0 \quad \Rightarrow \quad \max \tag{1}$$

(ii)   *Goodness of fit test*

We are testing the following hypotheses using a χ^2 goodness of fit test:

H_0 : the probabilities conform to a $Bin(3, p)$ distribution

vs H_1 : the probabilities do not conform to a $Bin(3, p)$ distribution

Using $\hat{p} = \dfrac{113}{459}$ from part (i), the probabilities for this binomial distribution are:

$$P(X = 0) = (1 - p)^3 = 0.4283$$

$$P(X = 1) = 3p(1 - p)^2 = 0.4197$$

$$P(X = 2) = 3p^2(1 - p) = 0.1371$$

[1]

$$P(X = 3) = p^3 = 0.0149$$

Multiplying these by 153 we obtain expected values of 65.54, 64.21, 20.97, 2.283.

Since the last one of these expected values is less than 5 we need to combine this with another group, say the third one. This gives:

Number of claims	0	1	2 and 3
Observed no. of policies	60	75	18
Expected no. of policies	65.54	64.21	23.25

[1]

The degrees of freedom $= 3 - 1 - 1 = 1$.   [1]

The statistic is:

$$\sum \frac{(O_i - E_i)^2}{E_i} = \frac{(60 - 65.54)^2}{65.54} + \frac{(75 - 64.21)^2}{64.21} + \frac{(18 - 23.25)^2}{23.25}$$

$$= 0.4683 + 1.813 + 1.185 = 3.47$$

[1]

Since this is less than the 5% critical value of 3.841, we have insufficient evidence at the 5% level to reject H_0. We therefore conclude that the model is a good fit.   [1]

9.14   (i)(a)   *Estimate the Poisson parameter*

The maximum likelihood estimator of the Poisson parameter (representing the average number of corpuscles in each square) is the sample mean, which is:

$$\hat{\mu} = \frac{0 \times 40 + 1 \times 66 + 2 \times 93 + \cdots + 8 \times 1}{400} = \frac{1{,}034}{400} = 2.585$$

[1]

(i)(b)   *Goodness of fit test*

The hypotheses are:

H_0 : The observed numbers conform to a Poisson distribution

vs H_1 : The observed numbers don't conform to a Poisson distribution.   [½]

We can use our estimate from part (i)(a) to calculate the expected numbers using the Poisson PF:

$$eg \ P(X=0)=e^{-2.585}=0.07540 \ \Rightarrow \ 30.16 \ cells \qquad [½]$$

Corpuscle count	0	1	2	3	4	5	6	7	≥ 8
Actual number	40	66	93	94	62	25	14	5	1
Expected number	30.2	78.0	100.8	86.8	56.1	29.0	12.5	4.6	2.0

[2]

If we pool the groups for counts of 7 or more, the value of the chi square statistic is:

$$\sum \frac{(O-E)^2}{E} = \frac{(40-30.2)^2}{30.2} + \frac{(66-78.0)^2}{78.0} + \cdots + \frac{(6-6.6)^2}{6.6}$$

$$= 3.180 + 1.846 + 0.604 + 0.597 + 0.620 + 0.552 + 0.18 + 0.055 = 7.63 \qquad [2]$$

The number of degrees of freedom is $8-1-1=6$.   [1]

Since this is less than the 5% critical value of 12.59, we have insufficient evidence at the 5% level to reject H_0 . We therefore conclude that the model is a good fit.   [1]

(ii)   *Test if patient has anaemia*

We are testing:

$$H_0 : \mu = 3 \quad vs \quad H_1 : \mu < 3 \qquad [½]$$

Let X be the count per cell, then:

$$\sum X \sim Poi(400\mu) \doteq N(400\mu, 400\mu) \ \Rightarrow \ \frac{X-400\mu}{\sqrt{400\mu}} \doteq N(0,1) \qquad [½]$$

Under H_0 , the statistic with continuity correction is:

$$z = \frac{1,034.5-1,200}{\sqrt{1,200}} = -4.78 \qquad [1]$$

This is less than the 1% critical value of -2.3263 , so there is sufficient evidence at the 1% level to reject H_0 . Therefore we conclude that $\mu < 3$, *ie* the patient does have anaemia.   [1]

Alternatively, using probability values, we have $P(Z < -4.78) < 0.0005\%$ *which is highly significant.*

9.15   Here we are testing:

H_0 :   The classification into the three AIDS statuses is independent of the presence or absence of the alleles

vs H_1 :   The classification into the three AIDS statuses is not independent of the presence or absence of the alleles.   [½]

The expected frequencies, calculated using $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$, are:

EXPECTED	Free of symptoms	Early symptoms	Suffering from AIDS	Total
Alleles present	10.7	8.7	28.6	48
Alleles absent	111.3	91.3	298.4	501
Total	122	100	327	549

[1]

The value of the chi square test statistic is:

$$\sum \frac{(O_i - E_i)^2}{E_i} = \frac{(24 - 10.7)^2}{10.7} + \cdots + \frac{(310 - 298.4)^2}{298.4} = 23.79$$ [2]

The test statistic is sensitive to rounding.

The number of degrees of freedom is given by $(2-1)(3-1) = 2$.   [½]

Since the test statistic is greater than the ½% χ_2^2 critical value of 10.60, we can reject H_0 at the ½% level, and conclude that the classification into the three AIDS statuses is not independent of the presence or absence of the alleles.   [1]

9.16   (i)(a)   **Sample mean**

The sample mean is:

$$\frac{250 \times 75 + 750 \times 51 + 1,750 \times 22 + 6,250 \times 5}{75 + 51 + 22 + 5} = \frac{126,750}{153} = 828.43$$

(i)(b)   **Goodness of fit of exponential distribution**

We are testing :

H_0 : the exponential is a suitable distribution

vs H_1 : the exponential is not a suitable distribution.

We first need to estimate the value of λ using the method of moments. The mean of the claim amount distribution is $1/\lambda$. Setting this equal to the sample mean gives a value of 0.0012071 for λ.   [1]

The probability that an exponential random variable lies between a and b is:

$$F(b) - F(a) = e^{-\lambda a} - e^{-\lambda b}$$

So if the claim amount is X we have:

$$P(0 < X < 500) = e^0 - e^{-500\lambda} = 1 - e^{-500\lambda} = 0.4531$$

$$P(500 < X < 1,000) = e^{-500\lambda} - e^{-1,000\lambda} = 0.2478$$

$$P(1,000 < X < 2,500) = e^{-1,000\lambda} - e^{-2,500\lambda} = 0.2502$$

$$P(2,500 < X < 10,000) = e^{-2,500\lambda} - e^{-10,000\lambda} = 0.0489 \qquad [1]$$

Multiplying these figures by 153, we obtain the expected values 69.33, 37.91, 38.27 and 7.48 respectively.   [1]

We then calculate the test statistic $\sum \dfrac{(O_i - E_i)^2}{E_i}$:

$$\frac{(75 - 69.33)^2}{69.33} + \frac{(51 - 37.91)^2}{37.91} + \frac{(22 - 38.27)^2}{38.27} + \frac{(5 - 7.48)^2}{7.48} = 12.7 \qquad [1]$$

The underlying distribution is χ^2 with $4 - 1 - 1 = 2$ degrees of freedom (since we have set the total and estimated the mean from the data).   [1]

The critical value of the χ_2^2 distribution is 5.991, so we have evidence to reject H_0 at the 5% level and conclude that the exponential is not an appropriate distribution.   [1]

(ii)   *Contingency table*

We are testing:

    H_0 :   the claim size is independent of postcode

vs H_1 :   the claim size is not independent of postcode

The observed values in each of the categories are:

Claim size, c	$0 \le c < 500$	$500 \le c < 1,000$	$1,000 \le c < 2,500$	$2,500 \le c < 10,000$	Total
Postcode 1	23	14	7	3	47
Postcode 2	30	16	11	1	58
Postcode 3	22	21	4	1	48
Total	75	51	22	5	153

[1]

We can calculate the expected frequencies in each category by multiplying the row and column totals, and dividing by 153:

Claim size, c	$0 \le c < 500$	$500 \le c < 1,000$	$1,000 \le c < 2,500$	$2,500 \le c < 10,000$
Postcode 1	23.04	15.67	6.76	1.54
Postcode 2	28.43	19.33	8.34	1.90
Postcode 3	23.53	16.00	6.90	1.57

[3]

Since there are three cells containing less than 5, we will combine the last two columns.

Claim size, c	$0 \le c < 500$	$500 \le c < 1,000$	$1,000 \le c < 10,000$
Postcode 1	23.04	15.67	8.29
Postcode 2	28.43	19.33	10.24
Postcode 3	23.53	16.00	8.47

[1]

So we can now calculate the value of the chi square statistic:

$$\chi^2 = \frac{(23 - 23.04)^2}{23.04} + \cdots + \frac{(5 - 8.47)^2}{8.47} = 4.58$$

[1]

The number of degrees of freedom is $(3-1)(3-1) = 4$.

[1]

Our observed value of the test statistic does not exceed 9.488, the upper 5% point of the χ^2_4 distribution. So we have insufficient evidence at the 5% level to reject H_0. Therefore we conclude that the claim size is independent of the postcode.

[1]

9.17   (i)   ***Test for association***

The hypotheses for the test are:

H_0:   There is no association between living in a single parent family and getting into trouble with the police

vs H_1:   There is an association between living in a single parent family and getting into trouble with the police.

[½]

The actual numbers in each category are:

ACTUAL	In trouble	Not in trouble	Total
Single parent	100	300	400
Two parent	240	960	1,200
Total	340	1,260	1,600

The expected numbers for each category are:

EXPECTED	In trouble	Not in trouble	Total
Single parent	85	315	400
Two parent	255	945	1,200
Total	340	1,260	1,600

[1]

The chi-square statistic can then be calculated:

$$\sum \frac{(O-E)^2}{E} = \frac{(100-85)^2}{85} + \frac{(300-315)^2}{315} + \frac{(240-255)^2}{255} + \frac{(960-945)^2}{945}$$

$$= 4.482$$ [2]

The number of degrees of freedom is $(2-1)(2-1)=1$.   [½]

Since the observed value of the test statistic exceeds 3.841, the upper 5% point of the χ_1^2 distribution, we can reject the null hypothesis and conclude that there is an association between single parent families and being in trouble with the police.   [1]

(ii)   *Comment*

However, the presence of an association does not justify the politician's assumption that single parents cause crime. There may be some other underlying causes (*eg* education levels, poverty) that influence family circumstances and crime rates together.   [1]

9.18   (i)   *Test for association*

The test required is a χ^2 contingency table test. The hypotheses are:

H_0 : There is no association between flower colour and leaf type

vs H_1 : There is some association between flower colour and leaf type.   [1]

The expected frequencies are:

	Red	*White*	*Pink*
Plain	87.3	82.1	46.7
Variegated	114.7	107.9	61.3

[2]

So the test statistic is:

$$\sum \frac{(O-E)^2}{E} = \frac{(97-87.3)^2}{87.3} + \cdots + \frac{(31-61.3)^2}{61.3} = 71.0$$ [2]

Comparing this with the figures in the *Tables* for the χ_2^2 distribution, we see that this figure is far larger than the ½% point of the distribution. We have overwhelming evidence against the null hypothesis, and we conclude that there is almost certainly some association between flower colour and leaf type. [1]

(ii)(a)   *Maximum likelihood estimate of q*

Assuming that this genetic model is correct, the likelihood function is:

$$L(q) = q^{97}\left(\frac{q}{2}\right)^{42}\left(\frac{1-3q}{2}\right)^{77} q^{105}\left(\frac{3q}{2}\right)^{148}\left(\frac{1-5q}{2}\right)^{31} \times constant$$

$$= q^{392}(1-3q)^{77}(1-5q)^{31} \times constant \qquad [1]$$

Taking logs:

$$\log L = 392\log q + 77\log(1-3q) + 31\log(1-5q) + constant \qquad [1]$$

Differentiating with respect to q:

$$\frac{d}{dq}\log L = \frac{392}{q} - \frac{231}{1-3q} - \frac{155}{1-5q} \qquad [1]$$

Setting this equal to zero, and multiplying through by $q(1-3q)(1-5q)$, we obtain:

$$392(1-3q)(1-5q) - 231q(1-5q) - 155q(1-3q) = 0 \qquad [1]$$

Gathering terms:

$$392 - 3522q + 7500q^2 = 0$$

Solving the quadratic equation:

$$q = \frac{3522 \pm \sqrt{3522^2 - 4 \times 7500 \times 392}}{15,000} = 0.18128 \text{ or } 0.28832$$

If $q = 0.28832$, then $\dfrac{1-5q}{2}$ is negative, so we can ignore the larger root. So the maximum likelihood estimate for q is $\hat{q} = 0.181$. [2]

We can check that this does indeed give a maximum:

$$\frac{d^2}{dq^2}\log L = -\frac{392}{q^2} - \frac{693}{(1-3q)^2} - \frac{775}{(1-5q)^2} < 0 \quad \Rightarrow \quad \text{max} \qquad [1]$$

(ii)(b)   **Test goodness of fit**

Using $\hat{q} = 0.181$, we can find the expected frequencies by multiplying the probabilities by 500. This gives the following table of expected frequencies:

	Red	White	Pink
Plain	90.6	45.3	114.0
Variegated	90.6	136.0	23.4

[2]

Using a chi-squared test, the hypotheses are:

H_0 : The probabilities of each plant type conform to this genetic model

vs H_1 : The probabilities of each plant type do not conform to this genetic model.

The test statistic is:

$$\sum \frac{(O-E)^2}{E} = \frac{(97-90.6)^2}{90.6} + \cdots + \frac{(31-23.4)^2}{23.4} = 18.5$$ [2]

Comparing this value with the appropriate points of the χ^2_4 distribution, we see that again we have strong evidence to reject H_0, and we conclude at the ½% level that this genetic model does not appear to fit the data well.   [1]

Note that this time we are not testing for association, ie it is an 'ordinary' chi-square goodness of fit test. So the number of degrees of freedom is the number of cells minus the number of estimated parameters minus 1. This gives us $6 - 1 - 1 = 4$ degrees of freedom here.

(iii)   **Comment**

None of the models suggested here appear to fit the data well. Of the pink flowers, there appear to be far too many with plain leaves and far too few with variegated leaves than we would expect under the assumption of independence. However, the genetic model in part (ii) appears to overcompensate for this, with the result that the actual number of pink flowers with plain leaves is smaller than that predicted by the model. A further model somewhere between the two models we have tried so far might give a better fit to the observed data.   [3]

9.19 (i)(a)   **State assumptions**

Each house independently must have the same probability of being burgled.   [1]

(i)(b)   **Derive the maximum likelihood estimator of p**

$$L(p) = [P(X=0)]^{39}[P(X=1)]^{38}[P(X=2)]^{18}[P(X=3)]^4 P(X=5)$$ [1]

Using a $Bin(6, p)$ distribution to calculate the probabilities:

$$L(p) = c[(1-p)^6]^{39}[p(1-p)^5]^{38}[p^2(1-p)^4]^{18}[p^3(1-p)^3]^4 p^5(1-p)$$

$$= cp^{91}(1-p)^{509}$$ [½]

$$\Rightarrow \quad \ln L(p) = \ln c + 91 \ln p + 509 \ln(1-p)$$ [½]

$$\Rightarrow \quad \frac{\partial}{\partial p} \ln L(p) = \frac{91}{p} - \frac{509}{1-p}$$ [1]

Setting the differential equal to zero to obtain the maximum:

$$\Rightarrow \quad \frac{91}{\hat{p}} - \frac{509}{1-\hat{p}} = 0 \quad \Rightarrow \quad \hat{p} = \frac{91}{600}$$ [1]

Checking it's a maximum:

$$\frac{\partial^2}{\partial p^2} \ln L(p) = -\frac{91}{p^2} - \frac{509}{(1-p)^2} < 0 \quad \Rightarrow \quad \text{max}$$ [1]

Alternatively, since the binomial distribution is additive, we could have looked at a single
$Bin(600, p)$ distribution instead.

(i)(c)   *Fit the binomial model and comment*

Using the estimate $\hat{p} = 91/600$ we get frequencies of 37.3, 40.0, 17.9, 4.3, 0.6, 0.0, 0.0, using

$$P(X = x) = \binom{6}{x} \hat{p}^x (1-\hat{p})^{6-x} .$$ [3½]

These are very similar to the observed frequencies – implying that it is a good fit.   [½]

(ii)   *Test whether binomial model with $p = 0.18$ is a good fit for the data*

Using $p = 0.18$ and $P(X = x) = \binom{6}{x} 0.18^x \times 0.82^{6-x}$ we get:

	0	1	2	3	4	5	6
observed	39	38	18	4	0	1	0
expected	30.40	40.04	21.97	6.43	1.06	0.09	0.00

[2]

Since the expected frequencies are less than five for 4, 5 and 6 houses burgled, we need to combine these columns together with the 3 column:

	0	1	2	3+
observed	39	38	18	5
expected	30.40	40.04	21.97	7.58

[1]

Calculating our statistic:

$$\chi^2 = \frac{(39-30.40)^2}{30.40} + \cdots + \frac{(5-7.58)^2}{7.58} = 4.13 \qquad [2]$$

There are now 4 groups so the number of degrees of freedom is $4-1=3$. Remember that the value for p of 0.18 was given and was not estimated using this data. [1]

We are carrying out a one-sided test. Our observed value of the test statistic is less than the 5% critical value of 7.815. So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the model is a good fit. [1]

9.20   (i)   *Illustrate data and comment on the assumptions*

We need a line plot showing the sample values for the two companies:

Company A

[½]

Company B

[½]

There is perhaps some very slight evidence of concentration at the centre of the distribution for A, but the sample sizes are small and it is difficult to tell whether an assumption of normality is reasonable. The variance of the data from Company B looks slightly smaller than that from Company A. However, it is unlikely that such a small difference is significant. There are no outliers in either distribution. [2]

(ii)   *Test whether appropriate to apply a two-sample t test*

We require the variances to be equal, so we are testing:

$$H_0 : \sigma_A^2 = \sigma_B^2 \quad vs \quad H_1 : \sigma_A^2 \neq \sigma_B^2 \qquad [½]$$

$$s_A^2 = \frac{1}{9}\left(494{,}126 - \frac{2{,}134^2}{10}\right) = 4{,}303.4 \qquad s_B^2 = \frac{1}{9}\left(541{,}463 - \frac{2{,}259^2}{10}\right) = 3{,}461.7 \qquad [1]$$

Using $\dfrac{S_A^2 \big/ S_B^2}{\sigma_A^2 \big/ \sigma_B^2} \sim F_{n_A-1, n_B-1}$ we obtain a test statistic of:

$$\frac{4,303.4 \big/ 3,461.7}{1} = 1.243 \tag{[1]}$$

We are carrying out a two-sided test. Comparing our statistic with the $F_{9,9}$ distribution, we see that it is less than the 5% critical value of 4.026. So we have insufficient evidence at the 5% level to reject the null hypothesis. Therefore it is reasonable to conclude that $\sigma_A^2 = \sigma_B^2$. [1½]

(iii)   *Test whether premiums charged by Company B was higher than those by Company A*

We are testing:

$$H_0 : \mu_B = \mu_A \quad vs \quad H_1 : \mu_B > \mu_A \tag{[½]}$$

Under this null hypothesis, we use:

$$\frac{\overline{x}_B - \overline{x}_A}{\sqrt{s_P^2 \left(\dfrac{1}{n_B} + \dfrac{1}{n_A} \right)}} \sim t_{n_A + n_B - 2}$$

Substituting in the values, we get a test statistic of:

$$\frac{225.9 - 213.4}{\sqrt{\dfrac{9 \times 4,303.4 + 9 \times 3,461.7}{18} \left(\dfrac{1}{10} + \dfrac{1}{10} \right)}} = 0.4486 \tag{[1]}$$

Comparing this with the t_{18} values gives a *p*-value of in excess of 30%. So we have insufficient evidence to reject our null hypothesis at the 30% level. Therefore it is reasonable to conclude that the level of premiums charged by Company B is the same as that charged by Company A. [1½]

(iv)   *Confidence interval for the difference between the proportions*

Using the pivotal value, from Chapter 8 of:

$$\frac{(\hat{p}_A - \hat{p}_B) - (p_A - p_B)}{\sqrt{\dfrac{\hat{p}_A \hat{q}_A}{n_A} + \dfrac{\hat{p}_B \hat{q}_B}{n_B}}} \overset{\cdot}{\sim} N(0,1) \tag{[½]}$$

We have:

$$\hat{p}_A = 0.5, \quad \hat{q}_A = 0.5, \quad \hat{p}_B = 0.6, \quad \hat{q}_B = 0.4, \quad n_A = n_B = 10 \tag{[½]}$$

We obtain a 95% confidence interval of:

$$-0.1 \pm 1.96 \sqrt{\frac{0.25}{10} + \frac{0.24}{10}} = (-0.53, 0.33) \tag{[1]}$$

Since this confidence interval contains zero, we cannot conclude that the proportions of premiums in excess of £200 are different for the two companies.   [1]

(v)   *Test whether Company A appears to have increased its premiums*

We now carry out a single sample *t*-test on the data for Company A. We are testing:

$$H_0 : \mu_A = 170 \quad vs \quad H_1 : \mu_A > 170 \qquad\qquad [\tfrac{1}{2}]$$

Our test statistic is:

$$\frac{213.4 - 170}{\sqrt{4,303.4 \, / \, 10}} = 2.092 \qquad\qquad [1]$$

Comparing this with values of the t_9 distribution, we find that we have a result that is significant at level somewhere between 2.5% and 5%. So we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that the company has increased its premiums since the previous year.   [1½]

10

Data Analysis

Syllabus objectives

2.1   Exploratory data analysis

   2.1.1   Describe the purpose of exploratory data analysis.

   2.1.2   Use appropriate tools to calculate suitable summary statistics and undertake exploratory data visualizations.

   2.1.3   Define and calculate Pearson's, Spearman's and Kendall's measures of correlation for bivariate data, explain their interpretation and perform statistical inference as appropriate.

   2.1.4   Use Principal Components Analysis to reduce the dimensionality of a complex data set.

0   Introduction

Actuaries, statisticians and many other professionals are increasingly engaged in analysing and interpreting large data sets, in order to determine whether there is any relationship between variables, and to assess the strength of that relationship. The methods in this and the following three chapters are perhaps more widely applied than any other statistical methods.

Exploratory data analysis (EDA) is the process of analysing data to gain further insight into the nature of the data, its patterns and relationships between the variables, before any formal statistical techniques are applied.

That is we approach the data free of any pre-conceived assumptions or hypotheses. We first see the patterns in the data *before* we impose any views on it and fit models.

In addition to discovering the underlying structure of the data and any relationships between variables, exploratory data analysis can also be used to:

- detect any errors (outliers or anomalies) in the data

- check the assumptions made by any models or statistical tests

- identify the most important/influential variables

- develop parsimonious models – that is models that explain the data with the minimum number of variables necessary.

For numerical data, this process will include the calculation of summary statistics and the use of data visualisations. Transformation of the original data may be necessary as part of this process.

For a single variable, EDA will involve calculating summary statistics (such as mean, median, quartiles, standard deviation, IQR and skewness) and drawing suitable diagrams (such as histograms, boxplots, quantile-quantile (Q-Q) plots and a line chart for time series/ordered data).

For bivariate or multivariate data, EDA will involve calculating the summary statistics for each variable and calculating correlation coefficients between each pair of variables. Data visualisation will typically involve scatterplots between each pair of variables.

Linear correlation between a pair of variables looks at the strength of the linear relationship between them. The diagrams below show the various degrees of positive correlation:

perfect positive   strong positive   weak positive
correlation   correlation   correlation

Recall that we met correlation in Chapter 3 and defined it for a population. In this chapter we will look at obtaining the sample correlation and then using this to make inferences about the population's correlation. This is similar to what we did with the sample mean, $\overline{X}$, and the population mean, μ, in Chapters 6 to 9.

For multivariate data sets with large dimensionality various techniques such as cluster analysis and principle components analysis (also called factor analysis) can be used to reduce the complexity of the data set.

Subject CS1 assumes that students can carry out EDA on univariate data sets.

This includes calculation of summary statistics (*eg* mean and variance) and construction of diagrams (*eg* histograms) which are assumed knowledge for Subject CS1.

This chapter covers three aspects of EDA:

- **using scatterplots to assess the shape of any correlation for bivariate data sets,**

- **calculating correlation coefficients to measure the strength of that correlation, and**

- **using principal components analysis (PCA) to identify the most important variables for multivariate data sets.**

A couple of results in this chapter are quoted without proof. Students are expected to memorise these and apply them in the exam.

1 Bivariate correlation analysis

In a bivariate correlation analysis the problem of interest is an assessment of the strength of the relationship between the two variables Y and X.

In any analysis, it is assumed that measurements (or counts) have been made, and are available, on the variables, giving us bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

1.1 Data visualisation

The starting point is always to visualise the data. For bivariate data, the simplest way to do this is to draw a scatterplot and get a feel for the relationship (if any) between the variables as revealed/suggested by the data.

> **R** **The R code to draw a scatterplot for a bivariate data frame,** `<data>`, **is:**
>
> ```
> plot(<data>)
> ```

We are particularly interested in whether there is a linear relationship between Y, the response (or dependent) variable, and X, the explanatory (or independent, or regressor) variable. That is the expected value of Y, for any given value x of X, is a linear function of that value x, *ie*:

$$E[Y \mid x] = \alpha + \beta x$$

Recall from Chapter 4 that $E[Y|x]$ is a conditional mean, which represents the average value of Y corresponding to a given value of x.

If a linear relationship (even a weak one) is indicated by the data, the methods of Chapter 11 (Linear Regression) can be used to fit a linear model, with a view to exploiting the relationship between the variables to help estimate the expected response for a given value of the explanatory variable.

We now look at two examples (one linear and one non-linear) which we will analyse throughout the chapter.

Question

A sample of ten claims and corresponding payments on settlement for household policies is taken from the business of an insurance company.

The amounts, in units of £100, are as follows:

Claim x	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
Payment y	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45

Draw a scatterplot and comment on the relationship between claims and payments.

Solution

The scatterplot for these data is as follows:

Settlement payments for claims

Here we can see that there appears to be a strong positive linear relationship. The plotted data points lie roughly in a straight line.

You can see from the graph that there appears to be a linear relationship between the claims and payments (*ie* the rate of change in payment is constant for a rate of change in the claim). So we will be able to use our linear regression work on these data values in the next chapter.

The next example contains a non-linear relationship between the variables.

A well-chosen transformation of *y* (or *x*, or even both) may however bring the data into a linear relationship.

This then allows us to use the linear regression techniques in the next chapter.

Question

The rate of interest of borrowing, over the next five years, for ten companies is compared to each company's leverage ratio (its debt to equity ratio).

The data is as follows:

Leverage ratio, x	0.1	0.4	0.5	0.8	1.0	1.8	2.0	2.5	2.8	3.0
Interest rate (%), y	2.8	3.4	3.5	3.6	4.6	6.3	10.2	19.7	31.3	42.9

Draw a scatterplot and comment on the relationship between company borrowing (leverage) and interest rate. Hence apply a transformation to obtain a linear relationship.

Solution

The scatterplot for these data is as follows:

Company borrowing and interest rates

It can clearly be seen that the data displays a non-linear relationship, since the rate of change in the interest rate increases with the leverage ratio.

In this case, the log of the interest rate against the leverage ratio produces a far more linear relationship:

Company borrowing and interest rates

1.2   Sample correlation coefficients

The degree of association between the *x* and *y* values is summarised by the value of an appropriate correlation coefficient each of which take values from –1 to +1.

The coefficient of linear correlation provides a measure of how well a linear regression model explains the relationship between two variables. The values of *r* can be interpreted as follows:

Value	Interpretation
$r = 1$	The two variables move together in the same direction in a perfect linear relationship.
$0 < r < 1$	The two variables tend to move together in the same direction but there is not a direct relationship.
$r = 0$	The two variables can move in either direction and show no linear relationship.
$-1 < r < 0$	The two variables tend to move together in opposite directions but there is not a direct relationship.
$r = -1$	The two variables move together in opposite directions in a perfect linear relationship.

In this section we look at three correlation coefficients: Pearson, Spearman's rank and Kendall's rank.

It is always important in data analysis to note that simply finding a mathematical relationship between variables tells one nothing in itself about the causality of that relationship or its continuing persistence through time. Qualitative as well as quantitative analysis is essential before making predictions or taking action.

Jumping to a 'cause and effect' conclusion — that a change in one variable causes a change in the other — is a common misinterpretation of correlation coefficients. For example, the correlation may be spurious, or there may be another variable not part of the analysis that is causal.

For some excellent examples of spurious correlations do visit tylervigen.com or read his book *'Spurious correlations: Correlation does not equal causation'.*

Pearson's correlation coefficient

Pearson's correlation coefficient *r* (also called Pearson's product-moment correlation coefficient) **measures the strength of *linear* relationship between two variables and is given by:**

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$$

where:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \left(\sum x_i\right)^2 / n$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2 / n$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right) / n$$

Note that S_{xx} and S_{yy}, the sum of squares of x and y respectively, are the sample variances of x and y except we don't divide by $(n-1)$. Similarly S_{xy} is the sample covariance except we don't divide by n.

Note also that $\sum_{i=1}^{n} x_i^2$ is often abbreviated to $\sum x^2$, *etc*.

Question

Show that:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - n\bar{x}^2$$

Solution

Expanding the bracket and splitting up the summation, we have:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x} x_i + \bar{x}^2)$$

$$= \sum x_i^2 - 2\bar{x} \sum x_i + \sum \bar{x}^2$$

$$= \sum x_i^2 - \frac{2\left(\sum x_i\right)^2}{n} + n\frac{\left(\sum x_i\right)^2}{n^2} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Now since $\sum x_i = n\bar{x}$, we get:

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - \frac{(n\bar{x})^2}{n} = \sum x_i^2 - n\bar{x}^2$$

These formulae are given on page 24 of the *Tables* in the $S_{xx} = \sum x_i^2 - n\bar{x}^2$ format.

Recall from Chapter 3 that the population correlation coefficient was defined to be:

$$\rho = \text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

You can see that Pearson's sample correlation coefficient, *r*, is an estimator of the population correlation coefficient, ρ, in the same way as $\bar{X}$ is an estimator of μ or S^2 is an estimator of σ^2.

The formula for the sample correlation coefficient, *r*, is given on page 25 of the *Tables*.

Let's now calculate this correlation coefficient for the examples we met earlier.

Question

For the claims settlement data we have:

Claim (£100's) *x*   2.10 2.40 2.50 3.20 3.60 3.80 4.10 4.20 4.50 5.00

Payment *(£100's) y*   2.18 2.06 2.54 2.61 3.67 3.25 4.02 3.71 4.38 4.45

Number of pairs of observations $n = 10$.

$$\sum x = 35.4, \ \sum x^2 = 133.76, \ \sum y = 32.87, \ \sum y^2 = 115.2025, \ \sum xy = 123.81$$

Calculate Pearson's correlation coefficient for the claims settlement data and comment.

Solution

$$S_{xx} = \sum x^2 - n\overline{x}^2 = 133.76 - \frac{35.4^2}{10} = 8.444$$

$$S_{yy} = \sum y^2 - n\overline{y}^2 = 115.2025 - \frac{32.87^2}{10} = 7.15881$$

$$S_{xy} = \sum xy - n\overline{x}\,\overline{y} = 123.81 - \frac{(35.4 \times 32.87)}{10} = 7.4502$$

$$\Rightarrow \quad r = \frac{7.4502}{\sqrt{8.444 \times 7.15881}} = 0.95824$$

As expected, this is high (close to +1), and indicates a strong positive linear relationship.

Question

For the original borrowing rate data:

Leverage ratio, x	0.1	0.4	0.5	0.8	1.0	1.8	2.0	2.5	2.8	3.0
Interest rate y	0.028	0.034	0.035	0.036	0.046	0.063	0.102	0.197	0.313	0.429

Number of pairs of observations $n = 10$.

$$\sum x = 14.9 \,, \quad \sum x^2 = 32.39 \,, \quad \sum y = 1.283 \,, \quad \sum y^2 = 0.341769 \,, \quad \sum xy = 3.082$$

Calculate Pearson's correlation coefficient for the borrowing rate data.

Solution

$$S_{xx} = \sum x^2 - n\overline{x}^2 = 32.39 - 10\left(\frac{14.9}{10}\right)^2 = 10.189$$

$$S_{yy} = \sum y^2 - n\overline{y}^2 = 0.341769 - 10\left(\frac{1.283}{10}\right)^2 = 0.1771601$$

$$S_{xy} = \sum xy - n\overline{x}\,\overline{y} = 3.082 - 10\left(\frac{14.9}{10}\right)\left(\frac{1.283}{10}\right) = 1.17033$$

$$\Rightarrow \quad r = \frac{1.17033}{\sqrt{10.189 \times 0.1771601}} = 0.87108$$

You may wish to try Q10.1(i) and (ii) to check you can calculate the correlation coefficient is on your own.

Since Pearson's correlation coefficient measures linear association it may give a low result when variables have a strong, but non-linear relationship. Whilst the value for the borrowing rate data is high, it is materially lower than in the first example, due to the non-linearity of the relationship.

However, the moral of the story is always to plot the data first. For example the following scatterplots (from the statistician Francis Anscombe) all have a correlation coefficient of 0.816:

Reference: Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician* **27** (1): 17–21. JSTOR 2682899

The R code for calculating a Pearson correlation coefficient for variables *x* and *y* is:

```
cor(x, y, method = "pearson")
```

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient r_s measures the strength of monotonic (but not necessarily linear) relationship between two variables.

So we are measuring how much they move together but the changes are not necessarily at a constant rate.

Formally, it is the Pearson correlation coefficient applied to the ranks, $r(X_i)$ and $r(Y_i)$, rather than the raw values, (X_i, Y_i), of the bivariate data.

So it just uses their relative sizes in relation to each other. We usually order them from smallest to largest.

If all the X_i's are unique, and separately all of the Y_i's are unique, *ie* there are no 'ties', then this calculation simplifies to:

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

where $d_i = r(X_i) - r(Y_i)$.

Note that since Spearman's rank correlation coefficient only considers ranks rather than the actual values, the value of the coefficient is less affected by extreme values/outliers in the data than Pearson's correlation coefficient. Hence this statistic is more robust.

Let's now calculate Spearman's rank correlation coefficient for the examples we met earlier.

Question

Calculate Spearman's rank correlation coefficient for the claims settlement data and comment.

Claim (£100's) *x*   2.10 2.40 2.50 3.20 3.60 3.80 4.10 4.20 4.50 5.00

Payment (£100's) *y*   2.18 2.06 2.54 2.61 3.67 3.25 4.02 3.71 4.38 4.45

Solution

For the claims settlement data:

Claim x	Payment y	Rank x	Rank y	d	d^2
2.1	2.18	1	2	−1	1
2.4	2.06	2	1	1	1
2.5	2.54	3	3	0	0
3.2	2.61	4	4	0	0
3.6	3.67	5	6	−1	1
3.8	3.25	6	5	1	1
4.1	4.02	7	8	−1	1
4.2	3.71	8	7	1	1
4.5	4.38	9	9	0	0
5	4.45	10	10	0	0

This gives:

$$r_s = 1 - \frac{6 \times 6}{10 \times (10^2 - 1)} = 0.9636$$

As expected, the Spearman's rank correlation coefficient is very high, since it is known from the calculation of the Pearson's correlation coefficient that there is a strong positive linear relationship (hence a strong monotonically increasing relationship).

Question

Calculate Spearman's rank correlation coefficient for the original borrowing rate data and comment.

Leverage ratio, x	0.1	0.4	0.5	0.8	1.0	1.8	2.0	2.5	2.8	3.0
Interest rate (%), y	2.8	3.4	3.5	3.6	4.6	6.3	10.2	19.7	31.3	42.9

Solution

For the corporate borrowing data, the ranks of the two data are exactly equal, hence Spearman's rank correlation coefficient is trivially equal to 1.

The reason that this is materially higher than the equivalent Pearson coefficient is because the non-linearity of the relationship does not feature in the calculation, only the fact that it is monotonically increasing.

R | **The R code for calculating a Spearman rank correlation coefficient for variables x and y is:**

```
cor(x, y, method = "spearman")
```

The Kendall rank correlation coefficient

Kendall's rank correlation coefficient τ measures the strength of dependence of rank correlation between two variables.

Like the Spearman rank correlation coefficient, the Kendall rank correlation coefficient considers only the relative values of the bivariate data, and not their actual values. It is far more intensive from a calculation viewpoint, however, since it considers the relative values of all possible pairs of bivariate data, not simply the rank of X_i and Y_i for a given *i*.

Despite the more complicated calculation it is considered to have better statistical properties than Spearman's rank correlation coefficient, particularly for small data sets with large numbers of tied ranks.

Any pair of observations $(X_i, Y_i); (X_j, Y_j)$ where $i \neq j$, is said to be *concordant* if the ranks for both elements agree, ie $X_i > X_j$ and $Y_i > Y_j$, or $X_i < X_j$ and $Y_i < Y_j$; otherwise they are said to be *discordant*.

Consider our settlement payments for claims example. Suppose claim A is greater than claim B. If the settlement for claim A is also greater than the settlement for claim B then they have the same relative rank orders, and we say that A and B are concordant pairs with respect to the random variables claims and settlement.

Let n_c be the number of concordant pairs, and let n_d be the number of discordant pairs. Assuming that there are no ties, the Kendall coefficient τ is defined as:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

The numerator is the difference in the number of concordant and discordant pairs.

The denominator is the total number of combinations of pairing each (X_i, Y_i) with each (X_j, Y_j). This could also be defined as $n_c + n_d$.

For example if there were 3 observations of X and Y then there would be $(3 \times 2)/2 = 3$ combinations:

- (X_1, Y_1) and (X_2, Y_2)

- (X_1, Y_1) and (X_3, Y_3)

- (X_2, Y_2) and (X_3, Y_3)

So τ can be interpreted as the difference between the probability of these objects being in the same order and the probability of these objects being in a different order.

Therefore, a value of -1 indicates all discordant pairs and $+1$ indicates all concordant pairs.

Intuitively, it is clear that if the number of concordant pairs is much larger than the number of discordant pairs, then the random variables are positively correlated. Whereas if the number of concordant pairs is much less than the number of discordant pairs, then the variables are negatively correlated.

Let's now calculate the Kendall rank correlation coefficient for the examples we met earlier.

Question

Calculate Kendall's rank correlation coefficient for the claims settlement data and comment.

Claim (£100's) *x*   2.10 2.40 2.50 3.20 3.60 3.80 4.10 4.20 4.50 5.00

Payment (£100's) *y*   2.18 2.06 2.54 2.61 3.67 3.25 4.02 3.71 4.38 4.45

Solution

For the example claims data

(x, y)	2.1,2.18	2.4,2.06	2.5,2.54	3.2,2.61	3.6,3.67	3.8,3.25	4.1,4.02	4.2,3.71	4.5,4.38	5.0,4.45
2.1,2.18		d	c	c	c	c	c	c	c	c
2.4,2.06			c	c	c	c	c	c	c	c
2.5,2.54				c	c	c	c	c	c	c
3.2,2.61					c	c	c	c	c	c
3.6,3.67						d	c	c	c	c
3.8,3.25							c	c	c	c
4.1,4.02								d	c	c
4.2,3.71									c	c
4.5,4.38										c
5.0,4.45										

Where *c* represents a concordant pair, and *d* represents a discordant pair.

Here $n_c = 42$, $n_d = 3$, so $\tau = (42-3)/(10 \times 9/2) = 0.8667$.

Alternatively using $\tau = (n_c - n_d)/(n_c + n_d)$ *gives* $\tau = (42-3)/(42+3) = 0.8667$.

Again the relatively high value demonstrates the strong correlation between the variables.

It's often easier to determine concordant and discordant pairs by using the ranks instead of the actual numbers.

First arrange the values in order of rank for x. Then the concordant pairs (C) are the number of observations below which have a higher rank for the y and the discordant pairs (D) are the number of observations below which have a lower rank for the y.

	Rank x	Rank y	C	D
2.1, 2.18	1	2	8	1
2.4, 2.06	2	1	8	0
2.5, 2.54	3	3	7	0
3.2, 2.61	4	4	6	0
3.6, 3.67	5	6	4	1
3.8, 3.25	6	5	4	0
4.1, 4.02	7	8	2	1
4.2, 3.71	8	7	2	0
4.5, 4.38	9	9	1	0
5.0, 4.45	10	10		
			42	3

Totalling the columns gives $n_c = 42$, $n_d = 3$ as before.

Question

Calculate Kendall's rank correlation coefficient for the original borrowing rate data and comment.

Leverage ratio, x	0.1	0.4	0.5	0.8	1.0	1.8	2.0	2.5	2.8	3.0
Interest rate (%), y	2.8	3.4	3.5	3.6	4.6	6.3	10.2	19.7	31.3	42.9

Solution

For the corporate borrowing data, clearly all the pairs are concordant, and so τ is trivially equal to 1.

The R code for calculating a Kendall rank correlation coefficient for variables *x* and *y* is:

```
cor(x, y, method = "kendall")
```

1.3   Inference

To go further than a mere description/summary of the data, a model is required for the distribution of the underlying variables (X, Y).

Inference under Pearson's correlation

The appropriate model is this: the distribution of (X, Y) is bivariate normal, with parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y$, and ρ.

Note that assuming a bivariate normal distribution means that we have continuous data, each (marginal) distribution is also normal, the variance is constant and we have a linear relationship between X and Y. If any of these assumptions are not met then inference will give misleading results.

In the bivariate normal model, both variables are considered to be random. However, they are correlated, so their values are 'linked'.

Here is a brief outline of the bivariate normal distribution for students who are interested.

The bivariate normal model assumes that the values of (X_i, Y_i) have a joint normal distribution with joint PDF $f_{X,Y}(x, y)$ $(-\infty < x, y < \infty)$ given by:

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\bigg/2(1-\rho^2)\right)$$

where ρ is the correlation parameter, where $-1 < \rho < 1$.

In the case where $\rho = 0$, the 'cross term' is zero and the PDF factorises into the product of the PDFs for two independent variables with a $N(\mu_X, \sigma_X^2)$ and a $N(\mu_Y, \sigma_Y^2)$ distribution. In the case where $\rho \to \pm 1$, the bivariate distribution degenerates into a single line $\dfrac{Y-\mu_Y}{\sigma_Y} = \pm\dfrac{X-\mu_X}{\sigma_X}$ *ie* the values of *X* and *Y* are directly linked.

If we integrate over all possible values of *y* to find the conditional expectation, we get the following result:

$$E(Y \mid X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

The important thing to note here is that the RHS is a linear function of *x*.

To assess the significance of any calculated *r*, the sampling distribution of this statistic is needed. The distribution of *r* is negatively skewed and has high spread/variability.

Two results are available (both of which are given on page 25 of the *Tables*)**.**

Result 1

Under $H_0 : \rho = 0$, $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ **has a *t* distribution with** $\nu = n - 2$ **degrees of freedom.**

From this result a test of $H_0 : \rho = 0$ **(the hypothesis of 'no *linear* relationship' between the variables) can be performed by working out the value of** *r* **which is 'significant' at a given level of testing, or by finding the probability value of the observed *r*.**

This result is given on page 25 of the *Tables*.

Question

Test $H_0 : \rho = 0$ *vs* $H_1 : \rho \neq 0$ for the claims settlement data. Recall that $r = 0.95824$.

Solution

For the given data $n = 10$ and $r = 0.958$ so our test statistic is:

$$\frac{0.95824\sqrt{8}}{\sqrt{1-0.95824^2}} = 9.478$$

Under H_0 this comes from a t_8 distribution. The *p*-value of $2 \times P(t_8 > 9.478)$ is less than 0.1%. We have extremely strong evidence to reject H_0 and conclude $\rho \neq 0$.

Result 2 (Fisher's transformation of *r*)

This is a more general result – it is not restricted to the case $\rho = 0$.

If $W = \dfrac{1}{2}\ln\dfrac{1+r}{1-r}$, **then *W* has (approximately) a normal distribution with mean** $\dfrac{1}{2}\ln\dfrac{1+\rho}{1-\rho}$ **and standard deviation** $\dfrac{1}{\sqrt{n-3}}$.

This is usually referred to as the Fisher *Z* transformation (because the resulting *z*-values are approximately normal). Accordingly, the letter *Z* is usually used.

Note that *W* can also be written as $\tanh^{-1} r$. This is the inverse hyperbolic tangent function, which, on modern Casio calculators, is accessed by pressing $\boxed{\text{hyp}}$ and then choosing Option 6 to get $\tanh^{-1}$.

From the result on *W*, tests of $H_0 : \rho = \rho_0$ **can be performed. Confidence intervals for** μ_w **and hence for** ρ **can also be found.**

This result is given on page 25 of the *Tables*.

Question

Considering the data on claims and settlements, carry out the test:

$$H_0 : \rho = 0.9 \quad vs \quad H_1 : \rho > 0.9$$

for the population of all claims/payments of this type.

Solution

For the given data:

$$n = 10 , \; r = 0.95824 , \; \text{observed value of } W = \tanh^{-1} 0.95824 = 1.9239$$

Under H_0, W has a normal distribution with mean $\tanh^{-1} 0.9 = 1.4722$ and standard deviation $1/\sqrt{10-3} = 0.37796$. So:

$$P(W > 1.921) = P\left(Z > \frac{1.9239 - 1.4722}{0.37796} \right) = P(Z > 1.195) \approx 0.12$$

So the *p*-value of $r = 0.958$ is about 0.12.

There is insufficient evidence to justify rejecting H_0 – which can stand.

Notes:

(a) **The bivariate normal assumption.**

The presence of 'outliers' – data points far away from the main body of the data – may indicate that the distributional assumption underlying the above methods is highly questionable.

(b) **Influence**

Just as a single observation can have a marked effect on the value of a sample mean and standard deviation, so a single observation separated from the bulk of the data can have a marked effect on the value of a sample correlation coefficient.

The R code for carrying out any hypothesis test using the Pearson correlation coefficient for variables *x* and *y* is:

```
cor.test(x, y, method = "pearson")
```

Inference under Spearman's rank correlation

Since we are using ranks rather than the actual data, no assumption is needed about the distribution of X, Y or (X,Y), *ie* it is a non-parametric test.

However, non-parametric tests are less powerful than parametric tests (*ie* ones that do assume a distribution) as we have less information. So we would need to obtain a more extreme result before we are able to reject H_0. On the plus side, the test is less affected by outliers.

Under a null hypothesis of no association/no monotonic relationship between X and Y the sampling distribution of r_s can (for small values of n) be determined precisely using permutations. This does not have the form of a common statistical distribution.

For example, if we had a sample size of 4, there would be $4! = 24$ ways of arranging the ranks of the Y variables, so each arrangement has a probability of $\frac{1}{24}$ of occurring. We then calculate Σd^2 for each arrangement and hence obtain the probabilities of getting each value of Σd^2.

We can then carry out a hypothesis test. For example, if we are testing $H_0 : \rho = 0$ vs $H_1 : \rho > 0$ with a 5% significance level, where the data values give $\Sigma d^2 = 3$, we can calculate $P\left(\Sigma d^2 \leq 3\right)$ using the probabilities obtained above and if we get less than 5% we would reject H_0. However, for large n this will be time consuming.

For larger values of n (> 20) we can use Results 1 and 2 from above. The limiting normal distribution will have a mean 0 and a variance of $1/(n-1)$.

Recall that Spearman's rank correlation coefficient is derived by applying Pearson's correlation coefficient to the ranks rather than the original data.

R | **The R code for carrying out a hypothesis test using Spearman's rank correlation coefficient for variables x and y is:**

```
cor.test(x, y, method = "spearman")
```

Inference under Kendall's rank correlation

Again, since we are using ranks, we have a non-parametric test.

Under the null hypothesis of independence of X and Y, the sampling distribution of τ can be determined precisely using permutations for small values of n.

We can carry out a hypothesis test in the same way as described above but calculating $n_c - n_d$ for each arrangement. However, again, for large n this will be time consuming.

For larger values of n (> 10), use of the Central Limit Theorem means that an approximate normal distribution can be used, with mean 0 and variance $2(2n+5)/9n(n-1)$.

R | **The R code for carrying out a hypothesis test using the Kendall rank correlation coefficient for variables x and y is:**

```
cor.test(x, y, method = "kendall")
```

Note that `cor.test` will determine exact p-values if $n < 50$ (ignoring tied values); for larger samples the test statistic is approximately normally distributed.

Question

An actuary wants to investigate if there is any correlation between students' scores in the CS1 mock exam and the CS2 mock exam. Data values from 22 students were collected and the results are:

Student	1	2	3	4	5	6	7	8	9	10	11
CS1 mock score	51	43	39	80	56	57	26	68	54	75	72
CS2 mock score	52	42	58	56	47	72	16	63	48	80	68

Student	12	13	14	15	16	17	18	19	20	21	22
CS1 mock score	85	48	27	63	76	64	55	78	82	52	60
CS2 mock score	82	54	38	57	71	50	45	60	59	49	61

You are given that $\Sigma d^2 = 494$, $n_c = 174$ and $n_d = 57$. Test $H_0 : \rho = 0$ vs $H_1 : \rho > 0$ for the mock score data using the Spearman's rank correlation coefficient and the Kendall's rank correlation coefficient along with normal approximations.

Solution

For the given data values:

$$n = 22$$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 494}{22(22^2 - 1)} = 0.72106$$

$$\tau = \frac{n_c - n_d}{n(n-1)/2} = \frac{174 - 57}{22 \times 21/2} = 0.50649$$

Under H_0, the Spearman's rank correlation coefficient has an approximately normal distribution with mean 0 and variance $\dfrac{1}{n-1}$. Our test statistic is:

$$\frac{0.72106 - 0}{\sqrt{\frac{1}{21}}} = 3.304$$

This exceeds the 5% critical value of 1.6449 so we have evidence at the 5% level to reject H_0. We conclude that the mock scores in CS1 and CS2 are positively correlated.

Under H_0, the Kendall's rank correlation coefficient has an approximately normal distribution with mean 0 and variance $\dfrac{2(2n+5)}{9n(n-1)}$. Our test statistic is:

$$\frac{0.50649-0}{\sqrt{\dfrac{2\times 49}{9\times 22\times 21}}}=3.299$$

This exceeds the 5% critical value of 1.6449 so we have evidence at the 5% level to reject H_0. We conclude that the mock scores in CS1 and CS2 are positively correlated.

2   Multivariate correlation analysis

So far, we have only considered bivariate data. In most practical applications, there are many variables to consider. We now consider the case $(\underline{X}, Y)$, where Y remains the variable of interest, but $\underline{X}$ is now a vector of possible explanatory variables.

For example, in motor insurance we may wish to see the connection between the claim amounts and a number of explanatory variables such as age, number of years driving, size of the engine and annual number of miles driven.

2.1   Data visualisation

Again, the starting point is always to visualise the data. For multivariate cases it is no bother for a computer package to plot a scattergraph matrix, *ie* scattergraphs between each pair of variables to make the relationships between them clear.

The R code to draw scatterplots for all pairs from a multivariate data frame, <data>, is:

```
plot(<data>)
```

or it is possible to use:

```
pairs(<data>)
```

Now let's look at a set of multivariate data.

Consider a set of equity returns from four different markets across 12 time periods $(\underline{X})$.

Market 1 (Mkt_1)	Market 2 (Mkt_2)	Market 3 (Mkt_3)	Market 4 (Mkt_4)
0.83%	4.27%	−1.79%	−0.39%
−0.12%	−3.72%	0.90%	−0.26%
−5.49%	−5.21%	−4.62%	−5.67%
2.75%	−6.26%	3.38%	1.40%
−5.68%	−7.37%	−5.21%	−5.05%
−3.70%	−1.60%	−2.34%	−2.66%
5.75%	5.08%	6.03%	5.48%
1.03%	−1.38%	2.37%	1.47%
0.69%	−0.17%	0.38%	−0.10%
−4.03%	−3.26%	−3.04%	−2.59%
0.54%	2.22%	1.42%	1.37%
3.03%	9.47%	2.95%	2.99%

We are going to use R to obtain a matrix of scatterplots, by plotting the market returns in pairs. We wish to consider the relationship between Market 4 and the other Markets.

This gives the following scattergraph matrix:

The bottom row has Market 4 as the response variable with the other three markets as the explanatory variables. We can see that there appear to be positive linear relationships between the response variable and explanatory variables.

Note that there are a strong positive linear relationships between Market 4 and explanatory variables Markets 1 and 3. Since Markets 1 and 3 move together there may be some 'overlap' between their influence on Market 4. We will look at how we can strip this 'overlap' out in the Principal Components Analysis (PCA) section later.

2.2   Sample correlation coefficient matrix

Similarly it is no bother for a computer package to calculate correlation coefficients between each pair of variables and display them in a matrix.

> **R** **The R code for calculation of a Pearson correlation coefficient matrix for a multivariate numeric data frame** `<data>` **is:**
>
> ```
>     cor(<data>, method = "pearson")
> ```

We can also use R on the equity return data to obtain the Pearson correlation coefficient matrix.

The Pearson correlation coefficient matrix for the four markets as produced in R output is:

```
            Mkt_1      Mkt_2      Mkt_3      Mkt_4
Mkt_1 1.0000000 0.6508163 0.9538019 0.9727972

Mkt_2 0.6508163 1.0000000 0.5321185 0.6893932

Mkt_3 0.9538019 0.5321185 1.0000000 0.9681911

Mkt_4 0.9727972 0.6893932 0.9681911 1.0000000
```

Notice that the diagonal elements are all 1. That's because there is perfect correlation between, say, Market 1 and Market 1. Notice also that it is symmetrical as $\text{corr}(X,Y) = \text{corr}(Y,X)$.

2.3   Inference

We can carry out tests on the correlation for each pair of variables using the methods described in Section 1.3.

> **R** **However, using the** `corr.test` **function from the** `psych` **package in R calculates the correlations and the *p*-values for every pair in one go.**

3 Principal component analysis

Principal component analysis is most easily tackled using a computer. In this section the Core Reading runs through the theory and gives an example, but this topic will be dealt with in more depth in the Paper B Online Resources (PBOR).

Until now we have considered the variables in separate pairs, but in practice the amount of analysis required in this approach grows exponentially with each additional variable.

Principal component analysis (PCA), also called factor analysis, provides a method for reducing the dimensionality of the data set, $\underline{X}$ – in other words, it seeks to identify the key components necessary to model and understand the data.

For many multivariate datasets there is correlation between each of the variables. This means there is some 'overlap' between the information that each of the variables provide. The technical phrase is that there is redundancy in the data. PCA gives us a process to remove this overlap.

The idea is that we create new uncorrelated variables, and we should find that only some of these new variables are needed to explain most of the variability observed in the data. The key thing is that each 'new' variable is a linear combination of the 'old' variables, so if we eliminate any of the new variables we are still retaining the most important bits of information.

We then rewrite the data in terms of these new variables, which are called principle components.

These components are chosen to be uncorrelated linear combinations of the variables of the data which maximise the variance.

The next section of Core Reading explains the process of how a PCA is carried out and contains some matrix theory. In parallel with the text, we will work through a simple matrix as an example so that you can see what is happening. The Core Reading starts with a reminder of how to determine eigenvectors and eigenvalues. This is assumed knowledge for the actuarial exams.

The eigenvalues of matrix $\underline{A}$ are the values λ such that $\det(\underline{A} - \lambda\,\underline{I}) = 0$ where $\underline{I}$ is the identity matrix. The corresponding eigenvector, $\underline{v}$, of an eigenvalue λ satisfies the equation $(\underline{A} - \lambda\,\underline{I})\underline{v} = 0$.

Consider an $n \times p$ centred data matrix $\underline{X}$. Using standard techniques from linear algebra, define $\underline{W}$ as a $p \times p$ matrix, whose columns are the eigenvectors of $\underline{X}^T\underline{X}$. The intuition for doing this is that $\underline{X}^T\underline{X}$ represents the (scaled) covariance of the data.

Here p represents the number of variables and n represents the number of observations of each variable. In a centred data matrix, the entries in each column have a mean of zero. We can obtain a centred matrix from the original matrix by subtracting the appropriate column mean from each entry. The sample variance/covariance matrix is $\mathbf{X}^T\mathbf{X}$ divided by $(n-1)$.

Suppose we are trying to model the chances of a student passing the CS1 exam. We are going to include in our model the average number of days per week each student does some studying (X_1) and the average number of hours each student studies at the weekend (X_2). The data values for one student are $x_1 = 2, x_2 = 10$ and for another student we have $x_1 = 4, x_2 = 6$. The original data matrix is therefore:

$$\begin{pmatrix} 2 & 10 \\ 4 & 6 \end{pmatrix}$$

The mean of the entries in the first column is 3 and the mean of the entries in the second column is 8, so the centred matrix is:

$$\mathbf{X} = \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}$$

We now need to calculate $\mathbf{X}^T \mathbf{X}$:

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} = \begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix}$$

We can see that this is the covariance matrix for the data in $\mathbf{X}$. The variance of the data set $(-1,1)$ is 2, the variance of the data set $(2,-2)$ is 8 and the covariance between the data sets is calculated as follows:

$$\frac{1}{n-1}\sum_{j=1}^{n}(x_{j1} - \overline{x}_1)(x_{j2} - \overline{x}_2) = \frac{1}{2-1}\sum_{j=1}^{2} x_{j1}x_{j2} = (-1)\times 2 + 1\times(-2) = -4$$

Here the scaled covariance is the same as the non-scaled covariance because the sample size is 2 so $n-1 = 1$.

Next we determine the eigenvalues, and from there the eigenvectors, for the matrix $\mathbf{X}^T\mathbf{X}$:

$$\begin{vmatrix} 2-\lambda & -4 \\ -4 & 8-\lambda \end{vmatrix} = 0 \;\Rightarrow\; (2-\lambda)(8-\lambda) - 16 = 0 \;\Rightarrow\; \lambda^2 - 10\lambda = 0 \;\Rightarrow\; \lambda = 0 \text{ or } 10$$

When $\lambda = 0$:

$$\begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \;\Rightarrow\; \left.\begin{matrix} 2x - 4y = 0 \\ -4x + 8y = 0 \end{matrix}\right\} \; x = 2y \;\Rightarrow\; \text{one eigenvector is } \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

When $\lambda = 10$:

$$\begin{pmatrix} -8 & -4 \\ -4 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \;\Rightarrow\; \left.\begin{matrix} -8x - 4y = 0 \\ -4x - 2y = 0 \end{matrix}\right\} \; y = -2x \;\Rightarrow\; \text{one eigenvector is } \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

The unit eigenvectors are:

$$\frac{1}{\sqrt{2^2+1^2}}\begin{pmatrix}2\\1\end{pmatrix}=\frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\end{pmatrix} \qquad\qquad \frac{1}{\sqrt{1^2+(-2)^2}}\begin{pmatrix}1\\-2\end{pmatrix}=\frac{1}{\sqrt{5}}\begin{pmatrix}1\\-2\end{pmatrix}$$

By definition:

$$\mathbf{W}=\frac{1}{\sqrt{5}}\begin{pmatrix}1 & 2\\-2 & 1\end{pmatrix}$$

The principal components decomposition $\underline{P}$ of $\underline{X}$ is then defined as $\underline{P}=\underline{XW}$.

$$\mathbf{P}=\mathbf{XW}=\frac{1}{\sqrt{5}}\begin{pmatrix}-1 & 2\\1 & -2\end{pmatrix}\begin{pmatrix}1 & 2\\-2 & 1\end{pmatrix}=\frac{1}{\sqrt{5}}\begin{pmatrix}-5 & 0\\5 & 0\end{pmatrix}$$

It is obvious that the second column doesn't provide any useful information but we consider the deletion of components below.

We have now transformed the data into a set of p orthogonal components. In order to make this more useful, ensure that the eigenvectors in $\underline{W}$ are ranked by the largest eigenvalue – *ie* the components which have the most explanatory power come first.

$\mathbf{W}$ is orthogonal if $\mathbf{W}^{-1}=\mathbf{W}^T$. It follows from this definition that the columns of $\mathbf{W}$ are orthogonal vectors, each with magnitude 1. In our example we did construct $\mathbf{W}$ ranked by the largest eigenvalue.

The goal is now to move on from simply transforming the data, and instead to use fewer than p components, so that we reduce the dimensionality of the problem. By eliminating those components with the least explanatory power we won't sacrifice too much information.

To assess the explanatory power of each component, consider $\underline{S}=\underline{P}^T\underline{P}$. This is a diagonal matrix where each diagonal element is the (scaled) variance of each component of the transformed data (the covariance between components is zero by construction).

$$\mathbf{P}^T\mathbf{P}=\left(\frac{1}{\sqrt{5}}\right)^2\begin{pmatrix}-5 & 5\\0 & 0\end{pmatrix}\begin{pmatrix}-5 & 0\\5 & 0\end{pmatrix}=\frac{1}{5}\begin{pmatrix}50 & 0\\0 & 0\end{pmatrix}=\begin{pmatrix}10 & 0\\0 & 0\end{pmatrix}$$

Recall that $\mathbf{X}^T\mathbf{X}$ gives the (scaled) covariance of the data using the original variables. Hence $\mathbf{P}^T\mathbf{P}$ gives the (scaled) covariance of the data using the new variables (components). Since the components are uncorrelated, the covariances between them are zero. The diagonal elements give the (scaled) variances of each component (the values in matrix $\mathbf{P}$). The sample variances are the diagonal elements divided by $(n-1)$, which in this example is just 1. Incidentally, it should be noted that the sample variances are equal to the corresponding eigenvalues.

For a given q, it is useful to consider the proportion of the variance that is explained by the first q principal components. This is given by the sum of the first q diagonal elements of $\underline{S}$ divided by the sum of all the diagonal elements. It is often the case that even for large data sets with many variables, the first two or three principal components explain 90% or even more of the total variance which is sufficient to model the situation.

In our example, 100% of the variance is explained by the first component.

There's no hard and fast rule for deciding exactly how many principal components we should use. One criterion is to keep as many components as necessary in order to 'explain' at least 90% of the total variance. Other criteria are covered after the Core Reading example on the following pages and will be considered further in PBOR.

Since $\underline{W}$ is orthogonal by construction, $\underline{X} = \underline{P}\underline{W}^T$.

This allows us to reconstruct the original (centred) data using all or just the reduced number of components.

In the general case, we would set the columns in **P** of the components we are eliminating to zero.

The R function for PCA on a numeric data frame <data> is:

```
prcomp(<data>)
```

Technically this uses the more numerically stable method of singular value decomposition (SVD) of the data matrix which obtains the same answers as using the eigenvalues of covariance matrix.

An alternative R function for PCA is `princomp(<data>)` which does use eigenvalues but also uses n rather than $n-1$ as the divisor in the variance/covariance matrix. Hence it will give slightly different results to `prcomp(<data>)`.

Notes:

1. **Since the principal components are linear combinations of the variables it is not useful for reducing the dimensionality where there are non-linear relationships. A suitable transformation (such as log) should be applied first.**

2. **Since the loadings of each variable that make up the components are chosen by maximising variance, variables that have the highest variance will be given more weight. It is often good practice (especially if different units of measurement are used for each variables) to scale the data before applying PCA.**

3. **No explanation has been provided for what these components represent in a practical, real-world sense. Intuitively, the first component is the overall trend of the data. For the second component onwards, intuition for this must be sought elsewhere. This is often done by regressing the components against variables external to the data which the statistical analyst has an *a priori* cause to believe may have explanatory power.**

There is now a Core Reading example based on the equity returns data from Section 2.1. It is very hard to check these figures manually due to the amount of data. We recommend you try to follow what is being done without attempting to check the numbers.

Consider our set of equity returns, $\underline{X}$, from four different markets across 12 time periods.

We obtain $\underline{X}$ by first centering the data (*ie* by subtracting the means of each market). Then:

$$\underline{X}^T\underline{X} = \begin{pmatrix} 0.01431 & 0.01310 & 0.01308 & 0.01249 \\ 0.01310 & 0.02830 & 0.01026 & 0.01245 \\ 0.01308 & 0.01026 & 0.01315 & 0.01192 \\ 0.01249 & 0.01245 & 0.01192 & 0.01153 \end{pmatrix}$$

This is the variance/covariance matrix of the centred data. Since we have 12 observations we need to divide each of these figures by $(n-1)=11$ to get the sample variance/covariances.

The eigenvectors are:

$$\underline{W} = \begin{pmatrix} -0.48118 & -0.33488 & 0.80202 & -0.11440 \\ -0.62118 & 0.77332 & -0.06531 & -0.10879 \\ -0.43394 & -0.47122 & -0.53559 & -0.55026 \\ -0.44078 & -0.26035 & -0.25621 & 0.81993 \end{pmatrix}$$

These eigenvectors have magnitude of 1.

Hence the principal component decomposition is:

$$\underline{P} = \underline{X}\underline{W} = \begin{pmatrix} -0.02822 & 0.04287 & 0.01630 & 0.00286 \\ 0.01374 & -0.02875 & -0.00084 & -0.00110 \\ 0.09663 & 0.01781 & 0.00049 & -0.00732 \\ -0.00237 & -0.07401 & 0.00630 & -0.00166 \\ 0.11079 & 0.00291 & 0.00195 & 0.00358 \\ 0.04243 & 0.02115 & -0.00744 & -0.00116 \\ -0.11673 & -0.01947 & -0.00169 & 0.00145 \\ -0.02033 & -0.02593 & -0.00545 & 0.00113 \\ -0.01066 & -0.00197 & 0.00571 & -0.00172 \\ 0.05706 & 0.01253 & -0.00543 & 0.00545 \\ -0.03578 & 0.00828 & -0.00639 & 0.00219 \\ -0.10657 & 0.04458 & -0.00350 & -0.00369 \end{pmatrix}$$

Now consider:

$$\underline{S} = \underline{P}^T\underline{P} = \begin{pmatrix} 0.05445 & 0 & 0 & 0 \\ 0 & 0.01218 & 0 & 0 \\ 0 & 0 & 0.00051 & 0 \\ 0 & 0 & 0 & 0.00013 \end{pmatrix}$$

This is the (scaled) variance/covariance matrix of the principal components. The diagonal elements are the scaled variances (the sample variances are these figures divided by 11) and the other elements are the scaled covariances (which are all zero as the components are uncorrelated).

The total (scaled) variance is the sum of the diagonals, which is 0.06727. We can now calculate how much of this total variance each principal component explains.

The first principal component explains 80.9% of the total variance, the first two 99.0%, and the first three 99.8%.

We obtain these figures as follows:

$$\frac{0.05445}{0.06727} = 80.9\%, \quad \frac{0.05445 + 0.01218}{0.06727} = 99.0\%, \quad \frac{0.05445 + 0.01218 + 0.00051}{0.06727} = 99.8\%,$$

It would therefore seem reasonable in this example to reduce the dimensionality to 2, using the first two components of P .

The decision criteria used here is to choose principal components that explain at least 90% of the total variance.

We would now continue our modelling using methods such as linear regression and GLMs on this reduced data set.

Whilst the first component is the trend, the second component will need to be regressed against one or several variables to determine an intuitive explanation.

To choose which components to keep we could also a *Scree test*. This involves the examination of a line chart of the variances of each component (called a Scree diagram). The Scree test retains only those principal components before the variances level off (which is easy to observe from the Scree diagram). For the Core Reading example, the Scree diagram is:

Since the scree plot levels off after the first two components this would imply that these two components are enough.

A further alternative is the *Kaiser test*. This suggests only using components with variances greater than 1. This method is only suitable if the data values are scaled and hence is not appropriate here as the data has only been centred (not scaled).

The chapter summary starts on the next page so that you can keep
all the chapter summaries together for revision purposes.

Chapter 10 Summary

Exploratory data analysis (EDA) is the process of analysing data to gain further insight into the nature of the data, its patterns and relationships between the variables, before any formal statistical techniques are applied.

Scatterplots are the first step to visualise the data and assess the shape of any correlation between a pair of variables. The strength of that correlation is measured by the sample correlation coefficient which takes a value from -1 to $+1$.

We can carry out hypothesis tests on the true population correlation coefficient, ρ, between two variables using the t result, the Fisher Z test or permutations.

Principal component analysis (PCA) is a method for reducing the dimensionality of a data set, $\underline{X}$ by identifying the key components necessary to model and understand the data. These components are chosen to be uncorrelated linear combinations of the variables of the data which maximise the variance.

The principal components decomposition $\mathbf{P}$ of $\mathbf{X}$ (an $n \times p$ centred data matrix) is defined to be $\mathbf{P} = \mathbf{XW}$, where $\mathbf{W}$ is a $p \times p$ matrix, whose columns are the eigenvectors of the matrix $\mathbf{X}^T \mathbf{X}$.

The (scaled) covariance $\mathbf{S} = \mathbf{P}^T \mathbf{P}$ gives the explanatory power of each component.

To choose which components to retain, we can keep as many components as necessary in order to 'explain' at least 90% of the total variance or use other criteria such as the Scree test or Kaiser test.

Pearson's correlation coefficient

Measures the strength of linear relationship between *x* and *y*.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Under $H_0 : \rho = 0$, $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$

Otherwise $\qquad \tanh^{-1} r \doteq N\left(\tanh^{-1} \rho, \dfrac{1}{n-3} \right)$

Spearman's rank correlation coefficient

Measures the strength of monotonic relationship. Use Pearson's formula but with ranks. If no ties then this simplifies to:

$$r_s = 1 - \frac{6\sum\limits_{i} d_i^2}{n(n^2-1)} \text{ where } d_i = r(X_i) - r(Y_i)$$

For inference use permutations or, for $n > 20$, Pearson's formulae with ranks. The limiting distribution is $N\left(0, \dfrac{1}{n-1}\right)$.

Kendall's rank correlation coefficient

Measures the strength of dependence of rank correlation. If no ties then:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

where n_c is the number of concordant pairs (where the ranks of both elements agree) and n_d is the number of discordant pairs.

For inference use permutations or, for $n > 10$, a $N\left(0, \dfrac{2(2n+5)}{9n(n-1)}\right)$ distribution.

Chapter 10 Practice Questions

10.1   A new computerised ultrasound scanning technique has enabled doctors to monitor the weights
        of unborn babies. The table below shows the estimated weights for one particular baby at
        fortnightly intervals during the pregnancy.

Gestation period (weeks)	30	32	34	36	38	40
Estimated baby weight (kg)	1.6	1.7	2.5	2.8	3.2	3.5

$$\sum x = 210 \quad \sum x^2 = 7{,}420 \quad \sum y = 15.3 \quad \sum y^2 = 42.03 \quad \sum xy = 549.8$$

(i)   Show that $S_{xx} = 70, S_{yy} = 3.015$ and $S_{xy} = 14.3$.

(ii)   Show that the (Pearson's) linear correlation coefficient is equal to 0.984 and comment.

(iii)   Explain why the Spearman's and Kendall's rank correlation coefficients are both equal to
        1.

(iv)   Carry out a test of $H_0 : \rho = 0$ vs $H_1 : \rho > 0$ using Pearson's correlation coefficient and:

        (a)   the t-test

        (b)   Fisher's transformation.

(v)   Test whether Pearson's sample correlation coefficient supports the hypothesis that the
        true correlation parameter is greater than 0.9.

10.2   A schoolteacher is investigating the claim that class size does not affect GCSE results. His

Exam style

        observations of nine GCSE classes are as follows:

Class	X1	X2	X3	X4	Y1	Y2	Y3	Y4	Y5
Students in class (c)	35	32	27	21	34	30	28	24	7
Average GCSE point score for class (p)	5.9	4.1	2.4	1.7	6.3	5.3	3.5	2.6	1.6

$$\sum c = 238 \quad \sum c^2 = 6{,}884 \quad \sum p = 33.4 \quad \sum p^2 = 149.62 \quad \sum cp = 983$$

(i)   (a)   Calculate Pearson's, Spearman's and Kendall's correlation coefficients.

        (b)   Use Pearson's correlation coefficient to test whether or not the data agrees with
                the claim that class size does not affect GCSE results.   [10]

(ii)   Following his investigation, the teacher concludes, 'bigger class sizes improve GCSE
        results'. Comment on this statement.   [2]

                                                                                          [Total 12]

10.3 A university wishes to analyse the performance of its students on a particular degree course. It records the scores obtained by a sample of 12 students at entry to the course, and the scores obtained in their final examinations by the same students. The results are as follows:

Student	A	B	C	D	E	F	G	H	I	J	K	L
Entrance exam score x (%)	86	53	71	60	62	79	66	84	90	55	58	72
Finals paper score y (%)	75	60	74	68	70	75	78	90	85	60	62	70

$$\sum x = 836 \quad \sum y = 867 \quad \sum x^2 = 60{,}016 \quad \sum y^2 = 63{,}603 \quad \sum (x - \bar{x})(y - \bar{y}) = 1{,}122$$

(i) (a) Explain why Spearman's and Kendall's rank correlation coefficients cannot be calculated here using the simplified formula.

(b) Calculate the Pearson's correlation coefficient. [3]

(ii) Test whether this data could come from a population with Pearson's correlation coefficient equal to 0.75. [3]

[Total 6]

Chapter 10 Solutions

10.1 **(i)**   ***Calculate Sxx, Syy and Sxy***

$$S_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 = 7{,}420 - \frac{1}{6} \times 210^2 = 70$$

$$S_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 = 42.03 - \frac{1}{6} \times 15.3^2 = 3.015$$

$$S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) = 549.8 - \frac{1}{6} \times 210 \times 15.3 = 14.3$$

(ii)   ***Calculate (Pearson's) linear correlation coefficient and comment***

Using the results from part (i):

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{14.3}{\sqrt{70 \times 3.015}} = 0.984336$$

There is a strong linear association between gestation period and foetal weight.

(iii)   ***Explain why the Spearman and Kendall rank correlation coefficients are both equal to 1***

The ranks of the two variables (gestation period and weight) are exactly equal, hence Spearman's rank correlation coefficient is equal to 1.

This means that all the pairs are concordant, and so τ is also equal to 1.

(iv)(a)   ***Test*** $\rho > 0$ ***using Pearson's correlation coefficient and the t-test***

We are testing $H_0 : \rho = 0$ vs $H_1 : \rho > 0$.

If H_0 is true, then the test statistic $\dfrac{r\sqrt{4}}{\sqrt{1-r^2}}$ has a t_4 distribution.

The observed value of this statistic is $\dfrac{0.984336 \times 2}{\sqrt{1-0.984336^2}} = 11.17$. This is much greater than 8.610, the upper 0.05% point of the t_4 distribution.

So, we reject H_0 at the 0.05% level and conclude that there is very strong evidence that $\rho > 0$ *ie* that there is a positive linear correlation between the baby's weight and gestation period.

(iv)(b)   **Test $\rho > 0$ using Pearson's correlation coefficient and Fisher's transformation**

If H_0 is true, then the test statistic $Z_r = \tanh^{-1} r$ has a $N(0, \frac{1}{3})$ distribution.

The observed value of this statistic is $\tanh^{-1} 0.984336 = 2.4208$, which corresponds to a value of $\dfrac{2.4208}{\sqrt{1/3}} = 4.193$ on the $N(0,1)$ distribution. This is much greater than 3.090, the upper 0.1% point of the standard normal distribution.

So, we reject H_0 at the 0.1% level and conclude that there is very strong evidence that $\rho > 0$ *ie* that there is a positive correlation between the baby's weight and gestation period.

(v)   **Test whether Pearson's sample correlation coefficient supports $\rho > 0.9$**

We are testing $H_0 : \rho = 0.9$ vs $H_1 : \rho > 0.9$.

If H_0 is true, then the test statistic Z_r has a $N(z_\rho, \frac{1}{3})$ distribution, where $z_\rho = \tanh^{-1} 0.9 = 1.4722$

The observed value of this statistic is $\tanh^{-1} 0.984336 = 2.4208$, which corresponds to a value of $\dfrac{2.4208 - 1.4722}{\sqrt{1/3}} = 1.643$ on the $N(0,1)$ distribution. This is just less than 1.645, the upper 5% point of the standard normal distribution.

So, we cannot reject H_0 at the 5% level *ie* the data does not provide enough evidence to conclude that the correlation parameter between the baby's weight and gestation period exceeds 0.9.

10.2   (i)(a)   **Calculate the correlation coefficients**

Pearson correlation coefficient

$$S_{cc} = \sum c^2 - \frac{\left(\sum c\right)^2}{n} = 6{,}884 - \frac{238^2}{9} = 590.2222 \qquad [½]$$

$$S_{cp} = \sum cp - \frac{\left(\sum c\right)\left(\sum p\right)}{n} = 983 - \frac{238 \times 33.4}{9} = 99.75556 \qquad [½]$$

$$S_{pp} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = 149.62 - \frac{33.4^2}{9} = 25.66889 \qquad [½]$$

$$\Rightarrow \quad r = \frac{S_{cp}}{\sqrt{S_{cc}S_{pp}}} = \frac{99.75556}{\sqrt{590.2222 \times 25.66889}} = 0.81045 \qquad [1½]$$

Spearman rank correlation coefficient

The ranks (from lowest to highest) and differences are as follows:

Class	X1	X2	X3	X4	Y1	Y2	Y3	Y4	Y5
Students in class (c)	9	7	4	2	8	6	5	3	1
Average GCSE point score for class (p)	8	6	3	2	9	7	5	4	1
Differences	1	1	1	0	−1	−1	0	−1	0

[1]

Hence

$$r_s = 1 - \frac{6 \times 6}{9(9^2 - 1)} = 0.95$$ [1]

Kendall rank correlation coefficient

Arranging in order of class rank:

Class	Y5	X4	Y4	X3	Y3	Y2	X2	Y1	X1
Students in class (c)	1	2	3	4	5	6	7	8	9
Average GCSE point score for class (p)	1	2	4	3	5	7	6	9	8
# concordant pairs	8	7	5	5	4	2	2	0	0
# discordant pairs	0	0	1	0	0	1	0	1	0

[1]

Totalling the rows gives $n_c = 33$, $n_d = 3$. Hence:

$$\tau = \frac{33 - 3}{9(9 - 1)/2} = 0.8\dot{3}$$ [1]

(i)(b)   *Test whether class size does not affect GCSE results*

We are testing:

$$H_0: \quad \rho = 0 \qquad vs \qquad H_1: \quad \rho \neq 0$$ [½]

The statistic is:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$ [½]

Under H_0 our t-statistic is:

$$\frac{0.81045\sqrt{7}}{\sqrt{1-0.81045^2}} = 3.660$$ [1]

This is greater than the 0.5% critical value of 3.499 for 7 degrees of freedom.   [½]

Therefore, we have sufficient evidence at the 1% level to reject H_0. Therefore we conclude that there *is* a correlation between class size and GCSE results (*ie* class size does affect GCSE results).   [½]

[Total 10]

We could use Fisher's transformation, but since this is only an approximation, it makes sense to use this accurate version instead when testing whether $\rho = 0$.

(ii)   **Comment**

There is strong positive correlation between class size and GCSE results (*ie* bigger classes have better GCSE results).   [1]

However, correlation does not necessarily imply causation, *ie* whilst bigger classes have better results, it is not necessarily the class size that causes the improvement.   [1]

[Total 2]

10.3   (i)(a)   **Why can't we use simplified formulae**

The ranks (from lowest to highest) and differences are as follows:

Student	A	B	C	D	E	F	G	H	I	J	K	L
Entrance exam score x (%)	11	1	7	4	5	9	6	10	12	2	3	8
Finals paper score y (%)	8	1	7	4	5	8	10	12	11	1	3	5

Since we have tied ranks we cannot use the simplified formula for Spearman or Kendall.   [1]

We would have to use a correction – which is best handled by a computer.

(i)(b)   **Pearson's correlation coefficient**

$$S_{xx} = 60,016 - \frac{836^2}{12} = 1,774.67$$

$$S_{yy} = 63,603 - \frac{867^2}{12} = 962.25$$ [1]

$$S_{xy} = 1,122$$

Therefore:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{1,122}{\sqrt{1,774.67 \times 962.25}} = 0.85860$$ [1]

                                                                                                    [Total 3]

(ii)   **Test whether** $\rho = 0.75$

We are testing $H_0 : \rho = 0.75$ *vs* $H_1 : \rho \neq 0.75$ [½]

If H_0 is true, then the test statistic Z_r has a $N\left(z_\rho, \frac{1}{9}\right)$ distribution, where

$z_\rho = \tanh^{-1} 0.75 = 0.97296$.   [½]

The observed value of this statistic is $\tanh^{-1} 0.85860 = 1.2880$, which corresponds to a value of

$\dfrac{1.2880 - 0.97296}{\sqrt{1/9}} = 0.945$ on the $N(0,1)$ distribution.   [1]

This is clearly less than 1.96, the upper 2.5% point of the standard normal distribution.   [½]

So, we have insufficient evidence at the 5% level to reject H_0 *ie* the data does not provide enough evidence to conclude that the correlation parameter is any different from 0.75.   [½]

                                                                                                    [Total 3]

11

Linear regression

Syllabus objectives

4.1   Linear regression

   4.1.1   Explain what is meant by response and explanatory variables.

   4.1.2   State the simple regression model (with a single explanatory variable).

   4.1.3   Derive the least squares estimates of the slope and intercept parameters in a simple linear regression model.

   4.1.4   Use appropriate statistical software to fit a simple linear regression model to a data set and interpret the output.

   - Perform statistical inference on the slope parameter.

   - Describe the use of various measures of goodness of fit of a linear regression model (R^2).

   - Use a fitted linear relationship to predict a mean response or an individual response with confidence limits.

   - Use residuals to check the suitability and validity of a linear regression model.

0   Introduction

In the last chapter we examined the correlation between two variables.

If there is a suitably strong enough correlation between the two variables (and there is cause and effect) we can justifiably calculate a 'regression line' which gives the mathematical form of this relationship:

Much of this chapter is concerned with obtaining estimates of the variables associated with this regression line and giving confidence intervals for our estimates using the methods from Chapter 8. Due to the mathematically rigorous nature of this work, there are a number of results that are quoted without proof and students are expected to memorise and apply these results in the exam.

This is a long chapter and will probably require two study sessions to cover it in detail. In the past, this material often formed one of the longer questions in the Subject CT3 exam.

In the previous chapter we carried out correlation analysis on bivariate and multivariate data to assess the strength of the relationship between variables.

In this unit we look at regression analysis to assess the nature of the relationship between Y, the response (or dependent) variable, and X, the explanatory (or independent, or regressor) variable(s).

The values of the response variable (our principal variable of interest) depend on, or are, in part, explained by, the values of the other variable(s), which is referred to as the explanatory variable(s).

Ideally, the values used for the explanatory variable(s) are controlled by the experimenter — (in the analysis they are in fact assumed to be error-free constants, as opposed to random variables with distributions).

Regression analysis consists of choosing and fitting an appropriate model — usually with a view to estimating the mean response (*ie* the mean value of the response variable) for specified values of the explanatory variable(s). A prediction of the value of an individual response may also be needed.

In this chapter only linear relationships will be considered which assume that the expected value of Y, for any given value x of X, is a linear function of that value x. For the bivariate case this simplifies to:

$$E[Y \mid x] = \alpha + \beta x$$

For the multivariate case with k explanatory variables, this is:

$$E\left[Y \mid x_1, x_2, ..., x_k \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Recall from Chapter 4 that $E[Y \mid x]$ is a conditional mean, which represents the average value of Y corresponding to a given value of x.

As always, before selecting and fitting a model, the data must be examined (*eg* in scatterplots) to see which types of model (and model assumptions) may or may not be reasonable.

Question

A sample of ten claims and corresponding payments on settlement for household policies is taken from the business of an insurance company.

The amounts, in units of £100, are as follows:

Claim x	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
Payment y	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45

The scatterplot from the previous unit was as follows:

Settlement payments for claims

Discuss whether a linear regression model is appropriate.

Solution

There appears to be a strong positive linear relationship and so fitting a linear regression model is appropriate.

If a non-linear relationship (or no relationship) between the variables is indicated by the data, then the methods of analysis discussed here are not applicable for the data as they stand. However a well-chosen transformation of y (or x, or even both) may bring the data into a form for which these methods are applicable.

The purpose of the transformation is to change the relationship into linear form, *ie* into the form $Y = a + bX$.

Question

Explain how to transform the relationship $Y = ab^x$ to a linear form.

Solution

If we take logs, the relationship becomes:

$$\log Y = \log a + x \log b$$

So if we work in terms of the variable $Y' = \log Y$, we have a linear relationship:

$$Y' = \log a + x \log b$$

---

1   The simple bivariate linear model

1.1   Model specification

Given a set of n pairs of data (x_i, y_i), $i = 1, 2, ..., n$, the y_i are regarded as observations of a response variable Y_i. For the purposes of the analysis the x_i, the values of an explanatory variable, are regarded as constant.

The simple linear regression model (with one explanatory variable):

The response variable Y_i is related to the value x_i by:

$$Y_i = \alpha + \beta x_i + e_i \quad i = 1, 2, ..., n$$

where the e_i are uncorrelated error variables with mean 0 and common variance σ^2.

So $E[e_i] = 0$, $var[e_i] = \sigma^2$, $i = 1, 2, ..., n$.

β is the slope parameter, α the intercept parameter.

This is equivalent to saying that $y = mx + c$, where *m* is the gradient or slope and *c* is the intercept *ie* where the line crosses the *y*-axis.

1.2   Fitting the model

We can estimate the parameters in a regression model using the 'method of least squares'.

Fitting the model involves:

(a)   estimating the parameters β and α, and

(b)   estimating the error variance σ^2.

The fitted regression line, which gives the estimated value of *Y* for a fixed *x*, is given by:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

where $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

These are the equations we use to calculate the 'best' values of α and β. They are given in the *Tables*.

Recall from the previous chapter that:

$$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 \big/ n = \sum x_i^2 - n\bar{x}^2$$

$$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 \big/ n = \sum y_i^2 - n\bar{y}^2$$

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i) \big/ n = \sum x_i y_i - n\overline{xy}$$

In regression questions, $\sum_{i=1}^{n} x_i^2$ is often abbreviated to $\sum x^2$, *etc* to simplify the notation.

Question

Show that the first of these relationships is true, *ie* that:

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - n\overline{x}^2$$

Solution

Expanding the bracket and splitting up the summation, we have:

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum (x_i^2 - 2\overline{x}\,x_i + \overline{x}^2)$$

$$= \sum x_i^2 - 2\overline{x}\sum x_i + \sum \overline{x}^2$$

$$= \sum x_i^2 - \frac{2\left(\sum x_i\right)^2}{n} + n\frac{\left(\sum x_i\right)^2}{n^2} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Now since $\sum x_i = n\overline{x}$, we get:

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - \frac{(n\overline{x})^2}{n} = \sum x_i^2 - n\overline{x}^2$$

These formulae are given on page 24 of the *Tables* in the $S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2$ format.

We are now going to look at how these estimates are derived.

This simply means that for a set of paired data $\{(x_i, y_i)\,;\; i = 1,\, 2,\dots,n\,\}$, the least squares estimates of the regression coefficients are the values $\hat{\alpha}$ and $\hat{\beta}$ for which:

$$q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\alpha + \beta x_i) \right]^2$$

is a minimum.

In fact for any model $Y_i = g(x_i) + e_i$, the least squares estimates of the regression coefficients can be determined as the values for which $q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - g(x_i)]^2$ is a minimum. The equations we need to solve to find the values of $\hat{\alpha}$ and $\hat{\beta}$ are sometimes called 'normal equations'.

Differentiating q partially with respect to α and β, and equating to zero, gives the normal equations:

$$\sum_{i=1}^{n} y_i = \hat{\alpha} n + \hat{\beta} \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2$$

Solving these equations by using determinants or the method of elimination then gives the least squares estimate of β as:

$$\hat{\beta} = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

This is just $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$.

The first of the two normal equations gives $\hat{\alpha}$ as:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \hat{\beta} \sum_{i=1}^{n} x_i}{n} = \overline{y} - \hat{\beta}\,\overline{x}$$

Being able to produce a full derivation of these results is important, as they were often tested in the past in the Subject CT3 exam.

Note that a fitted line will pass through the point $(\overline{x}, \overline{y})$.

Question

Show that the fitted line $\hat{y} = \hat{\alpha} + \hat{\beta} x$ passes through the point $(\overline{x}, \overline{y})$.

Solution

Substituting $\overline{x}$ into the RHS of $\hat{y} = \hat{\alpha} + \hat{\beta} x$ gives:

$$\hat{y} = \hat{\alpha} + \hat{\beta}\overline{x}$$

But $\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}$ hence:

$$\hat{y} = (\overline{y} - \hat{\beta}\,\overline{x}) + \hat{\beta}\,\overline{x} = \overline{y}$$

$\hat{\beta}$ is the observed value of a statistic $\hat{B}$ whose sampling distribution has the following properties:

$$E[\hat{B}] = \beta, \qquad \text{var}[\hat{B}] = \frac{\sigma^2}{S_{xx}}$$

The estimate of the error variance σ^2 is based on the sum of squares of the estimated errors:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$$

Alternatively, this can be calculated more easily using this equivalent formula:

$$\hat{\sigma}^2 = \frac{1}{n-2} \left(S_{yy} - S_{xy}^2 / S_{xx} \right)$$

This is given on page 24 of the *Tables*. We will see later that this is an unbiased estimate of σ^2.

The R code to fit a linear model to a bivariate data frame, `<data>=c(X,Y)` **and assign it to the object model, is:**

```
model <- lm(Y ~ X)
```

Then the estimates of the coefficients and error standard deviation can be obtained by:

```
summary(model)
```

The R code to draw the fitted regression line on an existing plot is:

```
abline(model)
```

We will cover the R for this chapter in the Paper B Online Resources (PBOR).

Question

The sample of ten claims and payments above (in units of £100) has the following summations:

$$\sum x = 35.4, \ \sum x^2 = 133.76, \ \sum y = 32.87, \ \sum y^2 = 115.2025, \ \sum xy = 123.81$$

Calculate the fitted regression line and the estimated error variance.

Solution

Number of pairs of observations $n = 10$.

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 133.76 - \frac{35.4^2}{10} = 8.444$$

$$S_{yy} = \sum y^2 - n\bar{y}^2 = 115.2025 - \frac{32.87^2}{10} = 7.1588$$

$$S_{xy} = \sum xy - n\bar{x}\bar{y} = 123.81 - \frac{(35.4 \times 32.87)}{10} = 7.4502$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{7.4502}{8.444} = 0.88231$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 3.287 - (0.88231 \times 3.54) = 0.164$$

$$\hat{\sigma}^2 = \frac{1}{n-2}(S_{yy} - S_{xy}^2/S_{xx}) = \frac{1}{8}\left(7.1588 - 7.4502^2/8.444\right) = 0.0732$$

So the fitted regression line is $\hat{y} = 0.164 + 0.8823x$ which is shown on the graph below.

Settlement payments for claims

You may wish to try Q11.1(i) to check you can calculate these on your own for a different data set.

Once we have worked out the estimates of α and β, we can calculate 'predicted' values of y corresponding to x_i using the formula $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$.

Question

For the claims settlement question above, calculate the expected payment on settlement for a claim of £350.

Solution

Since we are working in units of £100 a claim of £350 corresponds to $x = 3.5$.

Substituting this into the regression line gives:

$$\hat{y} = 0.164 + 0.8823 \times 3.5 = 3.25$$

So we would expect the settlement payment to be £325.

You may wish to try Q11.1(ii) to check you can calculate this on your own for a different data set.

1.3 Partitioning the variability of the responses

To help understand the 'goodness of fit' of the model to the data, the total variation in the responses, as given by $S_{yy} = \sum(y_i - \overline{y})^2$ should be studied.

Some of the variation in the responses can be attributed to the relationship with x (*eg y may tend to be high when x is high, low when x is low*) and some is random variation (unmodellable) above and beyond that. Just how much is attributable to the relationship – or 'explained by the model' – is a measure of the goodness of fit of the model.

We start from an identity involving y_i (the observed y value), $\overline{y}$ (the overall average of the y values) and $\hat{y}_i$ (the 'predicted' value of y).

Squaring and summing both sides of:

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

gives:

$$\sum(y_i - \overline{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \overline{y})^2$$

the cross-product term vanishing.

The sum on the left is the 'total sum of squares' of the responses, denoted here by SS_{TOT}.

The second sum on the right is the sum of the squares of the deviations of the fitted responses (the estimates of the conditional means) from the overall mean response (the estimate of the overall mean) – it summarises the variability accounted for, or 'explained' by the model. It is called the 'regression sum of squares', denoted here by SS_{REG}.

The first sum on the right is the sum of the squares of the estimated errors (response – fitted response, generally referred to in statistics as a 'residual' from the fit) – it summarises the remaining variability, that between the responses and their fitted values and so 'unexplained' by the model. It is called the 'residual sum of squares', denoted here by

SS_{RES}. **The estimate of** σ^2 **is based on it – it is** $\dfrac{SS_{RES}}{n-2}$.

So:

$$SS_{TOT} = SS_{RES} + SS_{REG}$$

Note that SS_{RES} is often also written as SS_{ERR} ('error').

For computational purposes $SS_{TOT} = S_{yy}$ **and:**

$$SS_{REG} = \sum \left[\left(\hat{\alpha} + \hat{\beta} x_i \right) - \left(\hat{\alpha} + \hat{\beta} \bar{x} \right) \right]^2 = \hat{\beta}^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

The last step uses the fact that $\hat{\beta} = S_{xy} / S_{xx}$.

So $SS_{RES} = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$.

Question

Determine the split of total variation in the claims and payments model between the residual sum of squares and the regression sum of squares.

Recall that:

$$n = 10 \,, \quad \sum x = 35.4 \,, \quad \sum y = 32.87$$

$$S_{xx} = 8.444 \,, \quad S_{yy} = 7.1588 \,, \quad S_{xy} = 7.4502$$

Solution

$$SS_{TOT} = S_{yy} = 7.1588$$

$$SS_{REG} = \frac{S_{xy}^2}{S_{xx}} = \frac{7.4502^2}{8.444} = 6.5734$$

$$\therefore SS_{RES} = SS_{TOT} - SS_{REG} = 0.5854$$

Note that $SS_{RES}/n-2 = 0.5854/8 = 0.0732$ gives the same value of $\hat{\sigma}^2$ that we obtained earlier using the alternative formula $(S_{yy} - S_{xy}^2/S_{xx})/(n-2)$.

You may wish to try Q11.1(iii)(a) to check you can partition the sum of squares on your own.

It can then be shown that:

$$E[SS_{TOT}] = (n-1)\sigma^2 + \beta^2 S_{xx} \qquad E[SS_{REG}] = \sigma^2 + \beta^2 S_{xx}$$

from which it follows that $E[SS_{RES}] = (n-2)\sigma^2$ **.**

Hence:

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)\right] = E\left[\frac{SS_{RES}}{n-2}\right] = \frac{1}{n-2}E[SS_{RES}] = \frac{(n-2)\sigma^2}{n-2} = \sigma^2$$

So $\hat{\sigma}^2$ is an unbiased estimator of σ^2 .

In the case that the data are 'close' to a line ($|r|$ high – a strong linear relationship) the model fits well, the fitted responses (the values on the fitted line) are close to the observed responses, and so SS_{REG} is relatively high with SS_{RES} relatively low.

r is referring to Pearson's correlation coefficient, which we calculated in Chapter 10.

In the case that the data are not 'close' to a line ($|r|$ low – a weak linear relationship) the model does not fit so well, the fitted responses are not so close to the observed responses, and so SS_{REG} is relatively low and SS_{RES} relatively high.

The proportion of the total variability of the responses 'explained' by a model is called the coefficient of determination, denoted R^2. Here, the proportion is:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

[The value of the proportion R^2 is usually quoted as a percentage].

R^2 can take values between 0% and 100% inclusive.

Question

Calculate the coefficient of determination for the claims and payments model and comment on it.

Recall that:

$$SS_{TOT} = 7.1588 \qquad SS_{REG} = 6.5734 \qquad SS_{RES} = 0.5854$$

Solution

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{6.5734}{7.1588} = 0.918 \ (91.8\%)$$

This value is very high and so indicates the overwhelming majority of the variation is explained by the model (and hence very little is left over in residual variation). Hence the linear regression model is a good fit to the data.

You may wish to try Q11.1(iii)(b) to check you can calculate the coefficient of determination on your own.

In this case (the simple linear regression model), note that the value of the coefficient of determination is the square of Pearson's correlation coefficient for the data – since:

$$r = \frac{S_{xy}}{(S_{xx}S_{yy})^{\frac{1}{2}}}$$

The Pearson's sample correlation coefficient was introduced in the previous chapter and is given on page 24 of the *Tables*.

Question

Calculate the correlation coefficient for the claims and payment data by using the coefficient of determination from the previous question.

Solution

$$r = \sqrt{0.918} = \pm 0.958$$

Since we saw earlier that there was a positive relationship between claims and the settlement payments we have a correlation coefficient of 0.958.

The R code to obtain the regression and residual sum of squares for a linear model assigned to the object `model`**, is:**

```
anova(model)
```

The coefficient of determination is given in the output of:

```
summary(model)
```

2   The full normal model and inference

2.1   The full normal model

The model must be specified further in order to make inferences concerning the responses based on the fitted model. In particular, information on the distribution of the Y_i's is required.

In the full model, we now assume that the errors, e_i, are independent and identically distributed as $N(0, \sigma^2)$ variables. This will then allow us to obtain the distributions for β and the Y_i's. We can then use these to construct confidence intervals and carry out statistical inference.

For the full model the following additional assumptions are made:

The error variables e_i are: (a) independent, and (b) normally distributed

Under this full model, the e_i's are independent, identically distributed random variables, each with a normal distribution with mean 0 and variance σ^2. It follows that the Y_i's are independent, normally distributed random variables, with $E[Y_i] = \alpha + \beta x_i$ and $var[Y_i] = \sigma^2$.

$\hat{B}$, being a linear combination of independent normal variables, itself has a normal distribution, with mean and variance as noted earlier.

The further results required are:

(1) $\hat{B}$ and $\hat{\sigma}^2$ are independent

(2) $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2}$ has a χ^2 distribution with $\upsilon = n - 2$.

Note: With the full model in place the Y_i's have normal distributions and it is possible to derive maximum likelihood estimators of the parameters α, , and (since maximum likelihood estimation requires us to know the distribution whereas least squares estimation does not)**.**

It is possible to show that the maximum likelihood estimators of and are the same as the least squares estimates, but the MLE of has a different denominator to the least squares estimate (see Question 11.11).

2.2   Inferences on the slope parameter β

To conform to usual practice the distinction between $\hat{B}$, the random variable, and its value $\hat{\beta}$, will now be dropped. Only one symbol, namely $\hat{\beta}$ will be used.

Using the fact that $E(\hat{\beta}) = \beta$ and $\text{var}(\hat{\beta}) = \sigma^2 / S_{xx}$ from Section 1.2:

$$A = \left(\hat{\beta} - \beta \right) \Big/ \left(\sigma^2 / S_{xx} \right)^{\frac{1}{2}} \text{ is a standard normal variable}$$

Repeating result (2) from Section 2.1:

$$B = (n-2)\hat{\sigma}^2 / \sigma^2 \text{ is a } \chi^2 \text{ variable with } \nu = n-2 \text{ degrees of freedom}$$

Now, since $\hat{\beta}$ and $\hat{\sigma}^2$ are independent, it follows that $A / \left\{ B / (n-2) \right\}^{\frac{1}{2}}$ has a *t* distribution with $\upsilon = n-2$, *ie*:

$$(\hat{\beta} - \beta) / \text{se}(\hat{\beta}) \text{ has a } t \text{ distribution with } \upsilon = n-2 \qquad\qquad (3)$$

where the symbol $\text{se}(\hat{\beta})$ denotes the estimated standard error of $\hat{\beta}$, namely $\left(\hat{\sigma}^2 / S_{xx} \right)^{\frac{1}{2}}$.

Result (3) can now be used for the construction of confidence intervals, and for tests, on the value of β, the slope coefficient in the model. $H_0 : \beta = 0$ is the 'no linear relationship' hypothesis.

Note that since $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$ and $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$, if $\hat{\beta} = 0$ then $S_{xy} = 0$ and $r = 0$ too.

This *t* distribution result for β is given on page 24 of the *Tables*.

Question

For the claims/settlements data:

(a)   calculate a two-sided 95% confidence interval for β, the slope of the true regression line

(b)   test the hypothesis $H_0 : \beta = 1$ vs $H_1 : \beta \neq 1$.

Recall that:

$$S_{xx} = 8.444 \,, S_{yy} = 7.1588 \,, S_{xy} = 7.4502 \,, \hat{\alpha} = 0.164 \,, \hat{\beta} = 0.88231 \,, \hat{\sigma}^2 = 0.0732$$

Solution

(a) $se(\hat{\beta}) = (0.0732 / 8.444)^{\frac{1}{2}} = 0.0931$

95% confidence interval for β is $\hat{\beta} \pm \left\{ t_{.025,8} \times se\left(\hat{\beta}\right) \right\}$

***ie* $0.8823 \pm (2.306 \times 0.0931)$ *ie* 0.8823 ± 0.2147**

So a 95% confidence interval is $(0.668, 1.10)$

(b)   **The 95% two-sided confidence interval in (a) contains the value '1', so the two-sided test in (b) conducted at the 5% level results in H_0 being accepted.**

You may wish to try Q11.1(iv) to check you can carry out a test for β on your own.

> **R**
>
> **In R, the statistic and *p*-value for the test of $H_0 : \beta = 0$ for a linear regression model are displayed in summary(model).**
>
> **In R, 95% confidence intervals for the parameters α and β from a linear regression model are given by:**
>
> ```
> confint(model,level=0.95)
> ```

2.3   Analysis of variance (ANOVA)

In Section 1.3 we partitioned the variability into that arising from the regression model (SS_{REG}) and the left-over residual (or error) variability (SS_{RES} or SS_{ERR}). We then calculated a ratio of the variances to obtain the proportion of variability that was determined (or explained) by the model (called the coefficient of determination, R^2). This gave us a crude measure of fit.

With the distributional assumptions underlying the full regression model we can now do a more formal test of fit. Recall from Chapter 6 that the variance of a sample taken from a normal distribution has a chi-square distribution, $(n-1)S^2 / \sigma^2 \sim \chi^2_{n-1}$, and the ratio of variances of two samples from a normal distribution has an *F* distribution, $(S_1^2 / \sigma_1^2) / (S_2^2 / \sigma_2^2) \sim F_{n_1-1, n_2-1}$. We can therefore use an *F* test to compare the regression variance to the residual variance.

Another method of testing the 'no linear relationship' hypothesis (*ie* $H_0 : \beta = 0$) is to analyse the sum of squares from Section 1.3.

In Section 2.1, we saw under the full normal model that $(n-2)\hat{\sigma}^2 / \sigma^2 \sim \chi^2_{n-2}$. Since $\hat{\sigma}^2 = SS_{RES} / (n-2)$ we have $SS_{RES} / (n-2) \sim \chi^2_{n-2}$.

When H_0 is true, $SS_{TOT}/(n-1)$ is the overall sample variance and so SS_{TOT}/σ^2 is χ^2_{n-1}. Since SS_{REG} and SS_{RES} are in fact independent and SS_{RES}/σ^2 is χ^2_{n-2} it follows that SS_{REG}/σ^2 is χ^2_1.

Therefore:

$$\frac{SS_{REG}}{SS_{RES}/(n-2)} = \frac{\text{regression mean sum of squares}}{\text{residual mean sum of squares}} = \frac{MSS_{REG}}{MSS_{RES}}$$

is $F_{1,n-2}$ and H_0 is rejected for 'large' values of this ratio.

The mean sum of squares (sometimes just called the mean square) is where we divide the sum of squares by the degrees of freedom. The sample variance, $s^2 = \sum(x_i - \overline{x})^2/(n-1)$ is actually a mean sum of squares.

Note that unlike the coefficient of determination we divide the regression variability by the residual rather than the total variability.

A large value of this ratio would mean that the majority of the variability is explained by the linear regression model. Therefore we would reject the null hypothesis of no linear relationship.

The results are usually set out in an ANOVA table:

Source of variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression **Residual**	**1** **$n-2$**	SS_{REG} SS_{RES}	$SS_{REG}/1$ $SS_{RES}/(n-2)$
Total	**$n-1$**	SS_{TOT}	

The test statistic is just the ratio of the values in the last column.

In R, the *F* statistic and its *p*-value for a regression model are given in the output of both `anova(model)` **and** `summary(model)`.

Question

For the data set of 10 claims and their settlement payments, we had:

$$SS_{TOT} = 7.1588 \qquad SS_{REG} = 6.5734 \qquad SS_{RES} = 0.5854$$

Construct the ANOVA table and carry out an *F* test to determine whether $\beta = 0$.

Solution

The ANOVA table is:

Source of variation	*d.f.*	*SS*	*MSS*
Regression	**1**	**6.5734**	**6.5734**
Residual	**8**	**0.5854**	**0.0732**
Total	**9**	**7.1588**	

Under $H_0 : \beta = 0$ we have $F = \dfrac{6.5734}{0.0732} = 89.8$ on (1, 8) degrees of freedom.

P-value of $F = 89.8$ is less than even 0.01, so H_0 is rejected at the 1% level.

You may wish to try Q11.1(v) to check you can construct an ANOVA table and carry out an *F*-test on your own.

2.4 Estimating a mean response and predicting an individual response

(a) Mean response

This is often the main issue – the whole point of the modelling exercise. For example, the expected settlement for claims of £460 can be estimated as follows:

If μ_0 is the expected (mean) response for a value x_0 of the explanatory variable (*ie* $\mu_0 = E[Y \mid x_0] = \alpha + \beta x_0$), μ_0 is estimated by $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta} x_0$, which is an unbiased estimator.

The variance of the estimator is given by:

$$\text{var}\left(\hat{\mu}_0\right) = \left\{ \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}} \right\} \sigma^2$$

This result is given on page 25 of the *Tables*.

The distribution actually used is a *t* distribution – the argument is similar to that described in Section 2.2:

$$(\hat{\mu}_0 - \mu_0) \,/\, \text{se}[\hat{\mu}_0] \text{ has a } t \text{ distribution with } \upsilon = n - 2 \tag{4}$$

where $\text{se}[\hat{\mu}_0]$ denotes the estimated standard error of the estimate, namely:

$$\text{se}\left[\hat{\mu}_0\right] = \left[\left\{ \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}} \right\} \hat{\sigma}^2 \right]^{\frac{1}{2}}$$

Result (4) can be used for the construction of confidence intervals for the value of the expected response when $x = x_0$.

(b)   Individual response

Rather than estimating an expected response $E[Y \mid x_0]$ an estimate, or prediction, of an individual response y_0 (for $x = x_0$) is sometimes required. The actual estimate is the same as in (a), namely:

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$$

but the uncertainty associated with this estimator (as measured by the variance) is greater than in (a) since the value of an individual response y_0 rather than the more 'stable' mean response is required. To cater for the extra variation of an individual response about the mean, an extra term σ^2 has to be added into the expression for the variance of the estimator of a mean response.

In other words, the variance of the individual response estimator is:

$$\text{var}(\hat{y}_0) = \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\} \sigma^2$$

The result is:

$$(\hat{y} - y_0) / \text{se}[\hat{y}_0] \text{ has a } t \text{ distribution with } \upsilon = n - 2 \qquad\qquad (5)$$

where $\text{se}[\hat{y}_0]$ denotes the estimated standard error of the estimate, namely:

$$\text{se}[\hat{y}_0] = \left[\left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\} \hat{\sigma}^2 \right]^{\frac{1}{2}}$$

Result (5) can then be used for the construction of confidence intervals (or prediction intervals) for the value of a response when $x = x_0$.

The resulting interval for an individual response y_0 is wider than the corresponding interval for the mean response μ_0.

Recall that for an individual response value we have $y_i = \alpha + \beta x_i + e_i$, which is the regression line $\alpha + \beta x_i$ plus an error term, e_i. Since $e_i \sim N(0, \sigma^2)$ an individual point is on the regression line on average – hence we have the same estimate $\hat{\alpha} + \hat{\beta}x_0$ as for the mean response. However, we can see that there is an additional σ^2 for the variance.

Question

Consider again the claims/settlements example. Calculate:

(a) **a 95% confidence interval for the expected payments on claims of £460.**

(b) **a 95% confidence interval for the predicted actual payments on claims of £460.**

Recall that $\hat{\alpha} = -0.164, \hat{\beta} = 0.88231, \hat{\sigma}^2 = 0.0732$ and $S_{xx} = 8.444$.

Solution

(a) **Estimate of expected payment $= 0.1636 + 0.88231(4.6) = 4.222$**

$$\text{se of estimate} = \sqrt{\left\{\frac{1}{10} + \frac{(4.6 - 3.54)^2}{8.444}\right\}0.0732} = 0.1306$$

$$t_{0.025,8} = 2.306$$

so confidence interval is $4.222 \pm (2.306 \times 0.1306)$ *ie* 4.222 ± 0.301

***ie* (3.921, 4.523) *ie* (£392, £452)**

(b) **Predicted payment $= 4.222$**

$$\text{se of estimate} = \sqrt{\left\{1 + \frac{1}{10} + \frac{(4.6 - 3.54)^2}{8.444}\right\}0.0732} = 0.3004$$

confidence interval is $4.222 \pm (2.306 \times 0.3004)$ *ie* 4.222 ± 0.693

***ie* (3.529, 4.915) *ie* (£353, £492)**

You may wish to try Q11.1(vi) and (vii) to check you can construct these confidence intervals on your own.

> **R**
>
> **In R, predicted *y* values for, say, $x_0 = 4$ in a linear regression model fitted to a data frame** `c(X,Y)` **can be obtained as follows:**
>
> ```
> newdata <-data.frame(X=4)
> predict(model,newdata)
> ```
>
> **The R code for 95% confidence intervals for the mean and individual response are:**
>
> ```
> predict(model,newdata,interval="confidence",level=0.95)
> predict(model,newdata,interval="predict",level=0.95)
> ```

2.5   Checking the model

The residual from the fit at x_i is the estimated error, the difference between the response y_i and the fitted value *ie*:

$$\text{residual at } x_i \text{ is } \hat{e}_i = y_i - \hat{y}_i$$

The R code for obtaining the fitted values and the residuals of a linear regression model is:

```
fitted(model)

residuals(model)
```

By examining the residuals it is possible to investigate the validity of the assumptions in the model about (i) the true errors e_i (which are assumed to be independent normal variables with means 0 and the same variance σ^2), and (ii) the nature of the relationship between the response and explanatory variables.

Plotting the residuals along a line may suggest a departure from normality for the error distribution. The sizes of the residuals should also be looked at, bearing in mind that the value of $\hat{\sigma}$ estimates the standard deviation of the error distribution.

Ideally we would expect the residuals to be symmetrical about 0 and no more than 3 standard deviations from it. So skewed residuals or outliers would indicate non-normality.

Alternatively a quantile-quantile (Q-Q) plot of the residuals against a normal distribution should form a straight line.

Recall that Q-Q plots were introduced in Chapter 5. They are far superior to dotplots, but will require the use of R to produce them using the function `qqnorm`.

Scatter plots of the residuals against the values of the explanatory variable (or against the values of the fitted responses) are also most informative. If the residuals do not have a random scatter – if there is a pattern – then this suggests an inadequacy in the model.

Question

The claims/settlement data values were as follows:

Claim *x*   2.10 2.40 2.50 3.20 3.60 3.80 4.10 4.20 4.50 5.00

Payment *y*   2.18 2.06 2.54 2.61 3.67 3.25 4.02 3.71 4.38 4.45

Calculate the residuals for the fitted regression model $\hat{y} = 0.164 + 0.8823x$.

Solution

The residuals $\hat{e}_i = y_i - \hat{y}_i$ are given in the table below:

x_i	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
$\hat{e}_i$	0.163	-0.221	0.171	-0.377	0.330	-0.266	0.239	-0.159	0.246	-0.125

The dotplot and the Q-Q plot of the residuals and the plot of the residuals against the explanatory variable are as follows:

There is nothing to suggest non-normality in the first diagram.

The dotplot is symmetrical about 0 and within $3\sigma = 0.811$ either side – so there are no outliers (though ideally we would ideally expect more values in the middle and less at the edge but this is unlikely with such a small data set).

Nor does there appear to be a pattern in the third diagram.

There appears to be no connection between the residuals and the explanatory variable (claims).

However, the Q-Q plot does possibly indicate some deficiency in at least one of the values.

If the residuals are normally distributed we would expect the Q-Q plot to be along the diagonal line whereas one of the values is some way from the line.

The Core Reading now considers a different set of data.

Suppose the plot of the residuals against the explanatory variable was as follows:

We can see that the size of the residuals tends to increase as *x* increases — this suggests that the error variance is not in fact constant, but is increasing with *x*. (A transformation of the responses may stabilise the error variance in a situation like this).

Typically we would log the data in such a situation.

You may wish to try Q11.1(viii) to check you can obtain residuals and interpret their plots.

2.6   Extending the scope of the linear model

In certain 'growth models' the appropriate model is that the expected response is related to the explanatory value through an exponential function – $E[Y_i \mid x_i] = \alpha \exp(\beta x_i)$.

In such a case the response data can be transformed using $w_i = \log y_i$ **and the linear model:**

$$W_i = \eta + \beta x_i + e_i \text{ (where } \eta = \log \alpha)$$

is then fitted to the data (x_i, w_i). **The fact that the error structure is additive in this representation implies that it plays a multiplicative role in the original form of the model. If such a structure is considered invalid, different methods from those covered in this chapter would have to be used.**

The concept of 'error structure' is touching on the subject of generalised linear models which we will study in Chapter 12.

In R we can apply a transformation at the model stage. For example:

```
model <- lm(Y ~ log(X))
```

Chapter 11 Summary

A regression model, such as the simple linear regression model, can be used to model the response when an explanatory variable operates at a given level, or to model bivariate data points.

The linear regression model is given by:

$$Y_i = \alpha + \beta x_i + e_i \qquad \text{where } e_i \sim N(0, \sigma^2)$$

The parameters α, β and σ^2 can be estimated using the formulae:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)$$

These are given on page 24 of the *Tables*.

Confidence intervals can be obtained for β using the result:

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$$

Prediction intervals for a mean response μ_0 or an individual response y_0 can be obtained using the results:

$$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{\left(\dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{S_{xx}} \right)\hat{\sigma}^2}} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{y}_0 - y_0}{\sqrt{\left(1 + \dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{S_{xx}} \right)\hat{\sigma}^2}} \sim t_{n-2}$$

These are also in the *Tables*.

Chapter 11 Summary (continued)

The fit of the linear regression model can be analysed by partitioning the total variance, SS_{TOT}, into that which is explained by the model, SS_{REG}, and that which is not, SS_{RES}. The formulae for these are as follows:

$$SS_{TOT} = \sum (y_i - \bar{y})^2 = s_{yy}$$

$$SS_{REG} = \sum (\hat{y}_i - \bar{y})^2 = \frac{s_{xy}^2}{s_{xx}}$$

$$SS_{RES} = \sum (\hat{y}_i - y_i)^2 = s_{yy} - \frac{s_{xy}^2}{s_{xx}}$$

The coefficient of determination, R^2, gives the percentage of this variance which *is* explained by the model:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{s_{xy}^2}{s_{xx}s_{yy}}$$

Note that $SS_{TOT} = SS_{RES} + SS_{REG}$.

Examining the residuals, $\hat{e}_i = y_i - \hat{y}_i$, we would expect them to be normally distributed about zero and to have no relationship with the *x* values. Both of these features can be examined using diagrams.

Chapter 11 Practice Questions

11.1   A new computerised ultrasound scanning technique has enabled doctors to monitor the weights of unborn babies. The table below shows the estimated weights for one particular baby at fortnightly intervals during the pregnancy.

Gestation period (weeks)	30	32	34	36	38	40
Estimated baby weight (kg)	1.6	1.7	2.5	2.8	3.2	3.5

$$\sum x = 210 \quad \sum x^2 = 7{,}420 \quad \sum y = 15.3 \quad \sum y^2 = 42.03 \quad \sum xy = 549.8$$

(i)   Show that:

(a) $S_{xx} = 70, S_{yy} = 3.015$ and $S_{xy} = 14.3$.

(b)   the fitted regression line is $\hat{y} = -4.60 + 0.2043x$.

(c) $\hat{\sigma}^2 = 0.0234$.

(ii)   Calculate the baby's expected weight at 42 weeks (assuming it hasn't been born by then).

(iii)   (a)   Calculate the residual sum of squares and the regression sum of squares for these data.

(b)   Calculate the coefficient of determination, R^2, and comment on its value.

(iv)   Carry out a test of $H_0 : \beta = 0$ vs $H_1 : \beta > 0$, assuming a linear model is appropriate.

(v)   Construct an ANOVA table for the sum of squares from part (iii)(a) and carry out an *F*-test stating the conclusion clearly.

(vi)   (a)   Estimate the mean weight of a baby at 33 weeks. Calculate the variance of this mean predicted response.

(b)   Hence, calculate a 90% confidence interval for the mean weight of a baby at 33 weeks.

(vii)   (a)   Estimate the actual weight of an individual baby at 33 weeks. Calculate the variance of this individual predicted response.

(b)   Hence, calculate a 90% confidence interval for the weight of an individual baby at 33 weeks.

[ctd.]

The table below shows some of the residuals:

Gestation period (weeks)	30	32	34	36	38	40
Residual	0.07			0.05	0.04	−0.07

(viii)   (a)   Calculate the missing residuals.

          (b)   Draw a dotplot of the residuals and comment.

          (c)   Plot the residuals against the x values and comment on the fit.

          (d)   Comment on the Q-Q plot of the residuals given below:

Normal Q-Q Plot

11.2   An analysis using the simple linear regression model based on 19 data points gave:

$$s_{xx} = 12.2 \qquad s_{yy} = 10.6 \qquad s_{xy} = 8.1$$

(i)   (a)   Calculate $\hat{\beta}$.

        (b)   Test whether β is significantly different from zero.

(ii)   (a)   Calculate r.

        (b)   Test whether ρ is significantly different from zero.

(iii)   Comment on the results of the tests in parts (i) and (ii).

11.3   The sums of the squares of the errors in a regression analysis are found to be:

$$SS_{REG} = \sum(\hat{y}_i - \overline{y})^2 = 6.4 \qquad SS_{RES} = \sum(y_i - \hat{y}_i)^2 = 3.6 \qquad SS_{TOT} = \sum(y_i - \overline{y})^2 = 10.0$$

Calculate the coefficient of determination and explain what this represents.

11.4   Explain how to transform the following models to linear form:

(i) $y_i = a + bx_i^2 + e_i$

(ii) $y_i = ae^{bx_i}$

11.5   A university wishes to analyse the performance of its students on a particular degree course. It
        records the scores obtained by a sample of 12 students at entry to the course, and the scores
        obtained in their final examinations by the same students. The results are as follows:

Student	A	B	C	D	E	F	G	H	I	J	K	L
Entrance exam score *x* (%)	86	53	71	60	62	79	66	84	90	55	58	72
Finals paper score *y* (%)	75	60	74	68	70	75	78	90	85	60	62	70

$$\sum x = 836 \quad \sum y = 867 \quad \sum x^2 = 60{,}016 \quad \sum y^2 = 63{,}603 \quad \sum(x - \overline{x})(y - \overline{y}) = 1{,}122$$

(i)   Calculate the fitted linear regression equation of *y* on *x*.   [3]

(ii)   Assuming the full normal model, calculate an estimate of the error variance σ^2 and
         obtain a 90% confidence interval for σ^2.   [3]

(iii)   By considering the slope parameter, formally test whether the data are positively
         correlated.   [3]

(iv)   Calculate a 95% confidence interval for the mean finals paper score corresponding to an
         individual entrance score of 53.   [3]

(v)   Calculate the proportion of variation explained by the model. Hence, comment on the fit
         of the model.   [2]
                                                                                                    [Total 14]

11.6 The share price, in pence, of a certain company is monitored over an 8-year period. The results are shown in the table below:

Time (years)	0	1	2	3	4	5	6	7	8
Price	100	131	183	247	330	454	601	819	1,095

$$\sum (x_i - \overline{x})^2 = 60 \qquad \sum (y_i - \overline{y})^2 = 925{,}262 \qquad \sum (x_i - \overline{x})(y_i - \overline{y}) = 7{,}087$$

An actuary fits the following simple linear regression model to the data:

$$y_i = \alpha + \beta x_i + e_i \qquad i = 0, 1, \ldots, 8$$

where $\{e_i\}$ are independent normal random variables with mean zero and variance σ^2.

(i) Determine the fitted regression line in which the price is modelled as the response and the time as an explanatory variable. [2]

(ii) Calculate a 99% confidence interval for:

   (a) β, the true underlying slope parameter

   (b) σ^2, the true underlying error variance. [5]

(iii) (a) State the 'total sum of squares' and calculate its partition into the 'regression sum of squares' and the 'residual sum of squares'.

   (b) Use the values in part (iii)(a) to calculate the 'proportion of variability explained by the model' and comment on the result. [5]

(iv) The actuary decides to check the fit of the model by calculating the residuals.

   (a) Complete the table of residuals (rounding to the nearest integer):

Time (years)	0	1	2	3	4	5	6	7	8
Residual	132		−21	−75		−104	−75	25	

   (b) Use a dotplot of the residuals to comment on the assumption of normality.

   (c) Plot the residuals against time and hence comment on the appropriateness of the linear model. [7]

[Total 19]

11.7   A schoolteacher is investigating the claim that class size does not affect GCSE results. His
        observations of nine GCSE classes are as follows:

Class	X1	X2	X3	X4	Y1	Y2	Y3	Y4	Y5
Students in class (c)	35	32	27	21	34	30	28	24	7
Average GCSE point score for class (p)	5.9	4.1	2.4	1.7	6.3	5.3	3.5	2.6	1.6

$$\sum c = 238 \quad \sum c^2 = 6{,}884 \quad \sum p = 33.4 \quad \sum p^2 = 149.62 \quad \sum cp = 983$$

(i)   Determine the fitted regression line for p on c .   [3]

(ii)   Class X5 was not included in the results above and contains 15 students. Calculate an
        estimate of the average GCSE point score for this individual class and specify the standard
        error for this estimate assuming the full normal model.   [4]

                                                                                                   [Total 7]

11.8   An actuary is fitting the following linear regression model through the origin:

$$Y_i = \beta x_i + e_i \qquad e_i \sim N(0, \sigma^2) \qquad i = 1, 2, \dots n$$

(i)   Show that the least squares estimator of β is given by:

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$$ [3]

(ii)   Derive the bias and mean square error of $\hat{\beta}$ under this model.   [4]

                                                                                                   [Total 7]

11.9   A life assurance company is examining the force of mortality, μ_x, of a particular group of policyholders. It is thought that it is related to the age, x, of the policyholders by the formula:

$$\mu_x = Bc^x$$

It is decided to analyse this assumption by using the linear regression model:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{where } \varepsilon_i \sim N(0, \sigma^2) \text{ are independently distributed}$$

The summary results for eight ages were as follows:

Age, x	30	32	34	36	38	40	42	44
Force of mortality, μ_x ($\times 10^{-4}$)	5.84	6.10	6.48	7.05	7.87	9.03	10.56	12.66
$\ln\mu_x$ (3 s.f.)	-7.45	-7.40	-7.34	-7.26	-7.15	-7.01	-6.85	-6.67

$$\sum x_i = 296 \quad \sum x_i^2 = 11{,}120 \quad \sum \ln\mu_{x_i} = -57.129 \quad \sum (\ln\mu_{x_i})^2 = 408.50 \quad \sum x_i \ln\mu_{x_i} = -2{,}104.5$$

(i)   (a)   Apply a transformation to the original formula, $\mu_x = Bc^x$, to make it suitable for analysis by linear regression. Hence, write down expressions for Y, α and β in terms of μ_x, B and c.

      (b)   Plot a graph of $\ln\mu_x$ against the age of the policyholder, x. Hence comment on the suitability of the regression model and state how this supports the transformation in part (a).   [4]

(ii)   Use the data to calculate least squares estimates of B and c in the original formula.   [3]

(iii) (a)   Calculate the coefficient of determination between $\ln\mu_x$ and x. Hence comment on the fit of the model to the data.

       (b)   Complete the table of residuals and use it to comment on the fit.   [5]

Age, x	30	32	34	36	38	40	42	44
Residual, $\hat{e}_i$	0.08		-0.03		-0.06		0.02	0.09

(iv)   Calculate a 95% confidence interval for the mean predicted response $\ln\mu_{35}$ and hence obtain a 95% confidence interval for the mean predicted value of μ_{35}.   [4]

[Total 16]

11.10   The government of a country suffering from hyperinflation has sponsored an economist to monitor
        the price of a 'basket' of items in the population's staple diet over a one-year period. As part of his
        study, the economist selected six days during the year and on each of these days visited a single
        nightclub, where he recorded the price of a pint of lager. His report showed the following prices:

Day (i)	8	29	57	92	141	148
Price (P_i)	15	17	22	51	88	95
$\ln P_i$	2.7081	2.8332	3.0910	3.9318	4.4773	4.5539

$$\sum i = 475 \quad \sum i^2 = 54{,}403 \quad \sum \ln P_i = 21.5953 \quad \sum (\ln P_i)^2 = 81.1584 \quad \sum i \ln P_i = 1{,}947.020$$

The economist believes that the price of a pint of lager in a given bar on day i can be modelled by:

$$\ln P_i = a + bi + e_i$$

where a and b are constants and the e_i's are uncorrelated $N(0, \sigma^2)$ random variables.

(i)   Estimate a, b and σ^2.   [5]

(ii)   Calculate the linear correlation coefficient r.   [1]

(iii)   Calculate a 99% confidence interval for b.   [2]

(iv)   Determine a 95% confidence interval for the average price of a pint of lager on day 365:

        (a)   in the country as a whole

        (b)   in a randomly selected bar.   [7]
                                                                                                    [Total 15]

11.11   (i)   Show that the maximum likelihood estimates (MLEs) of α and β in the simple linear
                regression model are identical to their least squares estimates.   [5]

        (ii)   Show that the MLE of σ^2 has a different denominator to the least squares estimate.   [4]
                                                                                                     [Total 9]

The solutions start on the next page so that you can
separate the questions and solutions.

Chapter 11 Solutions

11.1 **(i)(a)** *Calculate Sxx, Syy and Sxy*

$$s_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 = 7,420 - \frac{1}{6}\times 210^2 = 70$$

$$s_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 = 42.03 - \frac{1}{6}\times 15.3^2 = 3.015$$

$$s_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) = 549.8 - \frac{1}{6}\times 210 \times 15.3 = 14.3$$

(i)(b) *Fitted regression line*

Using the values from part (i)(a):

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{14.3}{70} = 0.2043$$

The mean values are:

$$\bar{x} = \frac{1}{n}\sum x = \frac{210}{6} = 35 \text{ and } \bar{y} = \frac{1}{n}\sum y = \frac{15.3}{6} = 2.55$$

So:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 2.55 - 0.2043 \times 35 = -4.60$$

Hence the fitted regression line is $\hat{y} = -4.60 + 0.2043x$.

(i)(c) *Error variance*

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{4}\left(3.015 - \frac{14.3^2}{70}\right) = 0.0234$$

(ii) *Estimated weight at 42 weeks*

Using the regression line from part (i)(b):

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = -4.60 + 0.2043 \times 42 = 3.98\text{kg}$$

(iii)(a) *Partition of the variability*

For the baby weights, $S_{xx} = 70$, $S_{yy} = 3.015$ and $S_{xy} = 14.3$. So:

$$SS_{TOT} = S_{yy} = 3.015$$

$$SS_{REG} = \frac{S_{xy}^2}{S_{xx}} = \frac{14.3^2}{70} = 2.921$$

$$SS_{RES} = SS_{TOT} - SS_{REG} = 3.015 - 2.921 = 0.094$$

(iii)(b) *Coefficient of determination*

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{2.921}{3.015} = 0.969 \text{ or } 96.9\%$$

So we see that, in this case, most of the variation is explained by the model. The model is an excellent fit to the data.

(iv) *Test for* β

If H_0 is true, then the test statistic $\dfrac{\hat{\beta} - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$ has a t_4 distribution.

The observed value of this statistic is $\dfrac{0.2043 - 0}{\sqrt{0.0234 / 70}} = 11.2$, which is much greater than 8.610, the

upper 0.05% point of the t_4 distribution.

So, we reject H_0 at the 0.05% level and conclude that there is extremely strong evidence that $\beta > 0$ *ie* that the baby's weight is increasing over time.

(v) *ANOVA*

The ANOVA table is:

Source of variation	d.f.	SS	MSS
Regression	1	2.921	2.921
Residual	4	0.0937	0.0234
Total	5	3.015	

Under $H_0 : \beta = 0$ we have $F = \dfrac{2.921}{0.0234} = 124.7$ on (1, 4) degrees of freedom.

The *p*-value of $F = 124.7$ is much less than even 0.01, so H_0 is rejected at the 1% level.

Therefore it is reasonable to assume that $\beta \neq 0$, *ie* there is linear relationship.

(vi)(a) *Estimate and variance of mean response*

Using the least squares regression line of $\hat{y} = -4.60 + 0.2043x$ when $x_0 = 33$ we have:

$$\hat{\mu}_0 = -4.60 + 0.2043 \times 33 = 2.141$$

ie the mean weight of a baby at 33 weeks is expected to be 2.141kg.

The variance of this estimator is calculated as:

$$\text{var}(\hat{\mu}_0) = \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\} \hat{\sigma}^2 = \left\{ \frac{1}{6} + \frac{(33-35)^2}{70} \right\} \times 0.0234 = 0.00524$$

(vi)(b) Confidence interval

The 90% confidence interval will be:

$$\hat{\mu}_0 \pm t_{0.05,4} \times s.e.(\hat{\mu}_0) = 2.141 \pm 2.132 \times \sqrt{0.00524} = (1.99, 2.30)$$

(vii)(a) Estimate and standard error of individual response

The *individual* predicted response is also $\hat{y}_0 = 2.141$ kg.

The variance of this estimator is calculated as:

$$\text{var}(\hat{y}_0) = \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\} \hat{\sigma}^2 = \left\{ 1 + \frac{1}{6} + \frac{(33-35)^2}{70} \right\} \times 0.0234 = 0.0287$$

(vii)(b) Confidence interval

The 90% confidence interval will be:

$$\hat{y}_0 \pm t_{0.05,4} \times s.e.(\hat{y}_0) = 2.141 \pm 2.132 \times \sqrt{0.0287} = (1.78, 2.50)$$

(viii)(a) Residuals

The completed table is:

Gestation period (weeks)	30	32	34	36	38	40
Residual	0.07	−0.24	0.15	0.05	0.04	−0.07

(viii)(b) *Dotplot and comment*

The dotplot is as follows:

All values are between $\pm 3\hat{\sigma} = \pm 3\sqrt{0.0234} = \pm 0.46$ so there appear to be no outliers.

There may be possible skewness but it's difficult to tell with such a small dataset.

(viii)(c) *Plot residuals against explanatory variable and comment*

The plot (on the next page) appears to be patternless which implies a good fit.

(viii)(d) *Interpret Q-Q plot*

One of the values is way off the diagonal line which indicates that the data set may be non-normal and hence the full normal linear regression model may not be appropriate.

11.2   (i)(a)   *Calculate slope parameter estimate*

Using the formula given on page 24 of the *Tables*:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{8.1}{12.2} = 0.66393$$

(i)(b)   *Test whether slope parameter is significantly different from zero*

We are testing:

$$H_0 : \beta = 0 \qquad vs \qquad H_1 : \beta \neq 0$$

Under H_0, $\dfrac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$ has a t_{n-2} distribution. Now:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \frac{1}{17}\left(10.6 - \frac{8.1^2}{12.2} \right) = 0.30718$$

So the observed value of the test statistic is:

$$\frac{0.66393 - 0}{\sqrt{0.30718 / 12.2}} = 4.184$$

Since this is much greater than 2.898, the upper 0.5% point of the t_{17} distribution, we have sufficient evidence to reject H_0 at the 1% level. Therefore it is reasonable to conclude that $\beta \neq 0$.

(ii)(a)   *Calculate the correlation coefficient*

Using the formula on page 25 of the *Tables*:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{8.1}{\sqrt{12.2 \times 10.6}} = 0.71228$$

(ii)(b)   *Test whether correlation coefficient is significantly different from zero*

We are testing:

$$H_0 : \rho = 0 \qquad vs \qquad H_1 : \rho \neq 0$$

Under H_0, $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ has a t_{n-2} distribution.

So the observed value of the test statistic is:

$$\frac{0.71228\sqrt{17}}{\sqrt{1-0.71228^2}} = 4.184$$

Since this is much greater than 2.898, the upper 0.5% point of the t_{17} distribution, we have sufficient evidence to reject H_0 at the 1% level. Therefore it is reasonable to conclude that $\rho \neq 0$.

(iii)   *Comment*

These tests are actually equivalent. Testing whether there is any correlation is equivalent to testing if the slope is not zero (*ie* it is sloping upwards and there is positive correlation or it is sloping downwards and there is negative correlation). So the tests give the same statistic and *p*-value.

11.3   The coefficient of determination is given by:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{6.4}{10.0} = 0.64$$

This gives the proportion of the total variance explained by the model. So 64% of the variance can be explained by the model, leaving 36% of the total variance unexplained.

11.4   (i)   *Transform quadratic to linear form*

Let $Y_i = y_i$ and $X_i = x_i^2$.

Then the model becomes $Y_i = a + bX_i + e_i$.

(ii)   *Transform exponential to linear form*

Taking logs gives:

$$\ln y_i = \ln a + bx_i$$

Let $Y_i = \ln y_i$ and $X_i = x_i$.

Then the model becomes $Y_i = \alpha + \beta X_i$ where $\alpha = \ln a$ and $\beta = b$.

11.5   (i)   *Fitted regression line*

Calculating the sums of squares:

$$S_{xx} = 60,016 - \frac{836^2}{12} = 1,774.67$$

[½]

$$S_{xy} = 1,122$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1,122}{1,774.67} = 0.63223$$

[1]

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 72.25 - 0.63223 \times 69.667 = 28.205$$

[1]

Hence, the fitted regression equation of *y* on *x* is $y = 28.205 + 0.63223x$.

[½]

(ii)   *Confidence interval for variance*

We have $S_{yy} = 63,603 - \frac{867^2}{12} = 962.25$, so:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{10}\left(962.25 - \frac{1,122^2}{1,774.\dot{6}}\right) = 25.289$$

[1]

Now $\dfrac{10\hat{\sigma}^2}{\sigma^2} \sim \chi_{10}^2$, which gives a confidence interval for σ^2 of:

$$\left(\frac{10 \times 25.289}{18.31}, \frac{10 \times 25.289}{3.94}\right) = (13.8, 64.2)$$

[2]

(iii)   *Test whether data are positively correlated*

We are testing $H_0 : \beta = 0$ *vs* $H_1 : \beta > 0$.

Now $\dfrac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{10}$. Our observed value here is:

$$\frac{0.63223 - 0}{\sqrt{25.289 / 1774.67}} = 5.296 \qquad\qquad\qquad [2]$$

This is a highly significant result, which exceeds the 0.5% critical value of the t_{10} distribution of 3.169. So we have sufficient evidence at the 0.5% level to reject H_0 and we conclude that $\beta > 0$ (*ie* the data are positively correlated). [1]

(iv)   *Confidence interval for the mean finals paper score*

The variance of the distribution of the mean finals score corresponding to an entrance score of 53 is:

$$\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{12} + \frac{(53 - 69.667)^2}{1{,}774.67}\right] \times 25.289 = 6.0657 \qquad\qquad [1]$$

The predicted value is $28.205 + 0.63223 \times 53 = 61.713$. [½]

We have a t_{10} distribution, so the 95% confidence interval is:

$$61.713 \pm 2.228 \times \sqrt{6.0657} = (56.2, 67.2) \qquad\qquad\qquad [1½]$$

(v)   *Calculate the proportion of variation explained by the model and comment*

The proportion of variability explained by the model is given by:

$$R^2 = \frac{S_{xy}^2}{S_{xx} s_{yy}} = \frac{1{,}122^2}{1{,}774.67 \times 962.25} = 73.7\% \qquad\qquad\qquad [1]$$

73.7% of the variation is explained by the model, which indicates that the fit is fairly good. It still might be worthwhile to examine the residuals to double check that a linear model is appropriate.

[1]

11.6   (i)   *Regression line*

We are given:

$$s_{xx} = 60 \qquad s_{yy} = 925{,}262 \qquad s_{xy} = 7{,}087$$

So:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{7{,}087}{60} = 118.117 \qquad\qquad\qquad [1]$$

Since $\bar{x} = \frac{36}{9} = 4$ and $\bar{y} = \frac{3,960}{9} = 440$, we get:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 440 - 118.117 \times 4 = -32.47$$ [1]

So the regression line is:

$$\hat{y} = -32.47 + 118.117x$$

(ii)(a)   *Confidence interval for slope parameter*

The pivotal quantity is given by:

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / s_{xx}}} \sim t_{n-2}$$

A 99% confidence interval is given by:

$$\hat{\beta} \pm t_{n-2;0.005} \sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}$$

From our data:

$$\hat{\sigma}^2 = \frac{1}{7}\left(925,262 - \frac{7,087^2}{60}\right) = 12,595.6$$ [1]

So the 99% confidence interval is given by:

$$118.117 \pm 3.499 \sqrt{\frac{12,595.6}{60}} = 118.117 \pm 50.696 = (67.4, 169)$$ [2]

(ii)(b)   *Confidence interval for variance*

The pivotal quantity is given by:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$ [1]

A 99% confidence interval is given by:

$$0.99 = P\left(\chi^2_{n-2;0.995} < \frac{(n-2)\hat{\sigma}^2}{\sigma^2} < \chi^2_{n-2;0.005}\right)$$

which gives a confidence interval of:

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;0.005}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;0.995}}\right)$$

Substituting in, the confidence interval (to 3 SF) is:

$$\left(\frac{7\times12{,}595.6}{20.28}, \frac{7\times12{,}595.6}{0.9893}\right) = (4350, 89100) \hspace{2cm} [1]$$

(iii)(a)   *Partition*

The total sum of squares, $SS_{TOT} = \sum(y_i - \overline{y})^2$ is given by s_{yy} which is 925,262.   [1]

The partition given at the bottom of page 25 in the *Tables* is:

$$\sum(y_i - \overline{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \overline{y})^2$$

ie $SS_{TOT} = SS_{RES} + SS_{REG}$

Now, modifying the $\hat{\sigma}^2$ formula on page 24 of the *Tables*, we have:

$$SS_{RES} = \sum(y_i - \hat{y}_i)^2 = s_{yy} - \frac{s_{xy}^2}{s_{xx}} = 925{,}262 - \frac{7{,}087^2}{60} = 88{,}169 \hspace{1cm} [1]$$

Alternatively, using $\hat{\sigma}^2$ from part (ii), we get $SS_{RES} = (n-2)\hat{\sigma}^2 = 7\times12{,}595.6$.

Hence:

$$SS_{REG} = 925{,}262 - 88{,}169 = 837{,}093 \hspace{3cm} [1]$$

Alternatively, this could be calculated as $SS_{REG} = \dfrac{s_{xy}^2}{s_{xx}} = \dfrac{7{,}087^2}{60} = 837{,}093$.

(iii)(b)   *Proportion of variability explained by the model*

This is the coefficient of determination, R^2, which is given by:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{837{,}093}{925{,}262} = 90.5\% \hspace{3cm} [1]$$

This tells us that 90.5% of the variation in the prices is explained by the model. Since this leaves only 9.5% from other non-model sources, it would appear that the model is a very good fit to the data.   [1]

(iv)(a)   *Residuals*

The residuals, e_i, be calculated from the actual prices, y_i, and the predicted prices, $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i$$

Using our regression line $\hat{y}_i = -32.47 + 118.117x_i$ from part (i), we get:

$$x = 1 \;\Rightarrow\; \hat{y} = -32.47 + 118.117 \times 1 \simeq 86 \;\Rightarrow\; \hat{e} = 131 - 86 \simeq 45 \qquad [1]$$

$$x = 4 \;\Rightarrow\; \hat{y} = -32.47 + 118.117 \times 4 \simeq 440 \;\Rightarrow\; \hat{e} = 330 - 440 \simeq -110 \qquad [1]$$

$$x = 8 \;\Rightarrow\; \hat{y} = -32.47 + 118.117 \times 8 \simeq 912 \;\Rightarrow\; \hat{e} = 1,095 - 912 \simeq 183 \qquad [1]$$

(iv)(b) *Dotplot of residuals*

The dotplot is:

[1]

Since $e_i \sim N(0, \sigma^2)$ we would expect the dotplot to be normally distributed about zero. This does not appear to be the case, but it is difficult to tell with such a small data set. [1]

(iv)(c) *Plot of residuals against time*

The graph is:

[1]

Clearly this is not patternless. The residuals are *not* independent of the time – this means that the linear model is definitely missing something and is not appropriate to these data. [1]

A plot of the original data (with the regression line) shows what's happening:

The price increases in an exponential (rather than linear) way. We should have used the log of the price against time instead.

11.7 (i) **Obtain the fitted regression line**

The regression line for p on c is given by:

$$p = \hat{\alpha} + \hat{\beta}c$$

where $\hat{\beta} = \dfrac{S_{cp}}{S_{cc}}$ and $\hat{\alpha} = \bar{p} - \hat{\beta}\bar{c}$.

$$S_{cc} = \sum c^2 - \frac{\left(\sum c\right)^2}{n} = 6{,}884 - \frac{238^2}{9} = 590.2222$$

[1]

$$S_{cp} = \sum cp - \frac{\left(\sum c\right)\left(\sum p\right)}{n} = 983 - \frac{238 \times 33.4}{9} = 99.75556$$

So:

$$\hat{\beta} = \frac{99.75556}{590.2222} = 0.16901$$

[½]

$$\hat{\alpha} = \frac{33.4}{9} - 0.16901 \times \frac{238}{9} = -0.75836$$

[½]

Hence, the fitted regression line is:

$$p = 0.16901c - 0.75836$$

[1]

(ii) *Estimate the GCSE score and its standard error*

The estimate of the average GCSE point score is obtained from the regression line:

$$\hat{P} = -0.75836 + 0.16901 \times 15 = 1.78 \qquad [1]$$

The standard error of this individual response is given by:

$$\sqrt{\left\{ 1 + \frac{1}{n} + \frac{(c_0 - \overline{c})^2}{S_{cc}} \right\} \hat{\sigma}^2} \qquad [1]$$

where $\hat{\sigma}^2 = \dfrac{1}{n-2} \left(S_{pp} - \dfrac{S_{cp}^2}{S_{cc}} \right) = \dfrac{1}{7} \left(25.66889 - \dfrac{99.75556^2}{590.2222} \right) = 1.25841$. $\qquad [1]$

Hence, the standard error is given by:

$$\sqrt{\left\{ 1 + \frac{1}{9} + \frac{(15 - \frac{238}{9})^2}{590.2222} \right\} 1.25841}$$

$$= \sqrt{1.33302 \times 1.25841}$$

$$= \sqrt{1.67748}$$

$$= 1.29518 \qquad [1]$$

11.8 (i) *Least squares estimate of slope parameter*

The least squares estimate minimises $\sum e_i^2$. Now:

$$q = \sum e_i^2 = \sum (Y_i - \beta x_i)^2 \qquad [\frac{1}{2}]$$

Differentiating this gives:

$$\frac{dq}{d\beta} = -2 \sum x_i (Y_i - \beta x_i) \qquad [1]$$

Setting this equal to zero gives:

$$\sum x_i (Y_i - \hat{\beta} x_i) = 0$$

$$\sum x_i Y_i - \hat{\beta} \sum x_i^2 = 0$$

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} \qquad [1]$$

The second derivative is $2 \sum x_i^2 > 0$, so we do have a minimum. $\qquad [\frac{1}{2}]$

(ii)   *Bias and mean square error*

The expectation of $\hat{\beta}$ is:

$$E(\hat{\beta}) = E\left(\frac{\sum x_i Y_i}{\sum x_i^2}\right) = \frac{\sum x_i E(Y_i)}{\sum x_i^2}$$ [½]

Now $E(Y_i) = E(\beta x_i + e_i) = \beta x_i + 0 = \beta x_i$. So: [½]

$$E(\hat{\beta}) = \frac{\beta \sum x_i^2}{\sum x_i^2} = \beta$$ [½]

Hence:

$$bias(\hat{\beta}) = E(\hat{\beta}) - \beta = 0$$ [½]

The variance of $\hat{\beta}$ is:

$$var(\hat{\beta}) = var\left(\frac{\sum x_i Y_i}{\sum x_i^2}\right) = \frac{\sum x_i^2 \, var(Y_i)}{\left(\sum x_i^2\right)^2}$$ [½]

Now $var(Y_i) = var(\beta x_i + e_i) = var(e_i) = \sigma^2$. So: [½]

$$var(\hat{\beta}) = \frac{\sigma^2 \sum x_i^2}{\left(\sum x_i^2\right)^2} = \frac{\sigma^2}{\sum x_i^2}$$

Hence:

$$MSE(\hat{\beta}) = var(\hat{\beta}) + bias^2(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}$$ [1]

11.9   (i)(a)   *Expressions for parameters*

Taking logs of the original expression gives:

$$\ln \mu_x = \ln B + x \ln c$$ [1]

This expression is now linear in x. Comparing the expression with $Y = \alpha + \beta x$ gives:

$$Y = \ln \mu_x \qquad \alpha = \ln B \qquad \beta = \ln c$$ [1]

(i)(b) *Scattergraph and comment*

The graph of $\ln \mu_x$ against x is shown below:

[1]

The graph appears to show an approximately linear relationship and this supports the transformation in part (i)(a). However, it does appear to have a slight curve and this would warrant closer inspection of the model to see if it is appropriate for the data. [1]

(ii) *Least squares estimates*

Obtaining the estimates of α and β using the formulae given on page 24 of the *Tables* with $y = \ln \mu_x$:

$$s_{xx} = \sum x^2 - n\bar{x}^2 = 11{,}120 - 8\left(\frac{296}{8}\right)^2 = 168$$

$$s_{xy} = \sum xy - n\bar{x}\bar{y} = -2{,}104.5 - 8\left(\frac{296}{8}\right)\left(\frac{-57.129}{8}\right) = 9.273$$

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{9.273}{168} = 0.055196 \qquad [1]$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{-57.129}{8} - 0.055196 \times \frac{296}{8} = -9.1834 \qquad [1]$$

Therefore, we obtain:

$$B = e^\alpha = e^{-9.1834} = 0.000103$$
$$c = e^\beta = e^{0.055196} = 1.06$$

[1]

(iii)(a) *Coefficient of determination and comment*

The coefficient of determination is given by:

$$R^2 = r^2 = \frac{s_{xy}^2}{s_{xx}s_{yy}} = \frac{9.273^2}{168 \times 0.53467} = 95.7\% \tag{[1]}$$

where $s_{yy} = \sum y^2 - n\bar{y}^2 = 408.50 - 8\left(\frac{-57.129}{8}\right)^2 = 0.53467$.

This tells us that 95.7% of the variation in the data can be explained by the model and so indicates an extremely good overall fit of the model. [1]

(iii)(b) *Calculate residuals and comment*

The completed table of residuals using $\hat{e}_i = y_i - \hat{y}_i$ is:

Age, x	30	32	34	36	38	40	42	44
Residual, $\hat{e}_i$	0.08	0.02	-0.03	-0.06	-0.06	-0.03	0.02	0.09

Age 32 yrs: $(-7.40) - (-9.1834 + 0.055196 \times 32) = 0.02$ [1]

Age 36 yrs: $(-7.26) - (-9.1834 + 0.055196 \times 36) = -0.06$

[1]

Age 40 yrs: $(-7.01) - (-9.1834 + 0.055196 \times 40) = -0.03$

The residuals should be patternless when plotted against x, however it is clear to see that some pattern exists – this indicates that the linear model is not a good fit and that there is some other variable at work here. [1]

(iv) *Confidence interval for mean predicted value*

Using the formula given on page 25 of the *Tables*, the variance of the mean predicted response is:

$$\left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right\}\hat{\sigma}^2 = \left\{\frac{1}{8} + \frac{(35-37)^2}{168}\right\} \times 0.0038056 = 0.0005663 \tag{[1]}$$

where $\hat{\sigma}^2 = \frac{1}{6}\left(0.53467 - \frac{9.273^2}{168}\right) = 0.0038056$. [1]

The estimate is $Y = \ln \mu_{35} = -9.1834 + 0.055196 \times 35 = -7.251$, so using the t_6 distribution the 95% confidence interval for $Y = \ln \mu_{35}$ is given by:

$$-7.251 \pm 2.447\sqrt{0.0005663} = (-7.309, -7.193) \tag{[1]}$$

Hence the 95% confidence interval for μ_{35} is given by:

$$(0.000669, 0.000752)$$ [1]

11.10   (i)   ***Estimate parameters***

Now using x for i and y for $\ln P_i$, we get:

$$s_{xx} = \sum x^2 - n\bar{x}^2 = 16,799$$

$$s_{xy} = \sum xy - n\bar{x}\bar{y} = 237.39$$

$$s_{yy} = \sum y^2 - n\bar{y}^2 = 3.4322$$ [2]

So the estimates for a, b and σ^2 are:

$$\hat{b} = \frac{s_{xy}}{s_{xx}} = \frac{237.39}{16,799} = 0.01413$$ [1]

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{21.5953}{6} - 0.01413\left(\frac{475}{6}\right) = 2.4805$$ [1]

$$\hat{\sigma}^2 = \frac{1}{n-2}(s_{yy} - \frac{s_{xy}^2}{s_{xx}}) = \frac{1}{4}(3.4322 - \frac{237.39^2}{16,799}) = 0.01940$$ [1]

(ii)   ***Correlation coefficient***

The correlation coefficient is:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{237.39}{\sqrt{16,799 \times 3.4322}} = 0.989$$ [1]

(iii)   ***Confidence interval for slope parameter***

Using the result given on page 24 of the *Tables*, we have:

$$\hat{b} \pm t_{4;0.005}\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}} = 0.01413 \pm 4.604\sqrt{\frac{0.01940}{16,799}}$$ [1]

This gives a confidence interval for b of $(0.00918, 0.0191)$. [1]

(iv)(a) ***Confidence interval for mean response***

If y_{365} denotes the log of the average price of a pint of lager in the country as a whole on day 365, the predicted value for y_{365} is:

$$\hat{y}_{365} = 2.4805 + 0.01413 \times 365 = 7.638$$ [1]

The distribution of $\dfrac{y_{365} - \hat{y}_{365}}{s_{365}}$ is t_4, where:

$$s_{365}^2 = \left[\frac{1}{n} + \frac{(365 - \bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{6} + \frac{[365 - (475/6)]^2}{16,799}\right] \times 0.01940 = 0.09758 \qquad [1]$$

So a symmetrical 95% confidence interval for y_{365} is:

$$y_{365} = 7.638 \pm 2.776\sqrt{0.09758} = 7.638 \pm 0.867 = (6.77, 8.51) \qquad [1]$$

and the corresponding confidence interval for P_{365} is:

$$(e^{6.771}, e^{8.505}) = (870, 4940) \qquad [1]$$

(iv)(b) *Confidence interval for individual response*

If y_{365}^* denotes the log of the observed price of a pint of lager in a randomly selected bar on day 365, then $\dfrac{y_{365}^* - \hat{y}_{365}}{s_{365}^*}$ has a t_4 distribution, where:

$$s_{365}^{*2} = \left[1 + \frac{1}{n} + \frac{(365 - \bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = s_{365}^2 + \hat{\sigma}^2 = 0.09758 + 0.01940 = 0.11698 \qquad [1]$$

This gives a confidence interval of:

$$y_{365}^* = 7.638 \pm 2.776\sqrt{0.11698} = 7.638 \pm 0.949 = (6.69, 8.59) \qquad [1]$$

So the confidence interval for P_{365}^* is:

$$(e^{6.689}, e^{8.587}) = (800, 5360) \qquad [1]$$

11.11 (i) *MLEs of α and β*

Each Y_i has a $N(\alpha + \beta x_i, \sigma^2)$ distribution, so the joint likelihood function is:

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right)^2\right] = \frac{1}{\sigma^n(\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2}\right] \qquad [1]$$

Taking logs we get:

$$\log L = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 + \text{constant} \qquad [\frac{1}{2}]$$

Differentiating with respect to β and then with respect to α:

$$\frac{\partial \log L}{\partial \beta} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i) \times (-x_i)$$ [1]

$$\frac{\partial \log L}{\partial \alpha} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i) \times (-1)$$ [½]

By setting $\dfrac{\partial \log L}{\partial \alpha}$ equal to 0 we get $\displaystyle\sum_{i=1}^{n} y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i = 0$. [½]

By setting $\dfrac{\partial \log L}{\partial \beta}$ equal to 0 we get $\displaystyle\sum_{i=1}^{n} y_i x_i - \hat{\alpha}\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} x_i^2 = 0$. [½]

These are the same normal equations that we got before, so the MLEs are as before, *ie*:

$$\hat{\beta} = \frac{n\left(\displaystyle\sum_{i=1}^{n} x_i y_i\right) - \left(\displaystyle\sum_{i=1}^{n} x_i\right)\left(\displaystyle\sum_{i=1}^{n} y_i\right)}{n\left(\displaystyle\sum_{i=1}^{n} x_i^2\right) - \left(\displaystyle\sum_{i=1}^{n} x_i\right)^2} = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\alpha} = \frac{\displaystyle\sum_{i=1}^{n} y_i - \hat{\beta}.\sum_{i=1}^{n} x_i}{n} = \overline{y} - \hat{\beta}\overline{x}$$ [1]

(ii)   *Show MLE of σ² has a different denominator to the least squares estimate*

Now differentiating the log likelihood with respect to σ:

$$\frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} - \frac{1}{2}\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \times -2\sigma^{-3}$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$ [1]

By setting $\dfrac{\partial \log L}{\partial \sigma}$ equal to 0 and substituting $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$, we obtain:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y} + \hat{\beta}\overline{x} - \hat{\beta}x_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \overline{y}) - \hat{\beta}(x_i - \overline{x})\right]^2$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\hat{\beta}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + \hat{\beta}^2\sum_{i=1}^{n}(x_i - \overline{x})^2\right]$$

$$= \frac{1}{n}\left[S_{yy} - 2\hat{\beta}S_{xy} + \hat{\beta}^2 S_{xx}\right]$$

$$= \frac{1}{n}\left[S_{yy} - 2\frac{S_{xy}}{S_{xx}}S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx}\right]$$

$$= \frac{1}{n}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) \qquad\qquad [3]$$

which has a different denominator from before (and therefore is a *biased*) estimator.

11b

Multiple linear regression

Syllabus objectives

4.1   Linear regression

    4.1.5   State the multiple linear regression model (with several explanatory variables).

    4.1.6   Use appropriate software to fit a multiple linear regression model to a data set and interpret the output.

    4.1.7   Use measures of model fit to select an appropriate set of explanatory variables.

0   Introduction

In the last chapter we carried out linear regression with only one explanatory variable. In this chapter we extend the linear model to many explanatory variables.

Fitting a multiple linear regression to data is difficult to do on a piece of paper, but is a straightforward application of the computer functions used in the previous chapter. Hence, much of this chapter is descriptive in nature with the calculations being done in R.

1   The multiple linear regression model

1.1   Introduction

We will now extend our linear regression model. Previously we examined the relationship between Y, the response (or dependent) variable and one explanatory (or independent or regressor) variable X. We now look at k explanatory variables, $X_1, X_2, \ldots, X_k$.

There are many problems where one variable can quite accurately be predicted in terms of another. However, the use of additional relevant information should improve predictions. There are many different formulae used to express regression relationships between more than two variables. Most are of the form:

$$E\left[Y|X_1, X_2 \ldots X_k\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

As with the simple linear regression model discussed earlier Y is a random variable whose values are to be predicted in terms of given data values $x_1, x_2, \ldots, x_k$.

$\beta_1, \beta_2 \ldots \beta_k$ are known as the multiple regression coefficients. They are numerical constants which can be determined from observed data.

1.2   Fitting the model

As for the simple linear model, the multiple regression coefficients are usually estimated by the method of least squares.

The response variable Y_i is related to the values $x_{i1}, x_{i2}, \ldots x_{ik}$ by

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \qquad\qquad i = 1, \ldots, n$$

and so the least squares estimates of $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ are the values $\hat\alpha, \hat\beta_1, \hat\beta_2, \ldots, \hat\beta_k$ for which:

$$q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left[y_i - \left(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k\ x_{ik}\right)\right]^2 \text{ is minimised}$$

As for the simple linear model, to find the estimates the above is differentiated partially with respect to α and $\beta_1, \beta_2, \ldots, \beta_k$ in turn and the results are equated to zero.

Question

A senior actuary wants to analyse the salaries of the 50 actuarial students employed by her company, using a linear model based on number of exam passes and years of experience. Express this model and the available data in terms of the notation given here.

Solution

The basic model would be:

$$E[Y \mid x_1, x_2] = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Here x_1 represents the number of exam passes, x_2 represents the number of years' experience and Y would represent the corresponding salary.

α, β_1 and β_2 are constants where:

- α reflects the average salary for a new student (with no exam passes or experience)

- β_1 and β_2 reflect the changes in pay associated with an extra exam pass and an extra year's experience, respectively.

Since the data relates to 50 $(=n)$ students, we need to introduce an extra subscript i corresponding to the i th student. So the actual salary for the i th student will be:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

where e_i is the difference between the student's actual salary and the theoretical salary for someone with the same number of exam passes and experience.

Manually solving the equations becomes complicated even with $k = 2$. As a result such multiple linear regressions are usually carried out using a computer package.

So on a paper-based exam we can only really test the general principles (as in the question above) rather than the actual modelling. To do actual modelling we'll use R.

The R code to fit a multiple linear model to a multivariate data frame is:

```
model <- lm(Y ~ X1+X2+…+Xk)
```

Then the estimates of the coefficients and error standard deviation can be obtained from:

```
summary(model)
```

Let us return now to the market equity returns data that we saw in Chapter 10.

Consider a set of equity returns from four different markets across 12 time periods ($\underline{X}$).

Mkt 1	Mkt 2	Mkt 3	Mkt 4
0.83%	4.27%	−1.79%	−0.39%
−0.12%	−3.72%	0.90%	−0.26%
−5.49%	−5.21%	−4.62%	−5.67%
2.75%	−6.26%	3.38%	1.40%
−5.68%	−7.37%	−5.21%	−5.05%
−3.70%	−1.60%	−2.34%	−2.66%
5.75%	5.08%	6.03%	5.48%
1.03%	−1.38%	2.37%	1.47%
0.69%	−0.17%	0.38%	−0.10%
−4.03%	−3.26%	−3.04%	−2.59%
0.54%	2.22%	1.42%	1.37%
3.03%	9.47%	2.95%	2.99%

The scatterplot from Chapter 10 was as follows:

Model the bottom row Mkt_4 as the response variable (Y) with the other three markets as the explanatory variables (X_1, X_2, X_3).

The basic form of the model will be:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where:

$x_i =$ return from Market i, $i = 1,2,3$

$Y =$ return from Market 4

We can use R to estimate the parameters α, $\beta_1, \beta_2, \beta_3$ for this model. For further details, see the CS1B PBOR course. We can also estimate the error variance; when we do so we obtain the following numbers.

Modelling this (using R) **gives:**

$$\hat{y} = -0.001954 + 0.211472 x_1 + 0.125051 x_2 + 0.598636 x_3 \text{ with } \hat{\sigma}^2 = 0.004928$$

Given the strong positive correlation between the first and third market – we may have been able to use principle components analysis (from Chapter 10) to reduce the number of variables before fitting our multiple linear regression model.

1.3 R^2 in the multiple regression case

In the bivariate case (Chapter 11 Section 1.3) we noted that the proportion of the total variation of the responses 'explained' by a model, called the coefficient of determination, denoted R^2, was equal to the square of the correlation coefficient between the dependent variable Y and the single independent variable x.

In the case of multiple regression with a single dependent variable, Y, and several independent variables, $x_1, x_2, ..., x_k$, R^2 measures the proportion of the total variation in Y 'explained' by the combination of explanatory variables in the model. The value of R^2 lies between 0 and 1. It will generally increase (and cannot decrease) as the number of explanatory variables k increases. If $R^2 = 1$ the model perfectly predicts the values of Y: 100% of the variation in Y is "explained" by variation in $x_1, x_2, ..., x_k$.

Because R^2 cannot decrease as more explanatory variables are added to the model, if it is used alone to assess the adequacy of the model, there will always be a tendency to add more explanatory variables. However, these may increase the value of R^2 by a small amount, while adding to the complexity of the model. Increased complexity is generally considered to be undesirable.

We prefer to use the principle of parsimony when fitting models – which means we choose the simplest model that does the job. So we need to introduce a new measure that prevents us from adding new variables unnecessarily.

To take account of the undesirability of increased complexity, computer packages will often quote an 'adjusted R^2' statistic. This is a correction of the R^2 statistic which is based on the mean square errors (*ie* the residual mean sum of squares, MSS_{RES}) **and takes account of**

the number of predictors, **k**, and the number of data points the model is based on. If we have **k** predictors, and **n** observations:

$$\text{'Adjusted } R^2 \text{'} = 1 - \frac{MSS_{RES}}{MSS_{TOT}} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2)$$

So MSS_{RES}/MSS_{TOT} would give a measure of how much variability is explained by the residuals (or errors) and takes values between 0 and 1. Hence $1-MSS_{RES}/MSS_{TOT}$ would give a measure of how much variability is explained by the regression model. Therefore it is similar measure to the original coefficient of determination, R^2.

Recall that the mean sum of squares (MSS) is the sum of squares divided by the degrees of freedom. So $MSS_{RES} = SS_{RES}/(n-k-1)$ and $MSS_{TOT} = SS_{TOT}/(n-1)$.

The model which maximises the 'adjusted R^2' statistics can be regarded in some sense as the 'best' model. Note, however, that the 'adjusted R^2' cannot be interpreted as the proportion of the variation in Y which is 'explained' by variation in the $x_1, x_2, ..., x_k$.

R | **The R code to obtain the regression and residual sum of squares for a linear model assigned to the object model, is:**

```
anova(model)
```

The adjusted R^2 is given in the output of:

```
summary(model)
```

Question

Calculate the adjusted R^2 **for the equity returns from four different markets**, given that $R^2 = 0.9831$.

Solution

We have 12 periods of data $(n=12)$ and we are modelling market 4 from the other 3 markets $(k=3)$. Hence, we have **an adjusted R^2 of:**

$$1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2) = 1 - \left(\frac{12-1}{12-3-1}\right)(1-0.9831) = \textbf{0.9768}$$

---

2   The full normal model and inference

2.1   The full normal model

Again, to make inferences concerning the responses based on the fitted model, we need to specify the model further. We make the same assumptions as for the linear model:

In the full model, we now assume that the errors, e_i, are independent and identically distributed $N(0, \sigma^2)$ random variables. This will then allow us to obtain the distributions for β and the Y_i's. We can then use these to construct confidence intervals and carry out statistical inference.

The error variables e_i are: (a) independent, and (b) normally distributed.

Under this full model, the e_i's are independent, identically distributed random variables, each with a normal distribution with mean 0 and variance σ^2. It follows that the Y_i's are independent, normally distributed random variables, with:

$$E[Y_i] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \text{ and } \text{var}[Y_i] = \sigma^2$$

This mimics the bivariate linear regression model but with the mean dependent on k explanatory variables.

2.2   Testing hypotheses on individual covariates

In multiple regression the coefficients $\beta_1, \beta_2, ..., \beta_k$ describe the effect of each explanatory variable on the dependent variable *Y* after controlling for the effects of other explanatory variables in the model.

Each coefficient β_j measures the increase in the value of the response variable y for a corresponding increase in the value of x_j *independent* of the other covariates.

As in the bivariate case, hypotheses about the values of $\beta_1, \beta_2, ..., \beta_k$ can be tested, notably the hypothesis $\beta_i = 0$ which states that, after controlling for the effects of other variables, the variable x_i has 'no linear relationship' with *Y*.

Recall that in the bivariate case a hypothesis of $\beta = 0$ was equivalent to $\rho = 0$.

Generally speaking, it is not useful to include in a multiple regression model a covariate x_i for which we cannot reject the hypothesis that $\beta_i = 0$.

R

In R, the statistic and *p*-value for the tests of $H_0 : \beta_i = 0$ are given in the output of `summary(model).`

Question

For our equity returns from four different markets, we have the following output from R:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001954   0.001529  -1.279  0.23689
Mkt_1        0.211472   0.165990   1.274  0.23842
Mkt_2        0.125051   0.041877   2.986  0.01744 *
Mkt_3        0.598636   0.155270   3.855  0.00484 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By considering the *p*-values given in the final column comment on the significance of the parameters.

Solution

A *p*-value of less than 0.05 (helpfully indicated by asterisks) indicates a significant result.

We can see that β_2 and β_3 are significantly different from zero.

2.3   Analysis of variance (ANOVA)

In Section 1.3 we partitioned the variability into that arising from the regression model (SS_{REG}) and the left-over residual (or error) variability (SS_{RES} or SS_{ERR}). We then calculated a ratio of the variances to obtain the proportion of variability that was determined (or explained) by the model (called the coefficient of determination, R^2). This gave us a crude measure of fit.

We can use ANOVA to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ against the alternative $H_1 : \beta_j \neq 0$ for at least one j .

The ANOVA table is now:

Source of variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Sum of Squares*
Regression	k	SS_{REG}	SS_{REG}/k
Residual	$n-k-1$	SS_{RES}	$SS_{RES}/(n-k-1)$
Total	$n-1$	SS_{TOT}	

On statistical computer packages the regression sum of squares is often subdivided into the sum of squares from each explanatory variable.

Our statistic is now:

$$\frac{SS_{REG}/k}{SS_{RES}/(n-k-1)} = \frac{\text{regression mean square}}{\text{residual mean square}}$$

which is $F_{k,n-k-1}$ where H_0 is rejected for 'large' values of this ratio.

The test statistic is just the ratio of the values in the last column.

Note that unlike the adjusted R^2 we divide the regression mean variability by the residual rather than the total mean variability.

A large value of this ratio would mean that the majority of the variability is explained by the multiple linear regression model. Therefore we would reject the null hypothesis of no linear relationship. At least one of the predictors must be explaining the variability.

> **In R, the *F* statistic and its *p*-value for a regression model are given in the output of both** `anova(model)` **and** `summary(model)`**.**

Question

For our equity returns from four different markets, where we model Mkt_4 using all of the other markets we have the following output:

```
F-statistic: 155.6 on 3 and 8 DF,  p-value: 1.972e-07
```

Explain this result.

Solution

So we can reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, since the p-value is extremely small.

So at least one of the coefficients is non-zero, *ie* there is some relationship with at least one of the covariates.

2.4 Estimating a mean response and predicting an individual response

The whole point of the modelling exercise is so that we can estimate values of the response variable Y given the input variables $x_1, x_2, \ldots, x_k$.

Mean response

As with the linear model we can estimate the expected (mean) response, μ_0, for a multiple linear regression model given a vector of explanatory variables, $\underline{x}_0$.

$$\mu_0 = E[Y \mid \underline{x}_0] = \alpha + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k}$$

μ_0 is estimated by $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}$ which is an unbiased estimator.

Recall that our multivariate linear regression model stated that the Y_i's are independent, normally distributed random variables, with $E[Y_i] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$. We have simply used this expected value to obtain an estimated mean response corresponding to the vector $\mathbf{x}_0$. We are using vector notation here:

$$\mathbf{x}_0 = \left(x_{01}, x_{02}, \ldots x_{0k} \right)$$

Individual response

Similarly, we could predict an individual response y_0 (for $\underline{x} = \underline{x}_0$) using the same estimate $\hat{y}_0 = \hat{\alpha} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}$ but with an extra σ^2 in the expression for the variance of the estimator compared to the mean response.

Recall that for an individual response value we have $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i$. Each individual response value is associated with an error term from the regression line. Since $e_i \sim N(0, \sigma^2)$ an individual point is on the regression line on average – hence we have the same estimate $\hat{\alpha} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}$ as for the mean response. However, there is an additional σ^2 for the variance.

Question

For our equity returns from four different markets, we had the following model:

$$\hat{y} = -0.001954 + 0.211472 x_1 + 0.125051 x_2 + 0.598636 x_3$$

where Market 4 is the response variable (Y) and the other three markets are the explanatory variables (X_1, X_2, X_3).

Use this model to construct an estimate for the return on Market 4 when the returns on Market 1, Market 2 and Market 3 are 8%, 4% and -1%, respectively.

Solution

Substituting these values into our equation gives:

$$\hat{y} = -0.001954 + 0.211472 \times 0.08 + 0.125051 \times 0.04 + 0.598636 \times -0.01 = 0.0140 \text{, ie } 1.40\%$$

We will use R to calculate confidence intervals for the mean and individual responses as they are beyond the scope of the written exam.

R | For our equity returns from four different markets, we can obtain a 95% confidence intervals for the mean and individual response when the returns on Market 1, Market 2 and Market 3 are 8%, 4% and -1%, respectively, using R as follows:

```
newdata <-data.frame(Mkt_1=0.08, Mkt_2=0.04 , Mkt_3=-0.01)

predict(model,newdata,interval="confidence",level=0.95)

predict(model,newdata,interval="predict",level=0.95)
```

These give (–1.95%,4.74%) and (–2.14%, 4.93%), respectively to 3 SF.

2.5 Checking the model

As we did for the linear regression model, we can also calculate the residual from the fit at x_i which is the estimated error, $\hat{e}_i = y_i - \hat{y}_i$ and then examine them to see if they are normally distributed and also independent of the explanatory variables.

Question

For our equity returns from four different markets, we had the following model:

$$\hat{y} = -0.001954 + 0.211472x_1 + 0.125051x_2 + 0.598636x_3$$

where Market 4 is the response variable (Y) and the other three markets are the explanatory variables (X_1, X_2, X_3).

The equity returns from four different markets for the first time period were:

Mkt 1	Mkt 2	Mkt 3	Mkt 4
0.83%	4.27%	−1.79%	−0.39%

Calculate the residual for this first time period.

Solution

Substituting the values of markets 1 to 3 during the first time period into our equation gives:

$$\hat{y} = -0.001954 + 0.211472 \times 0.0083 + 0.125051 \times 0.0427 + 0.598636 \times -0.0179 = -0.0056$$

So the residual is:

$$-0.0039 - (-0.0056) = 0.0017$$

A Q-Q plot can be used to test whether the residuals are normally distributed. A plot of the residuals against the fitted values can be used to determine if the variance is constant and whether they are independent of the explanatory variables.

Question

For our equity returns from four different markets, the Q-Q plot of the residuals and the plot of the residuals against the fitted values are given here.

Normal Q-Q Plot

Fitted values

Comment on these results.

Solution

The Q-Q plot suggests that the one of the data points does not fit the assumption of normality. The plot of residuals against fitted values possibly suggests greater variance for the lower fitted values. This may indicate dependency and thus imply the model has some deficiencies.

2.6   The process of selecting explanatory variables

Selecting the optimal set of explanatory variables is not always easy. Two general approaches can be outlined:

(1) Forward selection. Start with the single covariate that is most closely related to the dependent variable Y **. Add that to the model. Then search among the remaining covariates to find the one which improves the 'adjusted** R^2**' the most when added to the model. Continue adding covariates until adding any more causes the 'adjusted** R^2**' to fall.**

In Chapter 10 we saw that the Pearson correlation coefficient matrix for the returns on the four equity markets were:

```
             Mkt_1       Mkt_2       Mkt_3       Mkt_4

   Mkt_1 1.0000000  0.6508163  0.9538019  0.9727972

   Mkt_2 0.6508163  1.0000000  0.5321185  0.6893932

   Mkt_3 0.9538019  0.5321185  1.0000000  0.9681911

   Mkt_4 0.9727972  0.6893932  0.9681911  1.0000000
```

Let's use a forward selection approach to creating a multiple linear regression model. Mkt_4 is the response variable (Y) with the other three markets are the explanatory variables X_1, X_2, X_3.

<u>First covariate</u>

We would start with Mkt_1 as that has the highest correlation with Mkt_4.

Using R we get the model $\hat{y} = -0.000140 + 0.873309 x_1$ which has an adjusted R^2 of 0.941.

<u>Second covariate</u>

Using R, adding Mkt_2 gives the model $\hat{y} = -0.000063 + 0.816265 x_1 + 0.062317 x_2$ which has an adjusted R^2 of 0.9411.

Whereas adding Mkt_3 gives the model $\hat{y} = -0.001692 - 0.490675 x_1 - 0.418474 x_3$ which has an adjusted R^2 of 0.9564.

Hence, since adding the covariate Mkt_3 improves the adjusted R^2 the most – we would go for this model.

Third covariate

Now we have a model with both Mkt_1 and Mkt_3 as covariates, we will see if adding Mkt_2 produces an improvement.

Using R, adding Mkt_2 gives the model $\hat{y} = -0.001954 + 0.211472x_1 + 0.125051x_2 + 0.598636x_3$ which has an adjusted R^2 of 0.9768.

However, whilst this maximises the adjusted R^2, one of the coefficients of this model is not significantly different from zero.

(2) Backward selection. Start by adding all available covariates. Then remove covariates one by one for which the hypothesis that $\beta_i = 0$ cannot be rejected until the 'adjusted R^2' value reaches a maximum, and all the remaining covariates have a statistically significant impact on Y.

For our equity returns from four different markets, using R, the model with all covariates added is $\hat{y} = -0.001954 + 0.211472x_1 + 0.125051x_2 + 0.598636x_3$ which has an adjusted R^2 of 0.9768.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001954   0.001529  -1.279  0.23689
Mkt_1        0.211472   0.165990   1.274  0.23842
Mkt_2        0.125051   0.041877   2.986  0.01744 *
Mkt_3        0.598636   0.155270   3.855  0.00484 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the coefficient for Mkt_1 is not significantly different from zero. Removing this coefficient using R gives the model $\hat{y} = -0.002598 + 0.155102x_2 + 0.785578x_3$ which has an adjusted R^2 of 0.97652.

So removing this covariate causes the adjusted R^2 to *decrease* not increase. So we'd probably keep it.

The problem is the high correlation between Mkt_1 and Mkt_3 meaning that there is some overlap between them in descriptive ability. Ideally we would use Principal Components Analysis from Chapter 10 to reduce the number of covariates by removing this overlap.

2.7 Extending the scope of the multiple linear model

Interaction between terms

So far we have only considered each variable X_j as a *main effect*, that is where we incorporate each new variable via an additive term, $\beta_j X_j$. This means that a unit increase in X_j will increase the average response by β_j regardless of the other variables.

So in our multiple linear regression model with Mkt_4 as the response variable (Y), the three other markets (X_1, X_2, X_3) were included as main effects only:

$$\hat{y} = -0.001954 + 0.211472x_1 + 0.125051x_2 + 0.598636x_3$$

So an increase in, say Mkt_1, by 1% would lead to an increase in Mkt_4 of 0.211472×0.01.

However, it is often the case that the effect of one predictor variable, say X_1, on the response variable, Y, depends on the value of another predictor variable, say X_2. This is called *interaction*.

That is, we observe an additional effect when both predictors are present.

We model this by including an interaction term, denoted $X_1.X_2$ on paper which corresponds to the term $\gamma_{12}X_1X_2$, in the regression function.

The regression function for the two variables, X_1 and X_2 as main effects and their interaction is:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_{12} X_1 X_2$$

Note that when an interaction term is used in a model, both main effects must also be included.

R uses a colon to denote interaction, hence the code to fit the multiple linear model above is:

```
model <- lm(Y ~ X1+X2+X1:X2)
```

The shorthand notation for the main effects *and* the interaction between them is denoted $X_1 * X_2$ which corresponds to the whole model above.

So the equivalent way of specifying the above model in R is:

```
model <- lm(Y ~ X1*X2)
```

Interaction effects are described in greater detail in the Generalised Linear Models chapter.

Polynomial regression

Finally, the term 'linear' in linear regression means that the regression function is linear in the coefficients α and β_1, β_2, ..., β_k rather than linear in terms of the X_j's. For example, we could fit a quadratic model $Y = \alpha + \beta_1 X + \beta_2 X^2$.

Although this is a bivariate model (having only two measured variables X and Y) we model it as a multiple linear regression model treating X and X^2 as different variables.

We use the I( ) function in R to treat a term as a different variable. So the R code to fit this quadratic model is:

```
model <- lm(Y ~ X+I(X^2))
```

Chapter 11b Summary

The linear multiple regression model is given by:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \quad \text{where } e_i \sim N(0, \sigma^2)$$

The parameters $\alpha, \beta_1, \ldots, \beta_k$ and σ^2 can be estimated using a computer package.

Confidence intervals and tests can be carried out for $\beta_1, \ldots, \beta_k$.

ANOVA can be used to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ against the alternative $H_1 : \beta_j \neq 0$ for at least one j:

$$F_{k, n-k-1} = \frac{MSS_{REG}}{MSS_{RES}} = \frac{SS_{REG}/k}{SS_{RES}/(n-k-1)}$$

Confidence intervals can also be obtained for the predicted individual (or mean) response y.

We can partition the total variance, SS_{TOT}, into that which is explained by the model, SS_{REG} and that which is not, SS_{RES}. The coefficient of determination, R^2, gives the percentage of this variance which *is* explained by the combination of explanatory variables in the model.

However, since R^2 cannot decrease as more explanatory variables are added to the model, if it is used alone to assess the adequacy of the model, there will always be a tendency to add more explanatory variables which is undesirable. Hence, computer packages quote an 'adjusted R^2' statistic which is based on the mean square errors and takes account of the number of predictors, k, and the number of data points the model is based on.

$$\text{'Adjusted' } R^2 = 1 - \frac{MSS_{RES}}{MSS_{TOT}} = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)$$

If the model is a good fit then we would expect the residuals, $\hat{e}_i = y_i - \hat{y}_i$, to be normally distributed about zero, have constant variance and no relationship with the *x* values. These can be examined using diagrams.

We can use one of the following approaches to select an optimal set of explanatory variables:

(1) Forward selection. Start with the single covariate that is most closely related to the dependent variable Y. Add that to the model. Then search among the remaining covariates to find the one which improves the 'adjusted R^2' the most when added to the model. Continue adding covariates until adding any more causes the 'adjusted R^2' to fall.

(2) Backward selection. Start by adding all available covariates. Then remove covariates one by one for which the hypothesis that $\beta_i = 0$ cannot be rejected until the 'adjusted R^2' value reaches a maximum, and all the remaining covariates have a statistically significant impact on Y.

Interactive terms of the form $\gamma_{ab} x_{ai} x_{bi}$ should be added where the effect of one predictor variable on the response variable depends on the value of another predictor variable.

![masomomsingi.com watermark]

Chapter 11b Practice Questions

11b.1   The effectiveness of a tablet containing x_1 *mg* of drug 1 and x_2 *mg* of drug 2 was being tested. In trials the following results were obtained:

% effectiveness, y	x_1	x_2
92.5	50.9	20.8
94.9	54.1	16.9
89.3	47.3	25.2
94.1	45.1	49.7
98.9	37.6	95.2

$$\sum y = 469.7 \quad \sum x_1 = 235 \quad \sum x_2 = 207.8 \quad \sum x_1^2 = 11{,}202.68 \quad \sum x_2^2 = 12{,}886.42$$

$$\sum yx_1 = 22{,}028.78 \quad \sum yx_2 = 19{,}870.22 \quad \sum x_1 x_2 = 8{,}985.96$$

(i)   Using the multiple linear least square regression model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

(a)   Show that the least squares estimates of α, β_1 and β_2 satisfy:

$$\sum y_i = n\alpha + \beta_1 \sum x_{i1} + \beta_2 \sum x_{i2}$$

$$\sum y_i x_{i1} = \alpha \sum x_{i1} + \beta_1 \sum x_{i1}^2 + \beta_2 \sum x_{i2} x_{i1}$$

$$\sum y_i x_{i2} = \alpha \sum x_{i2} + \beta_1 \sum x_{i1} x_{i2} + \beta_2 \sum x_{i2}^2$$

(b)   Hence, using the above data, show that the fitted model is:

$$\hat{y} = 25.31 + 1.194 x_1 + 0.3015 x_2 \qquad\qquad\qquad [7]$$

(ii)   By considering the following output from R for this model, comment on the significance of the parameters.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.308441   2.002062   12.64 0.006200 **
drug_1       1.193671   0.036592   32.62 0.000938 ***
drug_2       0.301468   0.007048   42.77 0.000546 ***
```
[2]

(iii)   The coefficient of determination for the fitted model is $R^2 = 0.9992$. Calculate the adjusted R^2.   [2]

(iv)   The ANOVA table for the model is:

Source of variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression	2	49.1137	*
Residual	2	0.0383	*
Total	4	49.152	

Calculate the missing values, the *F* statistic and then carry out the *F* test stating the conclusion clearly.   [4]

(v)   Calculate the percentage effectiveness for a tablet containing 51.3 *mg* of drug x_1 and 18.3 *mg* of drug x_2.   [2]

(vi)   The plot of the residuals against the fitted values and the Q-Q plot of the residuals are given below. Use them to comment on the fit of the model.   [2]

(vii)   (a)   It is thought that the two drugs might have an interactive effect. Explain what this means.

         (b)   Write down the formula for the regression model containing the two drugs as main effects and also their interaction.

         (c)   The model in part (vii)(b) has an adjusted R^2 of 0.9969. Using the answer from part (iii) comment on whether the new model is an improvement.   [4]

[Total 23]

Chapter 11b Solutions

11b.1 (i)(a)   *Least squares estimates equations*

We need to minimise the expression $Q = \sum (y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2}))^2$.   [1]

To do this, we need to differentiate the expression with respect to the parameters and set the expressions equal to zero:

$$\frac{\partial Q}{\partial \alpha} = -2\sum (y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})) = 0$$

$$\Rightarrow \sum y_i = n\alpha + \beta_1 \sum x_{i1} + \beta_2 \sum x_{i2} \qquad\qquad \text{eqn (1)}$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum x_{i1}(y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})) = 0$$

$$\Rightarrow \sum y_i x_{i1} = \alpha \sum x_{i1} + \beta_1 \sum x_{i1}^2 + \beta_2 \sum x_{i2}x_{i1} \qquad\qquad \text{eqn (2)}$$

$$\frac{\partial Q}{\partial \beta_2} = -2\sum x_{i2}(y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})) = 0$$

$$\Rightarrow \sum y_i x_{i2} = \alpha \sum x_{i2} + \beta_1 \sum x_{i1}x_{i2} + \beta_2 \sum x_{i2}^2 \qquad\qquad \text{eqn (3)}$$

[3]

(i)(b)   *Evaluate the least squares estimates*

Substituting these values into the equations above, we get:

(1) $469.7 = 5\alpha + 235\beta_1 + 207.8\beta_2$

(2) $22,028.78 = 235\alpha + 11,202.68\beta_1 + 8,985.96\beta_2$

(3) $19,870.22 = 207.8\alpha + 8,985.96\beta_1 + 12,886.42\beta_2$

Solving these simultaneously:

$$47 \times (1) - (2) \quad\Rightarrow\quad 47.12 = -157.68\beta_1 + 780.64\beta_2 \qquad\qquad \text{eqn (4)}$$

$$5 \times (3) - 207.8 \times (1) \quad\Rightarrow\quad 1,747.44 = -3,903.2\beta_1 + 21,251.26\beta_2 \qquad\qquad \text{eqn (5)}$$

$$157.68 \times (5) - 3,903.2 \times (4) \quad\Rightarrow\quad \beta_2 = 0.301468 \qquad\qquad [1]$$

Substituting this back in, we get $\beta_1 = 1.19367$ and $\alpha = 25.3084$, which gives us a regression line of $y = 25.31 + 1.194x_1 + 0.3015x_2$.   [2]

(ii)   *Significance of the parameters*

The *p*-values for all the parameters are less than 0.05 and so they are all significantly different from zero.   [2]

(iii)   Adjusted R^2

We have $n=5$ trials and $k=2$ predictors. Hence:

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2) = 1 - \left(\frac{5-1}{5-2-1}\right)(1-0.9992) = 0.9984 \qquad [2]$$

(iv)   ANOVA

The completed ANOVA table for the model is:

Source of variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression	2	49.1137	24.5569
Residual	2	0.0383	0.0192
Total	4	49.152	

[1]

The *F* statistic is:

$$F = \frac{SS_{REG}/k}{SS_{RES}/(n-k-1)} = \frac{24.5569}{0.0192} = 1280 \quad \text{(3 SF)} \qquad [1]$$

This is far in excess of even the 1% $F_{2,2}$ critical value of 99.00. Hence we can reject the null hypothesis that $\beta_1 = \beta_2 = 0$. [2]

(v)   Predict the percentage effectiveness

Substituting in the values given:

$$\hat{y} = 25.31 + (1.194 \times 51.3) + (0.3015 \times 18.3) = 92.1\% \qquad [2]$$

(vi)   Interpret plots

The first plot appears to be random and there is no discernible increase in the variance – so this would imply that the model meets these assumptions. Point 1 (92.5%) does appear to be an outlier. But it is difficult to tell with such a small dataset. [1]

With the exception of point (1) the rest of the values lie along the diagonal line thus implying a normal distribution is appropriate. [1]

(vii)(a)   Interaction

If there is interaction between the two drugs then there is an additional effect caused when both are present compared with what would be expected if they were each administered singly. [1]

(vii)(b)   Formula

The formula is $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_{12} X_1 X_2$. [1]

(vii)(c) *Compare models*

The model with just the two drugs as main effects had an adjusted R^2 of 0.9984 in part (iii) whereas the new model with the interactive effect has an adjusted R^2 of 0.9969.

Since there is a decrease in the value of the adjusted R^2 the previous model would be considered the 'best' model as the interaction term does not improve the fit enough to justify the extra parameter.   [2]

End of Part 3

What next?

1.   Briefly **review** the key areas of Part 3 and/or re-read the **summaries** at the end of Chapters 9 to 12.

2.   Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 3. If you don't have time to do them all, you could save the remainder for use as part of your revision.

3.   Attempt **Assignment X3**.

Time to consider …

… 'revision and rehearsal' products

Revision Notes – Each booklet covers one main theme of the course and includes integrated questions testing Core Reading, relevant past exam questions and other useful revision aids. One student said:

> *'Revision books are the most useful ActEd resource.'*

ASET – This contains past exam papers with detailed solutions and explanations, plus lots of comments about exam technique. One student said:

> *'ASET is the single most useful tool ActEd produces. The answers do go into far more detail than necessary for the exams, but this is a good source of learning and I am sure it has helped me gain extra marks in the exam.'*

You can find lots more information, including samples, on our website at www.ActEd.co.uk.

Buy online at www.ActEd.co.uk/estore

12

Generalised linear models

Syllabus objectives

4.2   Generalised linear models

4.2.1   Define an exponential family of distributions. Show that the following distributions may be written in this form: binomial, Poisson, exponential, gamma, normal.

4.2.2   State the mean and variance for an exponential family, and define the variance function and the scale parameter. Derive these quantities for the distributions above.

4.2.3   Explain what is meant by the link function and the canonical link function, referring to the distributions above.

4.3.4   Explain what is meant by a variable, a factor taking categorical values and an interaction term. Define the linear predictor, illustrating its form for simple models, including polynomial models and models involving factors.

4.2.5   Define the deviance and scaled deviance and state how the parameters of a GLM may be estimated. Describe horw a suitable model may be chosen by using an analysis of deviance and by examining the significance of the parameters.

4.2.6   Define the Pearson and deviance residuals and describe how they may be used.

4.2.7   Apply statistical tests to determine the acceptability of a fitted model: Pearson's Chi-square test and the Likelihood ratio test.

4.2.8   Fit a generalised linear model to a data set and interpret the output.

0   Introduction

In Chapter 11 we introduced linear models. The multiple linear regression model built on the simple model by adding more explanatory variables and then we extended this further by allowing functions of these variables, including interaction.

Recall that the bivariate linear regression model was $Y_i = \alpha + \beta x_i + e_i$ and the multiple linear regression model with k explanatory variables was $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i$. In the full normal model we assumed the error terms were normally distributed with mean 0 and variance σ^2. Hence, the response variable, Y_i, is also normally distributed.

Generalised linear models (GLMs) extend this further by allowing the distribution of the data to be non-normal.

This is particularly important in actuarial work where the data very often do not have a normal distribution. For example, in mortality, the Poisson distribution is used in modelling the force of mortality, μ_x and the exponential is used for survival analysis. In general insurance, the Poisson distribution is often used for modelling the claim frequency and the gamma or lognormal distribution for the claim severity. Finally, in all forms of insurance, the binomial distribution is used to model propensity.

Claim severity is just another term for the size of a claim, claim frequency refers to the rate at which claims are received and propensity refers to the probability of an event happening.

In this chapter we also introduce the idea of categorical explanatory variables (sometimes called factors).

GLMs are widely used both in general and life insurance. They are used to:

- determine which rating factors to use (rating factors are measurable or categorical factors that are used as proxies for risk in setting premiums, *eg* age or gender)

- estimate an appropriate premium to charge for a particular policy given the level of risk present.

For example, in motor insurance, there are a large number of factors that may be used as proxies for the level of risk (type of car driven, age of driver, number of years past driving experience, *etc*). We can use a GLM both to determine which of these factors are *significant* to the assessment of risk (and hence which should be included) and to suggest an appropriate premium to charge for a risk that represents a particular combination of these factors.

Question

Suggest rating factors that an insurance company may consider in the pricing of a single life annuity contract.

Solution

Rating factors that might be used in the pricing of a single life annuity include:

- age

- sex (if permitted by legislation)

- size of fund with which to purchase an annuity

- postcode

- health status (for impaired life annuities).

We have only used continuous variables so far in linear regression, such as weight, height and size of claim. On the other hand, categorical explanatory variables can only take categories, such as gender, type of car driven and relationship status (*eg* single, married, *etc*).

1   Generalised linear models

Generalised linear models (GLMs) relate the response variable which we want to predict, to the explanatory variables or factors (called predictors, covariates or independent variables) about which we have information.

Just like for our linear models we have inputs (explanatory variables) and we use our model to predict the output (response variable).

To fully define a GLM, we need to specify the following three components.

1.   *A distribution of the response variable*

For linear models the response variable had a normal distribution, $Y \sim N(0, \sigma^2)$. We now extend this to a general form of distributions known as the exponential family.

For example, we might choose a gamma distribution to model the sizes of motor insurance claims or a Poisson distribution to model the number of claims, or a binomial distribution to model the probability of contracting a certain disease.

2.   *A 'linear predictor'*

The linear predictor, η, is a function of the covariates. For the bivariate linear regression model this was $\beta_0 + \beta_1 x$. For the multivariate linear regression model this was $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, which we then extended to functions of the explanatory variables.

Note that η is the Greek letter 'eta'.

For example, if the response variable is weight, a linear predictor of $\beta_0 + \beta_1 x$ would be appropriate for a model where we thought the only covariate was height, *x*.

Note that a linear predictor is linear in the parameters β_0 and β_1. It does not have to be linear in the covariates, for example $\eta = \beta_0 + \beta_1 x^2$ is also a linear predictor.

3.   *A 'link function'*

The link function connects the mean response to the linear predictor, $g(\mu) = \eta$, where $\mu = E(Y)$. For linear models the mean response was equal to the linear predictor, eg $\mu = E(Y) = \beta_0 + \beta_1 x$, so the link function is the identity function, $g(\mu) = \mu$.

The link function, like its name suggests, is the link between the linear predictor (input) and the mean of the distribution (output).

Remember that what we are trying to do in a GLM is determine a relationship between the mean of the response variable and the covariates. By setting the link function $g(\mu) = \eta$, then, assuming that the link function is invertible, we can make the mean μ the subject of the formula:

$$\mu = g^{-1}(\eta)$$

The notation is not straightforward to get to grips with. An example may help.

Example

Suppose that we are trying to model the number of claims on car insurance policies. The response variable, Y_i, is the number of claims on Policy i. We decide that a Poisson distribution would be appropriate:

$$Y_i \sim Poi(\mu_i)$$

Consider a model where we believe that the only covariate is the age, x_i, of the policyholder. The linear predictor is $\eta_i = \alpha + \beta x_i$.

A link function that is commonly used with the Poisson distribution (see Page 27 of the *Tables*) is:

$$g(\mu) = \log \mu$$

We set this equal to the linear predictor:

$$g(\mu_i) = \log \mu_i = \eta_i = \alpha + \beta x_i$$

Now we invert the formula so that μ_i is the subject of the formula:

$$\mu_i = \exp(\eta_i) = \exp(\alpha + \beta x_i)$$

We now have a relationship between the mean of the response variable and the covariate.

Key information

The three components of a GLM are:

1)   a distribution for the data (Poisson, exponential, gamma, normal or binomial)

2)   a linear predictor (a function of the covariates that is linear in the parameters)

3)   a link function (that links the mean of the response variable to the linear predictor).

In order to understand how these three components fit together, we give a couple of further examples below.

Example

Suppose that we are setting up a model to predict the pass rate for a particular student in a particular actuarial exam. We might expect there to be many factors that affect whether a student is likely to pass or not. We might decide to set up a three-factor model, so that the probability of passing is a function of:

(i)   the number of assignments N submitted by the student (a value from 0 to 4)

(ii)   the student's mark on the mock exam S (on a scale from 0 to 100)

(iii)   whether the student had attended tutorials or not (Yes/No).

We might then decide to use the linear predictor:

$$\eta = \alpha_i + \beta_1 N + \beta_2 S$$

where α_i takes one value for those attending tutorials and a different value for those who do not.

We now need a link function. η here will not necessarily take a value in the interval (0, 1). Depending on the values of α_i, β_1 and β_2, η might take any value. If we use the link function $g(\mu) = \log\left(\dfrac{\mu}{1-\mu}\right)$ and set this equal to the linear predictor η, we have $\log\left(\dfrac{\mu}{1-\mu}\right) = \eta$. We invert this function to make μ the subject to give $\mu = \dfrac{e^{\eta}}{1+e^{\eta}} = \dfrac{1}{e^{-\eta}+1} = \left(1+e^{-\eta}\right)^{-1}$. We can now see that μ will lie in the range from zero to one, and so can be used as a pass rate.

We now use maximum likelihood estimation to estimate the four parameter values: α_Y, α_N (the α parameters corresponding to having attended tutorials and not having attended tutorials, respectively), β_1 (the parameter for the number of assignments) and β_2 (the parameter for the mock mark). To do this we need (ideally) the actual exam results of a large sample of students who fall into each of the categories.

Having done this for a set of data, we might come up with the following parameter values for the linear predictor:

$$\alpha_Y = -1.501 \qquad \alpha_N = -3.196 \qquad \beta_1 = 0.5459 \qquad \beta_2 = 0.0251$$

We can now use the linear predictor and link function to predict pass rates for groups of students with a particular characteristic. For example, for a student who attends tutorials, submits three assignments and scores 65% on the mock, we have:

$$\eta = -1.501 + 0.5459 \times 3 + 0.0251 \times 65 = 1.7682$$

We now use the inverse of the link function to calculate μ:

$$\mu = \left(1 + e^{-1.7682}\right)^{-1} = 0.8542$$

So the model predicts an 85% probability of passing for a student in this situation. So in this particular situation, the linear predictor is $\eta = \alpha_i + \beta_1 N + \beta_2 S$ and the link function is $g(\mu) = \log\left(\dfrac{\mu}{1-\mu}\right)$.

Question

Using the model outlined above, answer the following questions.

(i)   Calculate the predicted pass probability for a student who attends tutorials, submits three assignments and scores 60% on the mock exam.

(ii)   Calculate how much the probability would go up if the fourth assignment were submitted.

(iii)   Calculate the highest pass probability for someone who does not attend tutorials.

(iv)   Determine whether anyone gets a probability of 0 or 1 under this model. If not, calculate the minimum and maximum pass rates.

(v)   State the underlying probability distribution.

Solution

(i)   Using the values $N = 3$, $S = 60$ and $\alpha_Y = -1.501$, we get $\eta = 1.6427$, so that $\mu = 0.83790$. So the model predicts an 84% pass rate.

(ii)   If the fourth assignment was submitted we use $N = 4$ instead of $N = 3$ and get $\eta = 2.1886$ so that $\mu = 0.8992$. So the pass rate goes up by about 6%.

(iii)   Using $\alpha_N = -3.196$, $N = 4$ and $S = 100$, we get $\eta = 1.4976$, so that $\mu = 0.8172$. So the highest possible pass rate for someone who does not attend tutorials is about 82%.

(iv)   No. The minimum probability (for someone who does not attend tutorials or submit assignments and who scores zero on the mock) is obtained from a value of $\eta = -3.196$ which gives a pass probability of about 4%. The maximum probability of passing (for someone who goes to tutorials, submits all the assignments and scores 100% on the mock) comes from a value of $\eta = 3.1926$ which gives a pass rate of about 96%. So these are the maximum and minimum pass rates predicted by the model.

(v)   In fact, what we are doing here is finding a parameter of a binomial distribution. For any group of students with the same characteristics (*ie* all having the same values for all of the 3 factors), the number who pass may be well-modelled using a binomial distribution. The parameter of the binomial distribution $Z \sim Bin(n, \mu)$ that we are trying to find is the

value of $\mu = \dfrac{e^{\eta}}{1 + e^{\eta}}$ that we found above.

Note that we are again using μ to denote a probability as well as the mean of the response variable $Y = Z / n$.

Example

A statistician is analysing data on truancy rates for different school pupils. She believes that the number of unexplained days off school in a year (*ie* those not due to sickness *etc*) for a particular pupil may have a Poisson distribution with parameter μ. However, she believes that there are a number of factors that may affect μ, for example: age of pupil, whether he/she lives within the catchment area, and sex.

She builds a generalised linear model based on these characteristics, using data from a large group of pupils. Her model will take the form:

$$\eta = \alpha_i + \beta_j + \gamma\, x$$

where $x = \text{age}$, and α and β are numerical variables corresponding to the different characteristics for location and sex respectively.

She has collected the data shown in the table below. Each figure gives the average number of unexplained absences in a year for 16 different groups of pupils, all the pupils within each group having the same characteristics.

Average number of unexplained absences per pupil in a year

Age last birthday

		8	10	12	14
Within catchment area	Male	1.8	2.0	6.3	14.1
	Female	0.5	1.6	5.0	16.2
Outside catchment area	Male	2.1	7.5	25.5	72.0
	Female	2.8	6.2	19.6	68.2

By carrying out a maximum likelihood estimation analysis, she calculates the values of the parameters that fit the model best. As a result she can find a value of η for any particular pupil, which she can use to find the appropriate Poisson parameter using the link function. In this case she needs a function that converts a number η that may take any value into a positive number (since the Poisson parameter μ must always be positive). So for example she could use the link function $g(\mu) = \log \mu$, so that when this is set equal to the linear predictor η and inverted, $\mu = e^{\eta}$. This will give her a positive value for μ, which she can use for her Poisson parameter.

So she might come up with the following values for the parameters:

$$\alpha_{WC} = -2.64 \qquad \alpha_{OC} = -1.14 \qquad \beta_M = -3.26 \qquad \beta_F = -3.54 \qquad \gamma = 0.64$$

where WC = Within catchment , OC = Outside catchment, M = Male , F = Female . She can now use the model to predict possible truancy rates for students with particular characteristics.

The link function $g(\mu) = \log \mu$ is called the canonical link function for the Poisson distribution. Canonical just means the accepted form of the function. It is a 'natural' function to use, and will often give sensible results.

In fact it is not compulsory to use the canonical link function and there may be situations where a different link function is more appropriate. Each case must be judged on its merits.

Question

Determine the expected number of unexplained days' absence for a female pupil living within the catchment area who is 12 years old.

Solution

For this combination of factors we have:

$$\eta = \alpha_W + \beta_F + 12\gamma = 1.5$$

Using the link function given, we have:

$$\mu = e^{1.5} = 4.48$$

So the expected number of days unexplained absence in this case is about 4.5.

We will consider each of the components of a GLM (distribution, linear predictor, link function) in the next three sections.

In practice, the distribution of the data is usually specified at the outset (often defined by the data), the linear predictor may be chosen according to what is thought appropriate or convenient, and then the best model structure is found by looking at a range of linear predictors. Of course, these are not rules which must be adhered to: it may be that it is possible that more than one distribution could be appropriate, and these should be investigated before making a final decision. It could be unclear which link function should be used, and again a range of functions can be investigated.

The R code to fit a generalised linear model to a multivariate data frame and assign it to the object model, is:

```
model <- glm(Y ~ …, family = … (link = … ))
```

We will specify the inputs for the blanks in the following three sections.

2   Exponential family

Recall that the distribution of the response variable, Y, in a GLM is a member of the exponential family.

The exponential family is the set of distributions whose probability function, or probability density function (PDF) can be written in the following form:

$$f_Y(y; \theta, \varphi) = \exp\left[\frac{(y\theta - b(\theta))}{a(\varphi)} + c(y, \varphi)\right] \tag{1}$$

where $a(\varphi)$, $b(\theta)$ and $c(y, \varphi)$ are specific functions.

This formula is given on Page 27 of the *Tables*. Note that φ is just another way of writing the Greek letter phi, usually written as ϕ.

There are two parameters in the above PDF. θ, which is called the 'natural' parameter, is the one which is relevant to the model for relating the response (Y) to the covariates, and φ is known as the scale parameter or dispersion parameter.

When trying to show that a distribution is a member of the exponential family, it is important to remember that θ is a function of $\mu = E(Y)$ only. We shall see later in the chapter exactly how θ is used to relate the response to the covariates.

Where a distribution has two parameters, such as the $N(\mu, \sigma^2)$, one approach to determining the scale parameter φ is to take φ to be the 'other' parameter in the distribution, *ie* the parameter other than the mean. For example, in the case of the normal distribution, we take $\varphi = \sigma^2$. Where a distribution has one parameter, such as $Poi(\lambda)$, we take $\varphi = 1$.

Consider the following statement about a continuous PDF:

$$\int_y f(y)\, dy = 1 \tag{2}$$

By substituting the expression from (1) and differentiating this with respect to θ, it can be shown that the mean and variance of Y are:

$$E[Y] = b'(\theta)$$

and

$$\text{var}(Y) = a(\varphi)b''(\theta).$$

These formulae can also be found on Page 27 of the *Tables*.

Question

Prove these two results for a member of the exponential family, using the results given above.

Solution

Mean

Differentiating both sides of equation (2) with respect to θ gives:

$$\int_y \frac{y - b'(\theta)}{a(\varphi)} f(y)\, dy = 0 \qquad\qquad (3)$$

Simplifying:

$$\frac{1}{a(\varphi)} \int_y y f(y)\, dy - \frac{b'(\theta)}{a(\varphi)} \int_y f(y)\, dy = 0$$

Since $\int y f(y)\, dy = E(Y)$, and $\int f(y)\, dy = 1$, we have:

$$\frac{1}{a(\varphi)} E(Y) - \frac{b'(\theta)}{a(\varphi)} = 0$$

Hence:

$$E(Y) - b'(\theta) = 0 \quad \Rightarrow \quad E(Y) = b'(\theta)$$

Variance

Using the product rule to differentiate equation (3) with respect to θ gives:

$$\int_y \frac{d^2}{d\theta^2} f(y)\, dy = \int_y \left\{ \left(\frac{y - b'(\theta)}{a(\varphi)} \right)^2 f(y) - \frac{b''(\theta)}{a(\varphi)} f(y) \right\} dy = 0$$

Splitting this into two separate integrals gives:

$$\frac{1}{\left[a(\varphi) \right]^2} \int_y \left(y - b'(\theta) \right)^2 f(y)\, dy - \frac{b''(\theta)}{a(\varphi)} \int_y f(y)\, dy = 0$$

Since $b'(\theta) = E(Y)$ then $\int (y - b'(\theta))^2 f(y)\, dy = \operatorname{var}(Y)$. Again $\int f(y)\, dy = 1$, so we have:

$$\frac{1}{\left[a(\varphi) \right]^2} \operatorname{var}(Y) - \frac{b''(\theta)}{a(\varphi)} = 0$$

Rearranging gives:

$$\operatorname{var}(Y) = a(\varphi) b''(\theta)$$

In general, note that the mean does not depend on φ, so when predicting Y it is θ which is of importance. Also, the variance of the data has two components: one which involves the scale parameter, and the other which determines the way the variance depends on the mean.

The variance of the normal distribution does not depend on the mean.

That's because μ and σ^2 are independent.

For other distributions, however, the variance does depend on the mean.

For example, the Poisson distribution has mean and variance both equal to the parameter μ. So knowing the mean of a Poisson distribution tells us the variance as well.

To emphasise this dependence on the mean the variance is often written as $\text{var}(Y) = a(\varphi)V(\mu)$, where the 'variance function' is defined as

$$V(\mu) = b''(\theta)$$

Note that the variance function does not give the variance directly, unless $a(\varphi) = 1$.

2.1 Normal distribution

To motivate these definitions and the subsequent developments, we consider first the normal distribution.

For members of an exponential family, we want to be able to find formulae for the mean and variance of the distribution from the general parameters.

First we will rewrite the normal distribution in the form of equation (1) and then consider other distributions as exponential families. Note that we use *f*, in a slight abuse of notation, for both continuous and discrete distributions.

We have seen this style of notation before in this subject. The alternative notation is to use $p(x)$ for a probability function and $f(x)$ for a density function. Provided that the method is clear, either notation is acceptable.

$$f_Y(y;\theta,\varphi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right]$$

$$= \exp\left[\frac{\left(y\mu - \mu^2/2\right)}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log 2\pi\sigma^2\right)\right]$$

which is in the form of (1), with:

$$\theta = \mu$$

$$\varphi = \sigma^2$$

$$a(\varphi) = \varphi$$

$$b(\theta) = \frac{\theta^2}{2}$$

$$c(y, \varphi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log 2\pi\phi\right)$$

Thus, the natural parameter for the normal distribution is μ **and the scale parameter is** σ^2.

Note that we could alternatively have taken $\varphi = \sigma$ and $a(\varphi) = \varphi^2$. There is no unique parameterisation.

Using the formulae above, the mean is $E(Y) = b'(\theta) = \theta = \mu$ **and the variance is**

$$\mathbf{var}(Y) = a(\varphi)b''(\theta) = \varphi = \sigma^2.$$

So these do give us the results that we expect for the normal distribution.

Question

Show that if we reparameterise the normal distribution using $\theta = 2\mu$, we still get the same results for the mean and variance of the distribution.

Solution

If we put $\theta = 2\mu$, we get the following expressions for the various functions:

$$a(\phi) = 2\phi \qquad\qquad b(\theta) = \theta^2 / 4 \qquad\qquad c(y, \phi) = -\tfrac{1}{2}(y^2 / \phi + \log 2\pi\phi)$$

Using the formulae for the mean and variance, as before:

$$E(Y) = b'(\theta) = 2\theta / 4 = \frac{2 \times 2\mu}{4} = \mu$$

and $\qquad \text{var}(Y) = b''(\theta)a(\phi) = 1 / 2 \times 2\phi = \phi = \sigma^2$

So the mean and variance are μ and σ^2, as before.

As mentioned above, for the normal distribution, the variance of *Y* does not depend on the mean (the variance function $V(\mu) = b''(\theta) = 1$**), whereas for other distributions the variance *does* depend on the mean.**

In R, to use a normal distribution in the `glm` **command, we set** `family=gaussian` **(or omit this option as this distribution is the default).**

2.2 Poisson distribution

For the Poisson distribution:

$$f_Y(y;\theta,\varphi) = \frac{\mu^y e^{-\mu}}{y!} = \exp[y\log\mu - \mu - \log y!]$$

which is in the form of (1), with:

$$\theta = \log\mu$$

$$\varphi = 1, \text{ so that } a(\varphi) = 1$$

$$b(\theta) = e^\theta$$

$$c(y,\varphi) = -\log y!$$

Thus, the natural parameter for the Poisson distribution is $\log\mu$**, the mean is**

$E(Y) = b'(\theta) = e^\theta = \mu$ **and the variance function is** $V(\mu) = b''(\theta) = e^\theta = \mu$**. The variance function tells us that the variance is proportional to the mean. We can see that the variance is actually *equal* to the mean since** $a(\varphi) = 1$**.**

Question

Comment on whether or not we can reparameterise the Poisson distribution using $\varphi = 2$, say.

Solution

Yes. Just as before with the normal distribution, there is more than one way to set up the parameters. However the natural approach is to use $\phi = 1$ rather than $\phi = 2$, and this is the most sensible approach to use in the exam.

In R, to use a Poisson distribution in the `glm` **command, we set** `family=poisson`**.**

2.3   Binomial distribution

This is slightly more awkward to deal with, since we have to first divide the binomial random variable by n. Thus, suppose $Z \sim Bin(n, \mu)$. Let $Y = Z/n$, so that $Z = nY$. The distribution of Z is $f_Z(z; \theta, \varphi) = \binom{n}{z} \mu^z (1 - \mu)^{n-z}$ and by substituting for z, the distribution of Y is:

$$f_Y(y; \theta, \varphi) = \binom{n}{ny} \mu^{ny} (1 - \mu)^{n-ny}$$

$$= \exp \left[n(y \log \mu + (1 - y) \log(1 - \mu)) + \log \binom{n}{ny} \right]$$

$$= \exp \left[n \left(y \log \left(\frac{\mu}{1 - \mu} \right) + \log(1 - \mu) \right) + \log \binom{n}{ny} \right]$$

which is in the form of (1), with:

$$\theta = \log \left(\frac{\mu}{1 - \mu} \right) \text{ (note that the inverse of this is } \mu = \frac{e^\theta}{1 + e^\theta} \text{)}$$

$$\varphi = n$$

$$a(\varphi) = \frac{1}{\varphi}$$

$$b(\theta) = \log(1 + e^\theta)$$

$$c(y, \varphi) = \log \binom{n}{ny}$$

The reason for all this is that θ is a function of μ, the distribution mean only. However, the binomial distribution as we typically quote it: $Bin(n, p)$ does not have μ as one of its parameters. So we will start by considering $Bin(n, \mu)$, which does have μ as a parameter, but has mean $n\mu$. We then divide this by n to get a distribution with μ in its probability function and which also has mean μ.

Note that $\varphi = n$, the 'other' parameter in the distribution (*ie* the parameter other than the mean).

Question

Verify that the formulae given in the Core Reading are correct.

Solution

If $\theta = \log \dfrac{\mu}{1-\mu}$, then to get n in the denominator we need $a(\phi) = 1/\phi$ with $\phi = n$. Similarly, $b(\theta)$ must be given by:

$$-b(\theta) = \log(1-\mu) = \log\left(1 - \frac{e^{\theta}}{1+e^{\theta}}\right) = \log\left(\frac{1}{1+e^{\theta}}\right) = -\log\left(1+e^{\theta}\right)$$

So $b(\theta) = \log\left(1+e^{\theta}\right)$ as required, and $c(y,\phi) = \log\begin{pmatrix} n \\ ny \end{pmatrix}$.

Thus, the natural parameter for the binomial distribution is $\log\left(\dfrac{\mu}{1-\mu}\right)$, the mean is:

$$\boldsymbol{E[Y] = b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \mu}$$

and the variance function is:

$$\boldsymbol{V(\mu) = b''(\theta) = \frac{e^{\theta}}{(1+e^{\theta})^2} = \mu(1-\mu)}$$

We can get the second derivative of $b(\theta)$ most easily by writing $b'(\theta) = 1 - \left(1+e^{\theta}\right)^{-1}$.

Question

Comment on whether these are the results we would expect.

Solution

Yes. Since Z is binomial with mean $n\mu$ and variance $n\mu(1-\mu)$ and $Z = nY$, we should have:

$$E(Y) = \frac{1}{n} \times E(Z) = \frac{1}{n} \times n\mu = \mu$$

and:

$$\text{var}(Y) = \frac{1}{n^2}\text{var}(Z) = \frac{n\mu(1-\mu)}{n^2} = \frac{\mu(1-\mu)}{n} = a(\phi)V(\mu)$$

These agree with the results that we actually got.

In R, to use a binomial distribution in the `glm` **command, we set** `family=binomial.`

2.4   Gamma distribution

The best way to consider the Gamma distribution is to change the parameters from α and λ to α and $\mu = \dfrac{\alpha}{\lambda}$, ie $\lambda = \dfrac{\alpha}{\mu}$.

Recall that θ must always be expressed as a function of μ, so the best way to start is to ensure that $\mu = \alpha / \lambda$ appears in the PDF formula. We can do this by replacing the λ:

$$f_Y(y;\theta,\varphi) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} = \frac{\alpha^\alpha}{\mu^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y\alpha/\mu}$$

$$= \exp\left[\left(-\frac{y}{\mu} - \log\mu\right)\alpha + (\alpha-1)\log y + \alpha\log\alpha - \log\Gamma(\alpha)\right]$$

which is in the form of (1), with:

$$\theta = -\frac{1}{\mu}$$

$$\varphi = \alpha$$

$$a(\varphi) = \frac{1}{\varphi}$$

$$b(\theta) = -\log(-\theta)$$

$$c(y,\varphi) = (\varphi-1)\log y + \varphi\log\varphi - \log\Gamma(\varphi)$$

Notice that θ here is negative, so that $\log(-\theta)$ is well defined.

Thus, the natural parameter for the gamma distribution is $\dfrac{1}{\mu}$, ignoring the minus sign. The mean is $E[Y] = b'(\theta) = -\dfrac{1}{\theta} = \mu$. The variance function is $V(\mu) = b''(\theta) = \dfrac{1}{\theta^2} = \mu^2$ and so the variance is $\dfrac{\mu^2}{\alpha}$.

Note that $\varphi = \alpha$, the 'other' parameter in the distribution (*ie* the parameter other than the mean).

> **In R, to use a normal distribution in the `glm` command, we set `family=Gamma`.**

Question

Show that the exponential distribution can be written in the form of a member of the exponential family.

Solution

We can write the PDF of an exponential distribution as:

$$f(y) = \lambda e^{-\lambda y} = e^{\log \lambda - \lambda y}$$

Since $E(Y) = \dfrac{1}{\lambda}$, this is in the appropriate form with:

$$\theta = -\lambda = -\frac{1}{\mu} \qquad\qquad b(\theta) = -\log(-\theta) \qquad\qquad \phi = 1 \qquad\qquad a(\phi) = \phi$$

and $\qquad c(y, \phi) = 0$

The $Exp(\lambda)$ distribution is just a $Gamma(1, \lambda)$ distribution, so our results are consistent with those for the gamma distribution.

2.5   Lognormal distribution

Finally, the lognormal distribution is often used, for example in general insurance to model the distribution of claim sizes. This can be incorporated in the framework of GLMs since if $Y \sim$ lognormal, $\log Y \sim$ normal. Thus, if the lognormal distribution is to be used, the data should first be logged and then the normal modelling distribution can be applied.

In R we could either set another variable, say `Z`, **equal to** `log(Y)` **and then model** `glm(Z ~ …, family = gaussian (link = … ))` **or we use** `glm(log(Y) ~ …, family = gaussian (link = … )).`

Syllabus objective 4.2.1 requires students to show that the binomial, Poisson, exponential, gamma and normal distributions are members of the exponential family.

Having the distribution of the GLM response variable belonging to the exponential family ensures the calculations are easier when estimating the parameters using maximum likelihood. It also ensures that the model possesses good statistical properties.

3 Linear predictor

Recall that the second component of a GLM is the linear predictor, η, which is a function of the covariates, *ie* the input variables to the model.

The covariates (also known as explanatory, predictor or independent variables), enter the model through the linear predictor. This is also where the parameters occur which have to be estimated. The requirement is that it is linear *in the parameters* that we are estimating.

There are two kinds of covariates used in GLMs: variables and factors.

3.1 Variables

In general, *variables* are covariates where the actual value of a variable enters the linear predictor. The age of the policyholder is an actuarial example of a variable. So far in our linear models we have only met continuous variables.

A variable is a type of covariate whose real numerical value enters the linear predictor directly, such as age (*x*). Other examples of variables in a car insurance context are annual mileage and number of years for which a driving licence has been held.

The bivariate linear model had a single continuous explanatory variable x with a linear predictor of $\beta_0 + \beta_1 x$. To fit this model it is necessary to estimate the parameters, β_0 and β_1 and so the actual value of x matters. For the multivariate linear regression model with k continuous main effect variables this was $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$. Again the values of $x_1, x_2, \ldots, x_k$ matter.

We use the same R formulae in the `glm` function as we did in the `lm` function:

    glm(Y ~ X, …)

    glm(Y ~ X1+X2+…+Xk, …)

As in the previous unit we can extend our models to include polynomials, to functions of the variable and to linear predictors including more than one variable.

Recall that the linear predictor is linear in the parameters (*eg* β_0 and β_1) and not necessarily linear in the covariates (*eg* $\eta = \beta_0 + \beta_1 x^2$ is also a linear predictor).

Some examples, where age (x_1) and duration (x_2) are treated as variables, are shown in the table below together with the formula used in the `glm` function in R.

model	linear predictor	R formula
1 (null model)	β_0	`Y ~ 1`
age	$\beta_0 + \beta_1 x_1$	`Y ~ X1`
age^2	$\beta_0 + \beta_1 x_1^2$	`Y ~ I(X^2)`
age + age^2	$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	`Y ~ X1 + I(X1^2)`
age + duration	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	`Y ~ X1 + X2`
log(age)	$\beta_0 + \beta_1 \log x_1$	`Y ~ log(X1)`

The null model is where there are no covariates and so there is just the intercept parameter. This will be estimated as the sample mean of the response values.

It's fairly easy to see that we start with an intercept parameter and then add a new term with a 'slope' parameter multiplied by the covariate. However, there is actually a little more happening before we get to this simplified linear predictor.

Suppose the linear predictor for age only is $\alpha_1 + \beta_1 x_1$ and the linear predictor for duration only is $\alpha_2 + \beta_2 x_2$ then we could obtain a linear predictor for both of these covariates by summing their individual linear predictors:

$$(\alpha_1 + \beta_1 x_1) + (\alpha_2 + \beta_2 x_2) = (\alpha_1 + \alpha_2) + \beta_1 x_1 + \beta_2 x_2$$

Now the final simplified version given in the table above, $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, combines the two constants together, *ie* $\beta_0 = \alpha_1 + \alpha_2$.

We can see that this simplified formula will give the same final values as the uncombined formula and so it makes sense to use the simplified version as it is more efficient and only requires us to estimate 3 instead of 4 parameters.

However, it is actually impossible to estimate α_1 and α_2 individually from any given data and hence we *have* to combine them in the linear predictor to overcome this issue. We can demonstrate this in the following question where we give 4 values of the linear predictor which should be sufficient to estimate 4 parameters.

Question

The table below shows the value of the linear predictor $\eta = (\alpha_1 + \alpha_2) + \beta_1 x_1 + \beta_2 x_2$ for different values of age (x_1) and duration (x_2).

Linear predictor, η	age (x_1)	duration (x_2)
35	20	0
37	20	1
45	30	0
55	30	5

Show that it is impossible to individually estimate all the parameters in the linear predictor.

Solution

Substituting the given values into the formula for the linear predictor gives the following four equations:

$$35 = (\hat{\alpha}_1 + \hat{\alpha}_2) + 20\hat{\beta}_1 \qquad\qquad (1)$$

$$37 = (\hat{\alpha}_1 + \hat{\alpha}_2) + 20\hat{\beta}_1 + \hat{\beta}_2 \qquad\qquad (2)$$

$$45 = (\hat{\alpha}_1 + \hat{\alpha}_2) + 30\hat{\beta}_1 \qquad\qquad (3)$$

$$55 = (\hat{\alpha}_1 + \hat{\alpha}_2) + 30\hat{\beta}_1 + 5\hat{\beta}_2 \qquad\qquad (4)$$

Subtracting equations (1) and (2) gives $\hat{\beta}_2 = 2$.

Subtracting equations (1) and (3) gives $10\hat{\beta}_1 = 10$ and hence $\hat{\beta}_1 = 1$.

However, substituting the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ into all four equations gives $(\hat{\alpha}_1 + \hat{\alpha}_2) = 15$. Hence we can only estimate their total $\hat{\beta}_0 = (\hat{\alpha}_1 + \hat{\alpha}_2) = 15$ and are unable to give individual estimates for α_1 and α_2.

Incidentally, the actual values used in the question above were $\alpha_1 = 5.5$ and $\alpha_2 = 9.5$. It is easy to see that the simplification above gives exactly the same answers as the original values.

Other models can also be fitted, including, for example, a model for age with no intercept term.

We omit the intercept in R by adding a $-\ 1$ to the formula.

It's unusual to have models with no intercept term as these would give a value of zero when a covariate is zero.

3.2 Interaction between variables

In addition to considering variables as the *main effect* we can include *interactions* between variables like we did in Chapter 11b, Section 2.7 (*ie* where the effect on the response variable of one predictor variable depends on the value of another predictor variable).

So far each covariate has been incorporated into the linear predictor through an additive term $\beta_i x_i$. Such a term is called the *main effect* for that covariate.

For a main effect, the covariate increases the linear predictor by β_i for each unit increase in x_i *independently* of all the other covariates.

When there is interaction between two variables, say x_i and x_j, then they are not independent and so the effect of one covariate, x_i, on the linear predictor depends on the value of the other covariate, x_j. We model this using an additive term of the form $\beta_{ij} x_i x_j$ in the linear predictor.

Recall that when an interaction term is used in a model, both main effects must also be included.

Otherwise we are saying that the variables don't contribute anything independently which implies that they are perfectly correlated and hence one of them is unnecessary.

model	linear predictor	R formula
age + duration + age.duration	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	`Y ~ X1+X2+X1:X2`
age * duration	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	`Y ~ X1*X2`

The two models in the table above are equivalent, and have been shown separately to illustrate the use of the dot and star model notation in R.

An interaction term is denoted using the dot notation. In the example above, 'age.duration' denotes the interaction between age and duration (although in R a colon is used to prevent confusion with a decimal point).

The star notation is used to denote the main effects *and* the interaction term. In the example above, age*duration = age + duration + age.duration.

3.3 Factors and interaction between factors

The other main type of covariate is a *factor*, which takes a categorical value. For example, the sex of the policyholder is either male or female, which constitutes a factor with 2 categories (or levels).

Other examples of factors in a car insurance context are postcode and car type.

This type of covariate can be parameterised so that the linear predictor has a term α_1 for a male, and a term α_2 for a female (*ie* α_i where $i = 1$ for a male and $i = 2$ for a female). In general, there is parameter for each level that the factor may take.

Factors are typically non-numerical (*eg* sex) and even those that are (*eg* vehicle rating group) it doesn't make any sense to include their value in the linear predictor. Instead we simply assign parameter values for each possible category it can take.

In the following table sex and vehicle rating group (`vrg`) are factors. If there is more than one factor in the model, then the inclusion of an interaction term implies that the effect of each factor depends on the level of the other factor.

model	linear predictor	R formula
sex	α_i	`Y ~ sex`
vehicle rating group	β_j	`Y ~ vrg`
sex + vehicle rating group	$\alpha_i + \beta_j$	`Y ~ sex + vrg`
sex + vehicle rating group + sex.vehicle rating group	$\alpha_i + \beta_j + \gamma_{ij}$	`Y ~ sex+vrg+sex:vrg`
sex*vehicle rating group	$\alpha_i + \beta_j + \gamma_{ij}$	`Y ~ sex * vrg`

Again, the last two models are identical.

As mentioned before sex is a factor with a parameter assigned to each of its two categories (α_i where $i = 1$ for a male and $i = 2$ for a female).

Similarly vehicle rating group is a factor with a parameter assigned to each of its categories. Suppose there were three categories (group 1, group 2 and group 3) then we would have $\beta_j, j = 1, 2, 3$. Note that we use a different subscript to sex since the main effects give the change to the linear predictor independently of all the other covariates, whereas using i as the subscript for both would mean that males would always be in group 1 and females would always be in group 2.

Again we can construct linear predictors that involve more than one covariate by summing the linear predictors for each individual covariate. Hence $\alpha_i + \beta_j$ for age + vrg. However, it is again impossible to estimate $\alpha_1, \alpha_2, \beta_1, \beta_2$ and β_3 individually from any given data and so we will have to combine constants together to overcome this issue. This will mean that one of those constants effectively becomes zero. This is called the base assumption in the model.

We can demonstrate this in the following question where we give 5 values of the linear predictor which should be sufficient to estimate 5 parameters.

Question

The table below shows the value of the linear predictor $\eta = \alpha_i + \beta_j$ for different values of i and j

Linear predictor, η	Sex, i	vrg, j
0.5	Male	1
0.48	Male	2
0.41	Male	3
0.6	Female	1
0.58	Female	2

Show that it is impossible to individually estimate all the parameters in the linear predictor.

Solution

Substituting the given values into the formula for the linear predictor gives the following five equations:

$$0.5 = \hat{\alpha}_1 + \hat{\beta}_1 \tag{1}$$

$$0.48 = \hat{\alpha}_1 + \hat{\beta}_2 \tag{2}$$

$$0.41 = \hat{\alpha}_1 + \hat{\beta}_3 \tag{3}$$

$$0.6 = \hat{\alpha}_2 + \hat{\beta}_1 \tag{4}$$

$$0.58 = \hat{\alpha}_2 + \hat{\beta}_2 \tag{5}$$

Subtracting (1) and (4) (or subtracting (2) and (5)) gives $\hat{\alpha}_2 = \hat{\alpha}_1 + 0.1$.

Subtracting (1) and (2) (or subtracting (4) and (5)) gives $\hat{\beta}_2 = \hat{\beta}_1 - 0.02$.

Subtracting (1) and (3) gives $\hat{\beta}_3 = \hat{\beta}_1 - 0.09$.

We can try other combinations but none of them will be able to give us estimates of all 5 parameters.

If, however, we set one constant to zero, say $\hat{\alpha}_1 = 0$ (*ie* our base assumption is that the policyholder is male) and all the other parameters are calculated relative to this base level, we obtain:

$$\hat{\alpha}_i = \begin{cases} 0 & i = 1 \, (\text{male}) \\ 0.1 & i = 2 \, (\text{female}) \end{cases} \text{ and } \hat{\beta}_j = \begin{cases} 0.5 & j = 1 \\ 0.48 & j = 2 \\ 0.41 & j = 3 \end{cases}$$

Incidentally the actual values used in creating the above question were:

$$\alpha_i = \begin{cases} 0.45 & i = 1\,(\text{male}) \\ 0.55 & i = 2\,(\text{female}) \end{cases} \quad \text{and} \quad \beta_j = \begin{cases} 0.05 & j = 1 \\ 0.03 & j = 2 \\ -0.04 & j = 3 \end{cases}$$

It might be worth spending a moment checking that every combination of sex and vehicle rating group gives exactly the same answer as the estimated values.

Finally, we consider an interaction between two factors.

sex*vehicle rating group = sex + vehicle rating group + sex . vehicle rating group

We start by summing the linear predictors for each of the three terms: sex, vehicle rating group and sex . vehicle rating group separately. We already know that the linear predictor for sex alone is $\eta = \alpha_i, i = 1, 2$. Similarly, the linear predictor for vehicle group alone is $\eta = \beta_j, j = 1, 2, 3$. We also need a linear predictor for the interaction effect $\eta = \alpha_i . \beta_j$ $i = 1, 2$ and $j = 1, 2, 3$. The dot notation here does not mean multiply. We have just written it in this format for now to indicate an interaction.

$$\eta = \alpha_i + \beta_j + \alpha_i . \beta_j$$

An alternative (and more commonly used) notation for the interaction term which depends on both i and j is γ_{ij}, so that:

$$\eta = \alpha_i + \beta_j + \gamma_{ij}$$

Interaction between sex and vehicle group indicates that the difference in risk levels for male and female drivers varies for different vehicle groups.

For example, if the response variable is the number of claims on a car insurance policy, the effect of being male might depend on whether the car being driven is a Porsche (where the driver might be tempted to show off) or a Mini (where the driver might drive more carefully).

However, it is again impossible to estimate all the parameters (α_1, α_2, β_1, β_2, β_3, γ_{11}, γ_{12}, γ_{13}, γ_{21}, γ_{22} and γ_{23}) individually from any given data and so we will have to combine constants together to overcome this issue. This time it will be harder as there are only 6 combinations that we can observe:

	group 1	group 2	group 3
male	$\alpha_1 + \beta_1 + \gamma_{11}$	$\alpha_1 + \beta_2 + \gamma_{12}$	$\alpha_1 + \beta_3 + \gamma_{13}$
female	$\alpha_2 + \beta_1 + \gamma_{21}$	$\alpha_2 + \beta_2 + \gamma_{22}$	$\alpha_2 + \beta_3 + \gamma_{23}$

But there are 11 parameters to estimate so we will have to set 5 of the parameters equal to zero to be able to solve the relevant equations.

For example, we might get the following:

$$\hat{\alpha}_i = \begin{cases} 0 & i=1 \,(\text{male}) \\ 0.66 & i=2 \,(\text{female}) \end{cases} \qquad \hat{\beta}_j = \begin{cases} 0 & j=1 \\ 0.5 & j=2 \\ 0.4 & j=3 \end{cases} \qquad \hat{\gamma}_{ij} = \begin{cases} & j=1 & j=2 & j=3 \\ i=1 & 0.55 & 0 & 0 \\ i=2 & 0 & -0.55 & -0.58 \end{cases}$$

when the true values might be:

$$\alpha_i = \begin{cases} 0.45 & i=1 \,(\text{male}) \\ 0.55 & i=2 \,(\text{female}) \end{cases} \qquad \beta_j = \begin{cases} 0.05 & j=1 \\ 0.03 & j=2 \\ -0.04 & j=3 \end{cases} \qquad \gamma_{ij} = \begin{cases} & j=1 & j=2 & j=3 \\ i=1 & 0.05 & 0.02 & -0.01 \\ i=2 & 0.06 & 0.03 & -0.03 \end{cases}$$

Again, it might be worth checking that every combination of sex and vehicle rating group gives exactly the same answer as the estimated values.

An alternative linear predictor that gives identical results to $\eta = \alpha_i + \beta_j + \gamma_{ij}$ would be:

$$\eta = \delta_{ij} \qquad \text{where } \delta_{ij} = \begin{cases} & j=1 & j=2 & j=3 \\ i=1 & 0.55 & 0.5 & 0.4 \\ i=2 & 0.66 & 0.61 & 0.48 \end{cases}$$

This gives the six possible combinations directly. Again, do spend a short while checking that every combination of sex and vehicle rating group gives exactly the same answer as before.

3.4 Predictors with variables and factors and interaction

Finally, we'll look at models which contain both variables and factors. For example, a model which includes an age effect and an effect for the sex of the policyholder could have a linear predictor

$$\alpha_i + \beta x$$

Again we can construct linear predictors that involve more than one covariate by summing the linear predictors for each individual covariate.

Here, the linear predictor for 'age' is $\eta = \beta_0 + \beta_1 x$ and the linear predictor for 'sex' is $\eta = \alpha_i, i = 1, 2$. Summing these gives:

$$\eta = \beta_0 + \beta_1 x + \alpha_i = (\beta_0 + \alpha_i) + \beta_1 x$$

Once again it will be impossible to estimate the parameters β_0 and α_i individually so we will have to combine them together:

$$\eta = \alpha_i' + \beta_1 x$$

Note that we have added a dash to indicate that the values here are not the same as in the original α_i whereas the Core Reading will just skip to the simplified result and write $\alpha_i + \beta_1 x$.

For example, suppose we have the following values of the linear predictor $\eta = (\beta_0 + \alpha_i) + \beta_1 x$ for different values of age (x) and different genders.

Linear predictor, η	age (x_1)	sex
1.45	20	Male
1.95	30	Male
1.55	20	Female
2.05	30	Female

Since there are four unknown parameters to estimate then four data points should be sufficient. Substituting the given values into the formula for the linear predictor gives the following four equations:

$$1.45 = (\hat{\beta}_0 + \hat{\alpha}_1) + 20\hat{\beta}_1 \qquad (1)$$

$$1.95 = (\hat{\beta}_0 + \hat{\alpha}_1) + 30\hat{\beta}_1 \qquad (2)$$

$$1.55 = (\hat{\beta}_0 + \hat{\alpha}_2) + 20\hat{\beta}_1 \qquad (3)$$

$$2.05 = (\hat{\beta}_0 + \hat{\alpha}_2) + 30\hat{\beta}_1 \qquad (4)$$

Subtracting equations (1) and (2) (or subtracting equations (3) and (4)) gives $10\hat{\beta}_1 = 0.5$ and hence $\hat{\beta}_1 = 0.05$.

Subtracting equations (1) and (3) (or subtracting equations (2) and (4)) gives $\hat{\alpha}_2 - \hat{\alpha}_1 = 0.1$ and hence $\hat{\alpha}_2 = \hat{\alpha}_1 + 0.1$.

However, substituting $\hat{\beta}_1 = 0.05$ and $\hat{\alpha}_2 = \hat{\alpha}_1 + 0.1$ into both the other equations gives $(\hat{\beta}_0 + \hat{\alpha}_2) = 0.55$ and so we are unable to estimate these parameters separately. Hence, we absorb the β_0 into the sex parameters to resolve this issue.

Notice that the parameter β_0 is redundant and has not been included (it could not be estimated separately from α_1 and α_2).

This gives

$$\hat{\beta}_1 = 0.05 \text{ and } \hat{\alpha}'_i = \begin{cases} 0.45 & \text{if } i = 1 \text{ (male)} \\ 0.55 & \text{if } i = 2 \text{ (female)} \end{cases}$$

Incidentally, the actual values used to construct the question above were:

$$\beta_0 = 0.5 , \ \beta_1 = 0.05 \text{ and } \alpha_i = \begin{cases} -0.05 & i = 1 \text{ (male)} \\ 0.05 & i = 2 \text{ (female)} \end{cases}$$

Again, it's probably worth spending a short while checking that every combination of age and sex gives exactly the same answer as the estimated values.

Notice also that the effect of the age of the policyholder is the same whether the policyholder is male or female.

In other words, age and sex are independent covariates. There is no *interaction* between them.

In this case if we were to draw a graph of the linear predictor, it would consist of two parallel straight lines (one for males and one for females).

Including the interaction between the age and sex would lead to a linear predictor of:

$$\alpha_i + \beta_i x \qquad (i = 1, 2)$$

Recall that an interaction is where the effect of one covariate (*eg* age) on the linear predictor depends on the value that another covariate (*eg* sex) takes.

In this case, the effect of the age of the policyholder is different for males and females.

Now the graph would be two straight lines, but they would no longer be parallel. So there is interaction between the effects of age and sex. For example, if the response variable was the number of accidents claimed for on a car insurance policy it might be the case that young men are more prone to accidents than young women but, as men get older, there is a steeper drop off in the number of accidents.

Let's now consider how we would construct the linear predictor $\alpha_i + \beta_i x$ for this model.

We start by summing the linear predictors for each of the three terms: age, sex and age.sex separately. We already know that the linear predictor for age alone is $\eta = \beta_0 + \beta_1 x$ and that the linear predictor for sex alone is $\eta = \alpha_i, i = 1,2$. We also need a linear predictor for the interaction effect $\eta = (\beta_0 + \beta_1 x).\alpha_i, \ i = 1,2$. Note that the dot notation here does not mean multiply. We have just written it in this format for now to indicate an interaction.

We add the three of these together:

$$\eta = \beta_0 + \beta_1 x + \alpha_i + (\beta_0 + \beta_1 x).\alpha_i$$

We could then use MLE on a set of past data to come up with estimates based on past data for each of the parameters. For example, these might be:

$$\beta_0 = 0.5 \ , \ \beta_1 = -0.05 \ , \ \alpha_i = \begin{cases} -0.05 & \text{if} \quad i = 1 \text{ (male)} \\ 0.05 & \text{if} \quad i = 2 \text{ (female)} \end{cases}$$

and interaction terms:

$$\beta_0 . \alpha_i = \begin{cases} 0.35 & \text{if} \quad i = 1 \text{ (male)} \\ 0.05 & \text{if} \quad i = 2 \text{ (female)} \end{cases} \text{ and } \beta_1 . \alpha_i = \begin{cases} -0.15 & \text{if} \quad i = 1 \text{ (male)} \\ -0.02 & \text{if} \quad i = 2 \text{ (female)} \end{cases}$$

Note that this approach, which is rather artificial, would involve estimating eight non-zero parameters. However, there is a more efficient way. We can combine the parameters β_0, α_i and $\beta_0 . \alpha_i$ (these terms are not attached to an x in the linear predictor). Similarly, we can combine the terms β_1 and $\beta_1 . \alpha_i$ (these terms are attached to an x in the linear predictor).

A linear predictor that gives identical results would be:

$$\eta = \alpha_i + \beta_i x$$

where:

$$\alpha_i = \begin{cases} 0.8 & \text{if} \quad i = 1 \text{ (male)} \\ 0.6 & \text{if} \quad i = 2 \text{ (female)} \end{cases} \text{ and } \beta_i = \begin{cases} -0.2 & \text{if} \quad i = 1 \text{ (male)} \\ -0.07 & \text{if} \quad i = 2 \text{ (female)} \end{cases}$$

You may want to spend a short while checking that for every combination of age and sex, the two linear predictors for age + sex + age.sex yield the same result.

The following table summarises the different models involving age (as a variable) and the factor of sex:

model	*linear predictor*	*R formula*
age	$\beta_0 + \beta_1 x_1$	`Y ~ X1`
sex	α_i	`Y ~ sex`
age + sex	$\alpha_i + \beta x_1$	`Y ~ X1 + sex`
age + sex + age.sex	$\alpha_i + \beta_i x_1$	`Y ~ X1+sex+X1:sex`
age*sex	$\alpha_i + \beta_i x_1$	`Y ~ X1 * sex`

Question

In UK motor insurance business, vehicle-rating group is also used as a factor. Vehicles are divided into twenty categories numbered 1 to 20, with group 20 including those vehicles that are most expensive to repair.

Suppose that we have a three-factor model specified as $age * (sex + vehicle\ group)$. Determine the linear predictor for a model of this type.

Solution

A helpful starting point is to consider the linear predictor for sex + vehicle group on its own. Summing the linear predictors for both of these main effects gives:

$$\eta = \alpha_i + \beta_j$$

Note that we don't attempt to simplify this to α_{ij} for example, as this notation is reserved for an interaction between sex and vehicle group, which we are not considering here.

Now we consider the linear predictor for age * (sex + vehicle group). Recall that this can also be written as:

age + (sex + vehicle group) + age . (sex + vehicle group)

We sum the linear predictors for each of these three components:

$$\eta = \left(\gamma_0 + \gamma_1 x\right) + \left(\alpha_i + \beta_j\right) + \left(\gamma_0 + \gamma_1 x\right).\left(\alpha_i + \beta_j\right)$$

Finally, we simplify by combining parameters:

$$\alpha_i + \beta_j + \gamma_i x + \delta_j x$$

Note that we have:

- combined γ_0, α_i and $\gamma_0 \cdot \alpha_i$ into a new α_i

- left β_j alone

- combined γ_1 and $\gamma_1 \cdot \alpha_i$ into γ_i

- renamed $\gamma_1 \cdot \beta_j$ as δ_j.

Note that in general, when we add a new main effect, we add $n-1$ parameters (or equivalently lose $n-1$ degrees of freedom), where n is the number of parameters that we would have used had the main effect stood on its own. In the case where the main effect is a factor, n is also the number of categories.

When we add an interactive factor, we add $(n-1)(m-1)$ parameters (or equivalently lose $(n-1)(m-1)$ degrees of freedom), where n and m are the number of parameters that we would have used had each of the main effects stood on their own. In the case where both these main effects are factors, n and m are also the number of possible categories for each factor.

4   Link functions

Recall that the link function connects the mean response to the linear predictor, $g(\mu) = \eta$, where $\mu = E(Y)$. Technically, it is necessary for the link function to be differentiable and invertible in order to fit a model.

An invertible function is one that is 'one-to-one', so that for any value of η there is a unique value of μ. We have seen already that it is important to be able to invert the link function in order to use the model to make predictions about the future.

Beyond these basic requirements, there are a number of functions which are appropriate for the distributions above. However, it is sensible to choose link functions to ensure our predicted response variables stay within sensible bounds and ideally minimise the residual variance.

In R we specify the link function by setting link equal to the appropriate function

  `identity, log, sqrt, logit, inverse, 1/mu^2,` **etc.**

For each distribution, the natural, or canonical, link function is defined $g(\mu) = \theta(\mu)$.

Remember that θ is the natural parameter for the exponential family form and that β is a function of the mean of the distribution μ.

If no link function is specified in R then these will be the default option. Hence the canonical link functions for the distributions given in Section 2 are:

normal	`identity`	$g(\mu) = \mu$
Poisson	`log`	$g(\mu) = \log \mu$
binomial	`logit`	$g(\mu) = \log\left(\dfrac{\mu}{1-\mu}\right)$
gamma	`inverse`	$g(\mu) = \dfrac{1}{\mu}$

Earlier, we showed that $\theta = -1/\mu$ for the gamma distribution. The minus sign is dropped in the canonical link function. This doesn't affect anything since constants will be absorbed into the parameters in the linear predictor.

The canonical link functions are given on Page 27 of the *Tables.*

These link functions work well for each of the above distributions, but it is not obligatory that they are used in each case. For example, we could use the identity link function in conjunction with the Poisson distribution, we could use the log link function for data which had a gamma distribution, and so on.

However, we need to consider the implications of the choice of the link function on the possible values for μ. For example, if the data have a Poisson distribution then μ must be

positive. **If we use the log link function, then** $\eta = \log(\mu)$ **and** $\mu = e^\eta$. **Thus,** μ **is guaranteed to be positive, whatever value (positive or negative) the linear predictor takes. The same is not true if we use the identity link function.**

Other link functions exist, and can be quite complex for specific modelling purposes. As a basis for actuarial applications, the above four functions are often sufficient.

Question

Determine the inverse of the link function $g(\mu) = \log\left(\dfrac{\mu}{1-\mu}\right)$ by setting it equal to η and

comment on why this might be an appropriate link function for the binomial distribution.

Solution

We used this inverse function in the actuarial exam pass rates example. It is:

$$\mu = \frac{e^\eta}{1+e^\eta} = \frac{1}{e^{-\eta}+1} = \left(1 + e^{-\eta}\right)^{-1}$$

It is an appropriate link function for the binomial distribution since it results in values of μ, the probability parameter, between 0 and 1.

5   Model fitting and comparison

The process of choosing a model also uses methods which are approximations, based on maximum likelihood theory, and this section outlines this process.

5.1   Obtaining the estimates

The parameters in a GLM are usually estimated using maximum likelihood estimation. The log-likelihood function, $\ell(y;\theta,\varphi) = \log(f_Y(y;\theta,\varphi))$, depends on the parameters in the linear predictor through the link function. Thus, maximum likelihood estimates of the parameters may be obtained by maximising ℓ with respect to the parameters in the linear predictor.

Notice that this depends on the invariance property from Chapter 7, that the MLE of a function is equal to the function of the MLE. We really want to find the MLE of the final parameter μ. However, because of the invariance property it is permissible to find the MLE of the linear predictor η, and translate this into the MLE for μ.

Question

Claim amounts for medical insurance claims for hamsters are believed to have an exponential distribution with mean μ_i:

$$f(y_i) = \frac{1}{\mu_i}e^{-y_i/\mu_i} = \exp\left\{-\frac{y_i}{\mu_i} - \log\mu_i\right\}$$

We have the following data for hamsters' medical claims, using the model above:

age x_i (months)	4	8	10	11	17
claim amount (£)	50	52	119	41	163

The insurer believes that a linear function of age affects the claim amount:

$$\eta_i = \alpha + \beta x_i$$

Using the canonical link function, write down (but do not try to solve) the equations satisfied by the maximum likelihood estimates for α and β, based on the above data.

Solution

The log of the likelihood function is:

$$\log L(\mu_i) = -\sum\frac{y_i}{\mu_i} - \sum\log\mu_i$$

The canonical link function for the exponential distribution is $g(\mu_i) = 1/\mu_i$. Recall that the link function connects the mean response to the linear predictor, $g(\mu_i) = \eta_i$.

Hence, we have:

$$\frac{1}{\mu_i} = \alpha + \beta x_i$$

Rearranging this gives:

$$\mu_i = \frac{1}{\alpha + \beta x_i}$$

This enables us to write the log-likelihood function in terms of α and β:

$$\log L(\alpha, \beta) = -\sum y_i(\alpha + \beta x_i) + \sum \log(\alpha + \beta x_i)$$

We can now differentiate this with respect to α and β:

$$\frac{\partial}{\partial \alpha} \log L(\alpha, \beta) = -\sum y_i + \sum \frac{1}{\alpha + \beta x_i}$$

$$\frac{\partial}{\partial \beta} \log L(\alpha, \beta) = -\sum x_i y_i + \sum \frac{x_i}{\alpha + \beta x_i}$$

So the equations satisfied by the MLEs of α and β are:

$$-\sum y_i + \sum \frac{1}{\hat{\alpha} + \hat{\beta} x_i} = 0$$

and: $$-\sum x_i y_i + \sum \frac{x_i}{\hat{\alpha} + \hat{\beta} x_i} = 0$$

Substituting in the given data values gives the following equations:

$$\frac{1}{\hat{\alpha} + 4\hat{\beta}} + \frac{1}{\hat{\alpha} + 8\hat{\beta}} + \frac{1}{\hat{\alpha} + 10\hat{\beta}} + \frac{1}{\hat{\alpha} + 11\hat{\beta}} + \frac{1}{\hat{\alpha} + 17\hat{\beta}} - 425 = 0$$

and: $$\frac{4}{\hat{\alpha} + 4\hat{\beta}} + \frac{8}{\hat{\alpha} + 8\hat{\beta}} + \frac{10}{\hat{\alpha} + 10\hat{\beta}} + \frac{11}{\hat{\alpha} + 11\hat{\beta}} + \frac{17}{\hat{\alpha} + 17\hat{\beta}} - 5{,}028 = 0$$

These are not particularly easy to solve without computer assistance. Using R to solve the equations gives $\hat{\alpha} = 0.160134$ and $\hat{\beta} = -0.000598$. We can then estimate the mean claim amounts for various ages using:

$$\hat{\mu}_i = \frac{1}{\hat{\alpha} + \hat{\beta} x_i}$$

Doing so gives estimates for the claim amounts of 6.34, 6.44, 6.49, 6.51 and 6.67 which are very poor indeed – so the model does not appear to be appropriate at all.

R **The R code to fit a generalised linear model to a multivariate data frame and assign it to the object model, is:**

```
model <- glm(Y ~ …, family = … (link = … ))
```

Then the estimates of the parameters and their approximate standard errors can be obtained by:

```
summary(model)
```

An example of a part of the summary output is shown below, which we can see is identical to the multivariate model output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.60859    2.43903   0.660    0.510
wt           1.62635    1.49068   1.091    0.275
disp        -0.03443    0.01536  -2.241    0.025 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.2 Significance of the parameters

As for the multiple linear regression model we can test whether each of the parameters is significantly different from zero. Generally speaking, it is not useful to include a covariate for which we cannot reject the hypothesis that $\beta = 0$.

Approximate standard errors of the parameters can be obtained using asymptotic maximum likelihood theory.

Recall from Chapter 7 that estimators are in general asymptotically normal and unbiased with variance equal to the Cramér-Rao lower bound:

$$\hat{\beta} \dot\sim N(\beta, CRLB) \qquad \text{large } n$$

Hence, when testing $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ we will have:

$$\frac{\hat{\beta} - 0}{s.e(\beta)} \dot\sim N(0,1)$$

For a two-tailed test the critical values are ± 1.96 and hence we will have a significant value if $\left|\hat{\beta}\right| > 1.96 \times s.e(\beta)$. We could approximate the 1.96 by 2 for simplicity.

As a rough guide, an indication of the significance of the parameters is given by twice the standard error. Thus, if:

$$\left|\hat{\beta}\right| \textbf{ > 2 standard error}(\hat{\beta}) \textbf{,}$$

the parameter is significant and should be retained in the model. Otherwise, the parameter is a candidate for being discarded.

It should be noted that in some cases, a parameter may appear to be unnecessary using this criterion, but the model without it does not provide a good enough fit to the data.

As with any model we need look at the whole situation and not just one aspect in isolation.

In R, the statistic and p-value for the tests of $H_0 : \beta = 0$ are given in the output of `summary(model)`.

So for the above printout we can see that the covariate `disp` **is significant, whereas the covariate** `wt` **is not. Therefore, we would remove** `wt` **from the model and see if it still provides a good enough fit to the data.**

Recall from Chapter 9 that a *p*-value is significant if it is less than 5% (*ie* 0.05).

5.3 The saturated model

To compare models we need a measure of the fit of a model to the data.

To do this we will compare our model to a model that is a perfect fit to the data, called the saturated model.

A saturated model is defined to be a model in which there are as many parameters as observations, so that the fitted values are equal to the observed values.

Key Information

In the saturated model we have $\hat{\mu}_i = y_i$, *ie* the fitted values are equal to the observed values.

Question

Claim amounts for medical insurance claims for hamsters are believed to have an exponential distribution with mean μ_i:

$$f(y_i) = \frac{1}{\mu_i} e^{-y_i/\mu_i} = \exp\left\{ -\frac{y_i}{\mu_i} - \log \mu_i \right\}$$

We are given the following data for hamsters' medical claims, using the model above:

age x_i (months)	4	8	10	11	17
claim amount (£)	50	52	119	41	163

The insurer believes that a model with 5 categories for age is sufficiently accurate:

$$\eta_i = \alpha_i \qquad i = 1,2,3,4,5$$

Using the canonical link function, show that the fitted values $(\hat{\mu}_i)$ are the observed claim amounts, y_i.

Solution

The log of the likelihood function is:

$$\log L(\mu_i) = -\sum \frac{y_i}{\mu_i} - \sum \log \mu_i$$

Setting the canonical link function for the exponential distribution to the linear predictor, $g(\mu_i) = 1/\mu_i = \eta_i$, gives:

$$\frac{1}{\mu_i} = \alpha_i \quad \Rightarrow \quad \mu_i = \frac{1}{\alpha_i}$$

This enables us to write the log-likelihood function in terms of α_i:

$$\log L(\alpha_i) = -\sum y_i \alpha_i + \sum \log(\alpha_i)$$

We can now differentiate this with respect to α_i:

$$\frac{\partial}{\partial \alpha_i} \log L(\alpha_i) = -y_i + \frac{1}{\alpha_i}$$

Note that all the terms other than those involving the specific α_i we are looking at disappear.

So the equations satisfied by the MLEs of α_i are:

$$-y_i + \frac{1}{\hat{\alpha}_i} = 0 \quad \Rightarrow \quad \hat{\alpha}_i = \frac{1}{y_i}$$

Hence, the fitted values are:

$$\hat{\mu}_i = \frac{1}{\hat{\alpha}_i} = y_i$$

The fitted values, $\hat{\mu}_i$, are equal to the observed values, y_i.

However, a model that fits the data perfectly is not necessarily a satisfactory model.

Suppose we were trying to model weight (*Y*) and height (*x*) using a linear model, $\beta_0 + \beta_1 x$. This is shown below on the left, whereas a graphical representation of the saturated model is on the right.

The saturated model *is* a perfect fit to the data, but since the fitted value is the observed value we cannot predict a value without first knowing what it is, *ie* it has no predictive ability for other heights.

However, the saturated model does provide an excellent benchmark against which to compare the fit of other models.

5.4　Scaled deviance (or likelihood ratio)

In order to assess the adequacy of a model for describing a set of data, we can compare the likelihood under this model with the likelihood under the saturated model.

The saturated model uses the same distribution and link function as the current model, but has as many parameters as there are data points. As such it fits the data perfectly. We can then compare our model to the saturated model to see how good a fit it is.

Suppose that L_S and L_M denote the likelihood functions of the saturated and current models, evaluated at their respective optimal parameter values. The likelihood ratio statistic is given by L_S / L_M. If the current model describes the data well then the value of L_M should be close to the value of L_S. If the model is poor then the value of L_M will be much smaller than the value of L_S and the likelihood ratio statistic will be large.

Alternatively, we could examine the natural log of the likelihood ratio statistic:

$$\log \frac{L_S}{L_M} = \ell_S - \ell_M$$

where $\ell_S = \log L_S$ and $\ell_M = \log L_M$.

The *scaled deviance* is defined as twice the difference between the log-likelihood of the model under consideration (known as the current model) and the saturated model.

Key Information

The scaled deviance for a particular model M is defined as:

$$SD_M = 2\left(\ell_S - \ell_M\right)$$

The deviance for the current model, D_M, is defined such that:

$$\text{scaled deviance} = \frac{D_M}{\varphi}$$

Remember that φ is a scale parameter, so it seems sensible that it should be used to connect the deviance with the scaled deviance. For a Poisson or exponential distribution, $\varphi = 1$ so the scaled deviance and the deviance are identical.

The smaller the deviance, the better the model from the point of view of model fit.

However, there will be a trade-off here. A model with many parameters will fit the data well. However a model with too many parameters will be difficult and complex to build, and will not necessarily lead to better prediction in the future. It is possible for models to be 'over-parameterised', *ie* factors are included that lead to a slightly, but not significantly, better fit. When choosing linear models, we will usually need to strike a balance between a model with too few parameters (which will not take account of factors that have a substantial impact on the data, and will therefore not be sensitive enough) and one with too many parameters (which will be too sensitive to factors that really do not have much effect on the results). We use the principle of parsimony here – that is we choose the simplest model that does the job.

This can be illustrated by considering the case when the data are normally distributed.

In this case, the log-likelihood for a sample of size *n* is:

$$\ell(y; \theta, \varphi) = \sum_{i=1}^{n} \log f_Y(y_i; \theta_i, \varphi)$$

$$= -\frac{n}{2}\log 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{(y_i - \theta_i)^2}{2\sigma^2}$$

The likelihood function for a random sample of size n is $f(y_1)f(y_2)...f(y_n)$. Note that when we take logs, we add the logs of the individual PDF terms to get the joint likelihood. Recall that for the normal distribution the natural parameter is just the mean, $\theta_i = \mu_i$.

For the saturated model, the parameter θ_i is estimated by y_i, and so the second term disappears. Thus, the scaled deviance (twice the difference between the values of the log-likelihood under the current and saturated models) is

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\theta}_i)^2}{\sigma^2}$$

where $\hat{\theta}_i$ is the fitted value for the current model.

The deviance (remembering that the scale parameter $\varphi = \sigma^2$), is the well-known residual sum of squares:

$$\sum_{i=1}^{n}(y_i - \hat{\theta}_i)^2$$

This is why the deviance is defined with a factor of two in it, so that for the normal model the deviance is equal to the residual sum of squares that we met in linear regression.

> **The residual deviance (*ie* the deviance after all the covariates have been included) is displayed as part of the results from** `summary(model)`**. For example:**
>
> ```
>         Null deviance: 43.86  on 31  degrees of freedom
>     Residual deviance: 21.40  on 29  degrees of freedom
>     AIC: 27.4
> ```
>
> **In R we can obtain a breakdown of how the deviance is reduced by each covariate added sequentially by using** `anova(model)`**. However, unlike for linear regression, this command does not automatically carry out a test.**

And recall that the smaller the residual (left over) deviance the better the fit of the model.

5.5 Using scaled deviance and Akaike's Information Criterion to choose between models

Adding more covariates will always improve the fit and thus decrease the deviance, however we need to determine whether adding a particular covariate leads to a significant decrease in the deviance.

For normally distributed data, the scaled deviance has a χ^2 distribution. Since the scale parameter for the normal $\varphi = \sigma^2$ must be estimated we would compare models by taking ratios of sum-of-squares and using *F*-tests (as in the analysis of variance for linear regression models).

We covered this in Section 2.3 from the previous chapter.

> **The code for comparing two normally distributed models,** `model1` **and** `model2`**, in R is:**
>
> ```
>     anova(model1, model2, test="F")
> ```

In the case of data which is not normally distributed, the scale parameter may be known (for example, for the Poisson distribution $\varphi = 1$), and the deviance is only asymptotically a χ^2 distribution. For these reasons, the common procedure is to compare two models by looking at the difference in the scaled deviance and comparing with a χ^2 distribution.

Since the distributions are only asymptotically normal the *F* test will not be very accurate. Hence, by simply comparing two approximate χ^2 distributions we will get a better result.

To be more precise, it's the absolute difference between the scaled deviances that is compared with χ^2.

Thus, if we want to decide if Model 2 (which has $p+q$ parameters and scaled deviance S_2) is a significant improvement over Model 1 (which has p parameters and scaled deviance S_1), we see if $S_1 - S_2$ is greater than the 5% value for the χ_q^2 distribution (see p169 of the Tables).

Recall that we subtract one degree of freedom for each extra parameter introduced. So it's the difference between p and $p+q$ that matters.

Since $\chi_p^2 + \chi_q^2 \sim \chi_{p+q}^2$ it makes sense to say that the difference in the scaled deviances has a χ_q^2 distribution.

What we are trying to do here is to decide whether the added complexity results in significant additional accuracy. If not, then it would be preferable to use the model with fewer parameters.

Alternatively, we could express this test in terms of the log-likelihood functions. If we let ℓ_p and ℓ_{p+q} denote the log-likelihoods of the models with p and $p+q$ parameters respectively, then the test statistic can be written as:

$$S_1 - S_2 = 2\left(\ell_S - \ell_p\right) - 2\left(\ell_S - \ell_{p+q}\right)$$
$$= -2\left(\ell_p - \ell_{p+q}\right)$$

This is the format given on page 23 of the *Tables* and will be used in Subject CS2 to compare Cox regression models.

Question

Explain why the test statistic will always be positive.

Solution

As we have mentioned before, adding more parameters will improve the fit of the model to the data. Therefore we would expect the value of the likelihood function to be larger for models with more parameters. Hence, $\ell_{p+q} > \ell_p$ and so the statistic will be positive.

The code for comparing these two (non-normally distributed) **models,** `model1` **and** `model2`, **in R is:**

```
anova(model1, model2, test="Chi")
```

A very important point is that this method of comparison can only be used for *nested* models. In other words, Model 1 must be a submodel of Model 2. Thus, we can compare two models for which the distribution of the data and the link function are the same, but the linear predictor has one extra parameter in Model 2. For example $\beta_0 + \beta_1 x$ and

$\beta_0 + \beta_1 x + \beta_2 x^2$ **. But we could not compare in this way if the distribution of the data or the link function are different, or, for example, when the linear predictors are $\beta_0 + \beta_1 x + \beta_2 x^2$ and $\beta_0 + \beta_3 \log x$. It should be clear that we *can* gauge the importance of factors by examining the scaled deviances, but we cannot use the testing procedure outlined above.**

In the first case the difference between the models is $\beta_2 x^2$ and so a significant difference between the models tells us that the quadratic term should be included. In the second case the difference between the models is $\beta_3 \log x - \beta_2 x^2$ and so a significant difference doesn't tell us *which* parameter is significant.

An alternative method of comparing models is to use Akaike's Information Criterion (AIC). Since the deviance will always decrease as more covariates are added to the model, there will always be a tendency to add more covariates. However this will increase the complexity of the model which is generally considered to be undesirable. To take account of the undesirability of increased complexity, computer packages will often quote the AIC, which is a penalised deviance:

$$\textbf{AIC} = \textbf{deviance} + \textbf{2} \times \textbf{number of parameters}$$

When comparing two models, the smaller the AIC, the better the fit. So if the change in deviance is more than twice the change in the number of parameters then it would give a smaller AIC.

This is approximately equivalent to checking whether the difference in deviance is greater than the 5% value of the χ^2 distribution for degrees of freedom between 5 and 15. However, it has the added advantage of being a simple way to compare GLMs without formal testing. This is similar to comparing the adjusted R^2 for multiple linear regression models in the previous chapter and hence is displayed as part of the output of a computer fitted GLM.

R | **In R the AIC is displayed as part of the results from** `summary(model).`

An example of this is given in the R box at the end of Section 5.4.

5.6   The process of selecting explanatory variables

As for multiple linear regression the process of selecting the optimal set of covariates for a GLM is not always easy. Again, we could use one of the two following approaches:

(1) Forward selection. Add the covariate that reduces the AIC the most or causes a significant decrease in the deviance. Continue in this way until adding any more causes the AIC to rise or does not lead to a significant improvement in the deviance. Note we should start with main effects before interaction terms and linear terms before polynomial.

Suppose we are modelling the number of claims on a motor insurance portfolio and we have data on the driver's age, sex and vehicle group. We would start with the null model (*ie* a single constant equal to the sample mean) then we would try each of single covariate models (linear function of age or the factors sex or vehicle group) to see which produces the most significant improvement in a χ^2 test or reduces the AIC the most. Suppose this was sex. Then we would try adding a second covariate (linear function of age or the factor vehicle group). Suppose this was age. Then we would try adding the third covariate (vehicle group). Then maybe we would try a quadratic function of the variable age (and maybe higher powers) or each of 2 term interactions (eg sex*age or sex*group or age*group). Finally we would try the 3 term interaction (*ie* sex*age*group).

(2) Backward selection. Start by adding all available covariates and interactions. Then remove covariates one by one starting with the least significant until the AIC reaches a minimum or there is no significant improvement in the deviance, and all the remaining covariates have a statistically significant impact on the response.

So with the last example we would start with the 3 term interaction sex*age*group and look at which parameter has the smallest *p*-value (in a test of it being zero) and remove that. We should see a significant improvement in a χ^2 test and the AIC should fall. Then we remove the next parameter with the smallest *p*-value and so on.

The Core Reading uses R to demonstrate this procedure. Whilst this will be covered in the CS1 PBOR, it's important to understand the process here.

Example

We demonstrate both of these methods in R using a binomial model on the `mtcars` dataset from the MASS package to determine whether a car has a V engine or an S engine (`vs`) using weight in 1000 lbs (`wt`) and engine displacement in cubic inches (`disp`) as covariates.

Forward selection

Starting with the null model:

```
model0 <- glm(vs ~ 1, data=mtcars, family=binomial)
```

The AIC of this model (which would be displayed using `summary(model0)`) is 45.86.

We have to choose whether we add `disp` or `wt` first. We try each and see which has the greatest improvement in the deviance.

```
model1 <- update(model0, ~.+ disp)

anova(model0, model1, test="Chi")
```

```
Model 1: vs ~ 1
Model 2: vs ~ disp
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        31     43.860
2        30     22.696  1   21.164 4.215e-06 ***
```

```
model2 <- update(model0, ~.+ wt)
```

```
        anova(model0, model2, test="Chi")
```

```
Model 1: vs ~ 1
Model 2: vs ~ wt
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        31     43.860
2        30     31.367  1   12.493 0.0004084 ***
```

So we can see that disp **has produced the more significant result – so we add that covariate first.**

The AIC of model 1 (adding disp) is 26.7 whereas the AIC of model 2 (adding wt) is 35.37. Therefore adding disp reduces the AIC more from model 0's value of 45.86.

Let us now see if adding wt **to** disp **produces a significant improvement:**

```
        model3 <- update(model1, ~.+ wt)

        anova(model1, model3, test="Chi")
```

```
Model 1: vs ~ disp
Model 2: vs ~ disp + wt
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        30     22.696
2        29     21.400  1   1.2954    0.255
```

This has not led to a significant improvement in the deviance so we would not add wt **(and therefore we definitely would not add an interaction term between** disp **and** wt**).**

The AIC of model 3 (adding wt) is 27.4 which is worse than model 1's AIC of 26.7. Therefore we would not add it.

Incidentally the AIC for models 0, 1, 2, 3 are 45.86, 26.7, 35.37 and 27.4. So using these would have given the same results (as Model 1 produces a smaller AIC than Model 2, and then Model 3 increases the AIC and so we would not have selected it).

Backward selection

Starting with all the possibilities:

```
        modelA <- glm(vs ~ wt * disp, data=mtcars, family=binomial)
```

The output is:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.308003   4.163350   0.554    0.579
wt          1.460010   1.689646   0.864    0.388
disp       -0.041214   0.035930  -1.147    0.251
wt:disp     0.001733   0.008023   0.216    0.829
```

None of these covariates is significant – so removing the interaction term wt:disp **which is the least significant.**

The parameter of the interaction term has the highest *p*-value of 0.829 and so is most likely to be zero.

```
        modelB <- update(model1, ~.-wt:disp)
```

```
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.60859     2.43903    0.660     0.510
wt           1.62635     1.49068    1.091     0.275
disp        -0.03443     0.01536   -2.241     0.025 *
```

The AIC has fallen from 29.361 to 27.4.

Alternatively, carrying out a χ^2 test using `anova(modelA, modelB, test="Chi")` would show that there is no significant difference between the models (*p*-value of 0.8417) and therefore we are correct to remove the interaction term between `wt` and `disp`.

The `wt` term is not significant so removing that:

```
        modelC <- update(modelB, ~.-wt)
```

```
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  4.137827    1.389354    2.978   0.00290 **
disp        -0.021600    0.007131   -3.029   0.00245 **
```

Both of these coefficients are significant and the AIC has fallen from 27.4 to 26.696.

Alternatively, carrying out a χ^2 test using `anova(modelB, modelC, test="Chi")` would show that there is no significant difference between the models (*p*-value of 0.255) and therefore we are correct to remove the `wt` covariate.

We would stop at this model. Had we removed the `disp` term (to give the null model) the AIC increases to 45.86.

Alternatively, carrying out a χ^2 test between these two models would show a very significant difference (*p*-value of less than 0.001) and therefore we should not remove the `disp` covariate.

We can see that both forward and backward selection lead to the same model being chosen.

5.7 Estimating the response variable

Once we have obtained our model and its estimates, we are then able to calculate the value of the linear predictor, η, and by using the inverse of the link function we can calculate our estimate of the response variable $\hat{\mu} = g^{-1}(\hat{\eta})$.

Substituting the estimated parameters into the linear predictor gives the estimated value of the linear predictor for different individuals. Now the link function links the linear predictor to the mean of the distribution. Hence we can obtain an estimate for the mean of the distribution of Y for that individual.

Let's now return to the Core Reading example on page 45.

Suppose, we wish to estimate the probability of having a V engine for a car with weight 2100 lbs and displacement 180 cubic inches.

Using our linear predictor $\beta_0 + \beta_1 \times \text{disp}$ (ie vs ~ disp) we obtained estimates of $\hat{\beta}_0 = 4.137827$ and $\hat{\beta}_1 = -0.021600$.

These coefficients displayed as part of the summary output of Model C in the example above.

Hence, for displacement 180 we have $\hat{\eta} = 4.137827 - 0.021600 \times 180 = 0.24983$. We did not specify the link function so we shall use the canonical binomial link function which is the logit function.

$$0.24983 = \log\left(\frac{\hat{\mu}}{1-\hat{\mu}}\right) \quad \Rightarrow \quad \hat{\mu} = \frac{e^{0.24983}}{1+e^{0.24983}} = 0.562$$

Recall that the mean for a binomial model is the probability. So the probability of having a V engine for a car with weight 2,100 lbs and displacement 180 cubic inches is 56.2%.

Note that because we removed the weight covariate, the figure 2,100 does not enter the calculation.

In R we can obtain this as follows:

```
newdata <-data.frame(disp=180)

predict(model,newdata,type="response")
```

---

6   Residuals analysis and assessment of model fit

Once a possible model has been found it should be checked by looking at the residuals. The residuals are based on the differences between the observed responses, y, and the fitted responses, $\hat{\mu}$. The fitted responses are obtained by applying the inverse of the link function to the linear predictor with the fitted values of the parameters.

We looked at how we could obtain predicted responses values in the previous section. The fitted values are the predicted Y values for the observed data set, x.

The R code for obtaining the fitted values of a GLM is:

```
fitted(model)
```

For example, in the actuarial pass rates model detailed on page 6, we could calculate from the model what the pass rate ought to be for students who have attended tutorials, submitted three assignments and scored 60% on the mock exam.

The difference between this theoretical pass rate and the actual pass rate observed for students who match the criteria exactly will give us the residuals.

Question

Draw up a table showing the differences between the actual and expected values of the truancy rates in the example on page 9.

Solution

Recall that the expected number of unexplained absences in a year were modelled by:

$$\eta = \alpha_i + \beta_j + \gamma x \qquad \text{where } x = \text{age}, \text{ and } \alpha \text{ and } \beta \text{ are as follows:}$$

$$\alpha_{WC} = -2.64 \qquad \alpha_{OC} = -1.14 \qquad \beta_M = -3.26 \qquad \beta_F = -3.54 \qquad \gamma = 0.64$$

where WC = Within catchment, OC = Outside catchment, M = Male, F = Female.

This gives expected values of:

		Age last birthday			
		8	10	12	14
Within catchment area	Male	0.46	1.65	5.93	21.33
	Female	0.35	1.25	4.48	16.12
Outside catchment area	Male	2.05	7.39	26.58	95.58
	Female	1.55	5.58	20.09	72.24

So the differences between the actual values (given on page 9) and expected values are:

		Age last birthday			
		8	10	12	14
Within catchment area	Male	1.34	0.35	0.37	−7.23
	Female	0.15	0.35	0.52	0.08
Outside catchment area	Male	0.05	0.11	−1.08	−23.58
	Female	1.25	0.62	−0.49	−4.04

The procedure here is a natural extension of the way we calculated residuals for linear regression models covered in the previous two chapters. However, because of the different distributions used we need to transform these 'raw' residuals so we are able to interpret them meaningfully.

There are two kinds of residuals: Pearson and deviance.

6.1 Pearson residuals

The Pearson residuals are defined as:

$$\frac{y - \hat{\mu}}{\sqrt{\text{var}(\hat{\mu})}}$$

The $\text{var}(\hat{\mu})$ in the denominator refers to the variance of the response distribution, $\text{var}(Y)$ using the fitted values, $\hat{\mu}$, in the formula. For example, since the variance of the exponential distribution is μ^2 we would have $\text{var}(\hat{\mu}) = \hat{\mu}^2$.

The Pearson residual, which is often used for normally distributed data, has the disadvantage that its distribution is often skewed for non-normal data. This makes the interpretation of residual plots difficult.

The R code for obtaining the Pearson residuals is:

```
residuals(model, type="pearson")
```

If the data comes from a normal distribution then the Pearson residuals will have a standard normal distribution. By comparing these residuals to a standard normal (*eg* by using a Q-Q plot) we can determine whether the model is a good fit. However, for non-normal data the Pearson residuals will not have a standard normal distribution and won't even be symmetrical. This makes it difficult to determine whether the model is a good fit. Hence we will need to use a different type of residual.

6.2   Deviance residuals

Deviance residuals are defined as the product of the sign of $y - \hat{\mu}$ and the square root of the contribution of y to the scaled deviance. Thus, the deviance residual is:

$$sign(y - \hat{\mu})d_i$$

where the scaled deviance is $\sum d_i^2$.

Recall that:

$$sign(x) = \begin{cases} +1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Deviance residuals are usually more likely to be symmetrically distributed and to have approximately normal distributions, and are preferred for actuarial applications.

The R code for obtaining the deviance residuals is:

```
residuals(model)
```

We can see that deviance residuals are more likely to be symmetrically distributed by considering the following result: If $\{x_i\}$ is a set of independent normal random variables then $Y = \sum X_i^2$ will have a χ^2 distribution. Therefore, since $\sum d_i^2$ (*ie* the scaled deviance) is approximately χ^2, it follows that d_i (and also the deviance residual) is likely to be approximately normal. We will cover this in more detail in Subject CM2.

Note that for normally distributed data, the Pearson and deviance residuals are identical.

Question

Show that, for normally distributed data, the Pearson and deviance residuals are identical.

Solution

If $Y_i \sim N(\mu_i, \sigma^2)$, then from Section 6.1 the Pearson residuals are:

$$\frac{y_i - \hat{\mu}_i}{\sqrt{var(\hat{\mu}_i)}} = \frac{y_i - \hat{\mu}_i}{\sigma}$$

In Section 5.4 we saw that the scaled deviance was:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2} = \sum_{i=1}^{n} d_i^2$$

So the deviance residuals are given by:

$$sign(y_i - \mu_i)d_i = sign(y_i - \mu_i)\left|\frac{y_i - \mu_i}{\sigma}\right| = \frac{y_i - \mu_i}{\sigma}$$

Hence the Pearson residuals and the deviance residuals are the same.

6.3   Using residual plots to check the fit

The assumptions of a GLM require that the residuals should show no patterns. The presence of a pattern implies that something has been missed in the relationship between the predictors and the response. If this is the case, other model specifications should be tried.

So in addition to the residuals being symmetrical we would expect no connection between the residuals and the explanatory covariates. Rather than plotting the residuals against each of the covariates we could just see if there is a pattern when plotted against the fitted values.

For our model above (on the `mtcars` dataset), **a plot of the residuals against the fitted values is as follows:**

There does appear to be some pattern here and there are three named outliers.

The line shows the trend, ideally this should be horizontal which indicates no pattern. Also the residuals should be small; if R names them then that means they are considered to be too large.

We could also plot a histogram of the residuals, or another similar diagnostic plot should also be examined in order to assess whether the distributional assumptions are justified.

Whilst a Q-Q plot is produced as an output of the GLMs process, there is some controversy over whether this is appropriate for non-normal distributions such as the binomial distribution in the Core Reading example above. Hence it has not been included in the Core Reading.

6.4   Acceptability of a fitted model

In addition to comparing models, statistical tests can be used to determine the acceptability of a particular model, once fitted. Pearson's chi-square test and the likelihood ratio test are typically used. These are described in Chapter 9, Sections 8 and 2 respectively. The tests for overall fit involve comparing the scaled deviance of the fitted model with the scaled deviance of the null model (with no covariates). The extent by which the fitted model reduces the scaled deviance (per additional parameter estimated) is a measure of how much the fitted model is an improvement on the null model.

Considerable flexibility in the interpretation of the tests based on statistical inference theory is sometimes necessary in order to arrive at a suitable model. Thus, the interpretation of deviances, residuals and significance of parameters given above should be viewed as useful guides in selecting a model, rather than strict rules which must be adhered to.

The chapter summary starts on the next page so that you can keep
all the chapter summaries together for revision purposes.

Chapter 12 Summary

Exponential family

There is a wide variety of distributions (normal, Poisson, binomial, gamma and exponential) that have a common form, called the exponential family.

If the distribution of Y is a member of the exponential family then the density function of Y can be written in the form:

$$f_Y(y;\theta,\varphi) = \exp\left[\frac{(y\theta - b(\theta))}{a(\varphi)} + c(y,\varphi)\right]$$

where θ is the natural parameter which is a function of the mean $\mu = E(Y)$ only of the distribution, and φ is a scale parameter. Where the distribution has two parameters (such as the normal, gamma and binomial distributions), we can take φ to be the parameter other than the mean. Where the distribution has one parameter (such as the Poisson and exponential distributions), we can take $\varphi = 1$. However, the parameterisations are not unique.

Mean, variance and variance function

$$E(Y) = b'(\theta) \qquad\qquad \text{var}(Y) = a(\varphi)b''(\theta) \qquad\qquad V(\mu) = b''(\theta)$$

The variance function is a function of the mean $\mu = E(Y)$ and gives a measure of how $\text{var}(Y)$ relates to μ.

Generalised linear models (GLMs)

A GLM takes multiple regression one step further by allowing the data to be non-normally distributed. Instead, we can use any of the distributions in the exponential family.

A GLM consists of three components:

1)   a distribution for the data (Poisson, exponential, gamma, normal or binomial)

2)   a linear predictor (a function of the covariates that is linear in the parameters)

3)   a link function (that links the mean of the response variable to the linear predictor).

Maximum likelihood estimation can be used to estimate the values of the parameters in the linear predictor.

Link functions

For each underlying distribution there is one link function that appears more natural to use than any other, usually because it will result in values for μ that are appropriate to the distribution under consideration. This is called the canonical link function, which means the 'accepted' link function. The canonical link functions are given on Page 27 of the *Tables*. They are equivalent to the natural parameter θ from the exponential family formulation of the PDF.

Covariates

A variable (*eg* age) is a type of covariate whose real numerical value enters the linear predictor directly, and a factor (*eg* sex) is a type of covariate that takes categorical values.

Linear predictors

Linear predictors are functions of the covariates. They are linear in the parameters and not necessarily in the covariates.

The simplest linear predictor is that for the constant model: $\eta = \alpha$, which is used if it is thought that the mean of the response variable is the same for all cases.

An interaction term is used in the predictor when two covariates are believed not to be independent. In other words, the effect of one covariate (*eg* the age of an individual) is thought to depend on the value of another covariate (*eg* whether the sex of an individual is male or female).

The dot '.' notation is used to indicate an interaction, *eg* age.sex is the interactive term between age and sex.

The star '*' notation is used to indicate the main effects as well as the interaction, *eg*:

> age*sex = age + sex + age.sex

An interaction (dot) term never appears on its own.

Saturated model

The model that provides the perfect fit to the data is called the saturated model. The saturated model has as many parameters as data points. The fitted values $\hat{\mu}_i$ are equal to the observed values y_i. The saturated model is not useful from a predictive point of view, however it is a good benchmark against which to compare the fit of other models via the scaled deviance.

Scaled deviance

The scaled deviance (or likelihood ratio) is used to compare the fit of the saturated model with the fit of another model. The scaled deviance of Model 1 is defined as:

$$SD_1 = 2\left(\ln L_S - \ln L_1\right)$$

where L_S is the likelihood of the saturated model.

The poorer the fit of Model 1, the bigger the scaled deviance will be.

Comparing models

Where the data are normally distributed, it can be shown that, for two *nested* models, Models 1 and 2 where Model 1 has p parameters and Model 2 has $p+q$ parameters:

$$SD_1 - SD_2 \sim \chi_q^2$$

For other distributions, the difference in the scaled deviances has an approximate (asymptotic) chi-square distribution with q degrees of freedom.

Alternatively, we can compare the reduction in the AIC of the two models.

The process of selecting explanatory variables

(1) Forward selection. Add the covariate that reduces the AIC the most or causes a significant decrease in the deviance. Continue in this way until adding any more causes the AIC to rise or does not lead to a significant improvement in the deviance. It is usual to consider main effects before interaction terms and linear terms before polynomials.

(2) Backward selection. Start by adding all available covariates and interactions. Then remove covariates one by one starting with the least significant until the AIC reaches a maximum or there is no significant improvement in the deviance, and all the remaining covariates have a statistically significant impact on the response.

Rules for determining the number of parameters in a model

The constant model has 1 parameter.

A model consisting of one main effect that is a variable (*eg* age) has two parameters (*eg* β_0 and β_1).

A model consisting of one main effect that is a factor (*eg* sex) has as many parameters as there are categories (*eg* α_i, i = 1 (male) and i = 2 (female)).

Rules for determining the number of parameters in a model (continued)

When a new main effect is added to a model (*eg* age + sex), we add on $n-1$ parameters where n is the number of parameters if the main effect were on its own (*eg* for age + sex, the number of parameters is 2 + (2 − 1) = 3).

When an interactive effect (a dot term) is added to a model (*eg* age + sex + age.sex), we add on $(m-1)(n-1)$ parameters for the interactive effect (*eg* for age + sex + age.sex, the number of parameters is 2 + (2 − 1) + (2 − 1)(2 − 1) = 4).

A model consisting of a star term only (*eg* age*sex) has mn parameters where *m* and *n* are the number of parameters if the main effects were on their own (*eg* for age*sex, the number of parameters is $2 \times 2 = 4$).

Residuals

A residual is a measure of the difference between the observed values y_i and the fitted values $\hat{\mu}_i$. Two commonly used residuals for GLMs are the Pearson residual and the deviance residual.

Pearson residual

This is $\dfrac{y - \hat{\mu}}{\sqrt{\text{var}(\hat{\mu})}}$ where $\text{var}(\hat{\mu})$ is $\text{var}(Y)$ with any values of μ replaced by their fitted values $\hat{\mu}$.

The Pearson residual, which is often used for normally distributed data, has the disadvantage that its distribution is often skewed for non-normal data. This makes the interpretation of residuals plots difficult.

Deviance residual

This is $sign(y - \hat{\mu})d_i$ where $\sum d_i^2$ is the scaled deviance of the model.

Deviance residuals are usually more likely to be symmetrically distributed and to have approximately normal distributions, and are preferred for actuarial applications.

For normally distributed data, the Pearson and deviance residuals are identical.

Testing whether a parameter is significantly different from zero

As a general rule, we can conclude that a parameter is significantly different from zero if it is at least twice as big in absolute terms as its standard error, *ie* if:

$$|\beta| > 2\,s.e(\beta)$$

Chapter 12 Practice Questions

12.1   Explain why the link function $g(\mu) = \log \mu$ is appropriate for the Poisson distribution by considering the range of values that it results in μ_i taking.

12.2   Explain the difference between the two types of covariate: a variable and a factor.

12.3   (i)   A random variable *Y* has density of exponential family form:

$$f(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

State the mean and variance of *Y* in terms of $b(\theta)$ and its derivatives and $a(\phi)$.   [1]

(ii)   (a)   Show that an exponentially distributed random variable with mean μ has a density that can be written in the above form.

(b)   Determine the natural parameter and the variance function.   [3]
[Total 4]

12.4   An insurer wishes to use a generalised linear model to analyse the claim numbers on its motor portfolio. It has collected the following data on claim numbers y_i, $i = 1, 2, ..., 35$ from three different classes of policy:

Class I	1	2	0	2	1	0	0	2	2	1
Class II	1	0	1	1	0					
Class III	0	0	0	0	0	1	0	1	0	0
	1	0	1	0	0	0	0	0	0	0

For these data values:

$$\sum_{i=1}^{10} y_i = 11 \qquad\qquad \sum_{i=11}^{15} y_i = 3 \qquad\qquad \sum_{i=16}^{35} y_i = 4$$

The company wishes to use a Poisson model to analyse these data.

(i)   Show that the Poisson distribution is a member of the exponential family of distributions.   [2]

(ii)   The insurer decides to use a model (Model A) for which:

$$\log \mu_i = \begin{cases} \alpha & i = 1, 2, \ldots, 10 \\ \beta & i = 11, 12, \ldots, 15 \\ \gamma & i = 16, 17, \ldots, 35 \end{cases}$$

where μ_i is the mean of the relevant Poisson distribution. Derive the likelihood function for this model, and hence find the maximum likelihood estimates for α, β and γ.   [4]

(iii)   The insurer now analyses the simpler model $\log \mu_i = \alpha$, for all policies. Calculate the maximum likelihood estimate for α under this model (Model B).   [2]

(iv)   Show that the scaled deviance for Model A is 24.93, and calculate the scaled deviance for Model B.   [5]

It can be assumed that $f(y) = y \log y$ is equal to zero when $y = 0$.

(v)   Compare Model A directly with Model B, by calculating an appropriate test statistic.   [2]

[Total 15]

12.5   In the context of Generalised Linear Models, consider the exponential distribution with density function $f(x)$, where:

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \qquad (x > 0).$$

(i)   Show that $f(x)$ can be written in the form of the exponential family of distributions.   [1]

(ii)   Show that the canonical link function, θ, is given by $\theta = -\dfrac{1}{\mu}$.   [1]

(iii)   Determine the variance function and the dispersion parameter.   [3]

[Total 5]

12.6   The random variable Z_i has a binomial distribution with parameters n and μ_i, where $0 < \mu_i < 1$. A second random variable, Y_i, is defined as $Y_i = Z_i / n$.

(i)   Show that Y_i is a member of the exponential family, stating clearly the natural and scale parameters and their functions $a(\varphi)$, $b(\theta)$ and $c(y, \varphi)$.   [4]

(ii)   Determine the variance function of Y_i.   [2]

[Total 6]

12.7   A statistical distribution is said to be a member of the exponential family if its probability function or probability density function can be expressed in the form:

$$f_Y(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

(i)   Show that the mean of such a distribution is $b'(\theta)$ and derive the corresponding formula for the variance by differentiating the following expression with respect to θ:

$$\int_y f(y)\,dy = 1$$ [4]

(ii)   Use this method to determine formulae for the mean and variance of the gamma distribution with density function

$$f(x) = \frac{\alpha^\alpha}{\mu^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x/\mu} \qquad (x > 0)$$ [3]

[Total 7]

12.8   Independent claim amounts $Y_1, Y_2, \ldots, Y_n$ are modelled as exponential distributions with mean

$\underset{\text{Exam style}}{}$ μ_i.

The fitted values for a particular model are denoted by $\hat{\mu}_i$. Derive an expression for the scaled deviance.   [5]

12.9   A small insurer wishes to model its claim costs for motor insurance using a simple generalised

$\underset{\text{Exam style}}{}$ linear model based on the three factors:

$$YO_i = \begin{cases} i = 1 & \text{for 'young' drivers} \\ i = 0 & \text{for 'old' drivers} \end{cases}$$

$$FS_j = \begin{cases} j = 1 & \text{for 'fast' cars} \\ j = 0 & \text{for 'slow' cars} \end{cases}$$

$$TC_k = \begin{cases} k = 1 & \text{for 'town' areas} \\ k = 0 & \text{for 'country' areas} \end{cases}$$

The insurer is considering three possible models for the linear predictor:

Model 1: $YO + FS + TC$

Model 2: $YO + FS + YO.FS + TC$

Model 3: $YO * FS * TC$

(i)   Write each of these models in parameterised form, stating how many non-zero parameter values are present in each model.   [6]

(ii)   Explain why Model 1 might not be appropriate and why the insurer may wish to avoid using Model 3.   [2]

(iii)   The student fitting the models has said 'We are assuming a normal error structure and we are using the canonical link function.' Explain what this means.   [3]

(iv)   The table below shows the student's calculated values of the scaled deviance for these three models and the constant model.

Model	Scaled Deviance	Degrees of freedom
1	50	7
$YO + FS + TC$	10	
$YO + FS + YO.FS + TC$	5	
$YO * FS * TC$	0	

Complete the table by filling in the missing entries in the degrees of freedom column and carry out the calculations necessary to determine which model would be the most appropriate.   [5]

[Total 16]

12.10   The following study was carried out into the mortality of leukaemia sufferers. A white blood cell count was taken from each of 17 patients and their survival times were recorded.

Exam style

Suppose that Y_i represents the survival time (in weeks) of the ith patient and x_i represents the logarithm (to the base 10) of the ith patient's initial white blood cell count ($i = 1, 2, \ldots, 17$).

The response variables Y_i are assumed to be exponentially distributed. A possible specification for $E(Y_i)$ is $E(Y_i) = \exp(\alpha + \beta x_i)$. This will ensure that $E(Y_i)$ is non-negative for all values of x_i.

(i)   Write down the natural link function associated with the linear predictor $\eta_i = \alpha + \beta x_i$. [2]

(ii)   Use this link function and linear predictor to derive the equations that must be solved in order to obtain the maximum likelihood estimates of α and β. [4]

(iii)   Given that the maximum likelihood estimate of α derived from the experimental data is $\hat{\alpha} = 8.477$ and $se(\hat{\alpha}) = 1.655$, construct an approximate 95% confidence interval for α and interpret this result. [2]

(iv)   The following two models are now to be compared:

Model 1: $E(Y_i) = \alpha$

Model 2: $E(Y_i) = \alpha + \beta x_i$

The scaled deviance for Model 1 is found to be 26.282 and the scaled deviance for Model 2 is 19.457. Test the null hypothesis that $\beta = 0$ against the alternative hypothesis that $\beta \neq 0$ stating any conclusions clearly. [3]

[Total 11]

Chapter 12 Solutions

12.1 When we set the link function $g(\mu) = \log \mu$ equal to the linear predictor η and then invert to

make μ the subject, we get $\mu = e^{\eta}$. This results in positive values only for μ, which is sensible

for a $Poi(\mu)$ distribution where μ is defined to be greater than 0.

12.2 A variable is a type of covariate (*eg* age) whose actual numerical value enters the linear predictor directly, and a factor is a type of covariate (*eg* sex) that takes categorical values.

12.3 *This question is taken from Subject 106, April 2003, Question 3.*

(i)   *Mean and variance*

We have:

$$E[Y] = b'(\theta) \qquad \text{var}[Y] = a(\phi)b''(\theta) \qquad\qquad\qquad [1]$$

(ii)(a)   *Exponential form*

The PDF of the exponential distribution with mean μ is:

$$f(y) = \frac{1}{\mu}\exp\left\{-\frac{y}{\mu}\right\}$$

This can be written as an exponential:

$$f(y) = \exp\left\{\ln\frac{1}{\mu} - \frac{y}{\mu}\right\} \qquad\qquad\qquad [\tfrac{1}{2}]$$

Comparing this to the standard form given in part (i), we can define:

$$\theta = -\frac{1}{\mu}, \quad a(\phi) = 1, \quad b(\theta) = -\ln\frac{1}{\mu} = -\ln(-\theta), \quad c(y,\phi) = 0 \qquad\qquad [1]$$

(ii)(b)   *Natural parameter and variance function*

The natural parameter is θ, so here the natural parameter is:

$$-\frac{1}{\mu} \qquad\qquad\qquad [\tfrac{1}{2}]$$

The variance function is (by definition) $b''(\theta)$, so here we find:

$$b'(\theta) = -\frac{1}{\theta} \qquad\qquad\qquad b''(\theta) = \frac{1}{\theta^2} = \mu^2 \qquad\qquad\qquad [1]$$

$$[\text{Total } 3]$$

12.4 (i) *Exponential family*

For the Poisson distribution, we have:

$$f(y) = e^{-\mu} \mu^y / y!$$

We wish to write this in the form:

$$g(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

So, rearranging the Poisson formula:

$$f(y) = \exp\left[\frac{y \log \mu - \mu}{1} - \log y! \right] \tag{1}$$

We can see that this has the correct form if we write:

$$\theta = \log \mu \qquad b(\theta) = \mu = e^{\theta} \qquad a(\phi) = \phi = 1 \qquad c(y, \phi) = -\log y! \tag{1}$$

(ii) *Maximum likelihood estimates*

Using the rearranged form for the Poisson distribution from part (i), we see that the log of the likelihood function can be written:

$$\log L(\mu_I, \mu_{II}, \mu_{III}) = \sum y_i \log \mu_i - \sum \mu_i - \sum \log y_i! \tag{*}$$

This now becomes, for Model A:

$$\log L = \alpha \sum_{i=1}^{10} y_i + \beta \sum_{i=11}^{15} y_i + \gamma \sum_{i=16}^{35} y_i - 10e^{\alpha} - 5e^{\beta} - 20e^{\gamma} - \sum_{i=1}^{35} \log y_i! \tag{1}$$

$$= 11\alpha + 3\beta + 4\gamma - 10e^{\alpha} - 5e^{\beta} - 20e^{\gamma} - \sum_{i=1}^{35} \log y_i! \tag{**}$$

Differentiating this log-likelihood function in turn with respect to α, β and γ, we get:

$$\frac{\partial}{\partial \alpha} \log L = 11 - 10e^{\alpha} \tag{½}$$

$$\frac{\partial}{\partial \beta} \log L = 3 - 5e^{\beta} \tag{½}$$

and:

$$\frac{\partial}{\partial \gamma} \log L = 4 - 20e^{\gamma} \tag{½}$$

Setting each of these expressions equal to zero in turn, we find that:

$$\hat{\alpha} = \log 1.1 = 0.09531$$ [½]

$$\hat{\beta} = \log 0.6 = -0.51083$$ [½]

and: $\hat{\gamma} = \log 0.2 = -1.60944$ [½]

These are our maximum likelihood estimates for α, β and γ.

(iii)   *Simpler model*

In this case the log-likelihood function reduces to:

$$\log L = \alpha \sum_{i=1}^{35} y_i - 35e^{\alpha} - \sum_{i=1}^{35} \log y_i! = 18\alpha - 35e^{\alpha} - \sum_{i=1}^{35} \log y_i!$$
(***)
[1]

Differentiating this with respect to α, and setting the result equal to zero, we find that:

$$18 - 35e^{\hat{\alpha}} = 0 \quad \Rightarrow \quad \hat{\alpha} = \log\left(\frac{18}{35}\right) = -0.66498$$ [1]

(iv)   *Scaled deviance for Model A and Model B*

The scaled deviance for Model A is given by:

$$\text{Scaled Deviance} = 2(\log L_S - \log L_A)$$

where $\log L_S$ is the value of the log likelihood function for the saturated model, and $\log L_A$ is the value of the log-likelihood function for Model A.

For the saturated model, we replace the μ_i's with the y_i's in Equation (*) – as it fits the observed data perfectly – so the expected results are the observed results. So:

$$\log L_S = \sum y_i \log y_i - \sum y_i - \sum \log y_i!$$
$$= 4 \times 2\log 2 - 18 - 4\log 2 = 4\log 2 - 18 = -15.2274$$ [1]

We use the hint in the question here. $y_i \log y_i$ is zero when $y = 0$, and also when $y = 1$. So the only contribution to the first term is when $y = 2$, giving 4 lots of $2\log 2$.

For the log likelihood for Model A, we replace the parameters α, β and γ with their estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ in Equation (**):

$$\log L_A = 11\hat{\alpha} + 3\hat{\beta} + 4\hat{\gamma} - 10e^{\hat{\alpha}} - 5e^{\hat{\beta}} - 20e^{\hat{\gamma}} - \sum_{i=1}^{35} \log y_i!$$

$$= 11\log 1.1 + 3\log 0.6 + 4\log 0.2 - 11 - 3 - 4 - 4\log 2 = -27.6944$$ [1]

The corresponding value for $\log L_A$ without the final term is –24.9218.

So the scaled deviance is twice the difference in the log likelihoods:

$$\text{Scaled Deviance} = 2(\log L_S - \log L_A) = 2\big((-15.2274) - (-27.6944)\big) = 24.93 \qquad [1]$$

as required.

We now repeat the process for Model B.

Using Equation (***), the log likelihood for Model B is:

$$\log L_B = 18\hat{\alpha} - 35e^{\hat{\alpha}} - \sum_{i=1}^{35} \log y_i !$$

$$= 18\log\left(\frac{18}{35}\right) - 18 - 4\log 2 = -32.7422 \qquad [1]$$

The value without the final term is –29.9696.

The scaled deviance is again twice the difference in the log likelihoods:

$$\text{Scaled Deviance} = 2(\log L_S - \log L_B) = 2\big((-15.2274) - (-32.7422)\big) = 35.03 \qquad [1]$$

(v)   **Comparing A with B**

We can use the chi-squared distribution to compare Model A with Model B. We calculate the difference in the scaled deviances (which is just $2(\log L_A - \log L_B)$):

$$35.03 - 24.93 = 10.10 \qquad [1]$$

This should have a chi-squared distribution with $3-1=2$ degrees of freedom, which has a critical value at the upper 5% level of 5.991. Our value is significant here, since $10.10 > 5.991$, so this suggests that Model A is a significant improvement over Model B. We prefer Model A here.   [1]

12.5   *This is Subject 106, September 2000, Question 2.*

(i)   **Exponential family**

We need to express the density function in the form $\exp\left\{\dfrac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$.

We can write the density function as:

$$f(y) = \exp\left\{ -\frac{y}{\mu} - \log \mu \right\}$$

So if we take:

$$\theta = -1/\mu \qquad b(\theta) = \log \mu = -\log(-\theta)$$

$$\phi = 1 \qquad a(\phi) = 1 \qquad c(y,\phi) = 0$$

then the density function will have the appropriate form.   [1]

(ii)   *Canonical link function*

We see from part (i) that $\theta = -1/\mu$.   [1]

(iii)   *Variance function and dispersion parameter*

The variance function is $b''(\theta)$. Differentiating $b(\theta)$ twice, we find that $b''(\theta) = 1/\theta^2 = \mu^2$. So the variance function is μ^2.   [2]

The dispersion parameter or scale parameter is $\phi = 1$.   [1]

12.6   (i)   **Show Y_i is member of the exponential family**

The PF of Z_i is:

$$f(z_i) = \binom{n}{z_i} \mu_i^{z_i} (1 - \mu_i)^{n - z_i}$$

The PF of Y_i can be obtained by replacing z_i with ny_i:

$$f(y_i) = \binom{n}{ny_i} \mu_i^{ny_i} (1 - \mu_i)^{n - ny_i}$$ [½]

This can be written as:

$$f(y_i) = \exp\left\{ \ln\binom{n}{ny_i} + ny_i \ln \mu_i + n\ln(1 - \mu_i) - ny_i \ln(1 - \mu_i) \right\}$$

$$= \exp\left\{ ny_i \ln\left(\frac{\mu_i}{1 - \mu_i}\right) + n\ln(1 - \mu_i) + \ln\binom{n}{ny_i} \right\}$$

$$= \exp\left\{ \frac{y_i \ln\left(\dfrac{\mu_i}{1 - \mu_i}\right) + \ln(1 - \mu_i)}{1/n} + \ln\binom{n}{ny_i} \right\}$$ [1]

Comparing this to the expression on page 27 of the *Tables*, we see that the natural parameter is:

$$\theta_i = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$$
[½]

Rearranging this gives $\mu_i = \dfrac{e^{\theta_i}}{1+e^{\theta_i}}$, so the function $b(\theta_i)$ is given by:

$$b(\theta_i) = -\ln(1-\mu_i) = -\ln\left(1-\frac{e^{\theta_i}}{1+e^{\theta_i}}\right) = -\ln\left(\frac{1}{1+e^{\theta_i}}\right) = \ln\left(1+e^{\theta_i}\right)$$
[1]

The scale parameter and its functions are:

$$\varphi = n, \qquad a(\varphi) = \frac{1}{\varphi}, \qquad c(y_i, \varphi) = \ln\binom{n}{ny_i} = \ln\binom{\varphi}{\varphi y_i}$$
[1]

(ii) *The variance function*

The variance function is given by $\text{var}(\mu) = b''(\theta)$. Differentiating $b(\theta)$ gives:

$$b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \mu_i$$
[½]

$$b''(\theta) = \frac{e^{\theta}\left(1+e^{\theta}\right) - e^{\theta}e^{\theta}}{\left(1+e^{\theta}\right)^2} = \frac{e^{\theta}}{\left(1+e^{\theta}\right)^2}$$
[½]

Substituting in $\theta = \ln\left(\dfrac{\mu_i}{1-\mu_i}\right)$ gives:

$$b''(\theta) = \frac{\dfrac{\mu_i}{1-\mu_i}}{\left(1+\dfrac{\mu_i}{1-\mu_i}\right)^2} = \frac{\mu_i}{1-\mu_i} \times (1-\mu_i)^2 = \mu_i(1-\mu_i)$$
[1]

[Total 6]

12.7 (i) *Derive mean and variance*

Mean

Differentiating both sides with respect to θ gives:

$$\int_y \frac{y - b'(\theta)}{a(\varphi)} f(y)\, dy = 0$$
[1]

Simplifying:

$$\frac{1}{a(\varphi)}\int_y yf(y)\,dy - \frac{b'(\theta)}{a(\varphi)}\int_y f(y)\,dy = 0$$

Since $\int yf(y)\,dy = E(Y)$, and $\int f(y)\,dy = 1$, we have:

$$\frac{1}{a(\varphi)}E(Y) - \frac{b'(\theta)}{a(\varphi)} = 0 \qquad\qquad [\frac{1}{2}]$$

Hence:

$$E(Y) - b'(\theta) = 0 \quad\Rightarrow\quad E(Y) = b'(\theta) \qquad\qquad [\frac{1}{2}]$$

Variance

Using the product rule to differentiate the above equation with respect to θ gives:

$$\int_y \frac{d^2}{d\theta^2}f(y)\,dy = \int_y \left\{\left(\frac{y-b'(\theta)}{a(\varphi)}\right)^2 f(y) - \frac{b''(\theta)}{a(\varphi)}f(y)\right\}dy = 0 \qquad\qquad [1]$$

Splitting this into two separate integrals gives:

$$\frac{1}{\left[a(\varphi)\right]^2}\int_y \left(y-b'(\theta)\right)^2 f(y)\,dy - \frac{b''(\theta)}{a(\varphi)}\int_y f(y)\,dy = 0$$

Since $b'(\theta) = E(Y)$ then $\int(y-b'(\theta))^2 f(y)\,dy = \text{var}(Y)$. Again $\int f(y)\,dy = 1$, so we have:

$$\frac{1}{\left[a(\varphi)\right]^2}\text{var}(Y) - \frac{b''(\theta)}{a(\varphi)} = 0 \qquad\qquad [\frac{1}{2}]$$

Rearranging gives:

$$\text{var}(Y) = a(\varphi)b''(\theta) \qquad\qquad [\frac{1}{2}]$$

(ii)   *Mean and variance of gamma*

The log of the PDF given is:

$$\log f(x) = \alpha(\log\alpha - \log\mu) - \log\Gamma(\alpha) + (\alpha-1)\log x - \frac{\alpha}{\mu}x$$

which can be written as:

$$f(x) = \exp\left\{\frac{-x/\mu - [-\log(1/\mu)]}{1/\alpha} + \alpha\log\alpha - \log\Gamma(\alpha) + (\alpha-1)\log x\right\}$$

This conforms to the definition of the exponential family, with:

$$\theta = -\frac{1}{\mu}, \quad b(\theta) = -\log(-\theta),$$

$$\phi = \alpha, \quad a(\phi) = 1/\phi,$$

$$c(x,\alpha) = \alpha\log\alpha - \log\Gamma(\alpha) + (\alpha - 1)\log x \qquad [1]$$

Applying the results in (i):

$$E(X) = b'(\theta) = -\frac{1}{-\theta} \times (-1) = -\frac{1}{\theta} = \mu \qquad [1]$$

and:

$$\text{var}(X) = a(\phi)b''(\theta) = \frac{1}{\phi} \times \frac{1}{\theta^2} = \frac{\mu^2}{\alpha} \qquad [1]$$

12.8   The scaled deviance is given by:

$$\text{scaled deviance} = 2\left[\ln L_S - \ln L_M\right] \qquad [\frac{1}{2}]$$

where L_S is the likelihood function for the saturated model, and L_M is the likelihood function for the fitted model.

First we need the log-likelihoods:

$$L(\mu_i) = f(y_1) \times \cdots \times f(y_n) = \frac{1}{\mu_1}e^{-\frac{1}{\mu_1}y_1} \times \cdots \times \frac{1}{\mu_n}e^{-\frac{1}{\mu_n}y_n} = \frac{1}{\mu_1 \times \cdots \times \mu_n}e^{-\sum \frac{1}{\mu_i}y_i} \qquad [1]$$

Taking logs:

$$\ln L(\mu_i) = -\sum \ln \mu_i - \sum \frac{1}{\mu_i}y_i \qquad [\frac{1}{2}]$$

So the log-likelihood of the fitted model, $\ln L_M$ is given by:

$$\ln L_M = -\sum \ln \hat{\mu}_i - \sum \frac{1}{\hat{\mu}_i}y_i \qquad [1]$$

For the log-likelihood of the saturated model, $\ln L_S$, the fitted values, $\hat{\mu}_i$ will just be the observed values, y_i. Hence:

$$\ln L_S = -\sum \ln y_i - \sum \frac{1}{y_i}y_i = -\sum \ln y_i - \sum 1 \qquad [1]$$

So the scaled deviance is:

$$\text{scaled deviance} = 2\left\{\left(-\sum \ln y_i - \sum 1\right) - \left(-\sum \ln \hat{\mu}_i - \sum \frac{1}{\hat{\mu}_i} y_i\right)\right\}$$

$$= 2\sum\left\{-\ln y_i - 1 + \ln \hat{\mu}_i + \frac{y_i}{\hat{\mu}_i}\right\}$$

$$= 2\sum\left\{\ln\left(\frac{\hat{\mu}_i}{y_i}\right) - 1 + \frac{y_i}{\hat{\mu}_i}\right\} \qquad\qquad [1]$$

12.9 (i) *Parameterised form*

In parameterised form, the linear predictors are (with i, j and k corresponding to the levels of *YO*, *FS* and *TC* respectively):

Model 1: $\alpha_i + \beta_j + \gamma_k$ (4 parameters)

There is one parameter to set the base level for the combination YO_0, FS_0, TC_0 and one additional parameter for each of the higher levels of the three factors.

Model 2: $\alpha_{ij} + \gamma_k$ (5 parameters)

There are four parameters for the 2×2 combinations of YO and FS (assuming TC_0) and one additional parameter for the higher level of TC.

Model 3: α_{ijk} (8 parameters)

There are eight parameters for the $2\times2\times2$ combinations of YO, FS and TC.

[2 for each model]

(ii) *Problems with Model 1 and Model 3*

Model 1 does not allow for the possibility that there may be interactions (correlations) between some of the factors. For example, it may be the case that young drivers tend to drive fast cars and to live in towns. [1]

With Model 3, which is a saturated model, it would be possible to fit the average values for each group exactly *ie* there are no degrees of freedom left. This defeats the purpose of applying a statistical model, as it would not 'smooth' out any anomalous results. [1]

The problem referred to with Model 3 corresponds to the idea of undergraduation in Subject CS2.

(iii) *Explaining 'normal error structure' and 'canonical link function'*

Normal error structure means that the randomness present in the observed values in each category (*eg* young/fast/town) is assumed to follow a normal distribution. [1]

The link function is the function applied to the linear estimator to obtain the predicted values. Associated with each type of error structure is a 'canonical' or 'natural' link function. In the case of a normal error structure, the canonical link function is the identity function . [2]

(iv)   ***Compare models***

The completed table, together with the differences in the scaled deviance and degrees of freedom, is shown below.

Model	Scaled Deviance	DF	Δ Scaled Deviance	Δ DF
Constant: 1	50	7		
Model 1: $YO + FS + TC$	10	4	40	3
Model 2: $YO + FS + YO.FS + TC$	5	3	5	1
Model 3: $YO * FS * TC$	0	0	5	3

[3]

Comparing the constant model and Model 1

The difference in the scaled deviances is 40.

This is greater than 7.815, the upper 5% critical value of a χ_3^2 distribution.

So Model 1 **is** a significant improvement over the constant model.   [½]

*Alternatively, using the AIC to compare models, since Δ(deviance) $> 2 \times \Delta$(parameters) Model 1 **is a** significant improvement over the constant model.*

Comparing Model 1 and Model 2

The difference in the scaled deviances is 5.

This is greater than 3.841, the upper 5% critical value of a χ_1^2 distribution.

So Model 2 **is** a significant improvement over Model 1.   [½]

*Alternatively, using the AIC to compare models, since Δ(deviance) $> 2 \times \Delta$(parameters) Model 2 **is a** significant improvement over Model 1.*

Comparing Model 2 and Model 3

The difference in the scaled deviances is 5.

This is less than 7.815, the upper 5% critical value of a χ_3^2 distribution.

So Model 3 is **not** a significant improvement over Model 2.   [½]

*Alternatively, using the AIC to compare models, since Δ(deviance) $\not> 2 \times \Delta$(parameters) Model 3 is **not** a significant improvement over Model 2.*

So Model 2 would be the most appropriate in this case.   [½]

12.10   (i)   **Natural link function**

Using the linear predictor $\eta_i = \alpha + \beta x_i$, we have $E(Y_i) = \mu_i = e^{\eta_i}$ so $\eta_i = g(\mu_i) = \ln \mu_i$ is the natural
link function.   [2]

(ii)   **Equations**

Assuming that $Y_i \sim Exp(\lambda_i)$, we have the following likelihood function:

$$L = \prod_{i=1}^{17} f(y_i) = \prod_{i=1}^{17} \lambda_i e^{-\lambda_i y_i}$$

Taking natural logs gives:

$$\ln L = \sum_{i=1}^{17} \ln \lambda_i - \sum_{i=1}^{17} \lambda_i y_i$$

$$= \sum_{i=1}^{17} \ln\left(\frac{1}{\mu_i}\right) - \sum_{i=1}^{17}\left(\frac{y_i}{\mu_i}\right) \text{ since } \mu_i = E(Y_i) = \frac{1}{\lambda_i}$$

$$= -\sum_{i=1}^{17}(\alpha + \beta x_i) - \sum_{i=1}^{17} y_i e^{-(\alpha + \beta x_i)}$$ [1]

Differentiating with respect to α gives:

$$\frac{\partial \ln L}{\partial \alpha} = -17 + \sum_{i=1}^{17} y_i e^{-(\alpha + \beta x_i)}$$ [1]

and differentiating with respect to β gives:

$$\frac{\partial \ln L}{\partial \beta} = -\sum_{i=1}^{17} x_i + \sum_{i=1}^{17} x_i y_i e^{-(\alpha + \beta x_i)}$$ [1]

Setting these expressions equal to 0, we obtain:

$$\sum_{i=1}^{17} y_i e^{-(\alpha + \beta x_i)} = 17$$

$$\sum_{i=1}^{17} x_i y_i e^{-(\alpha + \beta x_i)} = \sum_{i=1}^{17} x_i$$ [1]

(iii)   **Confidence interval**

An approximate 95% confidence interval for α is:

$$\hat{\alpha} \pm 1.96 \, s.e.(\hat{\alpha}) = 8.477 \pm (1.96 \times 1.655) = 8.477 \pm 3.244 = (5.233, 11.721)$$ [1]

Since this confidence interval does not contain zero we are 95% confident that the parameter is
non-zero and should be kept.   [1]

This is equivalent to the significance of a parameter test: $|\hat{\alpha}| > 2 \times se(\hat{\alpha})$.

(iv)   *Comparing models*

Test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. The test statistic is:

$$\Delta\text{dev} = 26.282 - 19.457 = 6.825 \hspace{3cm} [1]$$

Comparing with χ^2_1 we find that the value of the test statistic exceeds the upper 1% point (6.635) of this distribution. We therefore reject the null hypothesis and conclude that Model 2 significantly reduces the scaled deviance (*ie* it is significantly better fit to the data) so survival time is dependent on initial white blood cell count. [2]

13

Bayesian statistics

Syllabus objectives

5.1   Explain the fundamental concepts of Bayesian statistics and use these concepts to calculate Bayesian estimates.

   5.1.1   Use Bayes' theorem to calculate simple conditional probabilities.

   5.1.2   Explain what is meant by a prior distribution, a posterior distribution and a conjugate prior distribution.

   5.1.3   Derive the posterior distribution for a parameter in simple cases.

   5.1.4   Explain what is meant by a loss function.

   5.1.5   Use simple loss functions to derive Bayesian estimates of parameters.

0   Introduction

Earlier in this course we looked at the classical approach to statistical estimation, when we introduced the method of maximum likelihood and the method of moments. There we assumed that the parameters to be estimated were fixed quantities.

In this chapter we describe the Bayesian approach. This will also be used in .

The Bayesian philosophy involves a completely different approach to statistics, compared to classical statistical methods. The Bayesian version of estimation is considered here for the basic situation concerning the estimation of a parameter given a random sample from a particular distribution. Classical estimation involves the method of maximum likelihood.

The fundamental difference between Bayesian and classical methods is that the parameter θ is considered to be a random variable in Bayesian methods.

In classical statistics, θ is a fixed but unknown quantity. This leads to difficulties such as the careful interpretation required for classical confidence intervals, where it is the interval that is random. As soon as the data are observed and a numerical interval is calculated, there is no probability involved. A statement such as $P(10.45 < \theta < 13.26) = 0.95$ cannot be made because θ is not a random variable.

In classical statistics θ either lies within the interval or it does not. There can be no probability associated with such a statement.

In Bayesian statistics no such difficulties arise and probability statements can be made concerning the values of a parameter θ.

This means that it is quite possible to calculate a Bayesian confidence interval for a parameter. Although we shall not do this in this course, it is quite a common procedure in Bayesian statistics.

Another advantage of Bayesian statistics is that it enables us to make use of any information that we already have about the situation under investigation. Often researchers investigating an unknown population parameter have information available from other sources in advance of their study. This information might provide a strong indication of what values the parameter is likely to take. The classical statistical approach offers no scope for researchers to take this additional information into account. The Bayesian approach, however, does allow for the use of this information.

For example, suppose that an insurance company is reviewing its premium rates for a particular type of policy and has access to results from other insurers, as well as from its own policyholders. This information from other insurers cannot be taken into account directly because the terms and conditions of the policies for other companies may be slightly different. However, this additional information might be very useful, and hence should not be ignored.

1   Bayes' theorem

If $B_1, B_2, ..., B_k$ constitute a partition of a sample space S and $P(B_i) \neq 0$ for $i = 1, 2, ..., k$, then for any event A in S such that $P(A) \neq 0$:

$$P(B_r \mid A) = \frac{P(A \mid B_r)P(B_r)}{P(A)} \quad \text{where } P(A) = \sum_{i=1}^{k} P(A \mid B_i)P(B_i)$$

for $r = 1, 2, ..., k$.

A partition of a sample space is a collection of events that are mutually exclusive and exhaustive, *ie* they do not overlap and they cover the entire range of possible outcomes.

The result above is known as Bayes' theorem (or Bayes' formula) and is given on page 5 of the *Tables*. It follows easily from the result:

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

which rearranges to give the conditional probability formula:

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

However:

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

Now, replacing B by B_r, we have:

$$P(B_r \mid A) = \frac{P(B_r \cap A)}{P(A)} = \frac{P(B_r)P(A \mid B_r)}{P(A)}$$

and, from the law of total probability:

$$P(A) = \sum_i P(A \mid B_i) \, P(B_i)$$

Bayes' formula allows us to 'turn round' a conditional probability, *ie* it allows us to calculate $P(B \mid A)$ if we know $P(A \mid B)$.

Question

Three manufacturers supply clothing to a retailer. 60% of the stock comes from Manufacturer 1, 30% from Manufacturer 2 and 10% from Manufacturer 3. 10% of the clothing from Manufacturer 1 is faulty, 5% from Manufacturer 2 is faulty and 15% from Manufacturer 3 is faulty.

What is the probability that a faulty garment comes from Manufacturer 3?

Solution

Let A be the event that a garment is faulty and let B_i be the event that the garment comes from Manufacturer i.

Substituting the figures into the formula for Bayes' theorem:

$$P(B_3 \mid A) = \frac{(0.15)(0.1)}{(0.1)(0.6) + (0.05)(0.3) + (0.15)(0.1)} = \frac{0.015}{0.09} = 0.167$$

Although Manufacturer 3 supplies only 10% of the garments to the retailer, nearly 17% of the faulty garments come from that manufacturer.

An alternative way of tackling this question is to draw a tree diagram. There are 3 manufacturers so we start with 3 branches in our tree and mark on the associated probabilities:

Each garment is either faulty (event A) or perfect (event A'). These outcomes and their (conditional) probabilities are now added to the diagram:

The required probability is:

$$P(B_3 \mid A) = \frac{P(B_3 \cap A)}{P(A)}$$

From the diagram we can see that $P(B_3 \cap A) = 0.1 \times 0.15 = 0.015$. (This is obtained by multiplying the appropriate branch weights.) We can also see that there are three ways in which event A can occur. Since these are mutually exclusive, we can calculate $P(A)$ by summing the three associated probabilities. Hence:

$$P(A) = (0.6 \times 0.1) + (0.3 \times 0.05) + (0.1 \times 0.15) = 0.09$$

and it follows that:

$$P(B_3 \mid A) = \frac{0.015}{0.09} = 0.167$$

as before.

Bayes' theorem can be adapted to deal with continuous random variables. If X and Y are continuous, then the conditional PDF of Y given X is:

$$f_{Y|X}(x,y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x,y) f_Y(y)}{f_X(x)}$$

where:

$$f_X(x) = \int_y f_{X,Y}(x,y) \, dy = \int_y f_{X|Y}(x,y) f_Y(y) \, dy$$

2   Prior and posterior distributions

Suppose $\underline{X} = (X_1, X_2, ..., X_n)$ is a random sample from a population specified by the density or probability function $f(x; \theta)$ and it is required to estimate θ.

Recall that a random sample is a set of IID random variables. Here the Core Reading is using the letter f for both the density function of a continuous distribution and the probability function of a discrete distribution.

As a result of the parameter θ being a random variable, it will have a distribution. This allows the use of any knowledge available about possible values for θ before the collection of any data. This knowledge is quantified by expressing it as the prior distribution of θ.

The prior distribution summarises what we know about θ before we collect any data from the relevant population.

Then after collecting appropriate data, the posterior distribution of θ is determined, and this forms the basis of all inference concerning θ.

The Bayesian approach combines the sample data with the prior distribution. The conditional distribution of θ given the observed data is called the posterior distribution of θ.

2.1   Notation

As θ is a random variable, it should really be denoted by the capital Θ, and its prior density written as $f_{\Theta}(\theta)$. However, for simplicity no distinction will be made between Θ and θ, and the density will simply be denoted by $f(\theta)$. Note that referring to a density here implies that θ is continuous. In most applications this will be the case, as even when X is discrete (like the binomial or Poisson), the parameter (p or λ) will vary in a continuous space ($(0,1)$ or $(0, \infty)$, respectively).

Also the population density or probability function will be denoted by $f(x \mid \theta)$ rather than the earlier $f(x; \theta)$ as it represents the conditional distribution of X given θ.

The prior and posterior distributions of θ always have the same support (or domain). In other words, the set of possible values of θ is the same for both its prior and posterior distributions. So, if the prior distribution is continuous, then the posterior distribution is also continuous. Similarly, if the prior distribution is discrete, then the posterior distribution is also discrete.

Suppose, for example, that the parameter θ must take a value between 0 and 1. We might use a beta distribution as the prior distribution (as a beta random variable must take a value between 0 and 1). The posterior PDF of θ is then also non-zero for values of θ in the interval $(0,1)$ only.

2.2   Continuous prior distributions

Suppose that $\underline{X}$ is a random sample from a population specified by $f(x \mid \theta)$ and that θ has the prior density $f(\theta)$.

In other words, $X_1, \dots, X_n$ is a set of IID random variables whose distribution depends on the value of θ. Each of these random variables has PDF $f(x|\theta)$.

Determining the posterior density

The posterior density of $\theta|\underline{X}$ is determined by applying the basic definition of a conditional density:

$$f(\theta \mid \underline{X}) = \frac{f(\theta, \underline{X})}{f(\underline{X})} = \frac{f(\underline{X}|\theta)f(\theta)}{f(\underline{X})}$$

Note that $f(\underline{X}) = \int f(\underline{X}|\theta)f(\theta)d\theta$. This result is like a continuous version of Bayes' theorem.

We saw this result at the end of Section 1.

A useful way of expressing the posterior density is to use proportionality. $f(\underline{X})$ does not involve θ and is just the constant needed to make it a proper density that integrates to unity, so:

$$f(\theta|\underline{X}) \propto f(\underline{X}|\theta)f(\theta)$$

This formula is given on page 28 of the *Tables*.

Also, $f(\underline{X}|\theta)$, being the joint density of the sample values, is none other than the likelihood, making the posterior proportional to the product of the likelihood and the prior.

This idea is really the key to answering questions involving continuous prior distributions. The formula for the posterior PDF can also be expressed as follows:

$$f_{post}(\theta) = C \times f_{prior}(\theta) \times L$$

where:

- $f_{prior}(\theta)$ is the prior PDF of θ

- $f_{post}(\theta)$ is the posterior PDF of θ

- L is the likelihood function obtained from the sample data

- C is a constant that makes the posterior PDF integrate to 1.

Question

The annual number of claims arising from a particular group of policies follows a Poisson distribution with mean μ. The prior distribution of μ is exponential with mean 30.

In the previous two years, the numbers of claims arising from the group were 28 and 26, respectively.

Determine the posterior distribution of μ.

Solution

We are told that μ has an exponential distribution with a mean of 30. So $\mu \sim Exp(1/30)$ and the prior PDF of μ is:

$$f_{prior}(\mu) = \frac{1}{30}e^{-\mu/30} \ , \quad \mu > 0$$

Let X_j represent the number of claims in year j. Then $X_j \sim Poisson(\mu)$, and the likelihood function obtained from the sample data is:

$$L(\mu) = P(X_1 = 28)P(X_2 = 26) = \frac{e^{-\mu}\mu^{28}}{28!} \times \frac{e^{-\mu}\mu^{26}}{26!} = Ce^{-2\mu}\mu^{54}$$

where C is a constant.

Combining the prior distribution and the sample data, we see that the posterior PDF of μ is:

$$f_{post}(\mu) = Ke^{-61\mu/30}\mu^{54} \ , \quad \mu > 0$$

for some constant K.

Comparing this with the formula for the gamma PDF given on page 12 of the *Tables*, we see that the posterior distribution of μ is $Gamma\left(55, \frac{61}{30}\right)$.

The table in Section 4 of this chapter shows the posterior distribution for some common combinations of likelihoods and prior distributions. You do not need to learn this table. However, you should check that you can derive some of the results in the table, working along the lines shown in the solution above.

Conjugate priors

For a given likelihood, if the prior distribution leads to a posterior distribution belonging to the same family as the prior distribution, then this prior is called the conjugate prior for this likelihood.

The likelihood function determines which family of distributions will lead to a conjugate pair, *ie* a prior and posterior distribution that come from the same family. Conjugate distributions can be found by selecting a family of distributions that has the same algebraic form as the likelihood function, treating the unknown parameter as the random variable.

Question

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a Type 1 geometric distribution with parameter p, where p is a random variable.

Determine a family of distributions for p that would result in conjugate prior and posterior distributions.

Solution

Each of the random variables X_i has probability function:

$$P(X = x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \ldots$$

If the observed values of $X_1, X_2, \ldots, X_n$ are $x_1, x_2, \ldots, x_n$, then the likelihood function is:

$$L(p) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\sum x_i - n}$$

We know that p must take a value between 0 and 1. To result in a conjugate pair, the PDF of p must be of the form:

$$p^{something}(1-p)^{something}, \quad \text{for } 0 < p < 1$$

ie p must have a beta distribution.

Using conjugate distributions often makes Bayesian calculations simpler. Conjugate distributions may also be appropriate to use where there is a family of distributions that might be expected to provide a 'natural' model for the unknown parameter, *eg* in the previous example where the probability parameter p had to lie in the range $0 < p < 1$ (which is the range of values over which the beta distribution is defined).

Uninformative prior distributions

An uninformative prior distribution assumes that an unknown parameter is equally likely to take any value from a given set. In other words, the parameter is modelled using a uniform distribution.

As an example, suppose that we have a random sample $X_1, X_2, \ldots, X_n$ from a normal population with mean μ, but we have no prior information about μ. In this case it would be natural to model μ using a uniform distribution. Since μ can take any value between $-\infty$ and ∞, the appropriate uniform prior is $U(-\infty, \infty)$. This leads to a problem, however, since the PDF of this distribution is 0 everywhere.

We can get round this problem by using the distribution $U(-m, m)$ and then letting $m \to \infty$. If $\mu \sim U(-m, m)$, then the prior PDF of μ is:

$$f_{prior}(\mu) = \begin{cases} \dfrac{1}{2m} & \text{if } -m < \mu < m \\ 0 & \text{otherwise} \end{cases}$$

Also, since the data values come from a normal population, the likelihood function is:

$$L(\mu) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma} \right)^2 \right]$$

As usual, we are using σ to denote the population standard deviation. The formula for the PDF of $N(\mu, \sigma^2)$ is given on page 11 of the *Tables*.

The likelihood function can alternatively be expressed as:

$$L(\mu) = C \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma} \right)^2 \right]$$

where C is a constant that does not depend on μ.

Combining the prior PDF with the likelihood function gives:

$$f_{post}(\mu) = \begin{cases} K \exp\left[-\dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\left(\dfrac{x_i - \mu}{\sigma} \right)^2 \right] & \text{if } -m < \mu < m \\ 0 & \text{otherwise} \end{cases}$$

where K is also a constant that does not depend on μ. This constant is required to ensure that the PDF integrates to 1.

Letting $m \to \infty$, we see that the posterior PDF is proportional to:

$$\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma} \right)^2 \right], \quad \text{for } -\infty < \mu < \infty$$

Notice that the PDF of this posterior distribution is proportional to the likelihood function. This should be intuitive as, by definition, a posterior distribution is obtained by combining two pieces of information:

- prior knowledge of the parameter, and

- the sample data.

However, in this case we are using an uninformative prior as we have no prior knowledge of the parameter. The posterior distribution is therefore determined solely by the sample data.

2.3   Discrete prior distributions

When the prior distribution is discrete, the posterior distribution is also discrete. To determine the posterior distribution, we must calculate a set of conditional probabilities. This can be done using Bayes' formula.

Question

The number of claims received per week from a certain portfolio has a Poisson distribution with mean λ. The prior distribution of λ is as follows:

λ	1	2	3
Prior probability	0.3	0.5	0.2

Given that 3 claims were received last week, determine the posterior distribution of λ.

Solution

Let X be the number of claims received in a week. To determine the posterior distribution of λ, we must calculate the conditional probabilities $P(\lambda = 1 \,|\, X = 3)$, $P(\lambda = 2 \,|\, X = 3)$ and $P(\lambda = 3 \,|\, X = 3)$. The first of these is:

$$P(\lambda = 1 \,|\, X = 3) = \frac{P(\lambda = 1, X = 3)}{P(X = 3)} = \frac{P(X = 3 \,|\, \lambda = 1)P(\lambda = 1)}{P(X = 3)}$$

Since $X \sim Poisson(\lambda)$:

$$P(X = 3 \,|\, \lambda = 1) = \frac{e^{-1} \times 1^3}{3!} = \frac{1}{6}e^{-1}$$

and, from the given prior distribution, we know that:

$$P(\lambda = 1) = 0.3$$

So:

$$P(\lambda = 1 \,|\, X = 3) = \frac{\frac{1}{6}e^{-1} \times 0.3}{P(X = 3)} = \frac{\frac{1}{20}e^{-1}}{P(X = 3)}$$

Similarly:

$$P(\lambda = 2 \,|\, X = 3) = \frac{P(X = 3 \,|\, \lambda = 2)P(\lambda = 2)}{P(X = 3)} = \frac{\frac{e^{-2} \times 2^3}{3!} \times 0.5}{P(X = 3)} = \frac{\frac{2}{3}e^{-2}}{P(X = 3)}$$

and:

$$P(\lambda = 3 \mid X = 3) = \frac{P(X = 3 \mid \lambda = 3)P(\lambda = 3)}{P(X = 3)} = \frac{\dfrac{e^{-3} \times 3^3}{3!} \times 0.2}{P(X = 3)} = \frac{\dfrac{9}{10}e^{-3}}{P(X = 3)}$$

Since these conditional probabilities must sum to 1, the denominator must be the sum of the three numerators, *ie*:

$$P(X = 3) = \frac{1}{20}e^{-1} + \frac{2}{3}e^{-2} + \frac{9}{10}e^{-3} = 0.15343$$

This can also be seen using the law of total probability:

$$P(X = 3) = P(X = 3, \lambda = 1) + P(X = 3, \lambda = 3) + P(X = 3, \lambda = 3)$$

$$= P(X = 3 \mid \lambda = 1)P(\lambda = 1) + P(X = 3 \mid \lambda = 2)P(\lambda = 2) + P(X = 3 \mid \lambda = 3)P(\lambda = 3)$$

So the posterior probabilities are:

$$P(\lambda = 1 \mid X = 3) = \frac{\dfrac{1}{20}e^{-1}}{0.15343} = 0.11989$$

$$P(\lambda = 2 \mid X = 3) = \frac{\dfrac{2}{3}e^{-2}}{0.15343} = 0.58806$$

$$P(\lambda = 3 \mid X = 3) = \frac{\dfrac{9}{10}e^{-3}}{0.15343} = 0.29205$$

Alternatively, we could use a proportionality argument to determine the posterior probabilities. The posterior probabilities are proportional to the likelihood multiplied by the prior probability:

$$P(\lambda = 1 \mid X = 3) \propto P(X = 3 \mid \lambda = 1)P(\lambda = 1) = \frac{1}{6}e^{-1} \times 0.3 = \frac{1}{20}e^{-1} = 0.01839$$

$$P(\lambda = 2 \mid X = 3) \propto P(X = 3 \mid \lambda = 2)P(\lambda = 2) = \frac{e^{-2} \times 2^3}{3!} \times 0.5 = \frac{2}{3}e^{-2} = 0.09022$$

$$P(\lambda = 3 \mid X = 3) \propto P(X = 3 \mid \lambda = 3)P(\lambda = 3) = \frac{e^{-3} \times 3^3}{3!} \times 0.2 = \frac{9}{10}e^{-3} = 0.04481$$

Rescaling so the probabilities sum to 1 we get:

$$P(\lambda = 1 \mid X = 3) = \frac{0.01839}{0.01839 + 0.09022 + 0.04481} = 0.11989$$

$$P(\lambda = 2 \mid X = 3) = \frac{0.09022}{0.01839 + 0.09022 + 0.04481} = 0.58806$$

$$P(\lambda = 3 \mid X = 3) = \frac{0.04481}{0.01839 + 0.09022 + 0.04481} = 0.29205$$

The posterior probability that $\lambda = 1$ is lower than the corresponding prior probability. The other two posterior probabilities are higher than their corresponding prior probabilities. This is to be expected given that the observed number of claims was 3.

Once we have determined the posterior distribution of a parameter, we can use this distribution to estimate the parameter value. As we are about to see, the estimate will depend on the chosen loss function.

3   The loss function

To obtain an estimator of θ, a loss function must first be specified. This is a measure of the 'loss' incurred when $g(\underline{X})$ is used as an estimator of θ. A loss function is sought which is zero when the estimation is exactly correct, that is, $g(\underline{X}) = \theta$, and which is positive and does not decrease as $g(\underline{X})$ gets further away from θ. There is one very commonly used loss function, called quadratic or squared error loss. Two others are also used in practice.

Then the Bayesian estimator is the $g(\underline{X})$ that minimises the expected loss with respect to the posterior distribution.

The main loss function is quadratic loss defined by:

$$L(g(\underline{x}), \theta) = \left[g(\underline{x}) - \theta \right]^2$$

So, when using quadratic loss, the aim is to minimise:

$$E\left[(g(\underline{x}) - \theta)^2 \right] = \int_{\theta} (g(\underline{x}) - \theta)^2 \, f_{post}(\theta) d\theta$$

This is related to mean square error from classical statistics.

Recall that, if $\hat{\theta}$ is an estimator of θ, then:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2 \right] = \text{var}(\hat{\theta}) + \left[bias(\hat{\theta}) \right]^2$$

The formula for the squared error loss implies that as we move further away from the true parameter value, the loss increases at an increasing rate. The graph of the loss function is a parabola with a minimum of zero at the true parameter value.

A second loss function is absolute error loss defined by:

$$L(g(\underline{x}), \theta) = \left| g(\underline{x}) - \theta \right|$$

Here the graph of the loss function is two straight lines that meet at the point $(\theta, 0)$. As we move away from the true value in either direction, the loss increases at a constant rate.

A third loss function is '0/1' or 'all-or-nothing' loss defined by:

$$L(g(\underline{x}), \theta) = \begin{cases} 0 & \text{if } g(\underline{x}) = \theta \\ 1 & \text{if } g(\underline{x}) \neq \theta \end{cases}$$

In this case there is a constant loss of 1 for any parameter estimate that is not equal to the true underlying parameter value. If we hit the parameter value exactly, then the loss is zero.

The Bayesian estimator that arises by minimising the expected loss for each of these loss functions in turn is the mean, median and mode, respectively, of the posterior distribution, each of which is a measure of location of the posterior distribution.

We will prove these results shortly.

The expected posterior loss is:

$$EPL = E[L(g(\underline{x}), \theta)] = \int L(g(\underline{x}), \theta)\, f(\theta \mid \underline{x})\, d\theta$$

The lower limit of the integral is the smallest possible value of θ and the upper limit is the largest possible value of θ.

3.1 Quadratic loss

For simplicity, *g* will be written instead of *g*(*x*).

In other words, we are assuming that g is our estimate of θ.

So:

$$EPL = \int (g - \theta)^2 \, f(\theta | \underline{x}) \, d\theta$$

We want to determine the value of g that minimises the EPL, so we differentiate the EPL with respect to g. Using the formula for differentiating an integral (which is given on page 3 of the *Tables*), we see that:

$$\frac{d}{dg} EPL = 2 \int (g - \theta) \, f(\theta | \underline{x}) \, d\theta$$

Equating to zero:

$$g \int f(\theta | \underline{x}) \, d\theta = \int \theta \, f(\theta | \underline{x}) \, d\theta$$

But $\int f(\theta | \underline{x}) d\theta$ = 1.

This is because $f(\theta | \underline{x})$ is the PDF of the posterior distribution. Integrating the PDF over all possible values of θ gives the value 1.

So:

$$g = \int \theta \, f(\theta | \underline{x}) \, d\theta = E(\theta | \underline{x})$$

Clearly this minimises *EPL*.

We can see this from the graph of the loss function or by differentiating the EPL a second time:

$$\frac{d^2}{dg^2} EPL = 2 \int f(\theta | \underline{x}) \, d\theta = 2 > 0 \quad \Rightarrow \min$$

Therefore the Bayesian estimator under quadratic loss is the mean of the posterior distribution.

Question

For the estimation of a binomial probability θ from a single observation x of the random variable X with the prior distribution of θ being beta with parameters α and β, investigate the form of the posterior distribution of θ and determine the Bayesian estimate of θ under quadratic loss.

Solution

The proportionality argument will be used and any constants simply omitted as appropriate.

Prior:

$$f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

omitting the constant $\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ **.**

Likelihood:

$$f(x \mid \theta) \propto \theta^{x}(1-\theta)^{n-x}$$

omitting the constant $\begin{pmatrix} n \\ x \end{pmatrix}$ **.**

Combining the prior PDF with the likelihood function gives the posterior PDF:

$$f(\theta \mid x) \propto \theta^{x}(1-\theta)^{n-x} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

Now it can be seen that, apart from the appropriate constant of proportionality, this is the density of a beta random variable. Therefore the immediate conclusion is that the posterior distribution of θ given $X = x$ is beta with parameters $x+\alpha$ and $n-x+\beta$.

It can also be seen that the posterior density and the prior density belong to the same family of distributions. Thus the conjugate prior for the binomial distribution is the beta distribution.

The Bayesian estimate under quadratic loss is the mean of this distribution, that is:

$$\frac{x+\alpha}{(x+\alpha)+(n-x+\beta)} = \frac{x+\alpha}{n+\alpha+\beta}$$

We can use R to simulate this Bayesian estimate.

The R code to obtain the Monte Carlo Bayesian estimate of the above is:

```
x <- rep(0,n)

for (i in 1:n)

    {theta <- rbeta(1,alpha,beta)

    x[i] <- rbinom(1,1,theta)}

mean(x)
```

Alternatively, the R code to obtain the Monte Carlo estimate of the above based on *M* simulations is:

```
x <- rep(0,M)

for (i in 1:M)

        {theta <- rbeta(1,alpha,beta)

         x[i] <- rbinom(1,n,theta)}

mean(x)
```

Question

A random sample of size 10 from a Poisson distribution with mean λ yields the following data values:

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

The prior distribution of λ is *Gamma*$(5,2)$.

Calculate the Bayesian estimate of λ under squared error loss.

Solution

Using the formula for the PDF of the gamma distribution given on page 12 of the *Tables*, we see that the prior PDF of λ is:

$$f_{prior}(\lambda) = \frac{2^5}{\Gamma(5)} \lambda^4 e^{-2\lambda} , \quad \lambda > 0$$

Alternatively, we could say:

$$f_{prior}(\lambda) \propto \lambda^4 e^{-2\lambda} , \quad \lambda > 0$$

The likelihood function obtained from the data is:

$$L(\lambda) = P(X_1 = 3)P(X_2 = 4)\dots P(X_{10} = 2)$$

where $X_1,\dots,X_{10}$ are independent *Poisson*(λ) random variables. So:

$$L(\lambda) = \frac{e^{-\lambda}\lambda^3}{3!} \times \frac{e^{-\lambda}\lambda^4}{4!} \times \dots \times \frac{e^{-\lambda}\lambda^2}{2!} = Ce^{-10\lambda}\lambda^{31}$$

where C is a constant. (31 is the sum of the observed data values.)

Combining the prior distribution and the sample data, we see that:

$$f_{post}(\lambda) \propto \lambda^{35} e^{-12\lambda}, \quad \lambda > 0$$

Comparing this with the formula for the gamma PDF, we see that the posterior distribution of λ is *Gamma*(36,12). The Bayesian estimate of λ under squared error loss is the mean of this gamma distribution, *ie* $\frac{36}{12} = 3$.

3.2   Absolute error loss

Again, g will be written instead of $g(\underline{x})$.

So:

$$EPL = \int |g - \theta| \, f(\theta|\underline{x}) \, d\theta$$

Assuming the range for θ is (–∞, ∞), then:

$$EPL = \int_{-\infty}^{g} (g - \theta) f(\theta|\underline{x}) d\theta + \int_{g}^{\infty} (\theta - g) f(\theta|\underline{x}) d\theta$$

We need to split the integral into two parts so that we can remove the modulus sign. The first integral covers the interval where $\theta \leq g$. Here $|g - \theta| = g - \theta$. The second integral covers the interval where $\theta \geq g$. Here $|g - \theta| = \theta - g$.

Again, we want to determine the value of g that minimises the EPL, so we differentiate the EPL with respect to g.

[Recall that $\dfrac{d}{dy} \displaystyle\int_{a(y)}^{b(y)} f(x,y)dx = \displaystyle\int_{a(y)}^{b(y)} \dfrac{\partial}{\partial y} f(x,y)dx + b'(y)f(b(y),y) - a'(y)f(a(y),y)$]

(This is the formula for differentiating an integral, given on page 3 of the *Tables*.)

Replacing x by θ and y by g in the formula for differentiating an integral, we see that:

$$\frac{d}{dg} \int_{-\infty}^{g} (g - \theta) f(\theta|\underline{x}) d\theta = \int_{-\infty}^{g} f(\theta|\underline{x}) d\theta + (g-g)f(g \mid \underline{x}) - 0 = \int_{-\infty}^{g} f(\theta|\underline{x}) d\theta$$

and:

$$\frac{d}{dg} \int_{g}^{\infty} (\theta - g) f(\theta|\underline{x}) d\theta = \int_{g}^{\infty} (-1) \times f(\theta|\underline{x}) d\theta + 0 - (g-g)f(g \mid \underline{x}) = -\int_{g}^{\infty} f(\theta|\underline{x}) d\theta$$

So:

$$\frac{d}{dg} EPL = \int_{-\infty}^{g} f(\theta|\underline{x}) d\theta - \int_{g}^{\infty} f(\theta|\underline{x}) d\theta$$

Equating to zero:

$$\int_{-\infty}^{g} f(\theta | \underline{x}) d\theta = \int_{g}^{\infty} f(\theta | \underline{x}) d\theta$$

that is, $P(\theta \leq g) = P(\theta \geq g)$, which specifies the median of the posterior distribution.

Recall that the median of a continuous distribution is the value of M that divides the distribution into two parts, with a 50% probability of being less than (or equal to) M and a 50% probability of being greater than (or equal to) M.

Question

A random sample of size 10 from a Poisson distribution with mean λ yields the following data values:

> 3, 4, 3, 1, 5, 5, 2, 3, 3, 2

The prior distribution of λ is *Gamma*$(5,2)$. Calculate the Bayesian estimate of λ under absolute error loss.

Solution

From the solution to the previous question, we know that the posterior distribution of λ is *Gamma*$(36,12)$. The Bayesian estimate of λ under absolute error loss is the median of this distribution, which can be obtained very quickly using R. The command $qgamma(0.5,36,12)$ gives the answer to be 2.972268.

We use the R command q to calculate the percentiles of a distribution. We follow q with the name of the distribution. Here we want the median, or the 50th percentile, so the first argument is 0.5. The second and third arguments are the parameters of the gamma distribution.

Alternatively, the median could be calculated (approximately) using the *Tables*. To do this, we have to use the relationship between the gamma distribution and the chi-squared distribution (which is given in the *Tables* on page 12). Here we know that:

$$\lambda | \underline{x} \sim Gamma(36,12)$$

For notational convenience, let $W = \lambda | \underline{x}$. Then:

$$W \sim Gamma(36,12) \Leftrightarrow 2 \times 12W \sim \chi^2_{2\times36}$$

The median of the posterior distribution is the value of M such that:

$$P(W < M) = 0.5$$

or equivalently:

$$P(\chi^2_{72} < 24M) = 0.5$$

From page 169 of the *Tables*, we see that the 50th percentile of χ^2_{70} is 69.33 and the 50th percentile of χ^2_{80} is 79.33. Interpolating between these values, we find that the 50th percentile of χ^2_{72} is approximately:

$$(1-0.2)\times 69.33 + 0.2\times 79.33 = 71.33$$

So:

$$24M \approx 71.33$$

and hence:

$$M \approx 2.972$$

3.3 All-or-nothing loss

Here the differentiation approach cannot be used. Instead a direct approach will be used with a limiting argument.

Consider:

$$L(g(\underline{x}),\theta) = \begin{cases} 0 & \textbf{if } g-\varepsilon < \theta < g+\varepsilon \\ 1 & \textbf{otherwise} \end{cases}$$

so that, in the limit as $\varepsilon \to 0$, this tends to the required loss function.

Then the expected posterior loss is:

$$EPL = 1 - \int_{g-\varepsilon}^{g+\varepsilon} f(\theta|\underline{x})d\theta = 1 - 2\varepsilon.f(g\,|\,\underline{x}) \quad \textbf{for small } \varepsilon$$

This is saying that, for a narrow strip, the area under the function is approximately equal to the area of a rectangle whose width is 2ε and whose height is equal to the average value of the function over that strip.

Again, the Bayesian estimate is the value of g that minimises the EPL. To minimise the EPL, we need to maximise $2\varepsilon f(g|x)$. This occurs when $f(g|x)$ is maximised, *ie* at the mode of the posterior distribution.

The EPL is minimised by taking g to be the mode of $f(\theta\,|\,\underline{x})$.

Question

A random sample of size 10 from a Poisson distribution with mean λ yields the following data values:

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

The prior distribution of λ is $Gamma(5,2)$. Calculate the Bayesian estimate of λ under all-or-nothing loss.

Solution

From a previous question, we know that the posterior distribution of λ is $Gamma(36,12)$. The Bayesian estimate of λ under all-or-nothing loss is the mode of this distribution. To calculate the mode, we need to differentiate the posterior PDF (or the log of this PDF) and set the derivative equal to 0.

We have already seen that:

$$f_{post}(\lambda) = C\lambda^{35}e^{-12\lambda}$$

Taking logs (to make the differentiation easier):

$$\ln f_{post}(\lambda) = \ln C + 35\ln\lambda - 12\lambda$$

Differentiating:

$$\frac{d}{d\lambda}\ln f_{post}(\lambda) = \frac{35}{\lambda} - 12$$

The derivative is equal to 0 when $\lambda = \dfrac{35}{12}$.

Differentiating again:

$$\frac{d^2}{d\lambda^2}\ln f_{post}(\lambda) = -\frac{35}{\lambda^2} < 0$$

Since the second derivative is negative, the posterior PDF is maximised when $\lambda = \dfrac{35}{12}$. So the

Bayesian estimate of λ under all-or-nothing loss is $\dfrac{35}{12}$ or 2.917.

4   Some Bayesian posterior distributions

In this section we give a table of situations in which the Bayesian approach may work well. The likelihood function is given, together with the distributions of the prior and the posterior. Do not attempt to learn all the results given in this table. The results are here for reference purposes only, and you will not be expected to be able to quote all these results in the examination. However, you may like to select one or two of the results given here and check that you can prove that the distribution of the posterior is as stated. You could use the table as a way of generating extra questions on particular Bayesian results.

The negative binomial distribution referenced here is that described in the *Tables* as the Type 2 negative binomial distribution. The results for the Type 2 geometric distribution can be obtained from those for the Type 2 negative binomial by setting $k = 1$. You may like to work out the corresponding results for the Type 1 negative binomial and geometric distributions.

Notice that despite the large number of examples given, the posterior distribution in all these cases turns out to be gamma, beta or normal. So, in most Bayesian questions it is worth checking whether the posterior PDF takes the form of one of these three distributions before you start thinking about other possibilities.

Likelihood of IID random variables $X_1,...,X_n$	Unknown parameter	Distribution of parameter	
		Prior	Posterior
$Poisson(\lambda)$	$\lambda > 0$	$U(0,\infty)$	$Gamma(\sum x + 1, n)$
		$Exp(\lambda')$	$Gamma(\sum x + 1, n + \lambda')$
		$Gamma(\alpha', \lambda')$	$Gamma(\sum x + \alpha', n + \lambda')$
$Exp(\lambda)$	$\lambda > 0$	$U(0,\infty)$	$Gamma(n + 1, \sum x)$
		$Exp(\lambda')$	$Gamma(n + 1, \sum x + \lambda')$
		$Gamma(\alpha', \lambda')$	$Gamma(n + \alpha', \sum x + \lambda')$
$Gamma(\alpha, \lambda)$	$\lambda > 0$	$U(0,\infty)$	$Gamma(n\alpha + 1, \sum x)$
		$Exp(\lambda')$	$Gamma(n\alpha + 1, \sum x + \lambda')$
		$Gamma(\alpha', \lambda')$	$Gamma(n\alpha + \alpha', \sum x + \lambda')$
$N(\mu, \sigma^2)$	$-\infty < \mu < \infty$	$U(-\infty, \infty)$	$N\left(\dfrac{1}{n}\sum x, \dfrac{\sigma^2}{n}\right)$
		$N(\mu', \sigma'^2)$	$N\left(\dfrac{\dfrac{\sum x}{\sigma^2} + \dfrac{\mu'}{\sigma'^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma'^2}}, \dfrac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma'^2}}\right)$
$LogN(\mu, \sigma^2)$	$-\infty < \mu < \infty$	$U(-\infty, \infty)$	$N\left(\dfrac{1}{n}\sum \log x, \dfrac{\sigma^2}{n}\right)$
$Bin(m,p)$	$0 < p < 1$	$U(0,1)$	$Beta(\sum x + 1, nm - \sum x + 1)$
		$Beta(\alpha', \beta')$	$Beta(\sum x + \alpha', nm - \sum x + \beta')$
$NB(k,p)$	$0 < p < 1$	$U(0,1)$	$Beta(nk + 1, \sum x + 1)$
		$Beta(\alpha', \beta')$	$Beta(nk + \alpha', \sum x + \beta')$

Chapter 13 Summary

Bayesian estimation v classical estimation

A common problem in statistics is to estimate the value of some unknown parameter θ.

The classical approach to this problem is to treat θ as a fixed, but unknown, constant and use sample data to estimate its value. For example, if θ represents some population mean then its value may be estimated by a sample mean.

The Bayesian approach is to treat θ as a random variable.

Prior distribution

The prior distribution of θ represents the knowledge available about the possible values of θ before the collection of any sample data.

Likelihood function

A likelihood function, L, is then determined, based on a random sample $\underline{X} = (X_1, X_2, \dots, X_n)$. The likelihood function is the joint PDF (or, in the discrete case, the joint probability) of $X_1, X_2, \dots, X_n \mid \theta$.

Posterior distribution

The prior distribution and the likelihood function are combined to obtain the posterior distribution of θ.

When θ is a continuous random variable:

$$f_{post}(\theta) \propto f_{prior}(\theta) \times L$$

When θ is a discrete random variable, the posterior distribution is a set of conditional probabilities.

Conjugate distributions

For a given likelihood, if the prior distribution leads to a posterior distribution belonging to the same family as the prior, then this prior is called the conjugate prior for this likelihood.

Uninformative prior distributions

If we have no prior knowledge about θ, a uniform prior distribution should be used. This is sometimes referred to as an uninformative prior distribution. When the prior distribution is uniform, the posterior PDF is proportional to the likelihood function.

Loss functions

A loss function, such as quadratic (or squared) error loss, absolute error loss or all-or-nothing (0/1) loss gives a measure of the loss incurred when $\hat{\theta}$ is used as an estimator of the true value of θ. In other words, it measures the seriousness of an incorrect estimator.

Under squared error loss, the mean of the posterior distribution minimises the expected loss function.

Under absolute error loss, the median of the posterior distribution minimises the expected loss function.

Under all-or-nothing loss, the mode of the posterior distribution minimises the expected loss function.

Chapter 13 Practice Questions

13.1   The punctuality of trains has been investigated by considering a number of train journeys. In the sample, 60% of trains had a destination of Manchester, 20% Edinburgh and 20% Birmingham. The probabilities of a train arriving late in Manchester, Edinburgh or Birmingham are 30%, 20% and 25%, respectively.

A late train is picked at random from the group under consideration. Calculate the probability that it terminated in Manchester.

13.2   A random variable X has a Poisson distribution with mean λ, which is initially assumed to have a chi-squared distribution with 4 degrees of freedom.

Determine the posterior distribution of λ after observing a single value x of the random variable X.

13.3   The number of claims in a week arising from a certain group of insurance policies has a Poisson distribution with mean μ. Seven claims were incurred in the last week.

The prior distribution of μ is uniform on the integers 8, 10 and 12.

(i)   Determine the posterior distribution of μ.

(ii)   Calculate the Bayesian estimate of μ under squared error loss.

13.4   For the estimation of a population proportion p, a sample of n is taken and yields x successes. A suitable prior distribution for p is beta with parameters 4 and 4.

Exam style

(i)   Show that the posterior distribution of p given x is beta and specify its parameters.   [2]

(ii)   Given that 11 successes are observed in a sample of size 25, calculate the Bayesian estimate under all-or-nothing (0/1) loss.   [4]
                                                                                                          [Total 6]

13.5   The annual number of claims from a particular risk has a Poisson distribution with mean μ. The prior distribution for μ has a gamma distribution with $\alpha = 2$ and $\lambda = 5$.

Exam style

Claim numbers $x_1, \ldots, x_n$ over the last n years have been recorded.

(i)   Show that the posterior distribution is gamma and determine its parameters.   [3]

(ii)   Given that $n = 8$ and $\sum\limits_{i=1}^{8} x_i = 5$ determine the Bayesian estimate for μ under:

        (a)   squared-error loss

        (b)   all-or-nothing loss

        (c)   absolute error loss.   [5]
                                                                                                          [Total 8]

13.6   A single observation, x, is drawn from a distribution with the probability density function:

Exam style

$$f(x \mid \theta) = \begin{cases} \theta^{-1} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

The prior PDF of θ is given by:

$$f(\theta) = \theta \exp(-\theta), \qquad \theta > 0$$

Derive an expression in terms of x for the Bayesian estimate of θ under absolute error loss.   [4]

13.7   A proportion p of packets of a rather dull breakfast cereal contain an exciting toy (independently

Exam style   from packet to packet). An actuary has been persuaded by his children to begin buying packets of this cereal. His prior beliefs about p before opening any packets are given by a uniform distribution on the interval [0,1]. It turns out the first toy is found in the n_1 th packet of cereal.

   (i)   Specify the posterior distribution of p after the first toy is found.   [3]

A further toy was found after opening another n_2 packets, another toy after opening another n_3 packets and so on until the fifth toy was found after opening a grand total of $n_1 + n_2 + n_3 + n_4 + n_5$ packets.

   (ii)   Specify the posterior distribution of p after the fifth toy is found.   [2]

   (iii) Show the Bayes' estimate of p under quadratic loss is not the same as the maximum likelihood estimate and comment on this result.   [5]

[Total 10]

13.8   An actuary has a tendency to be late for work. If he gets up late then he arrives at work X

Exam style   minutes late where X is exponentially distributed with mean 15. If he gets up on time then he arrives at work Y minutes late where Y is uniformly distributed on [0,25]. The office manager believes that the actuary gets up late one third of the time.

Calculate the posterior probability that the actuary did in fact get up late given that he arrives more than 20 minutes late at work.   [5]

Chapter 13 Solutions

13.1   Let M denote the event 'a train chosen at random terminates in Manchester' (and let E and B have corresponding definitions). In addition, let L denote the event 'A train chosen at random runs late'. The situation can then be represented using the following tree diagram:

The required probability is:

$$P(M \mid L) = \frac{P(M \cap L)}{P(L)}$$

From the diagram, we see that:

$$P(M \cap L) = 0.6 \times 0.3 = 0.18$$

and:

$$P(L) = 0.6 \times 0.3 + 0.2 \times 0.2 + 0.2 \times 0.25 = 0.27$$

So:

$$P(M \mid L) = \frac{0.18}{0.27} = \frac{2}{3}$$

Alternatively, we can calculate the probability using Bayes' formula:

$$P(M \mid L) = \frac{P(M)P(L \mid M)}{P(M)P(L \mid M) + P(E)P(L \mid E) + P(B)P(L \mid B)}$$

$$= \frac{0.6 \times 0.3}{(0.6 \times 0.3) + (0.2 \times 0.2) + (0.2 \times 0.25)}$$

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

13.2   The prior distribution of λ is χ_4^2, which is the same as $Gamma(2, \frac{1}{2})$. So:

$$f_{prior}(\lambda) \propto \lambda e^{-\lambda/2}$$

The likelihood function for a single observation x from a $Poisson(\lambda)$ distribution is proportional to:

$$\lambda^x e^{-\lambda}$$

So:

$$f_{post}(\lambda) \propto \lambda e^{-\lambda/2} \times \lambda^x e^{-\lambda} = \lambda^{x+1} e^{-3\lambda/2}$$

Hence the posterior distribution of λ is $Gamma\left(x+2, \frac{3}{2}\right)$.

13.3   **(i)**   ***Posterior distribution***

Let X be the number of claims received in a week. To determine the posterior distribution of μ, we must calculate the conditional probabilities $P(\mu = 8 \mid X = 7)$, $P(\mu = 10 \mid X = 7)$ and $P(\mu = 12 \mid X = 7)$. The first of these is:

$$P(\mu = 8 \mid X = 7) = \frac{P(\mu = 8, X = 7)}{P(X = 7)} = \frac{P(X = 7 \mid \mu = 8)P(\mu = 8)}{P(X = 7)}$$

Since $X \sim Poisson(\mu)$:

$$P(X = 7 \mid \mu = 8) = \frac{e^{-8} \times 8^7}{7!}$$

and since the prior distribution is uniform on the integers 8, 10 and 12:

$$P(\mu = 8) = \frac{1}{3}$$

So:

$$P(\mu = 8 \mid X = 7) = \frac{\dfrac{e^{-8} \times 8^7}{7!} \times \dfrac{1}{3}}{P(X = 7)} = \frac{0.04653}{P(X = 7)}$$

Similarly:

$$P(\mu = 10 \mid X = 7) = \frac{P(X = 7 \mid \mu = 10)P(\mu = 10)}{P(X = 7)} = \frac{\dfrac{e^{-10} \times 10^7}{7!} \times \dfrac{1}{3}}{P(X = 7)} = \frac{0.03003}{P(X = 7)}$$

and: $$P(\mu = 12 \mid X = 7) = \frac{P(X = 7 \mid \mu = 12)P(\mu = 12)}{P(X = 7)} = \frac{\dfrac{e^{-12} \times 12^7}{7!} \times \dfrac{1}{3}}{P(X = 7)} = \frac{0.01456}{P(X = 7)}$$

Since these conditional probabilities must sum to 1, the denominator must be the sum of the numerators, *ie*:

$$P(X = 7) = 0.04653 + 0.03003 + 0.01456 = 0.09112$$

So the posterior probabilities are:

$$P(\mu = 8 \mid X = 7) = \frac{0.04653}{0.09112} = 0.51066$$

$$P(\mu = 10 \mid X = 7) = \frac{0.03003}{0.09112} = 0.32954$$

$$P(\mu = 12 \mid X = 7) = \frac{0.01456}{0.09112} = 0.15980$$

(ii)   ***Bayesian estimate under squared error loss***

The Bayesian estimate under squared error loss is the mean of the posterior distribution:

$$8 \times 0.51066 + 10 \times 0.32954 + 12 \times 0.15980 = 9.29830$$

13.4   (i)   ***Posterior distribution***

Since the prior distribution of p is *Beta*$(4, 4)$:

$$f_{prior}(p) \propto p^3 (1 - p)^3 \tag{½}$$

Now let X denote the number of successes from a sample of size n. Then $X \sim Binomial(n, p)$.

Since x successes have been observed, the likelihood function is:

$$L(p) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \propto p^x (1 - p)^{n-x} \tag{½}$$

Combining the prior PDF with the likelihood function gives:

$$f_{post}(p) \propto p^3 (1 - p)^3 \times p^x (1 - p)^{n-x} = p^{x+3} (1 - p)^{n-x+3} \tag{½}$$

Comparing this with the PDF of the beta distribution (given on page 13 of the *Tables*), we see that the posterior distribution of p is $Beta(x+4, n-x+4)$.   [½]

[Total 2]

(ii)   *Bayesian estimate under all-or-nothing loss*

The Bayesian estimate under all-or-nothing loss is the mode of the posterior distribution, *ie* the value of p that maximises the posterior PDF. To find the mode, we need to differentiate the PDF (or equivalently differentiate the log of the PDF) and equate it to zero.

Given that $x = 11$ and $n = 25$, the posterior of p is $Beta(15, 18)$ and:

$$f_{post}(p) = C\, p^{14}(1-p)^{17} \tag{1}$$

Taking logs (to make the differentiation easier):

$$\ln f(p) = \ln C + 14 \ln p + 17 \ln(1-p) \tag{1}$$

Differentiating:

$$\frac{d}{dp}\ln f(p) = \frac{14}{p} - \frac{17}{1-p} \tag{½}$$

The derivative is equal to 0 when:

$$14(1-p) = 17p$$

ie when:

$$p = \frac{14}{31} \tag{½}$$

Differentiating again:

$$\frac{d^2}{dp^2}\ln f(p) = -\frac{14}{p^2} - \frac{17}{(1-p)^2} < 0 \quad \Rightarrow \quad \text{max} \tag{½}$$

So the Bayesian estimate of p under all-or-nothing loss is $\frac{14}{31}$ or 0.45161.   [½]

[Total 4]

13.5   (i)   *Posterior distribution*

Since the prior distribution of μ is $Gamma(2,5)$:

$$f_{prior}(\mu) = \frac{5^2}{\Gamma(2)}\mu e^{-5\mu} \propto \mu e^{-5\mu} \tag{½}$$

The likelihood is the product of Poisson probabilities:

$$L(\mu) = \frac{\mu^{x_1}}{x_1!}e^{-\mu} \times \dots \times \frac{\mu^{x_n}}{x_n!}e^{-\mu} \propto \mu^{\sum x_i}e^{-n\mu}$$ [1]

So:

$$f_{post}(\mu) \propto \mu e^{-5\mu} \times \mu^{\sum x_i}e^{-n\mu} = \mu^{1+\sum x_i}e^{-(n+5)\mu}$$ [1]

Comparing this with the PDF of the gamma distribution (given on page 12 of the *Tables*), we see that the posterior distribution of μ is $Gamma(2+\sum x_i, n+5)$.   [½]

[Total 3]

(ii)(a)   *Squared-error loss*

When $n = 8$ and $\sum x_i = 5$, the posterior distribution of μ is $Gamma(7,13)$.   [½]

The Bayesian estimate of μ under squared error loss is the mean of the posterior distribution, *ie* $\frac{7}{13}$ or 0.538.   [½]

(ii)(b)   *All-or-nothing loss*

The Bayesian estimate of μ under all-or-nothing loss is the mode of the posterior distribution, *ie* the value of p that maximises the posterior PDF. To find the mode, we need to differentiate the PDF (or equivalently differentiate the log of the PDF) and equate it to zero.

Since the posterior distribution of μ is $Gamma(7,13)$:

$$f_{post}(\mu) = C\,\mu^6 e^{-13\mu}$$

where C is a constant.

Taking logs:

$$\ln f_{post}(\mu) = \ln C + 6\ln\mu - 13\mu$$ [½]

Differentiating:

$$\frac{d}{d\mu}\ln f_{post}(\mu) = \frac{6}{\mu} - 13$$ [½]

The derivative is equal to 0 when $\mu = \frac{6}{13}$.   [½]

Differentiating again:

$$\frac{d^2}{d\mu^2}\ln f_{post}(\mu) = -\frac{6}{\mu^2} < 0 \quad\Rightarrow\quad \text{max}$$ [½]

So the Bayesian estimate of μ under all-or-nothing loss is $\frac{6}{13}$ or 0.462.

The mode of Gamma(α, λ) is $\frac{\alpha - 1}{\lambda}$ provided that $\alpha > 1$.

(ii)(c)   *Absolute error loss*

The Bayesian estimate of μ under absolute error loss is the median of the posterior distribution.

Here we know that:

$$\mu \,|\, \underline{x} \sim Gamma(7,13)$$

For notational convenience, let $W = \mu \,|\, \underline{x}$. Then:

$$W \sim Gamma(7,13) \Leftrightarrow 2 \times 13W \sim \chi^2_{2 \times 7} \qquad \text{[½]}$$

The median of the posterior distribution is the value of M such that:

$$P(W < M) = 0.5$$

or equivalently:

$$P(\chi^2_{14} < 26M) = 0.5 \qquad \text{[½]}$$

From page 169 of the *Tables*, we see that the 50th percentile of χ^2_{14} is 13.34:

$$26M = 13.34 \quad \Rightarrow \quad M = \frac{13.34}{26} = 0.513 \qquad \text{[½]}$$

So the Bayesian estimate of μ under absolute error loss is 0.513. $\qquad$ [½]

$\hspace{10cm}$ [Total 5]

13.6   Since we only have a single observation, the likelihood function is equal to the PDF of the distribution from which the observation came, *ie*:

$$L(\theta) = \begin{cases} \theta^{-1} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases} \qquad \text{[½]}$$

Also, since $f_{prior}(\theta) = \theta \exp(-\theta)$ for $\theta > 0$, it follows that :

$$f_{post}(\theta) = C \times f_{prior}(\theta) \times L(\theta) = \begin{cases} Ce^{-\theta} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases} \qquad \text{[½]}$$

where C is a constant. This distribution is not in the *Tables*, so we will have to work from first principles to determine the value of the constant.

Integrating the posterior PDF over all possible values of θ gives 1:

$$\int_x^\infty Ce^{-\theta}\, d\theta = 1 \;\;\Rightarrow\;\; \left[-Ce^{-\theta} \right]_x^\infty = 1 \;\;\Rightarrow\;\; Ce^{-x} = 1 \;\;\Rightarrow\;\; C = e^x \qquad\qquad [1]$$

So the posterior PDF is:

$$f_{post}(\theta) = e^{-(\theta - x)}, \quad \theta > x \qquad\qquad [\tfrac{1}{2}]$$

The Bayesian estimate of θ under absolute error loss is the median of the posterior distribution. The median, m, satisfies the equation:

$$\int_m^\infty e^{-(\theta - x)}\, d\theta = \tfrac{1}{2} \qquad\qquad [\tfrac{1}{2}]$$

Integrating:

$$\left[-e^{-(\theta - x)} \right]_m^\infty = \tfrac{1}{2}$$

$$\Rightarrow e^{-(m - x)} = \tfrac{1}{2}$$

$$\Rightarrow \; m - x = \log 2$$

$$\Rightarrow \; m = x + \log 2 \qquad\qquad [1]$$

ie the Bayesian estimate of θ under absolute error loss is $x + \log 2$.   [Total 4]

13.7   *This question is Subject CT6, April 2012, Question 6.*

(i)   **Posterior distribution of p**

Let X be the number of packets of cereal that must be opened in order to find a toy. Then $X\,|\,p$ has a Type 1 geometric distribution with parameter p. The prior distribution for p is uniform over the interval $[0,1]$, so:

$$f_{prior}(p) = 1, \;\; 0 \le p \le 1$$

ie the prior PDF is constant over the interval $[0,1]$. $[\tfrac{1}{2}]$

The sample consists of one observation, n_1. So the likelihood function is:

$$L(p) = P(X = n_1) = (1 - p)^{n_1 - 1} p \qquad\qquad [1]$$

Combining the prior PDF and the likelihood function, we see that:

$$f_{post}(p) \propto f_{prior} \times L(p) = p(1 - p)^{n_1 - 1} \qquad\qquad [\tfrac{1}{2}]$$

So the posterior distribution of p is $Beta(2, n_1)$.   [1]

[Total 3]

(ii)   *Posterior distribution after the fifth toy is found*

The likelihood function is now:

$$L(p) = P(X_1 = n_1) P(X_2 = n_2) P(X_3 = n_3) P(X_4 = n_4) P(X_5 = n_5)$$

$$= (1-p)^{n_1-1} p \times (1-p)^{n_2-1} p \times (1-p)^{n_3-1} p \times (1-p)^{n_4-1} p \times (1-p)^{n_5-1} p$$

$$= (1-p)^{\sum n_i - 5} p^5 \qquad\qquad [1]$$

and hence:

$$f_{post}(p) \propto p^5 (1-p)^{\sum n_i - 5} \qquad\qquad [\frac{1}{2}]$$

So the posterior distribution of p is $Beta\left(6, \sum_{i=1}^{5} n_i - 4\right)$.   [½]

[Total 2]

(iii)   *Bayesian estimate under quadratic loss v maximum likelihood estimate*

The Bayesian estimate of p under quadratic loss is the mean of the posterior distribution:

$$\frac{6}{6 + \sum_{i=1}^{5} n_i - 4} = \frac{6}{\sum_{i=1}^{5} n_i + 2} \qquad\qquad [\frac{1}{2}]$$

The maximum likelihood estimate of p is the value of p that maximises the likelihood function:

$$L(p) = (1-p)^{\sum n_i - 5} p^5$$

This is the same as the value of p that maximises the log-likelihood function:

$$\log L(p) = 5 \ln p + \left(\sum_{i=1}^{5} n_i - 5\right) \ln(1-p)$$

Differentiating with respect to p:

$$\frac{d}{dp} \log L(p) = \frac{5}{p} - \left(\frac{\sum_{i=1}^{5} n_i - 5}{1-p}\right) \qquad\qquad [1]$$

The derivative is equal to 0 when:

$$\frac{5}{p} = \frac{\displaystyle\sum_{i=1}^{5} n_i - 5}{1 - p}$$

ie when:

$$5(1 - p) - p\left(\sum_{i=1}^{5} n_i - 5\right) = 0$$

The solution of this equation is:

$$p = \frac{5}{\displaystyle\sum_{i=1}^{5} n_i}$$ [1]

We can check this is a maximum by differentiating a second time with respect to p:

$$\frac{d^2}{dp^2} \log L(p) = -\frac{5}{p^2} - \frac{\left(\displaystyle\sum_{i=1}^{5} n_i - 5\right)}{(1 - p)^2}$$ [½]

Since each $n_i \geq 1$, both terms in the expression are negative, so we have a maximum. Hence the

maximum likelihood estimate of p is $\dfrac{5}{\displaystyle\sum_{i=1}^{5} n_i}$.   [½]

The two estimates are different. The Bayesian estimate of p under quadratic loss is the value of g that minimises the expected posterior loss:

$$\int_0^1 (g - p)^2 f_{post}(p)\, dp$$ [½]

The maximum likelihood estimate of p is the value of p that maximises the likelihood function.

[½]

We would expect the estimates to be different since they are calculated in different ways.   [½]

[Total 5]

13.8   *This question is Subject CT6, April 2013, Question 3.*

The required probability is:

$P(\text{up late} \mid >20 \text{ mins late})$

$$= \frac{P(>20 \text{ mins late} \mid \text{up late})P(\text{up late})}{P(>20 \text{ mins late} \mid \text{up late})P(\text{up late}) + P(>20 \text{ mins late} \mid \text{up on time})P(\text{up on time})}$$

[1]

Using the fact that when the actuary gets up late, he arrives at work X minutes late and when he gets up on time, he arrives at work Y minutes late, we have:

$$P(\text{up late} \mid >20 \text{ mins late}) = \frac{P(X>20)P(\text{up late})}{P(X>20)P(\text{up late}) + P(Y>20)P(\text{up on time})}$$ [1]

Since $X \sim Exp(1/15)$:

$$P(X>20) = 1 - F_X(20) = 1 - \left(1 - e^{-20 \times 1/15}\right) = e^{-4/3}$$ [1]

Also $Y \sim U(0,25)$, so:

$$P(Y>20) = 1 - F_Y(20) = 1 - \frac{20-0}{25-0} = \frac{1}{5}$$ [1]

Substituting these in gives:

$$P(\text{up late} \mid >20 \text{ mins late}) = \frac{e^{-4/3} \times \frac{1}{3}}{e^{-4/3} \times \frac{1}{3} + \frac{1}{5} \times \frac{2}{3}} = 0.39722$$ [1]

[Total 5]

14

Credibility theory

Syllabus objectives

5.1   Explain the fundamental concepts of Bayesian statistics and use these concepts to
       calculate Bayesian estimates.

   5.1.6   Explain what is meant by the credibility premium formula and describe the
            role played by the credibility factor.

   5.1.7   Explain the Bayesian approach to credibility theory and use it to derive
            credibility premiums in simple cases.

   5.1.9   Explain the differences between the two approaches (*ie* the Bayesian
            approach and the empirical Bayes approach) and state the assumptions
            underlying each of them.

0   Introduction

In this chapter we will discuss credibility theory and explain how it can be be used to calculate premiums or to estimate claim frequencies in general insurance. Here we will concentrate on the Bayesian approach to credibility. We will be using the theory of Bayesian estimation developed in Chapter 13 as well as some results from Chapter 4 involving conditional random variables.

1   Recap of conditional expectation results

Recall from Chapter 4 that if X and Y are discrete random variables, then:

$$E(X \mid Y = y) = \sum_x x \, P(X = x \mid Y = y)$$

Similarly, if X and Y are continuous random variables, then:

$$E(X \mid Y = y) = \int_x x \, f_{X \mid Y}(x, y) \, dx$$

Manipulation of conditional expectations is an important technique in credibility theory, as it is in many other areas of actuarial science. Some results are:

For any random variables X and Y (for which the relevant moments exist):

$$\boldsymbol{E[X] = E\big[E(X \mid Y)\big]} \tag{14.1.1}$$

This result is easy to demonstrate. If X and Y are discrete random variables, then:

$$E\big[E(X \mid Y)\big] = \sum_y E(X \mid Y = y) P(Y = y)$$

$$= \sum_y \left[\sum_x x \, P(X = x \mid Y = y) \right] P(Y = y)$$

$$= \sum_y \sum_x x \, P(X = x, Y = y)$$

$$= \sum_x \sum_y x \, P(X = x, Y = y)$$

$$= \sum_x x \sum_y P(X = x, Y = y)$$

$$= \sum_x x \, P(X = x)$$

$$= E(X)$$

A similar approach using integrals can be used if X and Y are continuous random variables.

Another important concept is that of conditional independence. If two random variables X_1 and X_2 are conditionally independent given a third random variable Y, then:

$$\boldsymbol{E[X_1 X_2 \mid Y] = E[X_1 \mid Y] E[X_2 \mid Y]} \tag{14.1.2}$$

Intuitively this says that both X_1 and X_2 depend on Y, but, if the value taken by Y is known, then X_1 and X_2 are independent. This does not imply that X_1 and X_2 are unconditionally independent, *ie* independent if the value taken by Y is not known. Hence, it may be the case that:

$$E[X_1 X_2] \neq E[X_1]E[X_2]$$

even though (14.1.2) holds.

2   Credibility

2.1   The credibility premium formula

The basic idea underlying the credibility premium formula is intuitively very simple and very appealing. Consider an extremely simple example.

Example

A local authority in a small town has run a fleet of ten buses for a number of years. The local authority wishes to insure this fleet for the coming year against claims arising from accidents involving these buses. The pure premium for this insurance needs to be calculated, *ie* the expected cost of claims in the coming year.

In order to make this calculation, the following data are available to you:

- For the past five years for this fleet of buses the average cost of claims per annum (for the ten buses) has been £1,600.

- Data relating to a large number of local authority bus fleets from all over the United Kingdom show that the average cost of claims per annum per bus is £250, so that the average cost of claims per annum for a fleet of ten buses is £2,500.

- However, while this figure of £2,500 is based on many more fleets of buses than the figure of £1,600, some of the fleets of buses included in this large data set operate under very different conditions (*eg* in large cities or in rural areas) from the fleet which is of concern here, and these different conditions are thought to affect the number and size of claims.

There are two extreme choices for the pure premium for the coming year:

(i)   £1,600 could be chosen as it is based on the most appropriate data

(ii)   £2,500 could be chosen because it is based on more data, so might be considered a more reliable figure.

The credibility approach to this problem is to take a weighted average of these two extreme answers, *ie* to calculate the pure premium as:

$$Z \times 1,600 + (1 - Z) \times 2,500$$

where Z is some number between zero and one. Z is known as the credibility factor. Purely for the sake of illustration, suppose Z is set equal to 0.6 so that the pure premium is calculated to be £1,960.

This example will be revisited to illustrate some points in the next section but now the above ideas will be expressed a little more formally. The problem is to estimate the expected aggregate claims, or, possibly, just the expected number of claims, in the coming year from a risk. By a risk we mean a single policy or a group of policies. These policies are, typically, short term policies and, for convenience, the term of the policies will be taken to be one year, although it could equally well be any other short period.

The following information is available:

- $\overline{X}$ **is an estimate of the expected aggregate claims / number of claims for the coming year based solely on data from the risk itself.**

We usually use lower case $\overline{x}$ to represent an estimate (which is a numerical value) and upper case $\overline{X}$ to represent the corresponding estimator (which is a random variable). However the Core Reading is using $\overline{X}$ to represent a numerical value here.

- μ **is an estimate of the expected aggregate claims / number of claims for the coming year based on collateral data, *ie* data for risks similar to, but not necessarily identical to, the particular risk under consideration.**

The credibility premium formula (or credibility estimate of the aggregate claims / number of claims) for this risk is:

$$Z\,\overline{X} + (1-Z)\mu \qquad\qquad\qquad\qquad \textbf{(14.2.1)}$$

where Z is a number between zero and one and is known as the credibility factor. The attractive features of the credibility premium formula are its simplicity and, provided $\overline{X}$ and μ are obviously reasonable alternatives, the ease with which it can be explained to a lay person.

Question

A specialist insurer that provides insurance against breakdown of photocopying equipment calculates its premiums using a credibility formula. Based on the company's recent experience of all models of copiers, the premium for this year should be £100 per machine. The company's experience for a new model of copier, which is considered to be more reliable, indicates that the premium should be £60 per machine.

Given that the credibility factor is 0.75, calculate the premium that should be charged for insuring the new model.

Solution

The premium based on the collateral data (including all machines) is:

$$\mu = 100$$

The premium based on the direct data (the new model) is:

$$\overline{X} = 60$$

So, using the credibility formula with $Z = 0.75$, the premium that should be charged is:

$$P = Z\,\overline{X} + (1-Z)\mu = 0.75 \times 60 + 0.25 \times 100 = £70$$

Examples of situations where an insurer might determine a premium rate by combining direct data for a risk with collateral data include the following:

- New type of cover

  An insurer offering a new type of cover (*eg* protection against damage caused by driverless vehicles) would not have enough direct data available initially from the claims from the new policies to judge the premium accurately. The insurer could use claims data from similar well-established types of cover (*eg* vehicles with drivers) as collateral data in the first few years. As the company sold more of the new policies, the pattern of claims arising from driverless vehicles would become clearer and the insurer could put more emphasis on the direct data.

- Unusual risk

  An insurer insuring a small number of yachts of a particular model would not have enough direct data for this model of yacht to set an appropriate premium rate. The insurer could use past claims experience from similar types of boats as collateral data. The insurer may never have enough experience for this particular model to assess the risk purely on the basis of the direct data.

- Experience rating

  An insurer insuring a fleet of motor vehicles operated by a medium sized company may wish to charge a premium that is based on the collateral data provided by motor fleets as a whole, but also takes into account the past experience provided by the direct data for this particular fleet. If the safety record for the company has been good, the company will pay a lower-than-average premium.

2.2 The credibility factor

The credibility factor Z is just a weighting factor. Its value reflects how much 'trust' is placed in the data from the risk itself, $\bar{X}$, compared with the data from the larger group, μ, as an estimate of next year's expected aggregate claims or number of claims – the higher the value of Z, the more trust is placed in $\bar{X}$ compared with μ, and vice versa. This idea will be clarified by going back to the simple example in Section 2.1.

Suppose that data from the particular fleet of buses under consideration had been available for more than just five years. For example, suppose that the estimate of the aggregate claims in the coming year based on data from this fleet itself were £1,600, as before, but that this is now based on ten years' data rather than just five. In this case, the figure of £1,600 is considered more trustworthy than the figure of £2,500, and this means giving the credibility factor a higher value, say 0.75 rather than 0.6. The resulting credibility estimate of the aggregate claims would be £1,825.

Now suppose the figure of £1,600 is based on just five years' data, but the figure of £2,500 is based only on data from bus fleets operating in towns of roughly the same size as the one under consideration, *ie* it no longer includes data from large cities or rural areas. (It is still assumed that the figure of £2,500 is based on considerably more data than the figure of £1,600.) In this case the collateral data would be regarded as more relevant than it was in Section 2.1 and so the credibility factor would be correspondingly reduced, for example to 0.4 from 0.6 giving a credibility premium of £2,140.

The models discussed in this chapter do not allow any scope for this kind of subjective adjustment.

Finally, suppose the situation is exactly as in Section 2.1 except that the figure of £2,500 is based only on data from bus fleets operating in London and Glasgow. In this case the collateral data might be regarded as less relevant than in Section 2.1 and so the credibility factor would be correspondingly increased, for example to 0.8 from 0.6, giving a credibility premium of £1,780.

So the amount of the collateral data is also a factor. If there is a great deal of (relevant) collateral data, the credibility factor may be reduced to allow for this.

From these simple examples it can be seen that, in general terms, the credibility factor in formula (14.2.1) would be expected to behave as follows:

- **The more data there are from the risk itself, the higher should be the value of the credibility factor.**

- **The more relevant the collateral data, the lower should be the value of the credibility factor.**

One final point to be made about the credibility factor is that, while its value should reflect the amount of data available from the risk itself, its value should not depend on the actual data from the risk itself, *ie* on the value of $\bar{X}$. If Z were allowed to depend on $\bar{X}$ then any estimate of the aggregate claims/number of claims, say ϕ, taking a value between $\bar{X}$ and μ could be written in the form of (14.2.1) by choosing Z to be equal to $(\phi - \mu)/(\bar{X} - \mu)$.

This is easily verified. Setting $Z = \dfrac{\phi - \mu}{\bar{X} - \mu}$, we see that:

$$
\begin{aligned}
Z\bar{X} + (1-Z)\mu &= \left(\frac{\phi - \mu}{\bar{X} - \mu}\right)\bar{X} + \left(1 - \frac{\phi - \mu}{\bar{X} - \mu}\right)\mu \\[2mm]
&= \left(\frac{\phi - \mu}{\bar{X} - \mu}\right)\bar{X} + \left(\frac{\bar{X} - \phi}{\bar{X} - \mu}\right)\mu \\[2mm]
&= \frac{\phi\bar{X} - \mu\bar{X}}{\bar{X} - \mu} + \frac{\bar{X}\mu - \phi\mu}{\bar{X} - \mu} \\[2mm]
&= \frac{\phi\bar{X} - \phi\mu}{\bar{X} - \mu} \\[2mm]
&= \phi
\end{aligned}
$$

The problems remain of how to measure the relevance of collateral data and how to calculate the credibility factor Z. There are two approaches to these problems: Bayesian credibility and empirical Bayes credibility theory.

The first of these is covered in the remainder of this chapter. The second is discussed in .

3   Bayesian credibility

3.1   Introduction

The Bayesian approach to credibility involves the same steps as Bayesian estimation, described in the last chapter:

- We start with a prior distribution for the unknown parameter under consideration (*eg* the claim frequency), which summarises any knowledge we have about its possible values. The form of the prior distribution should be derived from information provided by the collateral data.

- We then collect relevant data and use these values to obtain the likelihood function.

- The prior distribution and likelihood function are combined to produce the posterior distribution.

- A loss function is specified to quantify how serious misjudging the parameter value would be. The loss function should be based on commercial considerations of the financial effect on the insurer's business of incorrectly estimating the parameter (and hence the premium rates).

- The Bayesian estimate of the parameter value is then calculated.

This study of the Bayesian approach to credibility theory is illustrated by considering two models: the Poisson/gamma model and the normal/normal model.

The Poisson/gamma model can be used to estimate claim frequencies. The normal/normal model can be used to estimate aggregate (*ie* total) claim amounts. Other Bayesian models are included in the practice questions at the end of this chapter.

3.2   The Poisson/gamma model

Suppose the claim frequency for a risk, *ie* the expected number of claims in the coming year, needs to be estimated. The problem can be summarised as follows.

The number of claims each year is assumed to have a Poisson distribution with parameter λ.

The value of λ is not known, but estimates of its value are possible along the lines of, for example, 'there is a 50% chance that the value of λ is between 50 and 150'.

More precisely, before having available any data from this risk, the feeling about the value of λ is that it has a *Gamma*(α, β) distribution.

The gamma distribution is the conjugate prior for Poisson data. So, if the number of claims each year has a Poisson distribution with parameter λ and we use a gamma distribution as the prior distribution for λ, the posterior distribution for λ will also be a gamma distribution.

Data from this risk are now available showing the number of claims arising in each of the past *n* years.

This problem fits exactly into the framework of Bayesian statistics and can be summarised more formally as follows.

The random variable X represents the number of claims in the coming year from a risk.

The distribution of X depends on the fixed, but unknown, value of a parameter, λ.

The conditional distribution of X given λ is *Poisson*(λ).

The prior distribution of λ is *Gamma*(α, β).

So the PDF of the prior distribution is:

$$f_{prior}(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0$$

$x_1, x_2, \ldots, x_n$ are past observed values of X. (For convenience these data will be denoted $\underline{x}$.)

The likelihood function obtained from these data values is:

$$L = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \propto e^{-n\lambda} \lambda^{\sum x_i}$$

The problem is to estimate λ given the data $\underline{x}$, and the estimate wanted is the Bayes estimate with respect to quadratic loss, *ie* $E(\lambda \mid \underline{x})$.

Combining the prior distribution and the sample data, we see that:

$$f_{post}(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \times e^{-n\lambda} \lambda^{\sum x_i} = \lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda}, \quad \lambda > 0$$

The posterior distribution of λ given $\underline{x}$ is *Gamma*$\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)$.

This can be seen by comparing the formula for $f_{post}(\lambda)$ with the formula for the PDF of the gamma distribution, which is given on page 12 of the *Tables*.

Under quadratic loss, the Bayesian estimate of the unknown parameter λ is the mean of the posterior distribution, *ie*:

$$E(\lambda \mid \underline{x}) = \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + n} \tag{14.3.1}$$

Looking a little more closely at formula (14.3.1), the observed mean number of claims is $\sum_{i=1}^{n} x_i / n$ and the mean number based on prior beliefs is the mean of the prior gamma distribution, α/β.

We can rearrange the formula given in (14.3.1) for the posterior mean to show both the sample mean and the prior mean explicitly:

$$E(\lambda \mid \underline{x}) = \frac{\sum x_i + \alpha}{\beta + n} = \frac{\sum x_i}{\beta + n} + \frac{\alpha}{\beta + n} = \frac{n}{\beta + n} \times \frac{\sum x_i}{n} + \frac{\beta}{\beta + n} \times \frac{\alpha}{\beta}$$

In other words, **the expression for the posterior mean can be written in the credibility form using the rearrangement:**

$$E(\lambda \mid \underline{x}) = Z \frac{\sum\limits_{i=1}^{n} x_i}{n} + (1 - Z) \frac{\alpha}{\beta} \qquad\qquad \textbf{(14.3.2)}$$

where:

$$Z = \frac{n}{\beta + n} \qquad\qquad \textbf{(14.3.3)}$$

Recall that, if we have a random sample of size n from a Poisson distribution with parameter λ, the maximum likelihood estimate of λ is the sample mean, $\overline{x} = \dfrac{\sum x_i}{n}$. Formula (14.3.2) shows that the Bayesian estimate of λ is a weighted average of the maximum likelihood estimate of λ and the prior mean. The weighting factor Z applied to the maximum likelihood estimate is $\dfrac{n}{\beta + n}$. This is the credibility factor.

Now suppose no data were available from the risk itself (*ie* suppose $n = 0$) so that $Z = 0$. The only information available to help to estimate λ would be its prior distribution. The best estimate of λ would then be the mean of the prior distribution, which is α / β.

On the other hand, if only the data from the risk itself were available to estimate λ, the obvious estimate would be $\sum\limits_{i=1}^{n} x_i / n$. (This is the maximum likelihood estimate of λ.)

Notice that this estimate, which plays the role of $\overline{X}$ in the credibility premium formula (14.2.1), is a linear function of the observed values, $x_1, x_2, \dots, x_n$.

The value of Z depends on the amount of data available for the risk, n, and the collateral information, through β, which is consistent with what was said at the end of Section 2.2.

As n increases the sampling error of $\sum\limits_{i=1}^{n} x_i / n$ as an estimate for λ decreases.

Similarly, β reflects the variance of the prior distribution for λ. Thus Z reflects the relative reliability of the two alternative estimates of λ.

Bearing these three points in mind, the Bayesian estimate of λ given by (14.3.2) is in the form of a weighted average of the estimate of λ based solely on data from the risk itself and an estimate of λ based on some other information. This is precisely the form of the credibility estimate given by formula (14.2.1) (with 'collateral data' now being given a more precise interpretation as a prior distribution). Notice that for this model the credibility factor, Z, is no longer the rather vaguely defined quantity it was in Section 2.2; it is given precisely by formula (14.3.3).

3.3 Numerical illustrations of the Poisson/gamma model

In this section some simple numerical examples relating to the Poisson/gamma model will be considered which will help to make some further points about this model.

The set-up and the problem are exactly as in Section 3.2. The value of λ is 150, so that the number of claims arising from the risk each year has a *Poisson*(150) distribution. In practice the true value of λ will not be known. When working through this example, all that will be assumed to be known will be a prior distribution for λ, which, for the moment, will be taken to be *Gamma*(100,1). (Note that this distribution has mean 100 and standard deviation 10.) The actual number of claims arising each year from this risk are as follows:

Year	Number of Claims
1	144
2	144
3	174
4	148
5	151
6	156
7	168
8	147
9	140
10	161

Let's start by considering how the credibility factor changes from one year to the next. The formula for the credibility factor for this model is:

$$Z = \frac{n}{n+\beta} = \frac{n}{n+1}$$

where n represents the number of years of past data.

At the start of Year 1, we have no past data. So the credibility factor is zero at the outset.

At the start of Year 2, we have one year of past data. So:

$$Z = \frac{1}{1+1} = \frac{1}{2}$$

In subsequent years Z takes the values $\frac{2}{3}$, $\frac{3}{4}$ and so on.

Figure 1 below shows the credibility factor in successive years:

Figure 1

It can be seen from Figure 1 that the credibility factor increases with time. As time goes by, more data are collected from the risk itself and, the more data there are from the risk itself, the higher the credibility factor should be because of the increasing reliability of the data in estimating the true but unknown expected number of claims for the risk. (Mathematically, the fact that Z increases with time for this particular model is simply because $n/(\beta + n)$ is an increasing function of n for any positive value of β.)

Figure 2 below shows the credibility estimate of the number of claims in successive years for the *Gamma*(100,1) prior distribution for λ.

The estimated claim number figures are obtained by applying the relevant credibility factor to the data obtained to date.

At the start of Year 1, we have no past data. So the estimated number of claims is equal to the prior mean, which is 100.

At the start of Year 2, the estimated number of claims for the year is:

$$\frac{1}{2} \times 144 + \frac{1}{2} \times 100 = 122$$

Similarly, at the start of Year 3, the estimated number of claims for the year is:

$$\frac{2}{3} \times 144 + \frac{1}{3} \times 100 = 129.3$$

The figure of 144 is the mean number of claims per year in the first two years.

The credibility estimates for the other years are calculated in a similar way.

Figure 2 also shows the actual claim numbers (as given in the table at the start of this section).

Figure 2

An observation can be made about the estimates of the number of claims shown in Figure 2. The initial estimate is 100, the mean of the prior distribution of λ. However, this turns out to be a very poor estimate since all the actual claim numbers are around 150, and none, in the first ten years, is lower than 140. The graph shows the estimated number of claims increasing with time until it reaches the level of the actual claim numbers after eight years. This increase is due to progressively more weight, *ie* credibility, being given to the data from the risk itself and correspondingly less weight being given to the collateral data, *ie* the prior distribution of λ.

The estimated number of claims for Year 9 is slightly lower than the estimated number of claims for Year 8. This is because the actual figure for Year 8 is relatively small and so drags down the credibility estimate.

Now suppose the prior distribution of λ is *Gamma*(500,5) rather than *Gamma*(100,1).

The graphs below (Figures 3 and 4) show the credibility factors and the estimated numbers of claims in successive years, respectively, for the *Gamma*(500,5) prior, in each case with the values for the *Gamma*(100,1) prior shown for comparison.

Figure 3

Figure 4

The credibility factor and the estimated number of claims for the *Gamma*(500,5) prior have the same general features as for the *Gamma*(100,1) prior, *ie* an increasing credibility factor and increasing estimated number of claims.

The most obvious difference between these two cases is that the credibility factor increases more slowly for the *Gamma*(500,5) prior than it does for the *Gamma*(100,1) prior. (Mathematically, this is a simple consequence of a higher β value for the former than for the latter. See (14.3.3).)

Formula (14.3.3) says that $Z = \dfrac{n}{n+\beta}$. In the case of the *Gamma*(500,5) prior, $Z = \dfrac{n}{n+5}$. In the case of the *Gamma*(100,1) prior, $Z = \dfrac{n}{n+1}$.

This feature can be explained in terms of credibility theory as follows:

- The estimated number of claims based on collateral data, *ie* the mean of the prior distribution for λ, is the same for both prior distributions, *ie* 100.

- However, the standard deviation of the prior distribution is lower for the *Gamma*(500,5) prior, *ie* $\sqrt{20} = 4.472$, than for the *Gamma*(100,1) prior, *ie* 10.

- The size of the standard deviation of the prior distribution can be interpreted as an indication of how much confidence is placed in the initial estimate of the number of claims; the smaller the standard deviation of the prior distribution, the more reliable this initial estimate is believed to be.

- Since, in Bayesian credibility, the prior distribution represents the 'collateral data', the above statement can be expressed as 'the smaller the standard deviation of the prior distribution, the more relevant the collateral data are considered to be'.

- Given this interpretation, a smaller standard deviation for the prior distribution would be expected to result in a lower credibility factor since, as stated in Section 2.2, the more relevant the collateral data, the lower should be the value of the credibility factor.

There is one last, but not unimportant, general point to make about what has been done so far in Section 3. The problem has been to estimate the expected number of claims in the coming year. Since X represents the number of claims in the coming year, this could be naïvely interpreted as being required to estimate $E(X)$. This can be calculated very easily as follows:

$$E(X) = E\big[E(X \mid \lambda)\big] = E(\lambda) = \alpha / \beta$$

To see that α / β is not a sensible answer to the problem, look at Figure 2. To adopt α / β as the answer to the problem would mean estimating the number of claims each year as 100. This would obviously result in poorer estimates than the credibility estimates shown in Figure 2.

In fact, what is required is $E(X \mid \underline{x})$.

$E(X \mid \underline{x})$ is the expected number of claims next year given the past data. Using the conditional expectation formula:

$$E(X \mid \underline{x}) = E\big[E(X \mid \lambda) \mid \underline{x}\big] = E(\lambda \mid \underline{x})$$

since $X \mid \lambda \sim Poisson(\lambda)$. So $E(X \mid \underline{x})$ is equal to the mean of the posterior distribution of λ.

3.4 The normal/normal model

In this section another model is considered.

The problem in this section is to estimate the pure premium, *ie* the expected aggregate claims, for a risk. Let X be a random variable representing the aggregate claims in the coming year for this risk. The following assumptions are made.

The distribution of X depends on the fixed, but unknown, value of a parameter, θ.

θ is a random variable but its value, once determined, does not vary over time.

The conditional distribution of X given θ is $N(\theta, \sigma_1^2)$.

The uncertainty about the value of θ is modelled in the usual Bayesian way by regarding it as a random variable.

The prior distribution of θ is $N(\mu, \sigma_2^2)$.

So the PDF of the prior distribution is:

$$f_{prior}(\theta) = \frac{1}{\sigma_2\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma_2}\right)^2\right] \propto \exp\left[-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma_2}\right)^2\right], \quad -\infty < \theta < \infty$$

The values of μ, σ_1 and σ_2 are known.

n past values of X have been observed, which will be denoted $x_1, x_2, \ldots, x_n$, or, more briefly, $\underline{x}$.

The likelihood function obtained from these data values is:

$$L = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_i-\theta}{\sigma_1}\right)^2\right] \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma_1}\right)^2\right]$$

If the value of θ were known, the correct pure premium for this risk would be $E(X\,|\,\theta)$, which, from the fact that the conditional distribution of X given θ is $N(\theta, \sigma_1^2)$, is θ. The problem then is to estimate $E(X\,|\,\theta)$ given $\underline{x}$ and, as in the Poisson/gamma model, the Bayesian estimate with respect to quadratic loss will be used. This means that the estimate will be $E[E(X\,|\,\theta)\,|\,\underline{x}]$, which is the same as $E(\theta\,|\,\underline{x})$.

This is because $X\,|\,\theta$ has a $N(\theta, \sigma_1^2)$ distribution, so $E(X\,|\,\theta) = \theta$.

To determine the Bayesian estimate, we must first obtain the posterior distribution of θ. Combining the prior distribution and the sample data, we see that for $-\infty < \theta < \infty$:

$$f_{post}(\theta) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma_2}\right)^2\right] \times \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma_1}\right)^2\right]$$

$$= \exp\left[-\frac{1}{2}\left[\left(\frac{\theta-\mu}{\sigma_2}\right)^2 + \sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma_1}\right)^2\right]\right]$$

Since the exponent in this expression is a quadratic function of θ:

$$f_{post}(\theta) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta-\mu_*}{\sigma_*}\right)^2\right]$$

This has the form of the PDF of a $N(\mu_*, \sigma_*^2)$ distribution, for some values of μ_* and σ_*^2.

To determine the values of μ_* and σ_*^2, we compare the coefficients of both θ^2 and θ in the exponents. Because of the proportionality, any terms that don't involve θ are not important.

We have:

$$\left(\frac{\theta - \mu_*}{\sigma_*}\right)^2 = \left(\frac{\theta - \mu}{\sigma_2}\right)^2 + \sum_{i=1}^{n}\left(\frac{x_i - \theta}{\sigma_1}\right)^2 + C$$

for some constant C

Multiplying out both sides, this becomes:

$$\frac{\theta^2 - 2\theta\mu_* + \mu_*^2}{\sigma_*^2} = \frac{\theta^2 - 2\theta\mu + \mu^2}{\sigma_2^2} + \sum_{i=1}^{n}\left(\frac{x_i^2 - 2\theta x_i + \theta^2}{\sigma_1^2}\right) + C$$

Equating the coefficients of θ^2, we see that:

$$\frac{1}{\sigma_*^2} = \frac{1}{\sigma_2^2} + \frac{n}{\sigma_1^2} \qquad (1)$$

and equating the coefficients of θ gives:

$$\frac{\mu_*}{\sigma_*^2} = \frac{\mu}{\sigma_2^2} + \frac{1}{\sigma_1^2}\sum_{j=1}^{n}x_j = \frac{\mu}{\sigma_2^2} + \frac{n\bar{x}}{\sigma_1^2} \qquad (2)$$

It follows from (1) that:

$$\sigma_*^2 = \frac{1}{\dfrac{1}{\sigma_2^2} + \dfrac{n}{\sigma_1^2}} = \frac{\sigma_1^2\,\sigma_2^2}{\sigma_1^2 + n\,\sigma_2^2}$$

Also, dividing (2) by (1) gives:

$$\mu_* = \frac{\dfrac{\mu}{\sigma_2^2} + \dfrac{n\bar{x}}{\sigma_1^2}}{\dfrac{1}{\sigma_2^2} + \dfrac{n}{\sigma_1^2}} = \frac{\mu\sigma_1^2 + n\sigma_2^2\,\bar{x}}{\sigma_1^2 + n\sigma_2^2}$$

The formulae for μ_* and σ_*^2 are given on page 28 of the *Tables*, although slightly different notation is used there.

So the posterior distribution of θ given $\underline{x}$ is:

$$N\left(\frac{\mu\sigma_1^2 + n\sigma_2^2\,\bar{x}}{\sigma_1^2 + n\sigma_2^2}\,,\,\frac{\sigma_1^2\,\sigma_2^2}{\sigma_1^2 + n\sigma_2^2}\right)$$

where:

$$\bar{x} = \sum_{i=1}^{n} x_i\,/\,n$$

The Bayesian estimate of θ under quadratic loss is the mean of this posterior distribution:

$$E(\theta\mid\underline{x}) = \frac{\mu\sigma_1^2 + n\sigma_2^2\,\bar{x}}{\sigma_1^2 + n\sigma_2^2}$$

$$= \frac{\sigma_1^2}{\sigma_1^2 + n\sigma_2^2}\,\mu + \frac{n\sigma_2^2}{\sigma_1^2 + n\sigma_2^2}\,\bar{x}$$

or:

$$E(\theta\mid\underline{x}) = Z\,\bar{x} + (1-Z)\mu \qquad\qquad\qquad (14.3.4)$$

where:

$$Z = \frac{n}{n + \sigma_1^2\,/\,\sigma_2^2} \qquad\qquad\qquad (14.3.5)$$

Equation (14.3.4) is a credibility estimate of $E(\theta\mid\underline{x})$ since it is a weighted average of two estimates: the first, $\bar{x}$, is a maximum likelihood estimate based solely on data from the risk itself, and the second, μ is the best available estimate if no data were available from the risk itself.

Notice that, as for the Poisson/gamma model, the estimate based solely on data from the risk itself is a linear function of the observed data values.

There are some further points to be made about the credibility factor, Z, given by (14.3.5):

- **It is always between zero and one.**

- **It is an increasing function of n, the amount of data available.**

- **It is an increasing function of σ_2, the standard deviation of the prior distribution.**

These features are all exactly what would be expected for a credibility factor.

Notice also that, as σ_1^2 increases, the denominator increases, and so Z decreases. σ_1^2 denotes the variance of the distribution of the sample values. If this is large, then the sample values are likely to be spread over a wide range, and they will therefore be less reliable for estimation.

3.5   Further remarks on the normal/normal model

In Section 3.4 the normal/normal model for the estimation of a pure premium was discussed within the framework of Bayesian statistics. In this section the same model will be considered, without making any different assumptions, but in a slightly different way.

The reason for doing this is that some of the observations will be helpful when empirical Bayes credibility theory is considered in the next chapter.

In this section, as in Section 3.4, the problem is to estimate the expected aggregate claims produced each year by a risk. Let:

$$X_1, X_2, ..., X_n, X_{n+1}, ...$$

be random variables representing the aggregate claims in successive years. The following assumptions are made.

The distribution of each X_j depends on the value of a fixed, but unknown, parameter, θ.

Again, θ is a random variable whose value, once determined, does not change over time.

The conditional distribution of X_j given θ is $N(\theta, \sigma_1^2)$.

Given θ, the random variables $\{X_j\}$ are independent.

The prior distribution of θ is $N(\mu, \sigma_2^2)$.

The values of $X_1, X_2, ..., X_n$ have already been observed and the expected aggregate claims in the coming, *ie* $(n + 1)$th, year need to be estimated.

It is important to realise that the assumptions and problem outlined above are exactly the same as the assumptions and problem outlined in Section 3.4. Slightly different notation has been used in this section; in Section 3.4, $X_1, ..., X_n$ were denoted $x_1, ..., x_n$ since their values were assumed to be known, and X_{n+1} was denoted just X. The assumptions that the distribution of each X_j depends on θ, that the conditional distribution of X_j given θ is $N(\theta, \sigma_1^2)$, and that the prior distribution of θ is $N(\mu, \sigma_2^2)$ were all made in Section 3.4. The only assumption not made explicitly in Section 3.4 is that, given θ, the random variables $\{X_j\}$ are independent.

Having stressed that everything is the same as in Section 3.4, some consequences of the above assumptions will be considered. Some important consequences are:

Given θ, the random variables $\{X_j\}$ are identically distributed, as well as independent.

This is an immediate consequence of the assumption that given θ, each X_j has the same $N(\theta, \sigma_1^2)$ distribution.

The random variables $\{X_j\}$ are (unconditionally) identically distributed. The following formula can be written down for the unconditional distribution function of X_j:

$$P(X_j \le y) = \int_{-\infty}^{\infty} \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left\{ \frac{-(\theta - \mu)^2}{2\sigma_2^2} \right\} \Phi\left(\frac{y - \theta}{\sigma_1} \right) d\theta$$

This comes from the formula for calculating a marginal probability:

$$P(X_j \le y) = \int_{\theta} P(X_j \le y \mid \theta) f(\theta) d\theta$$

Since θ is assumed to have a $N(\mu, \sigma_2^2)$ distribution:

$$f(\theta) = \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\theta - \mu}{\sigma_2} \right)^2 \right]$$

Also, since $X_j \mid \theta$ is assumed to have a $N(\theta, \sigma_1^2)$ distribution:

$$P(X_j \le y \mid \theta) = P\left(N(0,1) \le \frac{y - \theta}{\sigma_1} \right) = \Phi\left(\frac{y - \theta}{\sigma_1} \right)$$

$\Phi()$ is the standardised normal distribution function.

This expression is the same for each value of j and hence the random variables $\{X_j\}$ are (unconditionally) identically distributed.

The random variables $\{X_j\}$ are not (unconditionally) independent. This can be demonstrated as follows.

Using Equation (14.1.1) and the fact that, given θ, X_1 and X_2 are conditionally independent:

$$
\begin{aligned}
E(X_1 X_2) &= E\big[E(X_1 X_2 \mid \theta) \big] \\
&= E\big[E(X_1 \mid \theta) E(X_2 \mid \theta) \big] \\
&= E(\theta^2) \qquad \text{(since } E(X_1 \mid \theta) = E(X_2 \mid \theta) = \theta) \\
&= \mu^2 + \sigma_2^2
\end{aligned}
$$

The idea used in this argument will be used repeatedly in the next chapter.

Now, if X_1 and X_2 were unconditionally independent:

$$E(X_1 X_2) = E(X_1) E(X_2)$$

However, using (14.1.1) again:

$$E(X_1) = E[E(X_1 \mid \theta)] = E(\theta) = \mu$$

Similarly, $E(X_2) = \mu$. Hence:

$$E(X_1 X_2) = \mu^2 + \sigma_2^2 \neq E(X_1)E(X_2)$$

This shows that X_1 and X_2 are not unconditionally independent. The relationship between X_1 and X_2 is that their means are chosen from a common distribution. If this mean, θ is known, then this relationship is broken and there exists conditional independence.

3.6   Discussion of the Bayesian approach to credibility

This approach has been very successful in the Poisson/gamma and normal/normal models. It has made the notion of collateral data very precise (by interpreting it in terms of a prior distribution) and has given formulae for the calculation of the credibility factor. What, then, are the drawbacks of this approach?

The first difficulty is whether a Bayesian approach to the problem is acceptable, and, if so, what values to assign to the parameters of the prior distribution. For example, although the Poisson/gamma model provides a formula (Equation 14.3.3) for the calculation of the credibility factor, this formula involves the parameter β. How a value for β might be chosen has not been discussed. The Bayesian approach to the choice of parameter values for a prior distribution is to argue that they summarise the subjective degree of belief about the possible values of the quantity to be estimated, for example, the mean claim number, λ, for the Poisson/gamma model.

The second difficulty is that even if the problem fits into a Bayesian framework, the Bayesian approach may not work in the sense that it may not produce an estimate which can readily be rearranged to be in the form of a credibility estimate. This point can be illustrated by using a uniform prior with a Poisson distribution for the number of claims.

Suppose that $X_1, \ldots, X_n$ is a random sample from a Poisson distribution with mean λ. If we assume that the prior distribution of λ is $U(0, \beta)$, then the prior PDF is constant for $0 < \lambda < \beta$, and the posterior PDF is proportional to the likelihood function, *ie*:

$$f_{post}(\lambda) \propto \prod_{i=1}^{n} P(X_i = x_i)$$

$$= \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i !}$$

$$\propto e^{-n\lambda} \lambda^{\sum x_i}$$

for $0 < \lambda < \beta$. The posterior PDF is 0 for other values of λ (since the prior PDF is 0 unless $0 < \lambda < \beta$). The posterior distribution of λ is not one of the standard distributions listed in the *Tables* (in fact it is a *truncated gamma distribution*) so we can't look up the formula for its mean or easily deduce whether it can be expressed in the form of a credibility estimate.

Chapter 14 Summary

Direct and collateral data

Claim numbers and aggregate claim amounts can be estimated using a combination of direct data (*ie* data from the risk under consideration) and collateral data (*ie* data from other similar, but not necessarily identical, risks).

Credibility premium and credibility factor

Let $\overline{x}$ be an estimate of the expected number of claims / aggregate claim amount for a particular risk for the coming year based on direct data.

Let μ be an estimate of the expected number of claims / aggregate claims for a particular risk for the coming year based on collateral data.

Then the credibility premium (or credibility estimate of the number of claims / aggregate claim amount) for this risk is:

$$Z\overline{x} + (1 - Z)\mu$$

where Z is a number between 0 and 1 and is known as the credibility factor. The closer it is to 1, the more weight is given to the direct data.

Bayesian credibility

In Bayesian credibility, the collateral data take the form of a prior distribution for a parameter, θ, say. As in Bayesian statistics, the problem is to calculate the Bayesian estimate of θ using quadratic loss, *ie* to estimate the mean of the posterior distribution of θ.

The resulting estimate is a credibility estimate if it can be written in the form:

$$Z \times \text{MLE of } \theta + (1 - Z) \times \text{prior mean}$$

Poisson/gamma model

The Poisson/gamma model can be used to estimate claim frequencies. The model assumes that the number of claims follows a Poisson distribution with parameter λ and the prior distribution for λ is $Gamma(\alpha, \beta)$. Given sample claims data $x_1, ..., x_n$, the resulting posterior distribution is $Gamma(\alpha + \sum x_i, \beta + n)$. The estimated claim frequency is the mean of the posterior distribution, which is easily shown to be a credibility estimate as:

$$\frac{\alpha + \sum x_i}{\beta + n} = \frac{n}{\beta + n} \times \frac{\sum x_i}{n} + \frac{\beta}{\beta + n} \times \frac{\alpha}{\beta}$$

is of the form $Z \times \text{MLE of } \theta + (1 - Z) \times \text{prior mean}$. The credibility factor, Z, is $\dfrac{n}{\beta + n}$.

Normal/normal model

The normal/normal model can be used to estimate aggregate claim amounts. The model assumes that the aggregate claim amount follows a normal distribution with mean μ and the prior distribution for μ is $N(\mu_0, \sigma_0^2)$. Given sample claims data $x_1, \ldots, x_n$, the resulting posterior distribution is another normal distribution. Formulae for the mean and variance of this posterior distribution are given on page 28 of the *Tables*. The estimated aggregate claim amount is the mean of the posterior distribution:

$$\mu_* = \frac{\dfrac{n\overline{x}}{\sigma^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}}$$

Again this clearly has the form of a credibility estimate. The credibility factor is the coefficient of $\overline{x}$ (the maximum likelihood estimate of μ):

$$Z = \frac{\dfrac{n}{\sigma^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}} = \frac{n}{n + \sigma^2/\sigma_0^2}$$

Chapter 14 Practice Questions

14.1   An insurer is setting the premium rate for the buildings in an industrial estate. Past experience for the estate indicates that a premium rate of £3 per £1,000 sum insured should be charged. The past experience of other similar estates for which the insurer provides cover indicates a premium rate of £5 per £1,000 sum insured. If the insurer uses a credibility factor of 75% for this risk, calculate the premium rate per £1,000 sum insured.

14.2   Claim amounts on a portfolio of insurance policies have an unknown mean μ. Prior beliefs about μ are described by a distribution with mean μ_0 and variance σ_0^2. Data are collected from n claims with mean claim amount $\bar{x}$ and variance s^2. A credibility estimate of μ is to be made, of the form:

$$Z\bar{x} + (1-Z)\mu_0$$

Suggestions for the choice of Z are:

A $\dfrac{n\sigma_0^2}{n\sigma_0^2 + s^2}$

B $\dfrac{n\sigma_0^2}{n\sigma_0^2 + n}$

C $\dfrac{\sigma_0^2}{n + \sigma_0^2}$

Explain whether each suggestion is an appropriate choice for Z.   [4]

14.3   The total claim amount per annum on a particular insurance policy follows a normal distribution with unknown mean θ and variance 200^2. Prior beliefs about θ are described by a normal distribution with mean 600 and variance 50^2. Claim amounts $x_1, x_2, \ldots, x_n$ are observed over n years.

(i)   State the posterior distribution of θ.   [2]

(ii)   Show that the mean of the posterior distribution of θ can be written in the form of a credibility estimate.   [3]

Now suppose that $n = 5$ and that total claims over the five years were 3,400.

(iii)   Calculate the posterior probability that θ is greater than 600.   [2]

[Total 7]

14.4   A statistician wishes to obtain a Bayesian estimate of the mean of an exponential distribution with

Exam style   density function $f(x) = \dfrac{1}{\mu} e^{-x/\mu}$. He is proposing to use a prior distribution with PDF:

$$f(\mu) = \frac{\theta^{\alpha} e^{-\theta/\mu}}{\mu^{\alpha+1} \Gamma(\alpha)} , \qquad \mu > 0$$

The mean of this distribution is $\dfrac{\theta}{\alpha - 1}$.

(i)   Write down the likelihood function for μ , based on observations $x_1, \ldots, x_n$ from an
        exponential distribution.   [1]

(ii)   Determine the form of the posterior PDF for μ , and hence show that an expression for
        the Bayesian estimate for μ under squared error loss is:

$$\hat{\mu} = \frac{\theta + \sum x_i}{n + \alpha - 1}$$ [3]

(iii)   Show that the Bayesian estimate for μ can be written in the form of a credibility
        estimate, giving a formula for the credibility factor.   [3]

(iv)   The statistician decides to use a prior distribution of this form with parameters $\theta = 40$
        and $\alpha = 1.5$. You are given the following summary statistics from the sample data:

$$n = 100, \quad \sum x_i = 9,826, \text{ and } \sum x_i^2 = 1,200,000$$

        Calculate the Bayesian estimate of μ and the value of the credibility factor.   [2]

(v)   Comment on the results obtained in part (iv).   [1]
                                                                                    [Total 10]

14.5   Let θ denote the proportion of insurance policies in a certain portfolio on which a claim is made.

Exam style   Prior beliefs about θ are described by a beta distribution with parameters α and β .

Underwriters are able to estimate the mean μ and variance σ^2 of θ .

(i)   Express α and β in terms of μ and σ .   [3]

A random sample of n policies is taken and it is observed that claims had arisen on d of them.

(ii)   (a)   Determine the posterior distribution of θ .

        (b)   Show that the mean of the posterior distribution can be written in the form of a
                credibility estimate.   [5]

(iii)   Show that the credibility factor increases as σ increases.   [3]

(iv)   Comment on the result in part (iii).   [1]
                                                                                    [Total 12]

Chapter 14 Solutions

14.1   The credibility premium (per £1,000 sum insured) is:

$$0.75 \times 3 + 0.25 \times 5 = £3.50$$

14.2   *This question is Subject CT6, September 2008, Question 1*

First, let's consider what happens when n increases. In A, Z increases. In B, Z is unaffected by n. In C, Z decreases. In practice, we want Z to increase as n increases, so B and C are inappropriate.   [1]

Now consider what happens when σ_0^2 (the variance of the prior distribution) increases. In all cases Z increases. In practice, we want Z to increase, so all expressions are appropriate in this respect.   [1]

Finally, consider what happens when s^2 (the sample variance) increases. In A, Z decreases. In B and C, Z is unaffected by s^2. In practice, we want Z to decrease, so B and C are inappropriate.
[1]

Overall, therefore, only the expression in A is an appropriate choice for Z.   [1]
[Total 4]

14.3   *This question is Subject CT6, April 2012, Question 5.*

(i)   *Posterior distribution*

The prior distribution of θ is $N(600, 50^2)$ and the sample data come from a $N(\theta, 200^2)$ distribution. From page 28 of the *Tables*, the posterior distribution of θ is:

$$N\left(\frac{\dfrac{n\overline{x}}{200^2} + \dfrac{600}{50^2}}{\dfrac{n}{200^2} + \dfrac{1}{50^2}}, \frac{1}{\dfrac{n}{200^2} + \dfrac{1}{50^2}} \right)$$ [2]

(ii)   *Credibility estimate*

The mean of the posterior distribution is:

$$\frac{\dfrac{n\overline{x}}{200^2} + \dfrac{600}{50^2}}{\dfrac{n}{200^2} + \dfrac{1}{50^2}} = \frac{\dfrac{n}{200^2}}{\dfrac{n}{200^2} + \dfrac{1}{50^2}} \overline{x} + \frac{\dfrac{1}{50^2}}{\dfrac{n}{200^2} + \dfrac{1}{50^2}} 600$$ [1]

This is a credibility estimate as it is of the form:

$$Z \times \text{MLE of } \theta + (1 - Z) \times \text{prior mean}$$ [1]

In this case:

- $$Z = \frac{\dfrac{n}{200^2}}{\dfrac{n}{200^2} + \dfrac{1}{50^2}}$$

- the maximum likelihood estimate of θ is $\overline{x}$

- the prior mean is 600.   [1]
                                                                                                                        [Total 3]

(iii)   *Posterior probability*

If $n = 5$ and $\displaystyle\sum_{i=1}^{5} x_i = 3,400$, then the posterior distribution of θ is $N(619.0476, 1904.7619)$.   [1]

So:

$$P(\theta \mid \underline{x} > 600) = P\left(Z > \frac{600 - 619.0476}{\sqrt{1,904.7619}} \right)$$

$$= 1 - P(Z < -0.4364)$$

$$= 1 - \Phi(-0.4364)$$

$$= \Phi(0.4364)$$

$$= 0.669$$ [1]
                                                                                                                        [Total 2]

14.4   (i)   *Likelihood function*

The likelihood function is:

$$L(\mu) = \frac{1}{\mu} e^{-x_1/\mu} \times \cdots \times \frac{1}{\mu} e^{-x_n/\mu} = \frac{e^{-\sum x_i/\mu}}{\mu^n}$$ [1]

(ii)   *Posterior PDF and Bayesian estimate*

The posterior PDF is proportional to the prior PDF multiplied by the likelihood function. So:

$$f_{post}(\mu) \propto \frac{e^{-\theta/\mu}}{\mu^{\alpha+1}} \times \frac{e^{-\sum x_i/\mu}}{\mu^n} = \frac{e^{-(\theta + \sum x_i)/\mu}}{\mu^{n+\alpha+1}}$$ [1]

This has the same form as the prior distribution, but with different parameters. So we have the same distribution as before, but with parameters:

$$\alpha^* = n + \alpha \quad \text{and} \quad \theta^* = \theta + \sum x_i$$ [1]

Using the formula for the mean given in the question:

$$E(\mu \mid \underline{x}) = \frac{\theta^*}{\alpha^* - 1} = \frac{\theta + \sum x_i}{n + \alpha - 1}$$ [1]

This is the Bayesian estimate for μ under squared error loss.   [Total 3]

(iii)   **Credibility estimate**

Splitting the formula for the posterior mean into two parts, we see that:

$$E(\mu \mid \underline{x}) = \frac{\theta + \sum x_i}{n + \alpha - 1}$$

$$= \frac{\theta}{n + \alpha - 1} + \frac{\sum x_i}{n + \alpha - 1}$$

$$= \frac{n}{n + \alpha - 1} \times \frac{\sum x_i}{n} + \frac{\alpha - 1}{n + \alpha - 1} \times \frac{\theta}{\alpha - 1}$$ [1]

This is a weighted average of the maximum likelihood estimate of μ (which is the sample mean $\frac{\sum x_i}{n}$) and the mean of the prior distribution ($\frac{\theta}{\alpha - 1}$). So it is a credibility estimate.   [1]

The credibility factor is:

$$Z = \frac{n}{n + \alpha - 1}$$ [1]

[Total 3]

(iv)   **Numerical values of estimate and credibility factor**

Using the given figures, the Bayesian estimate of μ is:

$$\frac{\theta + \sum x_i}{n + \alpha - 1} = \frac{40 + 9,826}{100 + 1.5 - 1} = 98.1692$$ [1]

and the value of the credibility factor is:

$$Z = \frac{n}{n + \alpha - 1} = \frac{100}{100 + 1.5 - 1} = 0.9950$$ [1]

[Total 2]

(v)   **Comment**

The value of Z is very close to 1. So the credibility estimate is very close to the sample mean (98.26), and takes little account of the prior mean (80). This is because n is much bigger than α.

[1]

14.5   *This question is Subject CT6, April 2014, Question 11.*

(i)   **Formulae for beta parameters**

We are told that:

$$\theta \sim Beta(\alpha, \beta) \qquad E(\theta) = \mu \qquad var(\theta) = \sigma^2$$

Using the formulae given on page 13 of the *Tables*, we see that:

$$\mu = \frac{\alpha}{\alpha + \beta} \qquad\qquad (1)$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \qquad\qquad (2)$$

Rearranging equation (1):

$$\alpha\mu + \beta\mu = \alpha \quad \Rightarrow \quad \beta\mu = \alpha(1 - \mu) \quad \Rightarrow \quad \beta = \frac{\alpha}{\mu}(1 - \mu) \qquad\qquad [½]$$

The RHS of equation (2) contains a factor of $\dfrac{\alpha}{\alpha + \beta}$, which is equal to μ, and also contains a

factor of $\dfrac{\beta}{\alpha + \beta}$, which is equal to $1 - \mu$. So:

$$\sigma^2 = \frac{\mu(1 - \mu)}{\alpha + \beta + 1} \qquad\qquad [1]$$

Substituting for β in the denominator then gives:

$$\sigma^2 = \frac{\mu(1 - \mu)}{\alpha + \dfrac{\alpha(1 - \mu)}{\mu} + 1} = \frac{\mu^2(1 - \mu)}{\alpha\mu + \alpha(1 - \mu) + \mu} = \frac{\mu^2(1 - \mu)}{\alpha + \mu} \qquad\qquad [½]$$

So: $\alpha + \mu = \dfrac{\mu^2(1 - \mu)}{\sigma^2}$

Hence:

$$\alpha = \frac{\mu^2(1 - \mu)}{\sigma^2} - \mu \qquad\qquad [½]$$

and:

$$\beta = \frac{\alpha}{\mu}(1 - \mu) = \left(\frac{\mu(1 - \mu)}{\sigma^2} - 1\right)(1 - \mu) \qquad\qquad [½]$$

                                                                                        [Total 3]

(ii)(a)   *Posterior distribution*

We know that the prior distribution of θ is *Beta*(α, β), so:

$$f_{prior}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1$$

Let X be the number of policies from a sample of size n on which a claim is made. Then $X \mid \theta \sim Binomial(n, \theta)$.

We have observed d policies with claims, so the likelihood function is:

$$L(\theta) = P(X = d \mid \theta) = \binom{n}{d} \theta^d (1-\theta)^{n-d} \propto \theta^d (1-\theta)^{n-d} \qquad [1]$$

Combining the prior distribution and the sample data, we see that:

$$f_{post}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^d (1-\theta)^{n-d}$$

$$= \theta^{\alpha+d-1}(1-\theta)^{\beta+n-d-1} \qquad [1]$$

for $0 < \theta < 1$. Comparing this with the PDF of the Beta distribution given on page 13 of the *Tables*, we see that the posterior distribution of θ is *Beta*$(\alpha + d, \beta + n - d)$. $\qquad [1]$

(ii)(b)   *Credibility estimate*

The mean of the posterior distribution is:

$$\frac{\alpha + d}{(\alpha + d) + (\beta + n - d)} = \frac{\alpha + d}{\alpha + \beta + n} \qquad [\tfrac{1}{2}]$$

Rearranging:

$$\frac{\alpha + d}{\alpha + \beta + n} = \frac{d}{\alpha + \beta + n} + \frac{\alpha}{\alpha + \beta + n}$$

$$= \left(\frac{n}{\alpha + \beta + n}\right) \times \frac{d}{n} + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \times \frac{\alpha}{\alpha + \beta}$$

This is a credibility estimate as it is of the form:

$$Z \times \text{MLE of } \theta + (1 - Z) \times \text{prior mean} \qquad [\tfrac{1}{2}]$$

In this case:

- $Z = \dfrac{n}{\alpha + \beta + n}$

- the maximum likelihood estimate of θ is $\dfrac{d}{n}$

- the prior mean is $\dfrac{\alpha}{\alpha + \beta}$.   [1]

[Total 5]

(iii)   *Behaviour of the credibility factor*

From equation (1) in part (i), we have:

$$\alpha + \beta = \frac{\alpha}{\mu}$$

So we can rewrite the credibility factor as:

$$Z = \frac{n}{\dfrac{\alpha}{\mu} + n}$$ [1]

Since $\alpha = \dfrac{\mu^2(1-\mu)}{\sigma^2} - \mu$, it follows that:

$$Z = \frac{n}{\dfrac{\mu(1-\mu)}{\sigma^2} - 1 + n}$$ [1]

Increasing σ will reduce the denominator, and hence increase the value of Z .   [1]

[Total 3]

(iv)   *Comment*

Increasing the standard deviation of the prior distribution means that there is greater uncertainty associated with the prior distribution, and hence it is less reliable. So, when estimating θ , we should put less weight on the prior mean and more weight on the maximum likelihood estimate of θ (which is calculated from the sample data alone). To achieve this the credibility factor, Z , must increase. The formula in (iii) illustrates this.   [1]

15

Empirical Bayes Credibility theory

Syllabus objectives

5.1   Explain the fundamental concepts of Bayesian statistics and use these concepts to calculate Bayesian estimates.

    5.1.8   Explain the empirical Bayes approach to credibility theory and use it to derive credibility premiums in simple cases.

    5.1.9   Explain the differences between the two approaches (*ie* the Bayesian approach and the empirical Bayes approach) and state the assumptions underlying each of them.

0   Introduction

In this chapter, we will study Empirical Bayes Credibility Theory (EBCT) Models 1 and 2.

The algebra for the EBCT models is fairly heavy going. Try not to get too bogged down with it. Concentrate on appreciating the differences between the models and being able to follow the numerical examples. Most of the formulae introduced in this chapter are given in the *Tables*.

1   Empirical Bayes Credibility Theory: Model 1

1.1   Introduction

In this chapter we will discuss two Empirical Bayes Credibility Theory (EBCT) models. As with the Poisson/gamma and normal/normal models discussed in the previous chapter, these models can be used to estimate the annual claim frequency or risk premium based on the values recorded over the last n years. Model 1 gives equal weight to each risk in each year. Model 2 is more sophisticated and takes into account the volume of business written under each risk in each year.

The main similarities and differences between the Bayesian models and the EBCT models are outlined below.

Risk parameter

Both approaches involve a risk parameter θ. With the Bayesian approach, the quantity we wish to estimate is θ, whereas, for the EBCT models, it is $m(\theta)$, *ie* a function of θ. Unlike the Bayesian approach, the EBCT models don't assume any specific statistical distribution for θ.

Conditional claim distribution

Like the Bayesian models in the previous chapter, EBCT Model 1 assumes that the random variables $X_j \mid \theta$ are independent and identically distributed. EBCT Model 2 also assumes that the random variables $X_j \mid \theta$ are independent (but does not assume that they are necessarily identically distributed).

Unlike the pure Bayesian approach, the EBCT models don't assume any specific statistical distribution for $X_j \mid \theta$. Instead, formulae are developed using the assumption that the mean $m(\theta)$ and variance $s^2(\theta)$ of $X_j \mid \theta$ can be expressed as functions of θ. The values of $m(\theta)$ and $s^2(\theta)$ are then estimated by the models.

Credibility formula

With both approaches, the resulting formula for estimating the claim frequency or risk premium for a given risk can be expressed using a credibility formula. In other words, the resulting estimate is a weighted average of:

- the mean value of the direct data, *ie* observations that have come from the risk under consideration, and

- an overall mean, which makes use of any collateral data.

Empirical Bayes credibility theory (EBCT) is the name given to a particular approach to the problems in Section 2 of the previous chapter. This approach has led to the development of a vast number of different models of varying degrees of complexity.

In this chapter, two of these models will be studied. In the remainder of Section 1, the simplest possible model will be studied. Although this model, which will be referred to as Model 1, is not very useful in practice, it does provide a good introduction to the principles underlying EBCT. In particular, it shows the similarities and the differences between the empirical Bayes and the pure Bayesian approaches to credibility theory. In Section 2 another model (which is an extension of Model 1) will be studied, which is far more useful in practice.

1.2   Model 1: specification

In this section the assumptions for Model 1 of EBCT will be set out. EBCT Model 1 can be regarded as a generalisation of the normal/normal model. This point will be considered in more detail later in this section.

The problem of interest is the estimation of the pure premium, or possibly the claim frequency, for a risk. Let $X_1, X_2, \ldots$ denote the aggregate claims, or the number of claims, in successive periods for this risk. A more precise statement of the problem is that, having observed the values of $X_1, X_2, \ldots, X_n$, the expected value of X_{n+1} needs to be estimated. From now on $X_1, X_2, \ldots, X_n$ will be denoted by $\underline{X}$.

Assumptions for EBCT Model 1

The following assumptions will be made for EBCT Model 1.

Assumption 1: The distribution of each X_j depends on a parameter, denoted θ , whose value is fixed (and the same for all the X_j s) but is unknown.

The parameter θ is actually a random variable. By saying that its value is fixed, the Core Reading means that the value of this unknown parameter does not change over time.

Assumption 2: Given θ , the X_j 's are independent and identically distributed.

The parameter θ is known as the risk parameter. It could, as in Section 3 of the previous chapter, be a real number or it could be a more general quantity such as a set of real numbers.

The risk parameter in the Poisson/gamma model is (the Poisson parameter) λ . The risk parameter for the normal/normal model is θ .

A consequence of these two assumptions is that the random variables $\left\{X_j\right\}$ are identically distributed.

An important point to note is that the X_j 's are not (necessarily) unconditionally independent.

The above assumptions and consequences were all either made or noted for the normal/normal model of Bayesian credibility in Section 3.4 of the previous chapter.

Next some notation is introduced. Define $m(\theta)$ and $s^2(\theta)$ as follows:

$$m(\theta) = E(X_j \mid \theta)$$

$$s^2(\theta) = \text{var}(X_j \mid \theta)$$

Two things should be noticed about $m(\theta)$ and $s^2(\theta)$. The first is that since, given θ, the X_j's are identically distributed, neither $m(\theta)$ nor $s^2(\theta)$ depends on j, as their notation suggests. The second is that since θ is regarded as a random variable, both $m(\theta)$ and $s^2(\theta)$ are random variables.

If the value of θ and the distribution of X_j given θ were known, the obvious estimate of the expected aggregate claims, or the expected number of claims, in any future year would be $m(\theta)$. Since it is assumed that θ is not known, the problem is to estimate $m(\theta)$ given $\underline{X}$.

Question

A specialist insurer concentrates on providing third party motor insurance to drivers of a certain type of car who are based in London. Explain what each of the variables and functions in EBCT Model 1 would represent if this model were used to describe the numbers of claims made by different drivers in different years. The only risk factor considered relevant is the safety standard adopted by each individual driver.

Solution

Here, the risk parameter θ represents the 'safety coefficient' of a particular driver. It is assumed that each driver has a constant inherent level of safety, which could (in theory) be measured by some means. The value of θ for a particular driver influences the likely number of claims that the driver will make. For example, a value of $\theta = 1$ may correspond to a driver who is '100% safe' (*ie* never makes any claims), while a value of $\theta = 0$ may correspond to a driver who is '0% safe' (*ie* makes a claim every time the car is used). The distribution of the values of θ, which will vary from one driver to the next, is assumed to be a definite (but unknown) probability distribution.

X_j is a random variable, representing the number of claims made by a particular driver in Year j.

$X_j \mid \theta$ represents the number of claims made in Year j by a particular driver whose safety coefficient is θ.

$m(\theta) = E(X_j \mid \theta)$ is the expected number of claims made in a year by a driver whose safety coefficient is θ. In this example, $m(\theta)$ will be a decreasing function of θ, since a higher safety coefficient will reduce the expected number of claims.

$s^2(\theta) = \text{var}(X_j \mid \theta)$ is the variance of the annual number of claims made by a driver whose safety coefficient is θ. $s^2(\theta)$ will take lower values for drivers with a high safety coefficient, since they are likely to make no claims, or possibly one claim, each year, whereas the numbers of claims made each year by drivers with low safety coefficients may range from none to five or more.

The similarities between EBCT Model 1 and the normal/normal model can be summarised as follows:

- **The role of θ is the same for both models: it characterises the underlying distributions of the processes being modelled, *eg* the aggregate claim distribution for each year of business.**

- **Assumptions concerning the unconditional distribution of the X_j's are the same: they are identically distributed in each case.**

- **Assumptions concerning the (conditional) distribution of the X_j's given θ are the same: they are conditionally independent in each case.**

EBCT Model 1 can be regarded as a generalisation of the normal/normal model. The particular points where it differs from, *ie* generalises, the normal/normal model are:

- **$E(X_j \mid \theta)$ is some function of θ, $m(\theta)$, for EBCT Model 1 but is simply θ for the normal/normal model. Hence:**

  $\text{var}[m(\theta)]$ {EBCT} corresponds to σ_2^2 ($=\text{var}(\theta)$) {normal/normal}

- **$\text{var}(X_j \mid \theta)$ is a function of θ, $s^2(\theta)$, for EBCT Model 1 but is a constant, σ_1^2, for the normal/normal model. Hence:**

  $E[s^2(\theta)]$ corresponds to σ_1^2 ($=\text{var}(X_j \mid \theta)$)

- **The normal/normal model makes very precise distributional assumptions about both X_j given θ, which is $N(\theta, \sigma_1^2)$, and θ, which is $N(\mu, \sigma_2^2)$. EBCT Model 1 makes no such distributional assumptions.**

- **The risk parameter, θ, is a real number for the normal/normal model but could be a more general quantity for EBCT Model 1.**

1.3   Model 1: the credibility premium

In the previous section the assumptions for Model 1 of EBCT were studied and the similarities between this model and the normal/normal model were emphasised. In this section a solution to the problem described in the previous section will be considered, that is, the estimation of $m(\theta)$ given the data $\underline{X}$. The derivation of the credibility premium under this model is beyond the scope of the CS1 syllabus and is not covered here.

The estimate of $m(\theta)$ given $\underline{X}$ under EBCT Model 1 is:

$$Z\,\overline{X} + (1-Z)E[m(\theta)]$$

where:

$$\overline{X} = \sum_{j=1}^{n} X_j / n \quad \text{and} \quad Z = \frac{n}{n + E[s^2(\theta)] / \text{var}[m(\theta)]} \tag{15.1.1}$$

The formulae for $\overline{X}$ and Z are given on page 29 of the *Tables*. However the formula $Z\,\overline{X} + (1-Z)E[m(\theta)]$ for estimating $m(\theta)$ given $\overline{X}$ is not given, and so it must be memorised.

The first, and most important, point to note about the solution is that it is in the form of a credibility estimate. In other words, it is in the form of Equation (14.2.1) from the previous chapter with:

$$E[m(\theta)] \text{ playing the role of } \mu$$

and:

$$\sum_{j=1}^{n} X_j / n \text{ playing the role of } \overline{X}$$

The second point to note is the similarity between the solution above and the solution in the normal/normal model, and, in particular, the similarity between the formulae for the credibility factors.

The formula for Z under the normal/normal model (Formula 14.3.5 from the previous chapter) is:

$$Z = \frac{n}{n + \dfrac{\sigma_1^2}{\sigma_2^2}}$$

where $\sigma_1^2 = \text{var}(X_j \mid \theta)$ and σ_2^2 is the variance of the prior distribution of θ.

Formula (15.1.1) is a straight generalisation of this since $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ can be regarded as generalisations of σ_1^2 and σ_2^2, respectively.

We saw this at the end of Section 1.2 of this chapter.

The similarities between EBCT Model 1 and the normal/normal model lead to the similarity in the resulting answers.

The final point to note is that the formula for the credibility estimate given by (15.1.1) involves three parameters, $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$, which have been treated so far as though they were known. EBCT makes no distributional assumptions about the risk parameter θ, unlike the Bayesian approach to credibility, but the values of these three parameters need to be known. It can be noted that these three parameters relate to first and second order moments as opposed to, say, higher order moments or some other quantities.

This is due to the derivation of the credibility premium and the detail behind this is beyond the scope of Subject CS1.

In practice, the true values of $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ will be unknown and will have to be estimated.

The way in which these parameters would be estimated is discussed in the next section.

1.4   Model 1: parameter estimation

In this section the solution to the problem of estimating $m(\theta)$ given $\underline{X}$ will be completed by showing how to estimate $E[m(\theta)]$, $\text{var}[m(\theta)]$ and $E[s^2(\theta)]$.

To do this a further assumption is needed: that there are available data from some other risks similar, but not identical, to the original risk. This requires the problem to be set up a little more generally, to make some extra assumptions and to change the notation slightly. An important distinction between a pure Bayes approach to credibility and EBCT is that no data are required to estimate parameters in the former case but data are required to estimate the parameters in the latter.

What is of interest is estimating the pure premium, or the expected number of claims, for a particular risk, which is one of N risks in a collective. By a collective is meant a collection of different risks that are related in a way to be made clearer later in this section. For simplicity, suppose the particular risk that is of interest is Risk Number 1 in this collective. It is assumed that the aggregate claims, or claim frequencies, for each of these N risks for each of the past n years have been observed. Let X_{ij} denote the aggregate claims, or number of claims, for Risk Number i, $i = 1, 2, …, N$, in Year j, $j = 1, 2, …, n$.

A reminder of this notation is given on page 29 of the *Tables*. It is important not to confuse N (the number of risks) and n (the number of years of data values).

These values are summarised in the following table:

		Year			
		1	2	..	n
	1	X_{11}	X_{12}	..	X_{1n}
	2	X_{21}	X_{22}	..	X_{2n}
Risk Number	.	.	.	..	.
	.	.	.	..	.
	N	X_{N1}	X_{N2}	..	X_{Nn}

Table 1

The risk numbers could refer to risks from different companies, insurers, shops, *etc*.

Each row of Table 1 represents observations for a different risk; the first row, relating to Risk Number 1, is the set of observed values denoted $\underline{X}$ in Section 1.2 and Section 1.3. (Notice that $X_1, X_2, …, X_n$ in Section 1.2 and Section 1.3 have now become $X_{11}, X_{12}, …, X_{1n}$ in this section.)

In Section 1.2, two assumptions were made about the connection between the observed values for the single risk then being considered. In this section exactly the same assumptions for each of the N risks in the collective are made.

For each risk i, $i = 1, 2, …, N$:

Assumption 1a: The distribution of each X_{ij}, $j = 1, 2, …, n$, depends on the value of a parameter, denoted θ_i, whose value is fixed (and the same for each value of j) but is unknown.

The Core Reading means that each θ_i is a random variable whose value does not change over time.

Assumption 2a: Given θ_i, the X_{ij}'s, $j = 1,2,…,n$, are independent and identically distributed.

Notice that the risk parameter, which was denoted θ in Section 1.2 and Section 1.3, is now denoted θ_i, and that an implication of these two assumptions is that the risk parameters for different risks have different values. (However, as in Section 1.2 and Section 1.3, the risk parameter for a given risk does not change value from year to year.)

These two assumptions show something about the relationships within each row of Table 1 but they do not show anything about the relationship between different rows, *ie* between different risks in the collective. The assumption that shows something about the relationship between different risks in the collective is the following.

Assumption 3: For $i \neq k$, the pairs (θ_i, X_{ij}) and (θ_k, X_{km}) are independent and identically distributed.

This assumption shows that the rows of the table are independent of each other.

Two immediate consequences of this assumption are:

- For $i \neq k$, X_{ij} and X_{km} are independent and identically distributed.

- The risk parameters $\theta_1, \theta_2, …, \theta_N$ are independent and identically distributed.

The connection between the different risks, *ie* rows of the table, is a result of the assumption that the risk parameters, $\theta_1, \theta_2, …, \theta_N$, are identically distributed. Intuitively, this means that if, by some means, the values of $\theta_2, \theta_3, …, \theta_N$ were known, then something about the common distribution of the θ_i's would be known and hence something about θ_1, or, at least, about the distribution it comes from, would be known.

The functions $m(\ )$ and $s^2(\ )$ were introduced in Section 1.2. Keeping the same definitions for these functions and applying them to all the risks in the collective:

$$m(\theta_i) = E(X_{ij} \mid \theta_i)$$

$$s^2(\theta_i) = \text{var}(X_{ij} \mid \theta_i)$$

Notice that, as in Section 1.2, neither $E(X_{ij} \mid \theta_i)$ nor $\text{var}(X_{ij} \mid \theta_i)$ depends on j since, given θ_i, the random variables $X_{i1}, X_{i2}, \ldots, X_{in}$ are identically distributed. Notice also that, since $\theta_1, \theta_2, \ldots, \theta_N$ are identically distributed, $E[m(\theta_i)]$, $E[s^2(\theta_i)]$ and $\text{var}[m(\theta_i)]$ do not depend on i. These are precisely the parameters denoted $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ in Section 1.2 and Section 1.3 and which the collective will be used to estimate.

More notation is needed. Denote:

$$\frac{1}{n}\sum_{j=1}^{n} X_{ij} \quad \text{by } \bar{X}_i$$

and:

$$\frac{1}{N}\sum_{i=1}^{N} \bar{X}_i \quad \left(= \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n} X_{ij} \right) \text{ by } \bar{X}$$

$\bar{X}_i$ is the mean of the values from Risk i and $\bar{X} = \dfrac{\bar{X}_1 + \cdots + \bar{X}_N}{N}$ is the overall mean, *ie* the average value of X_{ij} over all risks and all years. These formulae are also given on page 29 of the *Tables*.

Notice that what is now denoted $\bar{X}_i$ was denoted $\bar{X}$ in Section 1.2 and Section 1.3. It is important to recognise the difference between the meanings of the symbol $\bar{X}$ in this section and in earlier sections.

This new notation will be used to reformulate the credibility estimate of the pure premium, or the number of claims, for the coming year for Risk Number 1 in the collective as:

$$Z\,\bar{X}_1 + (1 - Z)E[m(\theta)]$$

where:

$$\bar{X}_1 = \frac{1}{n}\sum_{j=1}^{n} X_{1j} \quad \text{and} \quad Z = \frac{n}{n + E[s^2(\theta)] \big/ \text{var}[m(\theta)]} \tag{15.1.2}$$

It is important to realise that formula (15.1.2) is exactly the same as formula (15.1.1) but in this case the notation used for the data from the risk itself is $\bar{X}_1$ and X_{1j} rather than $\bar{X}$ and X_j.

This is because we are considering the credibility estimate for Risk 1 here, and so the direct data is the set of observations taken from Risk 1.

Estimators for $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ can now be produced. These estimators will be functions of $\left\{ \{X_{ij}\}_{j=1}^{n} \right\}_{i=1}^{N}$, whose values will be known when the credibility estimate of $m(\theta_1)$ is computed.

In other words, the values of $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ will be estimated using the data from all risks and all years.

Each row of the earlier table (Table 1) corresponds to a different value of θ. Bearing this and the definitions of $m(\theta_i)$ and $s^2(\theta_i)$ in mind, obvious estimators for $m(\theta_i)$ and $s^2(\theta_i)$ are:

$$\bar{X}_i \qquad \text{and} \qquad \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2$$

respectively.

So we estimate $m(\theta_i)$ by the sample mean of the values for Risk i and we estimate $s^2(\theta_i)$ by the sample variance of the values for Risk i.

Now, $E[m(\theta)]$ is the 'average' (over the distribution of θ) of the values of $m(\theta)$ for different values of θ. The obvious estimator for $E[m(\theta)]$ is the average of the estimators of $m(\theta_i)$ for $i = 1, 2, \ldots, N$. In other words, the estimator for $E[m(\theta)]$ is $\bar{X}$.

So we estimate $E[m(\theta)]$ by the average of the sample means $\bar{X}_1,\ldots,\bar{X}_N$. This is equivalent to the sample mean of the full set of (both direct and collateral) data.

Similarly, $E[s^2(\theta)]$ is the 'average' value of $s^2(\theta)$ and so an obvious estimator is the average of the estimators of $s^2(\theta_i)$, which is:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2\right] \qquad\qquad (15.1.3)$$

Alternatively, if we define:

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2 \ , \ i=1,2,\ldots,n$$

then S_i^2 is the variance of the observations recorded for Risk i and the estimator for $E[s^2(\theta)]$ can be expressed as:

$$\frac{S_1^2+\cdots+S_N^2}{N}$$

For each row of Table 1, $\bar{X}_i$ is an estimate of $m(\theta_i)$, $i=1,2,\ldots,N$. So it might be thought that the observed variance of these values, ie:

$$\frac{1}{N-1}\sum_{i=1}^{N}(\bar{X}_i-\bar{X})^2$$

would be an obvious estimator for $\text{var}[m(\theta)]$.

$\dfrac{1}{N-1}\sum\limits_{i=1}^{N}(\bar{X}_i - \bar{X})^2$ is the variance of the sample means, $\bar{X}_1, \ldots, \bar{X}_n$.

Unfortunately, this can be shown to be a biased estimator of $\text{var}[m(\theta)]$.

In fact it is positively biased, *ie* it tends to overestimate slightly.

It can be shown that an unbiased estimator of $\text{var}[m(\theta)]$ can be produced by subtracting a correction term from the above formula. An unbiased estimator for $\text{var}[m(\theta)]$ is:

$$\frac{1}{N-1}\sum_{i=1}^{N}(\bar{X}_i - \bar{X})^2 \; - \frac{1}{Nn}\sum_{i=1}^{N}\left[\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2\right] \qquad (15.1.4)$$

The correction term is $\dfrac{1}{n}$ times the estimator for $E[s^2(\theta)]$.

This last observation can simplify the working in numerical calculations. Just make sure that you use $1/n$, where n is the number of data values for each risk, rather than $1/N$ in your calculation.

These estimators can be summarised in the following table:

Parameter	Estimator
$E[m(\theta)]$	$\bar{X}$
$E[s^2(\theta)]$	$\dfrac{1}{N}\sum\limits_{i=1}^{N}\left[\dfrac{1}{n-1}\sum\limits_{j=1}^{n}(X_{ij} - \bar{X}_i)^2\right]$
$\text{var}[m(\theta)]$	$\dfrac{1}{N-1}\sum\limits_{i=1}^{N}(\bar{X}_i - \bar{X})^2 \; - \dfrac{1}{Nn}\sum\limits_{i=1}^{N}\left[\dfrac{1}{n-1}\sum\limits_{j=1}^{n}(X_{ij} - \bar{X}_i)^2\right]$

Table 2

These formulae are given on page 29 of the *Formulae and Tables for Examinations of the Faculty of Actuaries and the Institute of Actuaries.*

We will work through some numerical examples shortly to illustrate the use of these formulae.

Note that although $\text{var}[m(\theta)]$ is a non-negative parameter since it is a variance, the estimator given by (15.1.4) could be negative. Formula (15.1.4) is the difference between two terms. Each of these terms is non-negative but their difference need not be.

It could be that the second term works out to be bigger than the first.

If, in practice, (15.1.4) gives a negative value, the accepted procedure is to estimate $\text{var}[m(\theta)]$ as 0. Strictly speaking, this means that the estimator for $\text{var}[m(\theta)]$ is the maximum of 0 and the value given by (15.1.4). Although (15.1.4) gives an unbiased estimator of $\text{var}[m(\theta)]$, taking the maximum of 0 and (15.1.4) does not give an unbiased estimator. However, this pragmatic approach avoids a nonsensical estimate for $\text{var}[m(\theta)]$.

The parameter $E[s^2(\theta)]$ must also be non-negative, but its estimator, given by formula (15.1.3), will always be non-negative and so no adjustment to (15.1.3) is required.

It can be shown that the estimators for $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ are unbiased. The proof is beyond the scope of the syllabus.

Now consider formula (15.1.2) for the credibility factor for EBCT Model 1. The different 'ingredients' of this formula have the following definitions or interpretations:

n is the number of data values in respect of the risk

Usually this will correspond to the number of years of data available, as stated in the *Tables*.

$E[s^2(\theta)]$ is the average variability of data values from year to year for a single risk, *ie* the average variability within the rows of Table 1

$\text{var}[m(\theta)]$ is the variability of the average data values for different risks, *ie* the variability of the row means in Table 1.

Looking at formula (15.1.2) for Z :

$$Z = \frac{n}{n + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

the following observations can be made:

(i) Z is always between zero and one.

(ii) Z is an increasing function of n. This is to be expected – the more data there are from the risk itself, the more it will be relied on when the credibility estimate of the pure premium or number of claims is calculated.

(iii) Z is a decreasing function of $E[s^2(\theta)]$. This is to be expected – the higher the value of $E[s^2(\theta)]$, relative to $\text{var}[m(\theta)]$, the more variable, and hence less reliable, are the data from the risk itself relative to the data from the other risks in the collective.

(iv) Z is an increasing function of $\text{var}[m(\theta)]$. This is to be expected – the higher the value of $\text{var}[m(\theta)]$, relative to $E[s^2(\theta)]$, the more variability there is between the different risks in the collective and hence the less likely it is that the other risks in the collective will resemble the risk that is of interest, and the less reliance should be placed on the data from these other risks.

There appears to be a contradiction in this section. In Section 2.2 of the previous chapter it was stated that the credibility factor should not depend on the data from the risk being rated. However, these data have been used to estimate $\text{var}[m(\theta)]$ and $E[s^2(\theta)]$, whose values are then used to calculate Z. The explanation is that in principle the credibility factor, Z, as given by formula (15.1.2), does not depend on the actual data from the risk being rated. Unfortunately, the formula for Z involves two parameters, $\text{var}[m(\theta)]$ and $E[s^2(\theta)]$, whose values are unknown but which can, in practice, be estimated from data from the risk itself and from the other risks in the collective.

In other words, we include the data for the risk we're interested in when we estimate $\text{var}[m(\theta)]$ and $E[s^2(\theta)]$.

There is one further comment to be made about the assumptions for this model. Assumptions have been made concerning the identical distribution of the X_{ij}'s both from different risks and from the same risk. If the X_{ij}'s come from different risks it has been assumed that they are (unconditionally) identically distributed. If the X_{ij}'s come from the same risk, i, it has been assumed they are unconditionally identically distributed and conditionally, given θ_i, identically distributed. A consequence of these assumptions is that neither $E(X_{ij} \mid \theta_i)$ nor $\text{var}(X_{ij} \mid \theta_i)$ depends on j. This consequence was the only step in the derivation of the credibility estimate (15.1.2) that required assumptions about the identical distribution of the X_{ij}'s. In fact, Assumptions 1 and 2 in Section 1.2 and the corresponding Assumptions 1a and 2a in Section 1.4 could have been replaced by the following assumptions and it would still have been possible to derive the same credibility estimate.

Assumption 4: Given θ_i, the X_{ij}'s, $j = 1, 2, \dots, n$, are independent for each risk i, $i = 1, 2, \dots, N$.

Assumption 5: For $i \neq k$, the pairs (θ_i, X_{ij}) and (θ_k, X_{kl}) are independent and the risk parameters $\theta_1, \theta_2, \dots, \theta_N$, are also identically distributed.

Assumption 6: For each risk, i, $i = 1, 2, \dots, N$, neither $E(X_{ij} \mid \theta_i)$ nor $\text{var}(X_{ij} \mid \theta_i)$ depends on j.

Assumptions 4 and 5 are weakened versions of Assumptions 1 and 2 made earlier. Assumption 6 now has to be included as a separate assumption since it is not a consequence of Assumptions 4 and 5.

None of the results or formulae in Section 3 of the previous chapter would be altered in any way by making these weaker assumptions. The reason for making the slightly stronger assumptions, as was done in Section 1.2 and Section 1.4 of this chapter, is that they make the presentation a little easier. The reason for pointing out now that weaker assumptions could have been made is that this will help to link EBCT Model 1 with EBCT Model 2 in Section 2.

1.5   Example: Credibility premium using Model 1

We now consider a numerical example where we use EBCT Model 1 to estimate aggregate claim amounts. In the formulae given above, we used the notation X_{ij} to represent the *random* aggregate claim amount from Risk i in Year j. When talking about the corresponding *observed* data values, we should use lower case x_{ij}. However, some authors use X_{ij} to represent both the random variable and its observed values.

The following data represents the total claim amounts per year, x_{ij}, over a six-year period ($n = 6$) for five fleets of buses $(N = 5)$, A to E:

		Year, j					
		2005	**2006**	**2007**	**2008**	**2009**	**2010**
	A	1,250	980	1,800	2,040	1,000	1,180
	B	1,700	3,080	1,700	2,820	5,760	3,480
Fleet, i	**C**	2,050	3,560	2,800	1,600	4,200	2,650
	D	4,690	4,370	4,800	9,070	3,770	5,250
	E	7,150	3,480	5,010	4,810	8,740	7,260

The following functions can be determined from the data:

		$\sum_{j=1}^{6} x_{ij}/n = \bar{x}_i$	$\sum_{j=1}^{6} (x_{ij} - \bar{x}_i)^2$	$(\bar{x}_i - \bar{x})^2$
	A	1,375	973,150	5,569,600
	B	3,090	11,218,200	416,025
Fleet, i	**C**	2,810	4,562,000	855,625
	D	5,325	18,039,550	2,528,100
	E	6,075	19,130,550	5,475,600
Total		18,675	53,923,450	14,844,950

The column header row above spans "Year, j".

The estimator for $E[m(\theta)]$ is $\bar{X}$. Hence $E[m(\theta)]$ is estimated by:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} \bar{x}_i = \frac{18,675}{5} = 3,735$$

This is the overall mean, which we have obtained by calculating the average of the fleet means. Alternatively, we could sum the data values over all risks and all years, then divide by 30 (the total number of data values).

From Equation (15.1.3), $E[s^2(\theta)]$ **is estimated by:**

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right] = \frac{1}{5\times 5}\times 53{,}923{,}450 = 2{,}156{,}938$$

This is the average of the fleet variances.

From Equation (15.1.4), $\text{var}[m(\theta)]$ **is estimated by:**

$$\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i-\bar{x})^2 - \frac{1}{Nn}\sum_{i=1}^{N}\left[\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right]$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i-\bar{x})^2 - \frac{1}{n}\times \text{estimated value of } E[s^2(\theta)]$$

$$= \frac{1}{4}\times 14{,}844{,}950 - \frac{1}{6}\times 2{,}156{,}938$$

$$= 3{,}351{,}748$$

The first term in this expression, *ie* $\dfrac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i-\bar{x})^2$, is the variance of the fleet means.

Therefore, using Formula (15.1.1), the estimated value of Z **is:**

$$\frac{n}{n+\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} = 0.90313$$

For Fleet *A*, *ie* **Risk 1 in the table above, the credibility premium is the estimated value of:**

$$Z\overline{X}_1 + (1-Z)E[m(\theta)]$$

ie:

$$0.90313\times 1{,}375 + 0.09687\times 3{,}735 = 1{,}604$$

The credibility premiums, for fleets *B* **to** *E*, **calculated in a similar manner are:**

Fleet	B	C	D	E
Premium	3,152	2,900	5,171	5,848

R

Whilst these calculations can easily be performed in Excel, we can also use R as follows for data stored in a matrix `data`:

n	`n<-ncol(data)`
$E[m(\theta)]$	`m <-mean(rowMeans(data))`
$E[s^2(\theta)]$	`s <-mean(apply(data,1,var))`
var[$m(\theta)$]	`v<-var(rowMeans(data))-mean(apply(data,1,var))/n`
Z	`Z<-n/(n+s/v)`
premiums	`Z* rowMeans(data)+(1-Z)*m`

Question

The table below shows the aggregate claim amounts (in £m) for an international insurer's fire portfolio for a 5-year period, together with some summary statistics.

		Aggregate claim amount, Year j						
		1	2	3	4	5	$\overline{x}_i$	$\frac{1}{4}\sum_{j=1}^{5}(x_{ij}-\overline{x}_i)^2$
Country, i	1	48	53	42	50	59	50.4	39.3
	2	64	71	64	73	70	68.4	17.3
	3	85	54	76	65	90	74.0	215.5
	4	44	52	69	55	71	?	?

(i)   Fill in the missing entries in the last row of the table.

(ii)   Estimate the values of $E[m(\theta)]$, $E[s^2(\theta)]$ and var[$m(\theta)$] using EBCT Model 1, and hence estimate the credibility factor, Z.

(iii)   Calculate the credibility premium for each country using EBCT Model 1.

Solution

(i)   *Missing entries*

There are 4 risks (*ie* countries) and 5 data values for each risk. So $N=4$ and $n=5$.

$\bar{x}_i$ is the mean aggregate claim amount for Country i over the 5-year period. The missing entry in this column is:

$$\bar{x}_4 = \frac{44 + 52 + 69 + 55 + 71}{5} = 58.2$$

$\frac{1}{4}\sum_{j=1}^{5}(x_{ij} - \bar{x}_i)^2$ is the sample variance for Country i over the 5-year period. The missing entry in this column is:

$$\frac{1}{4}\sum_{j=1}^{5}(x_{4j} - \bar{x}_4)^2 = \frac{1}{4}\left[(44 - 58.2)^2 + \cdots + (71 - 58.2)^2\right] = 132.7$$

Alternatively, this can be calculated using the statistical functions on a calculator.

(ii)   **Estimated values**

The estimated value of $E[m(\theta)]$ is the overall mean:

$$\bar{x} = \frac{1}{4}\sum_{i=1}^{4}\bar{x}_i = \frac{50.4 + 68.4 + 74.0 + 58.2}{4} = 62.75$$

The estimated value of $E[s^2(\theta)]$ is the mean of the country variances:

$$\frac{39.3 + 17.3 + 215.5 + 132.7}{4} = 101.2$$

The estimated value of $\text{var}[m(\theta)]$ is given by:

$$\frac{1}{3}\sum_{i=1}^{4}(\bar{x}_i - \bar{x})^2 - \frac{1}{4 \times 5}\sum_{i=1}^{4}\frac{1}{4}\sum_{j=1}^{5}\left(x_{ij} - \bar{x}_i\right)^2$$

ie:

variance of the country means $-\frac{1}{5}\times$ estimated value of $E[s^2(\theta)]$

The variance of the country is:

$$\frac{1}{3}\sum_{i=1}^{4}\left(\bar{x}_i - \bar{x}\right)^2 = \frac{1}{3}\left[(50.4 - 62.75)^2 + \cdots + (58.2 - 62.75)^2\right] = 110.57$$

(Again, this can be calculated using the statistical functions on a calculator.)

So, the estimated value of $\text{var}[m(\theta)]$ is:

$$110.57 - \frac{1}{5}\times 101.2 = 90.33$$

The credibility factor is given by:

$$Z = \frac{n}{n + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

Using the estimated values of $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ calculated above, this is:

$$\frac{5}{5 + \dfrac{101.2}{90.33}} = 0.81695$$

(iii)   ***Credibility premiums***

The credibility premium for Country i is:

$$Z\,\overline{x}_i + (1 - Z)E[m(\theta)]$$

Using the values calculated earlier in this question, the credibility premiums are as follows:

Country 1: $0.81695 \times 50.4 + (1 - 0.81695) \times 62.75 = 52.66$

Country 2: $0.81695 \times 68.4 + (1 - 0.81695) \times 62.75 = 67.37$

Country 3: $0.81695 \times 74.0 + (1 - 0.81695) \times 62.75 = 71.94$

Country 4: $0.81695 \times 58.2 + (1 - 0.81695) \times 62.75 = 59.03$

2   Empirical Bayes Credibility Theory: Model 2

2.1   Introduction

EBCT Model 2 is a generalisation of Model 1. Although some of the formulae for Model 2 look similar to those for Model 1, it is important to appreciate the differences between the models (in terms of their assumptions and results).

In this section the techniques of Empirical Bayes credibility will be applied to a second, and slightly more complicated, model. The format will be exactly the same as in Section 1. In Section 2.2, the problem will be stated and the assumptions set out. The problem will be the same as in Section 1, *ie* to estimate the pure premium, or the expected number of claims, in the coming year for a risk.

Recall that the pure premium is equal to the expected total claim amount.

The assumptions, however, will be slightly different. In Section 2.3, the credibility estimate for the pure premium or expected number of claims will be considered. Finally, in Section 2.4, the method of estimating the values of the parameters that are part of the credibility estimate will be discussed.

The model in this section will be referred to as EBCT Model 2.

2.2   Model 2: specification

The problem is to estimate the expected aggregate claims, or the expected number of claims, in the coming year for a given risk. Let $Y_1, Y_2, \ldots, Y_n$ be random variables representing the aggregate claims or numbers of claims in successive years from this risk.

In Model 1 these quantities were denoted $X_1, X_2, \ldots, X_n$. We will see below why a different letter has been used here.

It is assumed that the values of $Y_1, Y_2, \ldots, Y_n$ have already been observed and the expected value of Y_{n+1} needs to be estimated.

So far the problem looks exactly like the problem in EBCT Model 1 outlined in Section 1. The important difference between EBCT Model 1 and EBCT Model 2 is that Model 2 involves an extra parameter known as the risk volume, P_j. Intuitively, the value of P_j measures the 'amount of business' in Year j. For example, P_j might represent the premium income for the risk in Year j or the number of separate policies comprising the risk in Year j. An important point to note is that the value of P_{n+1} at the start of Year $n+1$ is assumed to be known.

Next a new sequence of random variables, $X_1, X_2, \ldots,$ is defined as follows:

$$X_j = Y_j / P_j \qquad\qquad j = 1, 2, \ldots$$

The random variable X_j represents the aggregate claims, or the number of claims, in Year j standardised to remove the effect of different levels of business in different years.

So, depending on the context of the problem, X_j represents either:

- the aggregate claim amount in Year j per unit of risk volume, or

- the total number of claims in Year j per unit of risk volume.

In Model 1, we assume that P_j is always equal to 1, *ie* the volume of business is the same for each risk group.

Assumptions for EBCT Model 2

The assumptions that specify EBCT Model 2 are as follows.

Assumption 7: The distribution of each X_j depends on the value of a parameter, θ, whose value is the same for each j but is unknown.

Assumption 8: Given θ, the X_j's are independent (but not necessarily identically distributed).

Assumption 9: $E(X_j \mid \theta)$ does not depend on j.

Assumption 10: $P_j \text{var}(X_j \mid \theta)$ does not depend on j.

As in previous sections, θ is known as the risk parameter for the risk, and, as for EBCT Model 1, it could be just a single real valued number or a more general quantity such as a vector of real valued numbers. Assumption 7 is the standard assumption for all credibility models considered here. Assumption 8 corresponds to Assumption 2 in EBCT Model 1, but notice that Assumption 8 is *slightly weaker* than Assumption 2. Assumption 8 does not require the X_j's to be conditionally (given θ) identically distributed, but only to be conditionally independent. There is no assumption in EBCT Model 2 that the X_j's are unconditionally, or conditionally given θ, identically distributed.

If all the P_j's are equal to 1, then Assumptions 7-10, taken together, become the same as Assumptions 4, 5 and 6 (taken together) in EBCT Model 1. Thus, if all the P_j's are equal to 1, EBCT Model 2 is exactly the same as EBCT Model 1.

Having made Assumptions 9 and 10, $m(\theta)$ and $s^2(\theta)$ can be defined as follows:

$$m(\theta) = E(X_j \mid \theta)$$

$$s^2(\theta) = P_j \text{var}(X_j \mid \theta)$$

The definition of $m(\theta)$ corresponds exactly to the definition for EBCT Model 1 in Section 1 but the definition of $s^2(\theta)$ is slightly different.

In Model 2, there is a factor of P_j in the definition of $s^2(\theta)$. In Model 1, $P_j = 1$ and so $s^2(\theta) = \text{var}(X_j \mid \theta)$.

To gain a little more insight into Assumptions 9 and 10, consider the following example. Suppose the risk being considered is made up of a different number of independent policies each year and that the number of policies in Year j is P_j.

It is important to realise that P_j is a known quantity, not a random variable.

Suppose also that the aggregate claims in a single year from a single policy have mean $m(\theta)$ and variance $s^2(\theta)$, where $m(\ )$ and $s^2(\ )$ are functions of θ, and θ is the fixed, but unknown, risk parameter for all these policies. Now let Y_j denote the aggregate claims from all the policies in force in Year j.

Then $E(Y_j)$ is the expected aggregate claim amount from all policies in year j, and:

$$E(Y_j) = \sum_{k=1}^{P_j} \text{expected aggregate claim amount for policy } k = \sum_{k=1}^{P_j} m(\theta)$$

So:

$$E(Y_j) = P_j\, m(\theta)$$

Also, since the policies are assumed to be independent:

$$\text{var}(Y_j) = \sum_{k=1}^{P_j} \text{variance of aggregate claim amount for policy } k = \sum_{k=1}^{P_j} s^2(\theta)$$

ie:

$$\text{var}(Y_j) = P_j\, s^2(\theta)$$

Then, since $X_j = \dfrac{Y_j}{P_j}$:

$$E(X_j) = \frac{1}{P_j} E(Y_j) = m(\theta) \quad \text{and} \quad \text{var}(X_j) = \frac{1}{P_j^2} \text{var}(Y_j) = \frac{s^2(\theta)}{P_j}$$

So:

$$E(X_j) = m(\theta) \quad \text{and} \quad P_j\, \text{var}(X_j) = s^2(\theta)$$

2.3   Model 2: the credibility premium

The problem was stated rather loosely in the previous section as the estimation of the expected value of Y_{n+1}, given the values of $Y_1, Y_2, ..., Y_n$. We can now be rather more precise about this. The quantity to be estimated is the mean (given θ) of Y_{n+1}. This is given by $P_{n+1} m(\theta)$. Since in the model P_{n+1} is known at the start of Year $n+1$, the problem is to estimate $m(\theta)$. The data available are the values of each Y_j and its corresponding P_j for $j = 1, 2, ..., n$. The full derivation of the credibility premium under this model is beyond the scope of the Subject CS1 syllabus and is not covered here.

The solution to the problem, *ie* the 'best' linear estimator of $m(\theta)$ given $\underline{X}$ is given by:

$$\frac{E[m(\theta)]\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]} + \displaystyle\sum_{j=1}^{n} Y_j}{\displaystyle\sum_{j=1}^{n} P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

This can be written more attractively as:

$$Z\,\bar{X} + (1-Z)\,E[m(\theta)] \tag{15.2.1}$$

where:

$$\bar{X} = \frac{\displaystyle\sum_{j=1}^{n} P_j X_j}{\displaystyle\sum_{j=1}^{n} P_j} = \frac{\displaystyle\sum_{j=1}^{n} Y_j}{\displaystyle\sum_{j=1}^{n} P_j} \tag{15.2.2}$$

and:

$$Z = \frac{\displaystyle\sum_{j=1}^{n} P_j}{\displaystyle\sum_{j=1}^{n} P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} \tag{15.2.3}$$

This demonstrates the similarities to and the differences from the Model 1 result.

Question

Show that:

$$Z\,\bar{X} + (1-Z)\,E[m(\theta)] = \frac{E[m(\theta)]\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]} + \displaystyle\sum_{j=1}^{n} Y_j}{\displaystyle\sum_{j=1}^{n} P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

Solution

Using the given definitions of Z and $\overline{X}$:

$$Z\,\overline{X} + (1-Z)E[m(\theta)] = \frac{\sum\limits_{j=1}^{n}P_j}{\sum\limits_{j=1}^{n}P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} \times \frac{\sum\limits_{j=1}^{n}P_j X_j}{\sum\limits_{j=1}^{n}P_j} + \frac{\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}{\sum\limits_{j=1}^{n}P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} \times E[m(\theta)]$$

$$= \frac{\sum\limits_{j=1}^{n}P_j X_j}{\sum\limits_{j=1}^{n}P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} + \frac{\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}{\sum\limits_{j=1}^{n}P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} \times E[m(\theta)]$$

$$= \frac{\sum\limits_{j=1}^{n}Y_j}{\sum\limits_{j=1}^{n}P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} + \frac{\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}{\sum\limits_{j=1}^{n}P_j + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} \times E[m(\theta)]$$

as required.

The above result tells us how to estimate the value of X_{n+1}. If we want to estimate Y_{n+1}, the aggregate claim amount or claim frequency for the coming year, we have to multiply our estimate of X_{n+1} by P_{n+1}, the risk volume for that year.

Here are some additional points of note about this solution (given by Formulae 15.2.1-15.2.3):

(i)   If all the P_j 's are equal to 1, the solution given by (15.2.1) is exactly the same as the solution given in Section 1.3. This is as it should be since if all the P_j 's are equal to 1 then EBCT Model 2 is exactly the same as EBCT Model 1.

(ii)   Just as with EBCT Model 1, the solution given by (15.2.1) involves three parameters, $E[m(\theta)]$, $\text{var}[m(\theta)]$ and $E[s^2(\theta)]$. The way in which these parameters are estimated is explained in the next section.

2.4   Model 2: parameter estimation

The procedure for estimating the parameters $E[m(\theta)]$, $\text{var}[m(\theta)]$ and $E[s^2(\theta)]$ for EBCT Model 2 follows exactly the same steps as the procedure for EBCT Model 1, which was studied in Section 1.4.

As before, we are now moving to a two-dimensional data set with rows representing different risks.

It is now assumed that the risk that is of interest is one of a collective of N risks and that there exist data for each of these N risks for each of the past n years. These data consist of values for the aggregate claims, or the number of claims, and the corresponding risk volumes. Let Y_{ij} be a random variable denoting the aggregate claims, or the number of claims, for Risk Number i in Year j, $j = 1, 2, …, n$, $i = 1, 2, …, N$, and let P_{ij} be the corresponding risk volume.

For each i and j define:

$$X_{ij} = Y_{ij} / P_{ij}$$

The data are summarised in the following table, which corresponds to Table 1 for EBCT Model 1:

		Year			
		1	**2**	**…**	**n**
	1	Y_{11}, P_{11}	Y_{12}, P_{12}	$\cdots$	Y_{1n}, P_{1n}
	2	Y_{21}, P_{21}	Y_{22}, P_{22}	$\cdots$	Y_{2n}, P_{2n}
Risk Number	.	.	.	$\cdots$	.
	.	.	.	$\cdots$	.
	N	Y_{N1}, P_{N1}	Y_{N2}, P_{N2}	$\cdots$	Y_{Nn}, P_{Nn}

Table 3

For simplicity it is assumed, as was done in Section 1.4, that the risk that is of particular interest is Risk Number 1 in this collective. This means that what were denoted Y_j, P_j and X_j in Section 2.2 and Section 2.3 are now denoted Y_{1j}, P_{1j} and X_{1j} respectively in this section. The problem is to estimate the expected value of $X_{1,n+1}$ and the solution to this problem has already been given by Formulae (15.2.1)-(15.2.3), remembering that X_j and P_j are now denoted X_{1j} and P_{1j}. The purpose of the data from the other risks in the collective is purely to help to estimate the parameters $E[m(\theta)]$, $var[m(\theta)]$ and $E[s^2(\theta)]$ that appear in Formulae (15.2.1) and (15.2.3).

These other risks in the collective satisfy assumptions exactly the same as Assumptions 7-10 for Risk Number 1. These assumptions are as follows.

For each risk i, $i = 1, 2, …, N$:

Assumption 7a: The distribution of each X_{ij}, $j = 1, 2, …, n$, depends on a parameter, θ_i, whose value is the same for each j but is unknown.

Assumption 8a: Given θ_i, the X_{ij}'s are independent (but not necessarily identically distributed).

Assumption 9a: There exists a function $m(\ )$ such that $m(\theta_i) = E(X_{ij} | \theta_i)$.

Assumption 10a: There exists a function $s(\ )$ such that $s^2(\theta_i) = P_{ij} \, \text{var}(X_{ij} \mid \theta_i)$.

These four assumptions show that each risk in the collective satisfies the same assumptions as the particular risk that is of interest. The following two assumptions show the connection *between* different risks in the collective:

Assumption 11: The risk parameters $\theta_1, \theta_2, \ldots, \theta_N$, regarded as random variables, are independent and identically distributed.

Assumption 12: For $i \neq k$, the pairs (θ_i, X_{ij}) and (θ_k, X_{km}) are independent.

Notice that, since the θ_i 's are identically distributed, the values of $E[m(\theta_i)]$, $\text{var}[m(\theta_i)]$ and $E[s^2(\theta_i)]$ do not depend on i so that they can be denoted by $E[m(\theta)]$, $\text{var}[m(\theta)]$ and $E[s^2(\theta)]$, respectively, as in Section 2.2 and Section 2.3.

Denote:

$$\sum_{j=1}^{n} P_{ij} \qquad\qquad \text{by } \bar{P}_i$$

$$\sum_{i=1}^{N} \bar{P}_i \qquad\qquad \text{by } \bar{P}$$

$$\frac{1}{Nn-1} \sum_{i=1}^{N} \bar{P}_i \left(1 - \frac{\bar{P}_i}{\bar{P}} \right) \qquad\qquad \text{by } P^*$$

$$\sum_{j=1}^{n} \frac{P_{ij} X_{ij}}{\bar{P}_i} \qquad\qquad \text{by } \bar{X}_i$$

$$\sum_{i=1}^{N} \frac{\bar{P}_i \bar{X}_i}{\bar{P}} = \sum_{i=1}^{N} \sum_{j=1}^{n} \frac{P_{ij} X_{ij}}{\bar{P}} \qquad\qquad \text{by } \bar{X}$$

This notation is all given on page 30 of the *Formulae and Tables for Examinations of the Faculty of Actuaries and the Institute of Actuaries*.

Note that $\bar{X}_i$ and $\bar{X}$ are weighted averages of the X_{ij} 's, the weights being the risk volumes P_{ij}.

This is more obvious if we write the formulae for $\bar{X}_i$ and $\bar{X}$ as follows:

$$\bar{X}_i = \frac{\displaystyle\sum_{j=1}^{n} P_{ij} X_{ij}}{\displaystyle\sum_{j=1}^{n} P_{ij}} \quad \text{and} \quad \bar{X} = \frac{\displaystyle\sum_{i=1}^{N} \sum_{j=1}^{n} P_{ij} X_{ij}}{\displaystyle\sum_{i=1}^{N} \sum_{j=1}^{n} P_{ij}}$$

Alternatively, we could write:

$$\bar{X}_i = \frac{\sum\limits_{j=1}^{n} Y_{ij}}{\sum\limits_{j=1}^{n} P_{ij}} \quad \text{and} \quad \bar{X} = \frac{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{n} Y_{ij}}{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{n} P_{ij}}$$

With this new notation the credibility estimate of the pure premium, or number of claims, per unit of risk volume for the coming year for Risk Number 1 in the collective originally given by Formula (15.2.1)-(15.2.3) can be reformulated as:

$$Z_1 \bar{X}_1 + (1 - Z_1)E[m(\theta)]$$

where:

$$\bar{X}_1 = \frac{\sum\limits_{j=1}^{n} P_{1j} X_{1j}}{\sum\limits_{j=1}^{n} P_{1j}}$$

and:

$$Z_1 = \frac{\sum\limits_{j=1}^{n} P_{1j}}{\sum\limits_{j=1}^{n} P_{1j} + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} \qquad\qquad \textbf{(15.2.4)}$$

It is important to realise that these formulae are exactly the same as Formulae (15.2.1)-(15.2.3) but are written in the notation of this section rather than that of the previous section.

This credibility factor is specific to Risk 1, so is denoted by Z_1. More generally, the credibility factor for Risk i is given by:

$$Z_i = \frac{\sum\limits_{j=1}^{n} P_{ij}}{\sum\limits_{j=1}^{n} P_{ij} + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

The credibility factor varies from one risk to another because different risks have different risk volumes. The ratio $\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}$ is the same in each Z_i.

Question

Explain the effect each of the following changes (acting in isolation) would have on the value of the credibility factor for a given risk in EBCT Model 2.

(a) $E[s^2(\theta)]$ is increased.

(b) $\text{var}[m(\theta)]$ is reduced.

(c)   The unit of currency is changed.

(d)   All the P_j values for this risk are increased by the same factor.

Solution

(a) $E[s^2(\theta)]$ represents the average variability within each risk group. Increasing this would mean that claims experience tends to vary more from year to year. So we should have less confidence in the accuracy of an estimate based on past claims.

Since increasing $E[s^2(\theta)]$ increases the denominator, Z decreases, as expected.

(b) $\text{var}[m(\theta)]$ represents the variation between the average claim amount for different risks. Reducing this would mean that the various risks become more alike in terms of their claims experience. So the collateral data values become more credible and we should put more emphasis on the overall mean.

Since reducing $\text{var}[m(\theta)]$ increases the denominator, Z decreases, as expected.

(c)   Changing the unit of currency would not affect the credibility factor.

Since the quantities $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ are measured in squared units, their ratio is dimensionless. So changing the unit of currency would not affect Z .

(d)   The P_j 's specify the relative weightings to be put on the claims for each year. A uniform increase applied to all the weightings would not affect the credibility factor.

Since the definition $E[s^2(\theta)] = P_j \, \text{var}\left(X_j \big| \theta\right)$ includes a P_j factor, but $\text{var}[m(\theta)]$ doesn't, the ratio $\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}$ varies in proportion to the P_j 's. So any extra factor incorporated in the risk volumes would cancel out, leaving Z unchanged.

Unbiased estimators for $E[m(\theta)]$, $\text{var}[m(\theta)]$ and $E[s^2(\theta)]$ can be proposed based on the observed values $\left\{\{Y_{ij}, P_{ij}\}_{j=1}^{n}\right\}_{i=1}^{N}$.

These estimators are shown in the following table, and are also given on page 30 of the *Formulae and Tables for Examinations of the Faculty of Actuaries and the Institute of Actuaries.*

Parameter	Estimator
$E[m(\theta)]$	$\bar{X}$
$E[s^2(\theta)]$	$\dfrac{1}{N}\sum_{i=1}^{N}\left[\dfrac{1}{n-1}\sum_{j=1}^{n}P_{ij}\left(X_{ij}-\bar{X}_i\right)^2\right]$
$\text{var}[m(\theta)]$	$\dfrac{1}{P^*}\left\{\dfrac{1}{Nn-1}\sum_{i=1}^{N}\sum_{j=1}^{n}P_{ij}\left(X_{ij}-\bar{X}\right)^2-\dfrac{1}{N}\sum_{i=1}^{N}\left[\dfrac{1}{n-1}\sum_{j=1}^{n}P_{ij}\left(X_{ij}-\bar{X}_i\right)^2\right]\right\}$

Table 4

Points to note about these estimators are listed below.

(i)   They are in exactly the same form as the estimators for EBCT Model 1. Look back at Table 2. If all the P_{ij} 's were equal to 1, then the two sets of estimators would be identical.

(ii)   It can happen in practice that the formula for estimating $\text{var}[m(\theta)]$ gives a negative value even though $\text{var}[m(\theta)]$ must be non-negative. In such a situation the estimate of $\text{var}[m(\theta)]$ is taken to be zero.

(iii)   The proofs that the estimators are unbiased are beyond the scope of the Subject CS1 syllabus.

2.5   Example: Credibility premium using Model 2

This example extends the example in Section 1.5. We will now take into consideration the risk volume, represented by the number of buses in each fleet for each year. The table below shows the same claim amounts as in Section 1.5, Y_{ij}, but also the corresponding number of buses, P_{ij} over a six-year period $(n = 6)$ for the five fleets of buses $(N = 5)$, A to E:

		Year					
		2005	**2006**	**2007**	**2008**	**2009**	**2010**
	A	1,250 ; 5	980 ; 5	1,800 ; 4	2,040 ; 6	1,000 ; 5	1,180 ; 5
	B	1,700 ; 11	3,080 ; 13	1,700 ; 10	2,820 ; 12	5,760 ; 15	3,480 ; 14
Fleet	**C**	2,050 ; 3	3,560 ; 4	2,800 ; 4	1,600 ; 3	4,200 ; 3	2,650 ; 2
	D	4,690 ; 9	4,370 ; 9	4,800 ; 8	9,070 ; 8	3,770 ; 9	5,250 ; 10
	E	7,150 ; 7	3,480 ; 7	5,010 ; 8	4,810 ; 8	8,740 ; 9	7,260 ; 10

The following functions can be determined from the data:

		$\sum_{j=1}^{6} P_{ij} X_{ij}$	$\sum_{j=1}^{6} P_{ij} = \overline{P}_i$	$\sum_{j=1}^{6} P_{ij}(X_{ij} - \overline{X}_i)^2$	$\sum_{j=1}^{6} P_{ij}(X_{ij} - \overline{x})^2$
	A	8,250	30	217,910	1,680,442
	B	18,540	75	437,931	5,072,946
Fleet	C	16,860	19	1,812,785	4,726,028
	D	31,950	53	2,804,038	3,411,218
	E	36,450	49	1,706,731	4,722,398
Total		112,050	226	6,979,395	19,613,032

		$\overline{P}_i(1 - \overline{P}_i / \sum_{i=1}^{5} \overline{P}_i)$
	A	26.018
	B	50.111
Fleet	C	17.403
	D	40.571
	E	38.376
Total		172.478

Note that $X_{ij} = Y_{ij}/P_{ij}$.

So far we have been using X_{ij} and Y_{ij} to represent random variables. Their observed values are denoted by x_{ij} and y_{ij} .

Using the estimators listed in Table 4 and the notation given earlier, $E[m(\theta)]$ is estimated by:

$$\overline{x} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n} P_{ij} x_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{n} P_{ij}} = \frac{112,050}{226} = 495.796$$

This is the overall mean claim per bus, which we obtain by finding the total of the claims and dividing by the total number of buses. We are using lower case x in the formula above to represent observed values (rather than random variables).

$E[s^2(\theta)]$ **is estimated by:**

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{n-1}\sum_{j=1}^{n}P_{ij}(x_{ij}-\bar{x}_i)^2\right]=\frac{6,979,395}{25}=279,175.8$$

and:

$$P^*=\frac{1}{Nn-1}\sum_{i=1}^{N}\bar{P}_i\left(1-\bar{P}_i\Big/\sum_{i=1}^{N}\bar{P}_i\right)=\frac{172.478}{29}=5.9475$$

so that $\text{var}[m(\theta)]$ **is estimated by:**

$$\frac{1}{P^*}\left(\frac{1}{Nn-1}\sum_{i=1}^{N}\sum_{j=1}^{n}P_{ij}(\bar{x}_{ij}-\bar{x})^2-\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{n-1}\sum_{j=1}^{n}P_{ij}(x_{ij}-\bar{x}_i)^2\right]\right)$$

$$=\frac{1}{P^*}\left(\frac{1}{Nn-1}\sum_{i=1}^{N}\sum_{j=1}^{n}P_{ij}(\bar{x}_{ij}-\bar{x})^2-\text{estimated value of }E[s^2(\theta)]\right)$$

$$=\frac{1}{5.9475}\left(\frac{19,613,032}{29}-279,175.8\right)$$

$$=66,774$$

Therefore, using (15.2.4):

$$Z_i=\frac{\displaystyle\sum_{j=1}^{n}P_{ij}}{\displaystyle\sum_{j=1}^{n}P_{ij}+\frac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

ie **will be different for each fleet.**

For Fleet *A* **(Risk 1), the estimated value of** Z_1 **is:**

$$\frac{30}{30+\dfrac{279,175.8}{66,774}}=0.87768$$

The credibility factors, Z_i**, for fleets** *B* **to** *E***, estimated in a similar manner are:**

Fleet	B	C	D	E
Credibility factor	0.94720	0.81964	0.92688	0.92138

Using the estimated values of $E[m(\theta)]$ **and** *Z* **calculated above, the credibility premium per unit of risk volume for Fleet** *A* **(***ie* **Risk 1) is:**

$$0.12232\times495.796+0.87768\times275=302.0$$

The credibility premiums per unit of risk volume, for fleets *B* to *E*, calculated in similar manner are:

Fleet	B	C	D	E
Credibility premium	260.3	816.7	595.0	724.4

Multiplying these figures by the risk volumes for 2011 would then give the credibility premiums for 2011.

R

Again these calculations can easily be performed in Excel. We can also use R as follows for claims data stored in matrix `data` and risk volume stored in matrix `volume`:

n	`n<-ncol(data)`
N	`N<-nrow(data)`
X_{ij}	`X <- data/volume`
$\bar{X}_i$	`Xibar<-rowSums(data)/rowSums(volume)`
$\bar{P}_i$	`Pi <-rowSums(volume)`
$\bar{P}$	`P <-sum(Pi)`
P^*	`Pstar <-sum(Pi*(1-Pi/P))/(N*n-1)`
$E[m(\theta)]$	`m <- sum(data)/P`
$E[s^2(\theta)]$	`s<-mean(rowSums(volume*(X-Xibar)^2)/(n-1))`
$\text{var}[m(\theta)]$	`v<-(sum(rowSums(volume*(X-m)^2))/(n*N-1)-s)/Pstar`
Z_i	`Zi<-Pi/(Pi+s/v)`
premiums	`Zi* Xibar +(1-Zi)*m`

2+3

Question

The figures given in the table below are the aggregate claims (in £000s) for each of four risks over a period of four years.

	Year 1	Year 2	Year 3	Year 4
Risk 1	1,892	1,975	2,309	2,278
Risk 2	2,356	2,876	3,002	3,378
Risk 3	2,890	2,489	2,424	2,551
Risk 4	1,662	1,408	1,697	2,034

(i)   Assuming that the data satisfy the assumptions of EBCT Model 1, estimate the aggregate claim amount for Risk 1 in Year 5.

(ii)   We are now given the following risk volumes:

	Year 1	Year 2	Year 3	Year 4	Year 5
Risk 1	300	329	334	346	370
Risk 2	410	425	446	470	461
Risk 3	468	405	397	422	437
Risk 4	227	206	236	259	268

Use EBCT Model 2 to estimate the aggregate claim amount for Risk 1 in Year 5. The corresponding figure for Risk 2 is 3,050.8.

Solution

(i)   ***Estimate under EBCT Model 1***

The estimated aggregate claim amount for Risk 1 is:

$$Z\,\overline{x}_1 + (1 - Z)E[m(\theta)]$$

The sample mean and variance for Risk 1 are:

$$\overline{x}_1 = \frac{1,892 + 1,975 + 2,309 + 2,278}{4} = 2,113.5$$

$$s_1^2 = \frac{1}{3}\Big[(1,892 - 2,113.5)^2 + (1,975 - 2,113.5)^2 + (2,309 - 2,113.5)^2 + (2,278 - 2,113.5)^2\Big]$$

$$= 44,508.3333$$

The formula for the sample variance is given on page 22 of the *Tables*. Alternatively, this can be calculated using the statistical functions on a calculator.

The means and variances for the other risks are calculated in a similar way. The values are summarised in the table below:

	Sample mean	Sample variance
Risk 1	2,113.5	44,508.3333
Risk 2	2,903	178,454.6667
Risk 3	2,588.5	43,089.6667
Risk 4	1,700.25	66,090.9167

The estimated value of $E[m(\theta)]$ is the overall mean:

$$\bar{x} = \frac{2,113.5 + 2,903 + 2,588.5 + 1,700.25}{4} = 2,326.3125$$

The estimated value of $E[s^2(\theta)]$ is the mean of the sample variances:

$$\frac{44,508.3333 + 178,454.6667 + 43,089.6667 + 66,090.9167}{4} = 83,035.8958$$

The variance of the sample means (given in the table above) is 279,518.0573. So the estimated value of $var[m(\theta)]$ is:

$$279,518.0573 - \frac{1}{4} \times 83,035.8958 = 258,759.0833$$

The formula for the credibility factor is:

$$Z = \frac{n}{n + \dfrac{E[s^2(\theta)]}{var[m(\theta)]}}$$

Using the estimated values of $E[s^2(\theta)]$ and $var[m(\theta)]$, this is:

$$\frac{4}{4 + \dfrac{83,035.8958}{258,759.0833}} = 0.9257$$

So the estimated aggregate claim amount for Risk 1 in Year 5 is:

$$0.9257 \times 2,113.5 + (1 - 0.9257) \times 2,326.3125 = 2,129.3$$

or approximately £2,129,300.

(ii)   ***Estimate under EBCT Model 2***

The estimated aggregate claim amount per unit of risk volume for Risk 1 is:

$$Z_1 \bar{x}_1 + (1 - Z_1)E[m(\theta)]$$

The figures given in the first table in the question are the observed values of Y_{ij}. The figures given in the second table are the values of P_{ij}.

The total claim amount for Risk 1 over Years 1 to 4 is:

$$1,892 + 1,975 + 2,309 + 2,278 = 8,454$$

The total risk volume for Risk 1 over Years 1 to 4 is:

$$300 + 329 + 334 + 346 = 1,309$$

So:

$$\bar{x}_1 = \frac{8,454}{1,309} = 6.4584$$

The total claim amount for all risks over Years 1 to 4, is 37,221. The corresponding total risk volume is 5,680. So the estimated value of $E[m(\theta)]$ is:

$$\bar{x} = \frac{37,221}{5,680} = 6.5530$$

The credibility factor for Risk 1 is:

$$Z_1 = \frac{\displaystyle\sum_{j=1}^{4} P_{1j}}{\displaystyle\sum_{j=1}^{4} P_{1j} + \frac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} = \frac{1,309}{1,309 + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

We are told that the estimated aggregate claim amount for Risk 2 in Year 5 is 3,050.8. Since the risk volume for Risk 2 in Year 5 is known to be 461, the estimated aggregate claim amount per unit of risk volume is $\dfrac{3,050.8}{461}$. We can use this information to estimate the ratio $\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}$.

According to EBCT Model 2, the credibility estimate per unit of risk volume for Risk 2 is:

$$Z_2 \bar{x}_2 + (1 - Z_2)E[m(\theta)]$$

where:

$$\bar{x}_2 = \frac{2,356 + 2,876 + 3,002 + 3,378}{410 + 425 + 446 + 470} = \frac{11,612}{1,751}$$

and:

$$Z_2 = \frac{\displaystyle\sum_{j=1}^{4} P_{2j}}{\displaystyle\sum_{j=1}^{4} P_{2j} + \frac{E[s^2(\theta)]}{\text{var}[m(\theta)]}} = \frac{1,751}{1,751 + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

Using R to denote the estimated value of $\dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}$, we have:

$$\frac{1,751}{1,751 + R} \times \frac{11,612}{1,751} + \frac{R}{1,751 + R} \times 6.5530 = \frac{3,050.8}{461}$$

Multiplying through by $1,751 + R$ gives:

$$11,612 + 6.5530R = \frac{3,050.8}{461}(1,751 + R)$$

and solving for R, we find that:

$$R = 374.3257$$

So the estimated credibility factor for Risk 1 is:

$$\frac{1,309}{1,309 + 374.3257} = 0.7776$$

and the estimated aggregate claim amount per unit of risk volume is:

$$0.7776 \times 6.4584 + (1 - 0.7776) \times 6.5530 = 6.4794$$

Finally, since the risk volume for Risk 1 in Year 5 is known to be 370, the estimate total claim amount is:

$$6,4794 \times 370 = 2,397.4$$

or approximately £2,397,400.

Chapter 15 Summary

Empirical Bayes credibility

The empirical Bayes approach to credibility theory assumes that the number of claims or aggregate claim amount for each risk is dependent on an underlying risk parameter θ. However, no assumptions are made about the form of the distribution of θ.

The credibility premium can be expressed in terms of a credibility factor, which depends on the mean and variance of the conditional claim distribution. These quantities can be estimated from data derived from a number of different risks.

EBCT Model 1

Definitions:
X_{ij} represents the number of claims (or aggregate claim amount) for risk i $(i = 1, ..., N)$ in year j $(j = 1, ..., n)$.

Assumptions:
For each risk i, the distribution of X_{ij} depends on a parameter θ_i whose value is the same for each j but is unknown.

$X_{ij} \mid \theta_i$ are IID random variables.

θ_i are IID random variables.

For $i \neq k$, the pairs $\left(X_{ij}, \theta_i \right)$ and $\left(X_{km}, \theta_k \right)$ are IID.

There exist functions $m()$ and $s^2()$ such that $m(\theta_i) = E(X_{ij} \mid \theta_i)$ and $s^2(\theta_i) = \text{var}(X_{ij} \mid \theta_i)$.

Estimators:
Formulae for unbiased estimators for $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ are given on page 29 of the *Tables*.

Credibility factor:
$$Z = \frac{n}{n + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

Credibility premium:
The credibility premium for Risk i is $Z \bar{x}_i + (1 - Z)E[m(\theta)]$.

EBCT Model 2

Definitions:
Y_{ij} represents the number of claims (or aggregate claim amount) for risk i $(i = 1,...,N)$ in year j $(j = 1,...,n)$.

P_{ij} represents the corresponding risk volume (*eg* number of policies or premium income). The P_{ij}'s are assumed to be known.

$X_{ij} = Y_{ij} / P_{ij}$

Assumptions:
For each risk i, the distribution of X_{ij} depends on a parameter θ_i whose value is the same for each j but is unknown.

$X_{ij} \mid \theta_i$ are independent random variables.

θ_i are IID random variables.

For $i \neq k$, the pairs (X_{ij}, θ_i) and (X_{km}, θ_k) are independent random variables.

There exist functions $m()$ and $s^2()$ such that $m(\theta_i) = E(X_{ij} \mid \theta_i)$ and $s^2(\theta_i) = P_{ij} \operatorname{var}(X_{ij} \mid \theta_i)$.

Estimators:
Formulae for unbiased estimators for $E[m(\theta)]$, $E[s^2(\theta)]$ and $\operatorname{var}[m(\theta)]$ are given on page 30 of the *Tables*.

Credibility factors:
$$Z_i = \frac{\sum_{j=1}^{n} P_{ij}}{\sum_{j=1}^{n} P_{ij} + \dfrac{E[s^2(\theta)]}{\operatorname{var}[m(\theta)]}}$$

Credibility premium:
The credibility premium per unit of risk volume for Risk i is $Z_i \bar{x}_i + (1 - Z_i) E[m(\theta)]$.

📱 Chapter 15 Practice Questions

15.1   Consider the following statements made about EBCT Model 1.

(a) θ represents the 'true' risk premium for a given risk.

(b)   The variance of $X_j | \theta$ doesn't depend on θ.

(c)   None of the random variables or parameters in the model are assumed to have a normal distribution.

Explain whether each of these statements is true or false.

15.2   The table below shows the aggregate claim amounts (in £m) for an international insurer's fire portfolio for a 5-year period, together with some summary statistics.

		Aggregate claim amount, Year j				
		1	2	3	4	5
Country, i	1	48	53	42	50	59
	2	64	71	64	73	70
	3	85	54	76	65	90
	4	44	52	69	55	71

The volumes of business for each country for the insurer are as follows:

		Volume of business, Year j					
		1	2	3	4	5	6
Country, i	1	12	15	13	16	10	20
	2	20	14	22	15	30	25
	3	5	8	6	12	4	10
	4	22	35	30	16	10	12

Calculate the credibility premium for each country in Year 6 using EBCT Model 2.

15.3   An actuary has, for three years, recorded the volume of unsolicited advertising that he receives.

Exam style

He believes that the number of items that he receives follows a Poisson distribution with a mean which varies according to which quarter of the year it is. He has recorded Y_{ij} the number of items received in the i th quarter of the j th year ($i = 1, 2, 3, 4$ and $j = 1, 2, 3$). The actuary wishes to estimate the number of items that he will receive in the first quarter of year four. He has recorded the following data:

	Y_{i1}	Y_{i2}	Y_{i3}	$\bar{Y}_i = \frac{1}{3}\sum_j Y_{ij}$	$\sum_j \left(Y_{ij} - \bar{Y}_i\right)^2$
$i = 1$	98	117	124	113	362
$i = 2$	82	102	95	93	206
$i = 3$	75	83	88	82	86
$i = 4$	132	152	148	144	224

(i)   Estimate $Y_{1,4}$ the number of items that the actuary expects to receive in the first quarter of year four using the assumptions of EBCT Model 1.   [5]

The actuary believes that, in fact, the volume of items has been increasing at the rate of 10% per annum.

(ii)   Suggest how the approach in (i) can be adjusted to produce a revised estimate taking this growth into account.   [2]

(iii)   Calculate the maximum likelihood estimate of $Y_{1,4}$ (based on the quarter one data already observed and the 10% *pa* increase described above).   [5]

(iv)   Compare the assumptions underlying the approach in (i) and (ii) with those underlying the approach in (iii).   [2]
                                                                                              [Total 14]

15.4   An actuary wishes to analyse the amounts paid by a group of insurers on their respective portfolios of commercial property insurance policies using the models of Empirical Bayes Credibility Theory.

Exam style

The actuary obtains the following information about the amounts of claim payments made and the number of policies sold for each of three different insurers. The data obtained are as follows.

		Year 1	Year 2	Year 3	Year 4
Insurer A		£14.2m	£15.8m	£22.7m	£19.0m
		163	189	252	199
Insurer B		£58.6m	£63.1m	£81.0m	£64.2m
		4,435	4,761	5,576	4,581
Insurer C		£123m	£132m	£161m	£133m
		16,184	17,443	20,102	18,000

(i)   Analyse the data using EBCT Model 1, and calculate the expected total claim payment to be made by Insurer B in the coming year.   [6]

(ii)   Analyse the data using EBCT Model 2, and again calculate the expected payout amount for Insurer B in the coming year, assuming that the expected number of policies sold for the coming year for Insurer B is 4,800. You may use the summary statistics given below, which have been calculated using the formulae and notation given in the *Tables,* again working in millions of pounds. Subscripts 1, 2 and 3 refer to Insurers A, B and C respectively.   [8]

$$\sum P_{1j}(x_{1j} - \overline{x}_1)^2 = 0.014667 \qquad \sum P_{1j}(x_{1j} - \overline{x})^2 = 5.106461$$

$$\sum P_{2j}(x_{2j} - \overline{x}_2)^2 = 0.006103 \qquad \sum P_{2j}(x_{2j} - \overline{x})^2 = 0.336408$$

$$\sum P_{3j}(x_{3j} - \overline{x}_3)^2 = 0.003979 \qquad \sum P_{3j}(x_{3j} - \overline{x})^2 = 0.292641$$

(iii) Comment on your results.   [2]
                                                                                                                                       [Total 16]

15.5   An actuarial student is using Empirical Bayes Credibility Theory Model 2 to calculate credibility

Exam style   premiums for a group of insurers. The student has analysed the data for six different insurers,

using 10 years of past data for each insurer and has obtained the following figures:

$$\sum_{i=1}^{6} \sum_{j=1}^{10} P_{ij} = 1,498 \qquad\qquad P^* = 18.24$$

The estimated values of $E[m(\theta)]$, $E[s^2(\theta)]$ and $var[m(\theta)]$ based on the data from the six insurers are 4.00, 62.8 and 42.1, respectively.

The student has just received the following information relating to a seventh insurer (Insurer I), and he wishes to update the estimates of $E[m(\theta)]$, $E[s^2(\theta)]$ and $var[m(\theta)]$ using the claims data for Insurer I given in the below:

Year, j	1	2	3	4	5	6	7	8	9	10
Aggregate claim amount, y_{Ij}	100	85	90	102	109	106	128	132	150	131
Risk volume, P_{Ij}	22	24	26	20	25	30	29	35	40	36

(i)   Calculate the updated estimates for $E[m(\theta)]$, $E[s^2(\theta)]$ and $var[m(\theta)]$, and hence calculate the credibility premium for Insurer I for the coming year, given that Insurer I is expected to have a risk volume figure for the coming year of 38.   [18]

(ii)   The student also needs a credibility estimate for Insurer K, one of the six insurers included in the original analysis. He knows that, for Insurer K:

$$\sum_{j=1}^{10} y_{Kj} = 986 \text{ and } \sum_{j=1}^{10} P_{Kj} = 327$$

Explain whether the credibility premium for Insurer K (based on the full analysis of the seven insurers) will be greater or less than the corresponding figure for Insurer I (per unit of risk volume).   [2]
[Total 20]

Chapter 15 Solutions

15.1   (a)   This is false. θ is just a risk parameter that reflects the likelihood of claims. The true risk premium for a given risk is $E[m(\theta)|\underline{X}]$.

(b)   This is also false. The variance of $X_j|\theta$ is $s^2(\theta)$, which is a function of θ.

(c)   This is true. In fact, none of the quantities in the model are assumed to have any specific type of distribution.

15.2   The aggregate claim amounts per unit of risk volume are shown in the table below.

		Total claim amount per unit of risk volume, Year j				
		1	2	3	4	5
Country, i	1	4	3.533	3.231	3.125	5.9
	2	3.2	5.071	2.909	4.867	2.333
	3	17	6.75	12.667	5.417	22.5
	4	2	1.486	2.3	3.438	7.1

We can then calculate $\overline{P}_i$, $\overline{P}$ and P^*. The figures are given in the table below.

Country, i	$\overline{P}_i$	$\overline{P}_i\left(1-\dfrac{\overline{P}_i}{\overline{P}}\right)$
1	66	52.17
2	101	68.62
3	35	31.11
4	113	72.46
	$\overline{P}=315$	$P^*=11.81$

In addition, we have:

Country, i	$\overline{x}_i$	$\sum P_{ij}(x_{ij}-\overline{x}_i)^2$	$\sum P_{ij}(x_{ij}-\overline{x})^2$
1	3.818	57.13	58.94
2	3.386	111.59	147.71
3	10.571	1,237.82	2,756.56
4	2.575	267.73	492.04
	$\overline{x}=3.984$		

So the estimated value of $E[m(\theta)]$ is 3.984.

From the other columns in the table, we see that the estimated value of $E[s^2(\theta)]$ is:

$$\frac{1}{4}\sum_{i=1}^{4}\frac{1}{4}\sum_{j=1}^{5}P_{ij}\left(x_{ij}-\bar{x}_i\right)^2 = \frac{57.13+111.59+1,237.82+267.73}{16} = 104.64$$

and the estimated value of $\text{var}[m(\theta)]$ is:

$$\frac{1}{P^*}\left[\frac{1}{4\times5-1}\sum_{i=1}^{4}\sum_{j=1}^{5}P_{ij}\left(x_{ij}-\bar{x}\right)^2 - 104.64\right]$$

$$= \frac{1}{11.81}\left[\frac{58.94+147.71+2,756.56+492.04}{19} - 104.64\right]$$

$$= 6.539$$

The credibility factor for Country 1 is:

$$Z_1 = \frac{\displaystyle\sum_{j=1}^{n}P_{1j}}{\displaystyle\sum_{j=1}^{n}P_{1j}+\frac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

Using the estimated values calculated above, this is:

$$\frac{66}{66+\dfrac{104.64}{6.539}} = 0.8048$$

The risk premium per unit volume is:

$$Z_1\,\bar{x}_1 + (1-Z_1)E[m(\theta)]$$

Using the estimated values of Z_1 and $E[m(\theta)]$, this is:

$$0.8048\times3.818 + (1-0.8048)\times3.984 = 3.851$$

Since the volume for Country 1 for the coming year is 20 units, the credibility premium is:

$$20\times3.851 = 77.0$$

The table below shows the figures for all four countries:

Country	Estimated credibility factor	Risk premium per unit volume	EBCT premium
1	0.8048	3.851	77.0
2	0.8632	3.468	86.7
3	0.6862	8.504	85.0
4	0.8759	2.750	33.0

15.3   *This question is Subject CT6, April 2010, Question 11.*

(i)   **Estimate using EBCT Model 1**

We have $n = 3$ and $N = 4$. Using the formulae on page 29 of the *Tables*, the estimates are:

$E[m(\theta)]$ $\bar{y} = \frac{1}{4}\sum_{i=1}^{4}\bar{y}_i = \frac{1}{4}(113 + 93 + 82 + 144) = 108$ [½]

$E[s^2(\theta)]$ $\frac{1}{4}\sum_{i=1}^{4}\frac{1}{2}\sum_{j=1}^{3}(y_{ij} - \bar{y}_i)^2 = \frac{1}{8}(362 + 206 + 86 + 224) = 109.75$ [½]

$\text{var}[m(\theta)]$ $\frac{1}{3}\sum_{i=1}^{4}(\bar{y}_i - \bar{y})^2 - \frac{1}{3}\left[\frac{1}{4}\sum_{i=1}^{4}\frac{1}{2}\sum_{j=1}^{3}(y_{ij} - \bar{y}_i)^2\right]$

$= \frac{1}{3}[(113 - 108)^2 + (93 - 108)^2 + (82 - 108)^2 + (144 - 108)^2] - \frac{1}{3}(109.75)$

$= 704.08\dot{3}$ [2]

Using the formula on page 29 of the *Tables*, the estimated credibility factor is:

$\dfrac{3}{3 + \dfrac{109.75}{704.08\dot{3}}} = 0.95061$ [1]

Hence the estimate of $Y_{1,4}$ using EBCT Model 1 is:

$0.95061 \times 113 + (1 - 0.95061) \times 108 = 112.75$ [1]

[Total 5]

(ii)   **How to produce a revised estimate that takes volume into account**

We could use EBCT Model 2, with risk volumes $P_{i,j}$ for Quarter i of Year j of:

$P_{1,4} = 1$ $P_{1,3} = \dfrac{1}{1.1}$ $P_{1,2} = \dfrac{1}{1.1^2}$ $P_{1,1} = \dfrac{1}{1.1^3}$ [2]

Alternatively, we could adjust the given figures to Year 4 volumes by multiplying all the Year 3 figures by 1.1, multiplying all the Year 2 figures by 1.1^2 and multiplying all the Year 1 figures by 1.1^3. We could then apply EBCT Model 1 to the adjusted figures.

(iii) *Maximum likelihood estimate*

Assuming $Y_{1,1} \sim Poisson(\lambda)$, $Y_{1,2} \sim Poisson(1.1\lambda)$ and $Y_{1,3} \sim Poisson(1.1^2\lambda)$, the likelihood function based on the Quarter 1 data is:

$$L(\lambda) = P(Y_{1,1} = 98)P(Y_{1,2} = 117)P(Y_{1,3} = 124)$$

$$= \frac{e^{-\lambda}\lambda^{98}}{98!} \times \frac{e^{-1.1\lambda}(1.1\lambda)^{117}}{117!} \times \frac{e^{-1.1^2\lambda}(1.1^2\lambda)^{124}}{124!}$$

$$= Ce^{-3.31\lambda}\lambda^{339} \qquad\qquad [1]$$

Taking logs gives:

$$\ln L(\lambda) = \ln C - 3.31\lambda + 339\ln\lambda \qquad\qquad [\tfrac{1}{2}]$$

The derivative of the log-likelihood function is:

$$\frac{d}{d\lambda}\ln L(\lambda) = -3.31 + \frac{339}{\lambda} \qquad\qquad [1]$$

This is equal to 0 when:

$$\lambda = \frac{339}{3.31} = 102.417 \qquad\qquad [\tfrac{1}{2}]$$

Checking the second derivative:

$$\frac{d^2}{d\lambda^2}\ln L(\lambda) = -\frac{339}{\lambda^2} < 0 \Rightarrow max \qquad\qquad [1]$$

So, $\hat{\lambda}$, the maximum likelihood estimate of λ, is 102.417.

Assuming $Y_{1,4} \sim Poisson(1.1^3\lambda)$, the maximum likelihood estimate of $E(Y_{1,4})$ is $1.1^3\hat{\lambda} = 136.317$.

$$[1]$$
$$[Total\ 5]$$

(iv) *Comparison of assumptions*

In the EBCT approach of parts (i) and (ii), we are not assuming any particular distribution for the random variables $Y_{i,j}$. In the maximum likelihood approach in part (iii), we are explicitly assuming that each $Y_{i,j}$ follows a Poisson distribution. [1]

Also, the EBCT approach assumes that the data from all 4 quarters provide us with information about Quarter 1, whereas the maximum likelihood approach only considers the data from Quarter 1.   [1]
[Total 2]

15.4   (i)   *Analysis using EBCT Model 1*

Using EBCT Model 1, we obtain the following numerical values:

	Year 1	Year 2	Year 3	Year 4	$\bar{x}_i$	$\frac{1}{3}\sum(x_{ij}-\bar{x}_i)^2$
Insurer A	14.2	15.8	22.7	19.0	17.925	14.1158
Insurer B	58.6	63.1	81.0	64.2	66.725	96.4358
Insurer C	123	132	161	133	137.25	270.9167

Using the formulae given on page 29 of the *Tables*, we obtain the following estimates:

$E[m(\theta)]$ $\frac{1}{3}(17.925 + 66.725 + 137.25) = 73.967$ [1]

$E[s^2(\theta)]$ $\frac{1}{3}(14.1158 + 96.4358 + 270.9167) = 127.1561$ [2]

$\text{var}[m(\theta)]$ $\frac{1}{2}\Big[(17.925 - 73.967)^2 + (66.725 - 73.967)^2$

$\qquad\qquad + (137.25 - 73.967)^2\Big] - \frac{1}{4} \times 127.1561 = 3{,}567.1562$ [2]

The credibility factor for EBCT Model 1 is:

$$Z = \frac{n}{n + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

Using the estimated values of $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ given above, this is:

$$\frac{4}{4 + \dfrac{127.1561}{3{,}567.1562}} = 0.9912$$ [½]

So the EBCT premium for the coming year for Insurer B is:

$66.725 \times 0.9912 + 73.967 \times 0.0088 = £66.79m$ [½]
[Total 6]

(ii)   *Analysis using EBCT Model 2*

We first need to use the data to calculate the statistics we need. Using the notation given in the *Tables*, we have:

$$\bar{P}_1 = 163 + 189 + 252 + 199 = 803$$ [½]

The same approach for the other insurers gives the values:

$$\bar{P}_2 = 19,353$$ [½]

$$\bar{P}_3 = 71,729$$ [½]

For the whole portfolio we have:

$$\bar{P} = 803 + 19,353 + 71,729 = 91,885$$ [½]

and:

$$P^* = \frac{1}{11}\left[803\left(1 - \frac{803}{91,885}\right) + 19,353\left(1 - \frac{19,353}{91,885}\right) + 71,729\left(1 - \frac{71,729}{91,885}\right)\right]$$

$$= 2,891.5793$$ [½]

Now the x_i's in the question are actually the y_i's in the EBCT Model 2 notation. Using $w_i = x_i / P_i$ to stand for the premium per unit of risk volume (*ie* the x_i's in the Model 2 notation) we get:

$$\bar{w}_1 = \frac{14.2 + 15.8 + 22.7 + 19}{803} = 0.089290$$ [½]

$$\bar{w}_2 = 0.013791$$ [½]

$$w_3 = 0.007654$$ [½]

For the whole portfolio we have:

$$\bar{w} = \frac{14.2 + 15.8 + \cdots + 161 + 133}{91,885} = 0.009660$$ [½]

Using the summary statistics given in the question, we can now estimate for $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$:

$E[m(\theta)]$ 0.009660   [½]

$E[s^2(\theta)]$ $\dfrac{1}{3 \times 3}(0.014667 + 0.006103 + 0.003979) = 0.002750$ [½]

$$\text{var}[m(\theta)] \qquad \frac{1}{2,891.5793}\left[\frac{1}{11}(5.106461 + 0.336408 + 0.292641) - (0.002750)\right]$$

$$= 0.0001794 \qquad\qquad [1]$$

So the estimated credibility factor for Insurer B is:

$$\frac{19,353}{19,353 + \dfrac{0.002750}{0.0001794}} = 0.999208 \qquad\qquad [\frac{1}{2}]$$

Hence the credibility premium per unit of risk volume for Insurer B is:

$$0.999208 \times 0.013791 + 0.000792 \times 0.009660 = 0.013788 \qquad\qquad [\frac{1}{2}]$$

Assuming a risk volume in the coming year of 4,800, the risk premium for Insurer B is £66.18m.

$$[\frac{1}{2}]$$
$$[\text{Total 8}]$$

(iii) *Comment*

The two models give fairly similar results. The estimate in Model 2 will depend on the prediction of risk volume for the coming year. $\qquad [\frac{1}{2}]$

In both cases we have used a very high value for the credibility factor. So we are effectively ignoring the data from the other insurers, and are basing our estimate almost entirely on the data from Insurer B. $\qquad [\frac{1}{2}]$

This seems sensible, given that both the volume figures and the average claim amounts appear to be quite variable between the three different insurers. This suggests that we should not place too much emphasis on the data from Insurers A and C, and focus on the information that we have for Insurer B. $\qquad [1]$

$$[\text{Total 2}]$$

15.5 (i) *Updated estimates and the credibility premium*

Since $E[m(\theta)]$ is estimated by $\overline{x}$, we know that:

$$\frac{\displaystyle\sum_{i=1}^{6}\sum_{j=1}^{10} P_{ij}\, x_{ij}}{\displaystyle\sum_{i=1}^{6}\sum_{j=1}^{10} P_{ij}} = 4.00$$

So:

$$\sum_{i=1}^{6}\sum_{j=1}^{10} P_{ij}\, x_{ij} = 4 \times 1,498 = 5,992 \qquad\qquad [1]$$

From the data given for Insurer I:

$$\sum_{j=1}^{10} y_{Ij} = \sum_{j=1}^{10} P_{Ij} x_{Ij} = 1,133 \quad \text{and} \quad \sum_{j=1}^{10} P_{Ij} = 287$$
[1]

So the updated estimate of $E[m(\theta)]$ is:

$$\frac{\sum_{i=1}^{7} \sum_{j=1}^{10} P_{ij} x_{ij}}{\sum_{i=1}^{7} \sum_{j=1}^{10} P_{ij}} = \frac{5,992 + 1,133}{1,498 + 287} = 3.9916$$
[1]

Now consider $E[s^2(\theta)]$. The estimate of $E[s^2(\theta)]$ is calculated using the formula:

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{n-1} \sum_{j=1}^{n} P_{ij} (x_{ij} - \bar{x}_i)^2 \right\}$$

Since $N = 6$ (initially) and $n = 10$, we know that:

$$\sum_{i=1}^{6} \sum_{j=1}^{10} P_{ij} (x_{ij} - \bar{x}_i)^2 = 62.8 \times 9 \times 6 = 3,391.2$$
[1]

We now need to add on the contribution from Insurer I, *ie* $\sum_{j=1}^{10} P_{Ij} (x_{Ij} - \bar{x}_I)^2$. We can calculate this

by rewriting it as:

$$\sum_{j=1}^{10} P_{Ij} (x_{Ij} - \bar{x}_I)^2 = \sum_{j=1}^{10} P_{Ij} x_{Ij}^2 - 2\bar{x}_I \sum_{j=1}^{10} P_{Ij} x_{Ij} + \bar{x}_I^2 \sum_{j=1}^{10} P_{Ij}$$

Since $\bar{x}_I = \dfrac{\sum_{j=1}^{10} P_{Ij} x_{Ij}}{\sum_{j=1}^{10} P_{Ij}}$, the last two terms can be added together to give:

$$\sum_{j=1}^{10} P_{Ij} (x_{Ij} - \bar{x}_I)^2 = \sum_{j=1}^{10} P_{Ij} x_{Ij}^2 - \bar{x}_I^2 \sum_{j=1}^{10} P_{Ij}$$

$$= \left(22 \times \left(\frac{100}{22} \right)^2 + \cdots + 36 \times \left(\frac{131}{36} \right)^2 \right) - \left(\frac{1,133}{287} \right)^2 \times 287$$

$$= 4,539.0874 - 4,472.7840 = 66.3035$$
[2]

So:

$$\sum_{i=1}^{7} \sum_{j=1}^{10} P_{ij}(x_{ij} - \overline{x}_i)^2 = 3,391.2 + 66.3035 = 3,457.5035$$ [1]

and the updated estimate of $E[s^2(\theta)]$ is:

$$\frac{1}{7} \times \frac{1}{9} \times 3,457.5035 = 54.881$$ [1]

We are also asked for the updated estimate of $\text{var}[m(\theta)]$. The estimate of $\text{var}[m(\theta)]$ is calculated using the formula:

$$\frac{1}{P^*}\left(\frac{1}{Nn-1} \sum_{i=1}^{N} \sum_{j=1}^{n} P_{ij}(x_{ij} - \overline{x})^2 - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n-1} \sum_{j=1}^{n} P_{ij}(x_{ij} - \overline{x}_i)^2 \right)$$

We know that:

$$\frac{1}{18.24}\left(\frac{1}{59} \sum_{i=1}^{6} \sum_{j=1}^{10} P_{ij}(x_{ij} - \overline{x})^2 - 62.8 \right) = 42.1$$

So:

$$\begin{aligned}
\sum_{i=1}^{6} \sum_{j=1}^{10} P_{ij}(x_{ij} - \overline{x})^2 &= \sum_{i=1}^{6} \sum_{j=1}^{10} P_{ij}x_{ij}^2 - \overline{x}^2 \sum_{i=1}^{6} \sum_{j=1}^{10} P_{ij} \\
&= 59\,(42.1 \times 18.24 + 62.8) \\
&= 49,011.536
\end{aligned}$$ [1]

It follows that:

$$\sum_{i=1}^{6} \sum_{j=1}^{10} P_{ij}\, x_{ij}^2 = 49,011.536 + 4^2 \times 1,498 = 72,979.536$$ [1]

and adding the contribution from Insurer I:

$$\sum_{i=1}^{7} \sum_{j=1}^{10} P_{ij}\, x_{ij}^2 = 72,979.536 + 4,539.0874 = 77,518.6234$$ [1]

So:

$$\sum_{i=1}^{7} \sum_{j=1}^{10} P_{ij}(x_{ij} - \overline{x})^2 = 77,518.6234 - 3.9916^2 \times (1,498 + 287) = 49,078.449$$ [1]

We now need the updated value of P^*. We know that:

$$P^* = \frac{1}{Nn-1}\sum_{i=1}^{N}\overline{P}_i\left(1-\frac{\overline{P}_i}{\overline{\overline{P}}}\right) = \frac{1}{Nn-1}\left(\sum_{i=1}^{N}\overline{P}_i - \sum_{i=1}^{N}\frac{\overline{P}_i^2}{\overline{\overline{P}}}\right)$$

and:

$$\sum_{i=1}^{6}\overline{P}_i = \overline{\overline{P}} = 1,498$$

So:

$$\frac{1}{59}\left(1,498 - \frac{1}{1,498}\sum_{i=1}^{6}\overline{P}_i^2\right) = 18.24$$

Hence:

$$\sum_{i=1}^{6}\overline{P}_i^2 = 631,916.32$$

$$\sum_{i=1}^{7}\overline{P}_i^2 = 631,916.32 + 287^2 = 714,285.32$$

and:

$$P^* = \frac{1}{69}\left(\sum_{i=1}^{7}\overline{P}_i - \frac{1}{\overline{\overline{P}}}\sum_{i=1}^{7}\overline{P}_i^2\right) = \frac{1}{69}\left((1,498+287) - \frac{714,285.32}{(1,498+287)}\right) = 20.07015 \qquad [2]$$

So the updated estimate of $\text{var}[m(\theta)]$ is:

$$\frac{1}{20.07015}\left(\frac{1}{69}(49,078.449) - 54.881\right) = 32.70533 \qquad [1]$$

We can now find the EBCT premium for the coming year for Insurer I.

The estimated credibility factor for Insurer I is:

$$\frac{287}{287 + \dfrac{54.881}{32.705}} = 0.994187 \qquad [1]$$

So the credibility premium per unit volume is:

$$0.994187 \times \frac{1,133}{287} + 0.005813 \times 3.9916 = 3.94799 \qquad [1]$$

and the credibility premium for the coming year is:

$$3.94799 \times 38 = 150.024$$ [1]
                                                                    [Total 18]

(ii)   ***Insurer K***

The value of $\overline{x}_K$ is $\dfrac{986}{327} = 3.0153$, which is lower than $\overline{x}_I$, and is also lower than $\overline{x}$.   [½]

The estimated credibility factor for Insurer K is:

$$\frac{327}{327 + \dfrac{54.881}{32.705}} = 0.99489$$

This is bigger than the corresponding factor for Insurer I.   [½]

So we place more emphasis on the mean of the direct data for Insurer K, and this reduces the credibility estimate. As a result, the credibility premium per unit of volume will be lower for Insurer K than for Insurer I.   [1]
                                                                    [Total 2]

End of Part 4

What next?

1. Briefly **review** the key areas of Part 4 and/or re-read the **summaries** at the end of Chapters 13 to 15.

2. Ensure you have attempted some of the **Practice Questions** at the end of each chapter in Part 4. If you don't have time to do them all, you could save the remainder for use as part of your revision.

3. Attempt **Assignment X4**.

Time to consider …

… 'rehearsal' products

Mock Exam and Marking – You can attempt the Mock Exam and get it marked. Results of surveys have found that students who do a mock exam of some form have significantly higher pass rates. Students have said:

> *'I find the mock a useful tool in completing my pre-exam study. It helps me realise the areas I am weaker in and where I need to focus my study.'*

> *'Overall the marking was extremely useful and gave detailed comments on where I was losing marks and how to improve on my answers and exam technique. This is exactly what I was looking for – thank you!'*

You can find lots more information on our website at www.ActEd.co.uk.

Buy online at www.ActEd.co.uk/estore

And finally ...

Good luck!

Subject CS1: Assignment X1

2019 Examinations

Time allowed: 2¾ hours

Instructions to the candidate

1. *Please:*

    – *attempt all of the questions, as far as possible under exam conditions*

    – *begin your answer to each question on a new page*

    – *leave at least 2cm margin on all borders*

    – *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*

    – *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*

    – *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

2. *Please **do not**:*

    – *use headed paper*

    – *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow
the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an
electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist

- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com

- **do not submit a photograph of your script**

- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* 'CS1 Assignment X1 No. 12345', inserting your ActEd Student Number for 12345.

- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.

- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.

- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).

- Please include the 'feedback from marker' sheet when scanning.

- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS1: Assignment X1

2019 Examinations

Please complete the following information:

Name:

Number of following pages: _____

Please put a tick in this box if you have solutions and a cross if you do not: [ ]

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information): [ ]

ActEd Student Number (see Note below):

[ ] [ ] [ ] [ ] [ ]

Time to do assignment
(see Note below):   _____ hrs _____ mins

Under exam conditions
(delete as applicable):   yes / nearly / no

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

Note: If you take more than 2¾ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	**Total**
$\frac{}{2}$	$\frac{}{3}$	$\frac{}{3}$	$\frac{}{4}$	$\frac{}{5}$	$\frac{}{5}$	$\frac{}{5}$	$\frac{}{5}$	$\frac{}{6}$	$\frac{}{8}$	$\frac{}{9}$	$\frac{}{8}$	$\frac{}{17}$	$\frac{}{80}$ =_____%

Grade: A   B   C   D   E   **Marker's initials:** _____

Please tick the following checklist so that your script can be marked quickly. Have you:

[ ]   Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?

[ ]   Written your full name in the box above?

[ ]   Completed your ActEd Student Number in the box above?

[ ]   Recorded your attempt conditions?

[ ]   Numbered all pages of your script (excluding this cover sheet)?

[ ]   Written the total number of pages (excluding the cover sheet) in the space above?

[ ]   Included your Marking Voucher or ordered Series X Marking?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress   B = Good progress   C = Average progress
D = Below average progress   E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

X1.1 An actuarial student has said that the following three distributions are the same:

(i)   the chi square distribution with 2 degrees of freedom

(ii)   the exponential distribution with mean ½

(iii)   the gamma distribution with $\alpha = 1$ and $\lambda = ½$.

State with reasons whether the student is correct.   [2]

X1.2 The number of telephone calls per hour on a working day received at an insurance office follows a Poisson distribution with mean 2.5.

(i)   Calculate the probability that more than 7 telephone calls are received on a working day between 9am and 11am.   [1]

(ii)   Calculate the probability that, if the office opens at 8am, there are no telephone calls received until after 9am.   [2]
                                                                                    [Total 3]

X1.3 The random variable X has an exponential distribution with parameter λ.

Use the moment generating function to determine an expression for $E[X^4]$.   [3]

X1.4 A random variable X has probability density function:

$$f(x) = \frac{2 \times 5^2}{(5+x)^3}, \quad x > 0$$

(i)   Determine an expression for the distribution function, $F(x)$.   [1]

(ii)   Calculate two simulated observations from the distribution using the random numbers 0.656 and 0.285 selected from the $U(0,1)$ distribution.   [3]
                                                                                    [Total 4]

X1.5 The random variable X has a beta distribution with parameters $\alpha = 1$ and $\beta = 4$.

(i)   State the value of $E(X)$.   [1]

(ii)   Determine the median of X.   [3]

(iii)   Hence comment on the shape of this distribution.   [1]
                                                                                    [Total 5]

X1.6   A large life office has 1,000 policyholders, each of whom has a probability of 0.01 of dying during the next year (independently of all other policyholders).

      (i)   Derive a recursive relationship for the binomial distribution of the form:

$$P(X = x) = kg(x)P(X = x - 1)$$

            where k is a constant and $g(x)$ is a function of x.   [2]

      (ii)   Calculate the probabilities of the following events:

            (a)   there will be no deaths during the year

            (b)   there will be more than two deaths during the year

            (c)   there will be exactly twenty deaths during the year.   [3]

                                                                        [Total 5]

X1.7   On a portfolio of insurance policies, the claim size, Y is assumed to depend on the age of the policyholder, X. Suppose that the conditional mean and variance of Y are:

$$E(Y \mid X = x) = 2x + 400$$

$$\text{var}(Y \mid X = x) = \frac{x^2}{2}$$

The distribution of X over the portfolio is assumed to be normal with mean 50 and standard deviation 14.

Calculate the unconditional mean and standard deviation of Y.   [5]

X1.8   (i)   For a pair of jointly distributed random variables X and Y, derive the result:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$ [2]

      (ii)   The random variables *X* and *Y* are jointly distributed with standard deviations of 5 and 7 respectively and $\text{corr}(X, Y) = -3/7$. Calculate the standard deviation of $3X - 2Y + 5$.   [3]

                                                                              [Total 5]

X1.9   (i)   The random variables X and Y have a discrete joint distribution with joint probability function:

$$P(X = x, Y = y) = \begin{cases} c(x + 2y) & x = 0, 1, 2 \text{ and } y = 0, 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

where c is an appropriate constant.

Determine the conditional distribution of X given $Y = y$ for each value of y.   [3]

(ii)   It is subsequently discovered that the random variables, X and Y, are in fact continuous over the ranges $0 < x < 2$ and $0 < y < 2$ with the probability density function being the same as the probability function.

Determine the conditional distribution of X given $Y = y$.   [3]

[Total 6]

X1.10 (i)   For a lognormal distribution with mean m and standard deviation s, give an expression for μ, the mean of the underlying normal distribution.   [3]

(ii)   Claim amounts for a particular type of medical negligence are lognormally distributed with mean £15,000 and standard deviation £8,000. Calculate the probability that the next claim exceeds £20,000.   [3]

(iii)   An actuary is examining the number of large claims received by her company. To do this she counts the number of claims arriving until she receives one that exceeds £20,000. Calculate the mean number of claims that she will count (not including the £20,000 claim).   [2]

[Total 8]

X1.11 (i) Prove that if the random variable X has MGF $M_X(t)$, then:

(a) $E(X) = M_X'(0)$

(b) $\text{var}(X) = M_X''(0) - (M_X'(0))^2$. [3]

(ii) Let X be a random variable with probability density function:

$$f(x) = \begin{cases} \dfrac{1}{2}e^x & ; \quad x \le 0 \\[2mm] \dfrac{1}{2}e^{-x} & ; \quad x > 0 \end{cases}$$

(a) Show that the moment generating function of X is given by:

$$M_X(t) = (1 - t^2)^{-1}$$

for $|t| < 1$.

(b) Hence calculate the mean and the variance of X using the moment generating function in part (ii)(a). [6]

[Total 9]

X1.12 (i) Derive, from first principles, the cumulant generating function of a gamma distribution and show that it can be written as:

$$C_X(t) = -\alpha \ln\left(1 - \frac{t}{\lambda}\right) \qquad t < \lambda$$ [4]

(ii) Hence derive an expression for the coefficient of skewness of a gamma distribution. [4]

[Total 8]

X1.13 *X* and *Y* are discrete random variables. The only possible combinations of these two variables have the following probabilities:

		X	
	0	1	2
0	$\frac{1}{2}$	0	$\frac{1}{16}$
Y **1**	0	$\frac{1}{8}$	0
2	$\frac{1}{4}$	$\frac{1}{16}$	0

(i)   Show that *X* and *Y* are:

      (a)   *not* independent

      (b)   *not* uncorrelated.   [4]

(ii)   State the circumstances under which the result $E(X) = E[E(X \mid Y)]$ holds.   [1]

(iii)   Calculate:

      (a) $E(X + Y \mid X = 1)$

      (b) $E(X \mid Y = 2)$

      (c) $\text{var}(X \mid Y = 2)$.   [7]

(iv)   Determine the values of the random variable $E[Y^2 \mid X]$ and hence calculate $E[E(Y^2 \mid X)]$. [3]

(v)   Calculate $E[Y^2]$ and comment on your answer.   [2]

[Total 17]

END OF PAPER

All study material produced by ActEd is copyright and is sold for the exclusive use of the purchaser. The copyright is owned by Institute and Faculty Education Limited, a subsidiary of the Institute and Faculty of Actuaries.

Unless prior authority is granted by ActEd, you may not hire out, lend, give out, sell, store or transmit electronically or photocopy any part of the study material.

You must take care of your study material to ensure that it is not used or copied by anybody else.

Legal action will be taken if these terms are infringed. In addition, we may seek to take disciplinary action through the profession or through your employer.

These conditions remain in force after you have finished using the course.

Subject CS1: Assignment X2

2019 Examinations

Time allowed: 2¾ hours

Instructions to the candidate

1. *Please:*

   – *attempt all of the questions, as far as possible under exam conditions*

   – ***begin your answer to each question on a new page***

   – ***leave at least 2cm margin on all borders***

   – *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*

   – *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*

   – *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

2. *Please **do not**:*

   – *use headed paper*

   – *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow
the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an
electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist

- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com

- **do not submit a photograph of your script**

- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* 'CS1 Assignment X2 No. 12345', inserting your ActEd Student Number for 12345.

- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.

- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.

- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).

- Please include the 'feedback from marker' sheet when scanning.

- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS1: Assignment X2

2019 Examinations

Please complete the following information:

Name:

ActEd Student Number (see Note below):

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

Number of following pages: _____

Please put a tick in this box if you have solutions and a cross if you do not:

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information):

Time to do assignment
(see Note below):   _____ hrs _____ mins

Under exam conditions
(delete as applicable):   yes / nearly / no

Note: If you take more than 2¾ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	**Total**
$\frac{}{3}$	$\frac{}{3}$	$\frac{}{3}$	$\frac{}{3}$	$\frac{}{3}$	$\frac{}{4}$	$\frac{}{6}$	$\frac{}{6}$	$\frac{}{6}$	$\frac{}{7}$	$\frac{}{7}$	$\frac{}{7}$	$\frac{}{7}$	$\frac{}{15}$	$\frac{}{80}$ =_____%

Grade: A   B   C   D   E   **Marker's initials: _____**

Please tick the following checklist so that your script can be marked quickly. Have you:

[   ]   Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?

[   ]   Written your full name in the box above?

[   ]   Completed your ActEd Student Number in the box above?

[   ]   Recorded your attempt conditions?

[   ]   Numbered all pages of your script (excluding this cover sheet)?

[   ]   Written the total number of pages (excluding the cover sheet) in the space above?

[   ]   Included your Marking Voucher or ordered Series X Marking?

[   ]   Rated your X1 marker at **www.ActEd.co.uk/marking**?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress   B = Good progress   C = Average progress
D = Below average progress   E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

X2.1   Determine:

(i) $P(F_{9,24} < 3.256)$ [1]

(ii) $P(F_{3,5} < 0.18836)$ [1]

(iii)   the value of a such that $P(F_{8,6} > a) = 0.95$.   [1]

[Total 3]

X2.2   Suppose that a random sample of nine observations is taken from a normal distribution with mean $\mu = 0$. Let $\overline{X}$ and S^2 denote the sample mean and variance respectively.

Determine (to 2 decimal places) the probability that the value of $\overline{X}$ exceeds that of S, *ie* determine $P(\overline{X} > S)$.   [3]

X2.3   The waist measurements (in cm) of six male patients before and after undergoing a medically controlled diet are as follows:

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6
Before	106	98	110	100	105	96
After	98	97	82	89	80	90

Calculate a 90% confidence interval for the reduction in waist measurement following the diet. [3]

X2.4   A random sample of size $2n$ is taken from a geometric distribution for which:

$$P(X = x) = pq^{x-1} \qquad x = 1, 2, \ldots$$

Give an expression for the likelihood that the sample contains equal numbers of odd and even values of X.   [3]

X2.5   A random sample of 16 observations ($x_1, \ldots, x_{16}$) from a normal distribution gives:

$$\sum_{i=1}^{16} x_i = 128 \qquad \sum_{i=1}^{16} x_i^2 = 1{,}168$$

Calculate a 90% confidence interval for the population standard deviation.   [3]

X2.6   The following sample was taken from a normal distribution with mean μ and variance 20:

   56, 32, 49, 57, 44

   (i)   Calculate a symmetrical 95% confidence interval for μ.   [2]

   (ii)   Repeat part (i) for the situation where the population variance is unknown.   [2]
                                                                                                [Total 4]

X2.7   Two children play an 'incy-wincy' spider game. They take it in turns to roll two dice each and
           move their spiders up their drainpipes as follows:

Score	Movement
2, 3 or 4	Down 1
5, 6 or 7	Stay same
8, 9 or 10	Up 1
11 or 12	Up 2

   (i)   Using a normal approximation, calculate the probability that after 15 turns a child's spider
           will have moved up more than 8 squares from the start.   [5]

   (ii)   Comment briefly on the suitability of this approximation.   [1]
                                                                                                [Total 6]

X2.8   A sample of 50 independent and identically distributed observations from an $Exp(\lambda)$ distribution
           gave:

Range	$0 \le x < 1$	$1 \le x < 2$	$x \ge 2$
Frequency	30	15	5

   (i)   Show that the log-likelihood can be expressed as:

   $$\ln L(\lambda) = \text{constant} - 25\lambda + 45\ln(1 - e^{-\lambda})$$

   explaining clearly why the constant has arisen.   [3]

   (ii)   Hence calculate the maximum likelihood estimate of λ.   [3]
                                                                                                [Total 6]

X2.9   A random variable X has probability density function:

$$2e^{-2(x-\theta)} \qquad x \ge \theta$$

where the value of θ is unknown.

Five observations of X are:

1.90, 2.97, 1.88, 2.94 and 1.56.

(i)   Derive a formula for the maximum likelihood estimator of θ and obtain the maximum likelihood estimate for this sample.   [3]

(ii)   Show that $E(X) = \theta + \dfrac{1}{2}$ and hence calculate the method of moments estimate of θ .   [2]

(iii)   Comment briefly on your results.   [1]
                                                                                        [Total 6]

X2.10   A large life office has *n* policyholders, each with a probability of 0.01 of dying during the next year (independently of all other policyholders).

Calculate the approximate probability that there will be between 9 and 16 (both inclusive) deaths during the year, when:

(i) $n = 400$ [3]

(ii) $n = 3,000$.   [4]
                                                                                        [Total 7]

X2.11   The gamma distribution, with parameters α and λ , has moment generating function:

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-\alpha}$$

(i)   Show, using moment generating functions, that the sum of two independent gamma random variables, each with second parameter λ , is also a gamma random variable and state its parameters.   [2]

(ii)   A random sample $X_1, \dots, X_n$ is taken from a *Gamma*(α, λ) distribution. Derive the moment generating function of $2\lambda \sum X_i$, and hence show that it has a $\chi^2_{2n\alpha}$ distribution.   [3]

(iii)   Suppose that $\overline{X}$ is the mean of a random sample of size 5 taken from a *Gamma*$(2, 0.1)$ distribution. Use the result from part (ii) to calculate the probability that $\overline{X}$ exceeds 40.   [2]
                                                                                        [Total 7]

X2.12 The number of claims per annum from a certain type of medical insurance policy sold to policyholders over the age of 60 is believed to follow a *Poi*(λ) distribution, where the parameter λ is unknown. A sample of 10 policies gave rise to the following numbers of claims:

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

(i)   Use a normal approximation to calculate an approximate 99% confidence interval for the Poisson parameter λ.   [3]

(ii)   Comment on the accuracy of the interval obtained in part (i).   [2]

(iii)   Write down the equations that you would use to obtain the confidence interval for λ using an accurate method.   [2]

You are not required to solve these equations.

[Total 7]

X2.13 A random sample ($x_1, \dots, x_n$) is taken from a Poisson distribution, with parameter μ.

(i)   Show that the maximum likelihood estimator of μ is:

$$\hat{\mu} = \overline{X}$$ [3]

(ii)   Determine the bias and mean square error of $\hat{\mu}$.   [2]

(iii)   Show that the variance of $\hat{\mu}$ attains the Cramér-Rao lower bound.   [2]

[Total 7]

X2.14   A random sample of 10 pet insurance claims had an average size of £680. It is believed that claim amounts are exponentially distributed.

(i)   Using the fact that if $X_1, \ldots, X_n$ are exponentially distributed with parameter λ, then $2n\lambda\bar{X}$ has a χ^2_{2n} distribution, where $\bar{X}$ is the mean of $X_1, \ldots, X_n$, calculate an exact 90% confidence interval for the *mean* pet insurance claim size.   [3]

(ii)   Write down the likelihood function in terms of the mean μ of the exponential distribution and hence show that the maximum likelihood estimator of μ is $\bar{X}$.   [5]

(iii)   (a)   Show that the Cramér-Rao lower bound for estimators of the mean of the exponential distribution is given by:

$$\frac{\mu^2}{n}$$

(b)   Hence, calculate the estimated asymptotic standard error of the mean, $\bar{X}$.   [3]

(iv)   (a)   Use your results from (iii) and the asymptotic properties of estimators to calculate an approximate 90% confidence interval for the mean claim size.

(b)   Comment on the confidence intervals produced in (i) and (iv)(a).   [4]
                                                                                                                [Total 15]

END OF PAPER

Subject CS1: Assignment X3

2019 Examinations

Time allowed: 3¼ hours

Instructions to the candidate

1. *Please:*

    –   *attempt all of the questions, as far as possible under exam conditions*

    –   ***begin your answer to each question on a new page***

    –   ***leave at least 2cm margin on all borders***

    –   *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*

    –   *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*

    –   *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

2. *Please **do not**:*

    –   *use headed paper*

    –   *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow
the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an
electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist

- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com

- **do not submit a photograph of your script**

- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* 'CS1 Assignment X3 No. 12345', inserting your ActEd Student Number for 12345.

- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.

- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.

- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).

- Please include the 'feedback from marker' sheet when scanning.

- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS1: Assignment X3

2019 Examinations

Please complete the following information:

Name:

Number of following pages: _____

Please put a tick in this box if you have solutions and a cross if you do not:

ActEd Student Number (see Note below):

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information):

Time to do assignment
(see Note below):   _____ hrs _____ mins

Under exam conditions
(delete as applicable):   yes / nearly / no

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

Note: If you take more than 3¼ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	**Total**	
_	_	_	_	_	_	_	_	__	__	__	__	__	____	=____%
1	1	2	6	3	4	6	6	12	10	15	15	19	100	

Grade: A   B   C   D   E   **Marker's initials: _____**

Please tick the following checklist so that your script can be marked quickly. Have you:

[   ]   Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?
[   ]   Written your full name in the box above?
[   ]   Completed your ActEd Student Number in the box above?
[   ]   Recorded your attempt conditions?
[   ]   Numbered all pages of your script (excluding this cover sheet)?
[   ]   Written the total number of pages (excluding the cover sheet) in the space above?
[   ]   Included your Marking Voucher or ordered Series X Marking?
[   ]   Rated your X2 marker at **www.ActEd.co.uk/marking**?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress   B = Good progress   C = Average progress
D = Below average progress   E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

Questions 1 to 4 are based on the data set given in the table below:

x	2	4	5	9
y	3.0	6.8	8.2	15.1

$$\sum x = 20 \quad \sum x^2 = 126 \quad \sum xy = 210.1 \quad \sum y = 33.1 \quad \sum y^2 = 350.49$$

It is proposed to fit a simple linear regression model to these data:

$$Y_i = a + bx_i + e_i \text{ where } e_i \sim N(0, \sigma^2)$$

X3.1   Calculate the least squares estimate for the slope parameter b.   [1]

X3.2   Calculate an unbiased estimate of the variance parameter.   [1]

X3.3   Determine a 95% confidence interval for b based on the sample data.   [2]

X3.4   (i)   Show that the sample correlation coefficient is 0.9995 to 4 significant figures.   [1]

       (ii)   Hence obtain:

           (a)   a 95% confidence interval for ρ, the underlying correlation coefficient

           (b)   the coefficient of determination, R^2, and comment on your result.   [5]

                                                                                                      [Total 6]

X3.5   It is desired to test the value of the parameter p for a random variable that has a binomial distribution. In order to test the null hypothesis $H_0 : p = 0.4$ against the alternative hypothesis $H_1 : p = 0.6$, the following test is devised:

The number of successes, X, in a sample of size 50 is determined. If $X \geq 25$, then H_0 is rejected.

Calculate the approximate size of this test.   [3]

X3.6   Define the following terms:

       (i)   a Type I error   [1]

       (ii)   a Type II error   [1]

       (iii)   the size of a test   [1]

       (iv)   the power of a test.   [1]

                                                                                                       [Total 4]

X3.7   A random sample from a $N(\mu, \sigma^2)$ distribution, where both parameters are unknown, gave the following values:

11.8, 5.4, 8.2, 4.6, 13.6, 10.1, 10.4, 11.2, 12.2, 17.5

Test each of the hypotheses:

(i) $\mu = 9$ [3]

(ii) $\sigma^2 = 8$ [3]

against an appropriate two-sided alternative.   [Total 6]

X3.8   Support for the current government is assessed by means of a survey of 5,000 people. Of those questioned 2,185 said that they would vote for the current government in the next election.

(i)   Test whether this proportion is greater than 42%.   [3]

Following a rather embarrassing scandal a second survey is commissioned. This time, 1,191 out of 3,000 people said that they would vote for the current government in the next election.

(ii)   Test to see if there has been any significant change in the proportion supporting the current government.   [3]

[Total 6]

X3.9   A study was carried out into the effects of smoking on life expectancy. The average number (x) of cigarettes smoked per day from age 50 by 11 individuals was calculated and the number (y) of years from age 50 until their deaths was recorded. The results were as follows:

x	0.0	1.1	17.3	10.6	25.1	5.2	11.8	40.0	15.6	13.8	3.6
y	42.3	30.7	26.3	36.8	8.9	25.1	10.8	10.0	25.2	17.2	29.1

For these data values:

$$\sum_{i=1}^{11} x_i = 144.1, \sum_{i=1}^{11} y_i = 262.4$$

$$\sum_{i=1}^{11} x_i^2 = 3,255.91, \sum_{i=1}^{11} y_i^2 = 7,481.26, \sum_{i=1}^{11} x_i y_i = 2,495.43$$

(i)   Calculate Pearson's correlation coefficient. Comment on the value obtained.   [2]

(ii)   Calculate Spearman's rank correlation coefficient and comment on the value obtained.
                                                                                                                                                  [3]

(iii)   State a general advantage of using Spearman's rank correlation coefficient.   [1]

(iv)   Carry out a test to determine if Spearman's rank correlation coefficient is significantly different from zero assuming it is appropriate to use a normal approximation and comment on this assumption.   [3]

(v)   Calculate the Kendall's rank correlation coefficient.   [3]
                                                                                                                                          [Total 12]

X3.10   1,000 male and 1,000 female subjects were chosen at random by a researcher and cross-classified according to sex and to whether or not they were colour-blind, giving the following table:

	male	female
normal	908	993
colour-blind	92	7

(i)   Perform a χ^2 test on this contingency table to show that there is overwhelming evidence against the hypothesis that there is no association between an individual's sex and whether or not the individual is colour-blind.   [5]

(ii)   A genetic model states that the human population is split in the proportions illustrated in the following table, where q $(0 < q < 1)$ is a parameter relating to the distribution of the colour-blindness defect among the relevant genes.

	male	female
normal	$\dfrac{1-q}{2}$	$\dfrac{1-q^2}{2}$
colour-blind	$\dfrac{q}{2}$	$\dfrac{q^2}{2}$

The maximum likelihood estimate of q calculated on this data is 0.0895. Test the goodness-of-fit of this model to the data.   [5]
                                                                                    [Total 10]

X3.11 Following archaeological excavations at a site in Egypt, ten samples of wood were carbon-dated and their ages x (years) estimated as:

4,900	4,750	4,820	4,710	4,760
4,570	4,300	4,680	4,800	4,670

$$\sum x = 46{,}960 \qquad \sum x^2 = 220{,}772{,}800$$

(i)   Calculate a 95% confidence interval for the true mean age of the wood found at this site.   [3]

(ii)   Present these data values graphically and comment on the validity of the confidence interval calculated in part (i).   [2]

(iii)   Ideally the archaeologist would like the 95% confidence interval for the true mean age, calculated in (i) above, to have a width of no more than 200 years.

Calculate the minimum sample size needed.   [3]

(iv)   At a second site, eight samples of wood gave the following results:

$$\sum y = 36{,}000 \quad \sum y^2 = 162{,}280{,}000$$

Calculate a 95% confidence interval for the difference between the mean ages of the wood found at the two sites.   [3]

(v)   Obtain a 90% confidence interval for the ratio of the underlying variances in the ages of the two samples of wood. Hence comment on the validity of the confidence interval given in part (iv).   [4]
[Total 15]

X3.12   A research chemist thinks he has discovered a new desiccant which is more efficient at extracting moisture from chemicals than the existing one. In order to test the claim, equal amounts of a homogeneously mixed compound are put into each of sixteen desiccators. These are divided into two batches of eight, labelled A and B, and in each batch the desiccators are numbered 1 to 8. Into each desiccator is also put a standard amount of the respective desiccant under test. Batch A contains the existing desiccant whilst the new desiccant is placed in Batch B. The desiccators are sealed for 24 hours and then the increase in weight in grams of each of the sixteen samples of desiccant is measured. The results are:

Sample number		1	2	3	4	5	6	7	8
Existing desiccant	(A)	4.59	5.05	4.49	5.33	4.66	4.98	5.67	5.23
New desiccant	(B)	4.75	5.03	4.66	5.56	4.90	4.88	5.80	5.33

$$\sum A = 40.0 \quad \sum A^2 = 201.1574 \quad \sum B = 40.91 \quad \sum B^2 = 210.3659$$

(i)   (a)   (1)   Draw a plot of the data and comment briefly.

                (2)   Perform a test to verify that the variances arising from the use of each desiccant are not significantly different and comment briefly in relation to your plot of the data.

        (b)   Use a *t* test to investigate the claim that the new desiccant extracts more moisture than the existing one.   [8]

(ii)   It was subsequently discovered that eight different compounds had been used in the above test. The *i* th pair of desiccators A and B had contained equal weights of compound $i, i = 1, 2, \ldots, 8$. Perform a new analysis with the same aim, as in part (i)(b) above, again using a *t* test.   [5]

(iii)   Comment on any difference found between the analyses, and the cause.   [2]

                                                                                          [Total 15]

X3.13 It is thought that a plumber charges £22 per hour plus an administrative charge of £15 per call-out.

A sample of eight invoices was obtained corresponding to jobs with durations of 1 hour, 2 hours, ... , 8 hours. For each invoice the total cost of the job was noted with the following results:

Time x (hours):	1	2	3	4	5	6	7	8
Cost y (£):	40	50	81	89	122	128	151	179

$$\sum(x-\overline{x})^2 = 42 \quad \sum(y-\overline{y})^2 = 16{,}492 \quad \sum(x-\overline{x})(y-\overline{y}) = 826$$

The following model is used to represent the data:

$$Y_i = a + bx_i + e_i$$

where Y_i ($i = 1, 2, \ldots, 8$) are the costs, x_i ($i = 1, 2, \ldots, 8$) are the fixed times and e_i ($i = 1, 2, \ldots, 8$) are independent errors with a $N(0, \sigma^2)$ distribution.

(i) (a) Derive formulae for the least squares estimators of a and b.

   (b) Explain how your answer to part (i)(a) would have differed if you had been asked to calculate the maximum likelihood estimators and justify your answer.   [6]

(ii) Calculate the regression coefficients $\hat{a}$ and $\hat{b}$.   [2]

(iii) Carry out a test to establish whether or not the slope in the model agrees with the suggested £22 per hour.   [4]

(iv) Calculate a 90% confidence interval for the:

   (a) average cost of a job lasting 4 hours

   (b) cost of an individual job lasting 6 hours.   [6]

(v) Comment on relative widths of the two intervals calculated in part (iv).   [1]

[Total 19]

END OF PAPER

Subject CS1: Assignment X4

2019 Examinations

Time allowed: 3¼ hours

Instructions to the candidate

1. *Please:*

   – *attempt all of the questions, as far as possible under exam conditions*

   – ***begin your answer to each question on a new page***

   – ***leave at least 2cm margin on all borders***

   – *write in black ink using a medium-sized nib because we will be unable to mark illegible scripts*

   – *note that assignment marking is not included in the price of the course materials. Please purchase Series Marking or a Marking Voucher before submitting your script.*

   – *note that we only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2019 exams.*

2. *Please **do not**:*

   – *use headed paper*

   – *use highlighting in your script.*

At the end of the assignment

If your script is being marked by ActEd, please follow
the instructions on the reverse of this page.

In addition to this paper, you should have available actuarial tables and an
electronic calculator.

Submission for marking

You should aim to submit this script for marking by the recommended submission date. The recommended and deadline dates for submission of this assignment are listed on the summary page at the back of this pack and on our website at **www.ActEd.co.uk**.

Scripts received after the deadline date will not be marked, unless you are using a Marking Voucher. *It is your responsibility to ensure that scripts reach ActEd in good time*. If you are using Marking Vouchers, then please make sure that your script reaches us by the Marking Voucher deadline date to give us enough time to mark and return the script before the exam.

When submitting your script, please:

- complete the cover sheet, including the checklist

- scan your script, cover sheet (and Marking Voucher if applicable) and save as a pdf document, then email it to: ActEdMarking@bpp.com

- **do not submit a photograph of your script**

- **do not include the question paper in the scan.**

In addition, please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* 'CS1 Assignment X4 No. 12345', inserting your ActEd Student Number for 12345.

- The assignment should be scanned the **right way up** (so that it can be read normally without rotation) and as a single document. We cannot accept individual files for each page.

- Please set the resolution so that the script is legible and the resulting PDF **is less than 4 MB** in size.

- Do not protect the PDF in any way (otherwise the marker cannot return the script to ActEd, which causes delays).

- Please include the 'feedback from marker' sheet when scanning.

- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

Subject CS1: Assignment X4

2019 Examinations

<table>
<tr><td colspan="2">Please complete the following information:</td></tr>
<tr>
<td>

Name:

ActEd Student Number (see Note below):

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting it will help us to process your scripts quickly. If you do not know your ActEd Student Number, please email us at ActEd@bpp.com.

Your ActEd Student Number is not the same as your IFoA Actuarial Reference Number or ARN.

</td>
<td>

Number of following pages: _____

Please put a tick in this box if you have solutions and a cross if you do not: ☐

Please tick here if you are allowed extra time or other special conditions in the profession's exams (if you wish to share this information): ☐

Time to do assignment
(see Note below):   _____ hrs _____ mins

Under exam conditions
(delete as applicable):   yes / nearly / no

Note: If you take more than 3¼ hours, you should indicate how much you completed within this exam time so that the marker can provide useful feedback on your progress.

</td>
</tr>
</table>

Score and grade for this assignment (to be completed by marker):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total
$\frac{}{4}$	$\frac{}{5}$	$\frac{}{7}$	$\frac{}{8}$	$\frac{}{9}$	$\frac{}{9}$	$\frac{}{10}$	$\frac{}{5}$	$\frac{}{8}$	$\frac{}{17}$	$\frac{}{18}$	$\frac{}{100}$ =_____%

Grade: A   B   C   D   E   **Marker's initials: _____**

Please tick the following checklist so that your script can be marked quickly. Have you:

[   ]   Checked that you are using the latest version of the assignments, *ie* 2019 for the sessions leading to the 2019 exams?

[   ]   Written your full name in the box above?

[   ]   Completed your ActEd Student Number in the box above?

[   ]   Recorded your attempt conditions?

[   ]   Numbered all pages of your script (excluding this cover sheet)?

[   ]   Written the total number of pages (excluding the cover sheet) in the space above?

[   ]   Included your Marking Voucher or ordered Series X Marking?

[   ]   Rated your X3 marker at **www.ActEd.co.uk/marking**?

Please follow the instructions on the previous page when submitting your script for marking.

Feedback from marker

Notes on marker's section

The main objective of marking is to provide specific advice on how to improve your chances of success in the exam. The most useful aspect of the marking is the comments the marker makes throughout the script, however you will also be given a percentage score and the band into which that score falls. Each assignment tests only part of the course and hence does not give a complete indication of your likely overall success in the exam. However it provides a good indicator of your understanding of the material tested and the progress you are making with your studies:

A = Excellent progress   B = Good progress   C = Average progress
D = Below average progress   E = Well below average progress

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

X4.1 The number of claims per month has a type 2 negative binomial distribution with parameters k and p. The number of claims observed over n months are $x_1, \ldots, x_n$ and it is desired to estimate p.

(i) Show that the beta distribution is the conjugate prior for p. [3]

(ii) The prior distribution for p is a beta distribution with $\alpha = 3$ and $\beta = 4$. The total number of claims over 12 months is 8 and $k = 2$. Obtain the Bayesian estimate for p under quadratic loss. [1]

[Total 4]

X4.2 Show that, given a random sample of size n from a $\log N(\mu, \sigma^2)$ distribution, if an uninformative prior is used for μ, then the posterior distribution for μ is $N\left(\dfrac{1}{n} \sum \ln x_i, \dfrac{\sigma^2}{n} \right)$. [5]

X4.3 Aggregate claims, Y_j, and the policies sold, P_j, have been observed over the past $j = 1, \ldots, n$ years. Let $X_j = Y_j / P_j$ be the average amount claimed per policy.

(i) State carefully the assumptions made for a single risk by Model 2 of Empirical Bayes Credibility. [2]

Let $m(\theta) = E[X_j \mid \theta]$ and $s^2(\theta) = P_j \text{var}[X_j \mid \theta]$ where θ is a risk parameter.

(ii) Show that:

(a) $E(X_j) = E[m(\theta)]$

(b) $\text{var}(X_j) == \dfrac{1}{P_j} E[s^2(\theta)] + \text{var}[m(\theta)]$ [2]

A company has insured 3 similar risks over the past 4 years and has estimated $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ to be 2.275, 2.870 and 0.1172 respectively.

The total amount claimed (in £000's) and the number of policies sold for Risk 1 over the past 4 years are given below:

Risk 1		Year		
	1	2	3	4
Total amount claimed	200	230	260	290
Number of policies sold	100	110	120	130

(iii) Calculate the total credibility premium next year for Risk 1 using EBCT Model 2 assuming that 140 policies are sold next year. [3]

[Total 7]

X4.4 The number of claims per year from individual policies in a portfolio is believed to follow a *Poi*(λ) distribution. The *Gamma*(5,2) distribution is chosen as a prior distribution for λ. A random sample of 10 policies is observed over the last year, and the numbers of claims are found to be as follows:

4, 1, 0, 0, 2, 0, 0, 1, 3, 1

(i)   Derive the posterior distribution for λ.   [3]

(ii)   Hence, determine the Bayesian estimate for λ:

   (a)   under squared error loss

   (b)   under zero-one error loss

   (c)   under absolute error loss.   [5]

                                                                                          [Total 8]

X4.5 A random variable Y has a gamma distribution with PDF:

$$f(y) = \frac{\alpha^\alpha}{\mu^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{\alpha}{\mu}y} \qquad y > 0$$

(i)   (a)   Show that $f(y)$ can be written in the form of a member of the exponential family and identify the natural parameter and the dispersion parameter.

   (b)   Use the properties of exponential families to obtain the mean and variance of this distribution.

   (c)   Hence, or otherwise, determine the variance function.   [7]

(ii)   Another random variable, X, has a negative binomial distribution with PF:

$$P(X = x) = \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} p^k (1-p)^x \qquad x = 0,1,2,\dots \quad (k \neq 1)$$

Explain clearly why this negative binomial distribution is **not** a member of the exponential family.   [2]

                                                                                          [Total 9]

X4.6 (i) In the context of Empirical Bayes Credibility Theory Model 1, the credibility factor Z is given by:

$$Z = \frac{n}{n + \dfrac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

Explain how changes in the values of n, $E[s^2(\theta)]$ and $\text{var}[m(\theta)]$ affect the value of Z and comment on why Z behaves in this way. [3]

(ii) The table below shows the aggregate claim statistics for each of four risks over three years:

Risk i	$\sum\limits_{j=1}^{3} X_{ij}$	$\sum\limits_{j=1}^{3} (X_{ij} - \bar{X}_i)^2$
1	2,184	22,344
2	2,721	20,294
3	3,450	21,800
4	3,099	23,994

Use EBCT Model 1 to calculate the credibility premium for the coming year for Risk 3. [4]

(iii) Without carrying out any calculations, determine with reference to part (i) what would happen to the credibility factor if a fifth risk were added with:

$$\sum_{j=1}^{3} X_{5j} = 3,999 \quad \text{and} \quad \sum_{j=1}^{3} (X_{5j} - \bar{X}_5)^2 = 21,734 \qquad [2]$$

[Total 9]

X4.7   An analyst at a general insurance company is examining claims data on a portfolio of home insurance policies in a particular region. An exponential distribution models the claim amounts and the following rating factors are used:

SA   sum assured, x (as a continuous variable)

PT   property type, T_i (as a factor with $i = 1, 2, \ldots, 10$)

NB   number of bedrooms, B_j (as a factor with $j = 1, 2, \ldots, 6$)

The table below shows 4 models considered by the analyst and their scaled deviances for the data set.

Model	Parameterised form of the linear predictor	Number of parameters	Scaled deviance
SA	$\alpha + \beta x$	2	238.4
SA + PT			206.7
SA + PT + SA • PT			178.3
SA ∗ PT + NB			166.2
SA ∗ PT ∗ NB	$\alpha_{ij}'' + \beta_{ij}'' x$	120	58.9

(i)   Complete the table.   [3]

(ii)   Determine the model the analyst should choose on the basis of scaled deviance.   [5]

(iii)   Describe the further information the analyst should consider before making her recommendation about an appropriate choice of model.   [2]

[Total 10]

X4.8 A life insurance company runs a statistical analysis of mortality rates. The company considers a population of 100,000 individuals. It assumes that the number of deaths, X, during one year has a Poisson distribution with expectation $E[X] = \mu$. Over four years the company has observed the following realisations of X (number of deaths).

Year	1	2	3	4
Number of deaths (per 100,000 lives)	1,140	1,200	1,170	1,190

The maximum likelihood estimator for the parameter μ of the Poisson distribution is given by $\overline{X}$.

(i) Obtain the maximum likelihood estimate of the parameter μ using these data values. [1]

To obtain a more realistic model, it is proposed that the number of deaths should depend on the age of the population. To this end the total population is divided into four age groups of equal size and the number of deaths in each group during the following year is counted. The observed values are given in the following table.

Middle age (t) in group	25	35	45	55
Number of deaths (x) in age group	84	113	255	727

For these data values we have:

$$\Sigma t = 160, \ \Sigma t^2 = 6{,}900, \ \Sigma x = 1{,}179, \ \Sigma x^2 = 613{,}379, \ \Sigma xt = 57{,}515$$

(ii) (a) Calculate the correlation coefficient between the middle age t in a group and the number of deaths x in that group, and comment briefly on its value.

(b) Perform a linear regression of the number of deaths, x, as a function of the middle age, t, of the group. [4]

[Total 5]

X4.9 Y_1, Y_2, ..., Y_n are independent observations from a normal distribution with $E[Y_i] = \mu_i$ and $\text{var}[Y_i] = \sigma^2$.

(i) Write the density of Y_i in the form of an exponential family of distributions. [2]

(ii) Identify the natural parameter and derive the variance function. [2]

(iii) Show that the Pearson residual is the same as the deviance residual. [4]

[Total 8]

X4.10 For each of m independent policies, the probability of one claim in a year is θ ($0 < \theta < 1$) and the probability of no claims in a year is $1 - \theta$. The total number of claims in one year is a random variable X. Independent observations $x_1, ..., x_n$ of X are available. The prior distribution of θ has density $f(\theta) \propto \{\theta(1-\theta)\}^{\beta-1}$, $0 < \theta < 1$, for some constant $\beta > 0$.

(i)   (a)   Derive the posterior distribution of θ given $x_1, ..., x_n$.

      (b)   Derive the maximum likelihood estimate of θ, $g(\underset{\sim}{x})$.

      (c)   Derive the Bayesian estimate of θ under quadratic loss, and show that it takes the form of a credibility estimate:

            $$Z g(\underset{\sim}{x}) + (1-Z)\mu$$

            where μ is a quantity you should specify in terms of the prior distribution of θ.

      (d)   Explain what happens to Z as the number of data points increases.   [11]

(ii) Calculate the Bayesian estimate of θ and the value of Z if $n = 6$, $m = 10$ and $x_1 = 1$, $x_2 = 4$, $x_3 = 2$, $x_4 = 1$, $x_5 = 1$, $x_6 = 3$, when:

      (a) $\beta = 1$

      (b) $\beta = 4$.

      By considering the prior variance, comment on the effect on Z of increasing β, and relate this effect to the quality of the prior information about θ in each case.   [6]

                                                                                                    [Total 17]

X4.11 $Y_1, ..., Y_{15}$ are independent claim amounts which are being modelled as follows:

$$Y_i \sim Exp\left(\frac{1}{\mu_i}\right)$$

where:

$$\frac{1}{\mu_i} = \begin{cases} \alpha & \text{for } i = 1, ..., 10 \\ \alpha + \beta & \text{for } i = 11, ..., 15 \end{cases}$$

(i)   (a)   Show that the log-likelihood is given by:

$$\ln L(\alpha, \beta) = \sum_{i=1}^{10} \{\ln \alpha - \alpha y_i\} + \sum_{i=11}^{15} \{\ln(\alpha + \beta) - (\alpha + \beta)y_i\}$$

(b)   Derive the maximum likelihood estimates of α and β.

(c)   Show that the scaled deviance for this model is:

$$2\left\{ \sum_{i=1}^{10} \left(-\ln y_i - 1 + \ln \bar{y}_1 + \frac{y_i}{\bar{y}_1} \right) + \sum_{i=11}^{15} \left(-\ln y_i - 1 + \ln \bar{y}_2 + \frac{y_i}{\bar{y}_2} \right) \right\}$$

where $\bar{y}_1 = \dfrac{1}{10} \displaystyle\sum_{i=1}^{10} y_i$ and $\bar{y}_2 = \dfrac{1}{5} \displaystyle\sum_{i=11}^{15} y_i$.   [8]

The observed values of $Y_1, ..., Y_{10}$ have a mean of 430 and the observed values of $Y_{11}, ..., Y_{15}$ have a mean of 520. The first observed value is $y_1 = 425$.

(ii)   (a)   Show that the deviance residual for y_1 is -0.0117.

(b)   Obtain the Pearson residual for y_1.

(c)   Comment on which is the appropriate residual for this distribution and describe how is it used to determine the fit of this model to the data.   [6]

It has been argued that the model used is too complicated and that a simplified model with $\beta = 0$ should be used, *ie*:

$$\frac{1}{\mu_i} = \alpha \qquad \text{for } i = 1, ..., 15$$

For this simplified model the scaled deviance is 0.135, and for the original model the scaled deviance is 0.0120.

(iii)   (a)   Test whether the original model is a significant improvement over the simplified
                model.

         (b)   The standard error of the estimator of β is 0.000769. Use the maximum
                likelihood estimate of β (from part (i)(b)) and the given standard error to test if
                the parameter β is significantly different from zero.   [4]

                                                                                                    [Total 18]

END OF PAPER

For the session leading to the April 2019 exams – CS1 & CM2 Subjects

Marking vouchers

Subjects	Assignments	Mocks
CS1	6 March 2019	13 March 2019
CM2	20 March 2019	27 March 2019

Series X and Y Assignments

Subjects	Assignment	Recommended submission date	Final deadline date
CS1	**X1**	**5 December 2018**	9 January 2019
CM2		**19 December 2018**	23 January 2019
CS1	**X2**	**19 December 2019**	23 January 2019
CS1	**Y1**	**2 January 2019**	30 January 2019
CM2	**X2**	**9 January 2019**	6 February 2019
CM2	**Y1**	**16 January 2019**	13 February 2019
CS1	**X3**	**16 January 2019**	13 February 2019
CM2		**30 January 2019**	27 February 2019
CS1	**X4**	**30 January 2019**	27 February 2019
CS1	**Y2**	**13 February 2019**	6 March 2019
CM2	**X4**	**13 February 2019**	13 March 2019
CM2	**Y2**	**27 February 2019**	20 March 2019

Mock Exams

Subjects	Recommended submission date	Final deadline date
CS1 (Paper A/B)	**27 February 2019**	13 March 2019
CM2 (Paper A/B)	**13 March 2019**	27 March 2019

We encourage you to work to the recommended submission dates where possible.
If you submit your mock on the final deadline date you are likely to receive your script back less than a week before your exam.

For the session leading to the September 2019 exams – CS1 & CM2 Subjects

Marking vouchers

Subjects	Assignments	Mocks
CS1	21 August 2019	28 August 2019
CM2	28 August 2019	4 September 2019

Series X and Y Assignments

Subjects	Assignment	Recommended submission date	Final deadline date
CS1	**X1**	**22 May 2019**	17 July 2019
CM2		**29 May 2019**	24 July 2019
CS1	**X2**	**5 June 2019**	24 July 2019
CM2		**12 June 2019**	31 July 2019
CS1	**Y1**	**19 June 2019**	31 July 2019
CM2		**26 June 2019**	7 August 2019
CS1	**X3**	**3 July 2019**	7 August 2019
CM2		**10 July 2019**	14 August 2019
CS1	**X4**	**17 July 2019**	14 August 2019
CM2		**24 July 2019**	21 August 2019
CS1	**Y2**	**31 July 2019**	21 August 2019
CM2		**7 August 2019**	28 August 2019

Mock Exams

Subjects	Recommended submission date	Final deadline date
CS1 (Paper A/B)	**14 August 2019**	28 August 2019
CM2 (Paper A/B)	**21 August 2019**	4 September 2019

We encourage you to work to the recommended submission dates where possible.

If you submit your mock on the final deadline date you are likely to receive your script back less than a week before your exam.

Assignment X1 Solutions

Markers: This document sets out one approach to solving each of the questions (sometimes with alternatives). Please give credit for any other valid approaches.

Solution X1.1

The gamma, exponential and χ^2 distributions are covered in Chapter 1.

By definition, the χ^2_2 distribution is the same as the $Gamma(1, \tfrac{1}{2})$ distribution.   [1]

The exponential distribution with mean ½ has parameter 2. This is a $Gamma(1, 2)$ distribution, and so is not equivalent to the other two.   [1]

Therefore the student is wrong.   [Total 2]

Be careful to distinguish between the parameter λ and the mean $1/\lambda$ for the exponential distribution.

Solution X1.2

The Poisson process is covered in Chapter 1.

(i)   *Probability of more than 7 calls*

The number of telephone calls, N, in a two hour period also follows a Poisson distribution:

$$N \sim Poi(2 \times 2.5) = Poi(5)$$ [½]

Using the Poisson tables:

$$P(N > 7) = 1 - P(N \le 7) = 1 - 0.86663 = 0.13337$$ [½]
                                                                                                   [Total 1]

(ii)   *Probability of no calls until after 9am*

The waiting time, T, in hours for the first telephone call has an exponential distribution:

$$T \sim Exp(2.5)$$ [1]

Using the cumulative distribution function of the exponential distribution:

$$P(T > 1) = 1 - P(T \le 1) = 1 - F(1) = 1 - (1 - e^{-2.5 \times 1}) = e^{-2.5} = 0.08208$$ [1]
                                                                                                   [Total 2]

Solution X1.3

Generating functions are covered in Chapter 2.

If $X \sim Exp(\lambda)$ then $M_X(t) = \left(1 - \dfrac{t}{\lambda}\right)^{-1}$. In order to determine $E[X^4]$ we can expand the MGF,

since:

$$E[X^4] = 4! \times \text{ the coefficient of } t^4 \text{ in the expansion of the MGF} \qquad [1]$$

The expansion is:

$$\left(1 - \frac{t}{\lambda}\right)^{-1} = 1 + \frac{t}{\lambda} + \frac{(-1)(-2)}{2!}\left(-\frac{t}{\lambda}\right)^2 + \frac{(-1)(-2)(-3)}{3!}\left(-\frac{t}{\lambda}\right)^3 + \frac{(-1)(-2)(-3)(-4)}{4!}\left(-\frac{t}{\lambda}\right)^4 + \cdots$$

$$[1]$$

Extracting the coefficient of t^4, we get:

$$E[X^4] = 4! \times \frac{(-1)(-2)(-3)(-4)}{4!}\left(-\frac{1}{\lambda}\right)^4 = \frac{24}{\lambda^4} \qquad [1]$$

$$[\text{Total } 3]$$

Solution X1.4

The material for this question is covered in Chapter 1.

(i)   ***Distribution function***

$$F(x) = P(X \le x) = \int_0^x f(t)\, dt = \int_0^x \frac{2 \times 5^2}{(5+t)^3}\, dt = \left[-\frac{5^2}{(5+t)^2} \right]_0^x = 1 - \left(\frac{5}{5+x} \right)^2 \qquad [1]$$

Alternatively, we could just recognise that it's a Pareto distribution with $\alpha = 2$ and $\lambda = 5$ and quote the distribution function from page 14 of the Tables.

(ii)   ***Simulation***

We need to rearrange to get $x = F^{-1}(u)$:

$$u = 1 - \left(\frac{5}{5+x} \right)^2 \quad \Rightarrow \quad x = \frac{5}{\sqrt{1-u}} - 5 \qquad [1]$$

Substituting in our random numbers, we obtain:

$$x = \frac{5}{\sqrt{1-0.656}} - 5 = 3.52 \qquad [1]$$

$$x = \frac{5}{\sqrt{1-0.285}} - 5 = 0.913 \qquad [1]$$

[Total 3]

Solution X1.5

> *The beta distribution is covered in Chapter 1.*

(i)   Mean

Using the formula from page 13 the *Tables*:

$$E(X) = \frac{\alpha}{\alpha + \beta} = \frac{1}{1+4} = \frac{1}{5} \qquad [1]$$

(ii)   Median

The median M is the value that is halfway through the distribution:

$$P(X \leq M) = 0.5 \qquad [1]$$

The $Beta(1,4)$ distribution has PDF $f(x) = 4(1-x)^3$, so:

$$P(X \leq M) = \int_0^M 4(1-x)^3 \, dx = \left[-(1-x)^4 \right]_0^M = 1 - (1-M)^4 \qquad [1]$$
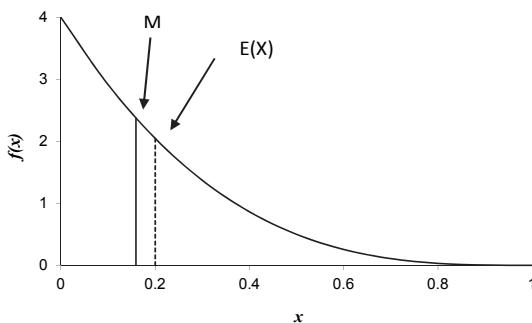
Hence:

$$1 - (1-M)^4 = 0.5 \quad \Rightarrow \quad M = 0.159 \qquad [1]$$

$$[\text{Total } 3]$$

(iii)   Distribution shape

Using the result from part (i), the mean is 0.2. Since the mean is to the right of the median this suggests that the distribution is positively skewed.   [1]

Solution X1.6

The binomial distribution is covered in Chapter 1.

(i)   **Derive a recursive relationship for the binomial distribution**

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$

$$P(X = x-1) = \binom{n}{x-1} p^{x-1} (1-p)^{n-x+1} = \frac{n!}{(n-x+1)!(x-1)!} p^{x-1} (1-p)^{n-x+1} \qquad [1]$$

$$\Rightarrow \quad \frac{P(X = x)}{P(X = x-1)} = \frac{n-x+1}{x} \times \frac{p}{1-p}$$

$$\Rightarrow \quad P(X = x) = \frac{n-x+1}{x} \times \frac{p}{1-p} P(X = x-1) \qquad [1]$$

So $k = \dfrac{p}{1-p}$ and $g(x) = \dfrac{n-x+1}{x}$.   [Total 2]

(ii)(a)   **Probability of no deaths during the year**

Let X be the random variable 'number of deaths during the next year':

$$X \sim Bin(1000, 0.01) \quad \Rightarrow \quad P(X = 0) = 0.99^{1,000} = 0.0000432 \qquad [1]$$

(ii)(b)   **Probability of more than two deaths during the year**

Using the recursive relationship gives:

$$P(X = 1) = 1,000 \times \frac{0.01}{0.99} \times 0.0000432 = 0.000436$$

$$P(X = 2) = \frac{999}{2} \times \frac{0.01}{0.99} \times 0.000436 = 0.00220$$

$$\Rightarrow \quad P(X > 2) = 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)] = 0.997321 \qquad [1]$$

(ii)(c)   **Probability of exactly twenty deaths during the year**

$$P(X = 20) = \binom{1,000}{20} \times 0.01^{20} \times 0.99^{980} = 0.00179 \qquad [1]$$

[Total 3]

Using a Poisson approximation $X \sim Poi(1,000 \times 0.01) = Poi(10)$:

$$\Rightarrow \quad P(X = 20) = \frac{e^{-10} 10^{20}}{20!} = 0.00187$$

Markers, please award full marks for this approach.

Solution X1.7

This question applies knowledge of conditional means and variances. The key results needed to answer this question can be found on Page 16 of the Tables.

Using $E(Y) = E\big[E(Y|X)\big]$ from Page 16 of the *Tables*, we get:

$$E(Y) = E\big[E(Y|X)\big] = E(2X + 400) = 2E(X) + 400 \tag{[1]}$$

But $E(X) = 50$, so:

$$E(Y) = 2 \times 50 + 400 = 500 \tag{[1]}$$

Using $\operatorname{var}(Y) = E\big[\operatorname{var}(Y|X)\big] + \operatorname{var}\big[E(Y|X)\big]$ from page 16 of the *Tables*, we get:

$$\operatorname{var}(Y) = E\left(\frac{X^2}{2}\right) + \operatorname{var}(2X + 400)$$

$$= \tfrac{1}{2}\big[E(X^2)\big] + 4\operatorname{var}(X) \tag{[1]}$$

$$= \tfrac{1}{2}\big(14^2 + 50^2\big) + 4 \times 14^2 = 2{,}132 \tag{[1]}$$

So the standard deviation of Y is 46.17.   [1]

[Total 5]

Alternatively, students may use $\operatorname{var}(Y) = E(Y^2) - \big[E(Y)\big]^2$, to give the following solution:.

Using $E(Y) = E\big[E(Y|X)\big]$ from Page 16 of the Tables, we get:

$$E(Y) = E\big[E(Y|X)\big] = E(2X + 400) = 2E(X) + 400 \tag{[1]}$$

But $E(X) = 50$, so:

$$E(Y) = 2 \times 50 + 400 = 500 \tag{[1]}$$

Similarly:

$$E(Y^2) = E\big[E(Y^2|X)\big] = E\Big[\operatorname{var}(Y|X) + \big[E(Y|X)\big]^2\Big] \tag{[½]}$$

Substituting in the results for $\text{var}(Y|X)$ *and* $E(Y|X)$ *gives:*

$$E\left(Y^2\right) = E\left[\frac{X^2}{2} + (2X + 400)^2\right] = E\left[\frac{9}{2}X^2 = 1{,}600X + 160{,}000\right]$$

$$= \frac{9}{2}E\left(X^2\right) + 1{,}600E(X) + 160{,}000$$

$$= \frac{9}{2}\left(14^2 + 50^2\right) + 1{,}600 \times 50 + 160{,}000 = 252{,}132 \qquad \textit{[1]}$$

Finally:

$$\text{var}(Y) = E\left(Y^2\right) - \left[E(Y)\right]^2 = 252{,}132 - 500^2 = 2{,}132 \qquad \textit{[½]}$$

So the standard deviation of Y is 46.17. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ *[1]*

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ *[Total 5]*

Solution X1.8

The variance of a sum is covered in Chapter 3.

(i)   **Proof**

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

$$= \text{cov}(X, X+Y) + \text{cov}(Y, X+Y) \qquad [1]$$

$$= \text{cov}(X, X) + \text{cov}(X, Y) + \text{cov}(Y, X) + \text{cov}(Y, Y)$$

$$= \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y) \qquad [1]$$
$$[\text{Total 2}]$$

Alternatively:

$$\text{var}(X+Y) = E[\{(X+Y) - E(X+Y)\}^2] = E[\{(X - E(X)) + (Y - E(Y))\}^2]$$

$$= E[(X - E(X))^2] + E[(Y - E(Y))^2] + 2E[(X - E(X))(Y - E(Y))]$$

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

or:

$$\text{var}(X+Y) = E[(X+Y)^2] - E^2[X+Y]$$

$$= E[X^2 + 2XY + Y^2] - [E(X) + E(Y)]^2$$

$$= E(X^2) + 2E(XY) + E(Y^2) - E^2(X) - 2E(X)E(Y) - E^2(Y)$$

$$= [E(X^2) - E^2(X)] + [E(Y^2) - E^2(Y)] + 2[E(XY) - E(X)E(Y)]$$

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

(ii)   **Standard deviation**

Using the definition of the correlation coefficient gives:

$$\text{corr}(X, Y) = -\frac{3}{7} = \frac{\text{cov}(X, Y)}{\sqrt{5^2 \times 7^2}} \quad \Rightarrow \quad \text{cov}(X, Y) = -15 \qquad [1]$$

Using the result from part (i):

$$\text{var}(3X - 2Y + 5) = \text{var}(3X - 2Y)$$

$$= \text{var}(3X) + \text{var}(-2Y) + 2\text{cov}(3X, -2Y) \qquad [1]$$

$$= 9\text{var}(X) + 4\text{var}(Y) - 12\text{cov}(X, Y) = 9 \times 5^2 + 4 \times 7^2 - 12 \times -15 = 601 \qquad [\tfrac{1}{2}]$$

Hence the standard deviation is $\sqrt{601} = 24.5$. $[\tfrac{1}{2}]$
$$[\text{Total 3}]$$

Solution X1.9

Conditional moments are covered in Chapter 4.

(i)   ***Discrete conditional distribution of X given Y = y***

If we draw up a table of possible values for X and Y, we have:

		X		
		0	1	2
	0	0	c	$2c$
Y	1	$2c$	$3c$	$4c$
	2	$4c$	$5c$	$6c$

Since $\sum_x \sum_y P(X=x, Y=y) = 1$, we have $c = \dfrac{1}{27}$.   [1]

Now $P(X=x \mid Y=y) = \dfrac{P(X=x, Y=y)}{P(Y=y)}$.

$$P(Y=y) = \sum_x P(X=x, Y=y) = \sum_{x=0}^{2} \frac{1}{27}(x+2y)$$

$$= \frac{1}{27}[2y + (1+2y) + (2+2y)] = \frac{1}{27}[3+6y] = \frac{3}{27}[1+2y] \qquad [1]$$

$$P(X=x \mid Y=y) = \frac{\dfrac{1}{27}(x+2y)}{\dfrac{3}{27}(1+2y)} = \frac{x+2y}{3(1+2y)} \qquad [1]$$

[Total 3]

Note: Students do not actually need to calculate the value of c to get the answer and full marks.

Alternatively, we could calculate the three conditional distributions:

$$P(X = 0 \mid Y = 0) = 0 \qquad P(X = 1 \mid Y = 0) = \frac{1}{3} \qquad P(X = 2 \mid Y = 0) = \frac{2}{3}$$

$$P(X = 0 \mid Y = 1) = \frac{2}{9} \qquad P(X = 1 \mid Y = 1) = \frac{3}{9} \qquad P(X = 2 \mid Y = 1) = \frac{4}{9}$$

$$P(X = 0 \mid Y = 2) = \frac{4}{15} \qquad P(X = 1 \mid Y = 2) = \frac{5}{15} \qquad P(X = 2 \mid Y = 2) = \frac{6}{15} \qquad [2]$$

or we could give the answer as:

$$P(X = x \mid Y = 0) = \frac{x}{3}$$

$$P(X = x \mid Y = 1) = \frac{x + 2}{9}$$

$$P(X = x \mid Y = 2) = \frac{x + 4}{15}$$

for $x = 1, 2, 3$.   [2]

(ii)   ***Continuous conditional distribution of X given Y = y***

Using $\iint\limits_{x \, y} f(x, y) \, dx dy = 1$ gives:

$$\int\limits_{x=0}^{2} \int\limits_{y=0}^{2} c(x + 2y) \, dx dy = \int\limits_{x=0}^{2} c \left[xy + y^2 \right]_{y=0}^{2} dx = \int\limits_{x=0}^{2} c(2x + 4) \, dx = 1$$

So:

$$c \left[x^2 + 4x \right]_{x=0}^{2} = 12c = 1 \quad \Rightarrow \quad c = \frac{1}{12} \tag{[1]}$$

Now $f_{X \mid Y = y}(x, y) = \dfrac{f_{X,Y}(x, y)}{f_Y(y)}$:

$$f_Y(y) = \int\limits_{x=0}^{2} \frac{1}{12}(x + 2y) \, dx = \frac{1}{12} \left[\frac{1}{2} x^2 + 2xy \right]_{x=0}^{2} = \frac{2}{12}(1 + 2y) \tag{[1]}$$

$$f_{X \mid Y = y}(x, y) = \frac{\frac{1}{12}(x + 2y)}{\frac{2}{12}(1 + 2y)} = \frac{x + 2y}{2(1 + 2y)} \qquad 0 < x < 2 \tag{[1]}$$

[Total 3]

Note: Again students do not actually need to calculate the value of c to get the answer and full marks.

Solution X1.10

The lognormal distribution is covered in Chapter 1.

(i)   **Expression for μ**

Using the formulae for the mean and variance of the lognormal distribution:

$$e^{\mu + \frac{1}{2}\sigma^2} = m \quad \text{and} \quad e^{2\mu + \sigma^2}(e^{\sigma^2} - 1) = s^2 \qquad \text{[½]}$$

Eliminating μ, we obtain:

$$m^2(e^{\sigma^2} - 1) = s^2 \qquad \text{[½]}$$

$$\Rightarrow \qquad \sigma^2 = \ln\left(1 + \frac{s^2}{m^2}\right) \qquad \text{[1]}$$

and, from the equation for m:

$$\mu = \ln m - \tfrac{1}{2}\ln\left(1 + \frac{s^2}{m^2}\right) \qquad \text{[1]}$$

We can also express this in an alternative form:

$$\mu = \ln m - \tfrac{1}{2}\ln\left(1 + \frac{s^2}{m^2}\right) = \tfrac{1}{2}\ln m^2 - \tfrac{1}{2}\ln\left(\frac{m^2 + s^2}{m^2}\right) = \tfrac{1}{2}\ln\left(\frac{m^4}{m^2 + s^2}\right) \qquad \text{[Total 3]}$$

(ii)   **Probability that the next claim exceeds £20,000**

We have $m = 15,000$ and $s = 8,000$, using the equations above gives:

$$\mu = 9.4906 \qquad \sigma^2 = 0.25033 \qquad \text{[1]}$$

So we have $X \sim \log N(9.4906, 0.25033) \Rightarrow \ln X \sim N(9.4906, 0.25033)$:

$$P(X > 20,000) = P(\ln X > \ln 20,000) = P\left(Z > \frac{\ln 20,000 - 9.4906}{\sqrt{0.25033}}\right) = P(Z > 0.825) \qquad \text{[1]}$$

$$= 1 - P(Z < 0.825) = 1 - 0.79531 = 0.20469 \qquad \text{[1]}$$

$$\text{[Total 3]}$$

(iii)   **Mean number of claims**

We are counting the number of failures before the first success. Therefore this is a Type 2 geometric distribution, with $p = 0.20469$. $\qquad$ [1]

Hence, the mean is $\frac{q}{p} = \frac{0.79531}{0.20469} = 3.9$ claims. $\qquad$ [1]

$$\text{[Total 2]}$$

Alternatively, we could find the mean of a Type 1 geometric distribution and subtract 1.

Solution X1.11

MGFs are covered in Chapter 2.

(i)(a)   *Proof of mean result*

From the definition of the MGF:

$$M_X(t) = E[e^{tX}]$$

Differentiating with respect to *t* gives:

$$M'_X(t) = E[Xe^{tX}]$$

Putting $t = 0$ gives $M'_X(0) = E[Xe^0] = E[X]$.   [1]

(i)(b)   *Proof of variance result*

Differentiating again gives $M''_X(t) = E[X^2 e^{tX}]$.

Putting $t = 0$ gives $M''_X(0) = E[X^2]$.   [1]

$$\Rightarrow \quad var(X) = E[X^2] - E^2[X] = M''_X(0) - \left[M'_X(0)\right]^2$$ [1]

[Total 3]

Alternatively, parts (i)(a) and (i)(b) can be proved using the expansion:

$$M_X(t) = E\left(e^{tX}\right) = E\left(1 + tX + \frac{t^2 X^2}{2!} + \frac{t^3 X^3}{3!} + \cdots\right)$$

$$= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots$$

$$\Rightarrow \quad M'_X(t) = E(X) + tE(X^2) + \frac{t^2}{2!}E\left(X^3\right) + \cdots \quad \Rightarrow \quad M'_X(0) = E(X)$$ *[1]*

$$\Rightarrow \quad M''_X(t) = E(X^2) + tE\left(X^3\right) + \cdots \quad \Rightarrow \quad M''_X(0) = E(X^2)$$

$$\Rightarrow \quad var(X) = E[X^2] - E^2[X] = M''_X(0) - \left[M'_X(0)\right]^2$$ *[2]*

[Total 3]

(ii)(a)   **Derive the MGF**

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{0} e^{tx} \tfrac{1}{2} e^{x} \, dx + \int_{0}^{\infty} e^{tx} \tfrac{1}{2} e^{-x} \, dx$$

$$= \tfrac{1}{2} \int_{-\infty}^{0} e^{(t+1)x} \, dx + \tfrac{1}{2} \int_{0}^{\infty} e^{(t-1)x} \, dx$$

$$= \frac{1}{2} \left[\frac{e^{(t+1)x}}{t+1} \right]_{-\infty}^{0} + \frac{1}{2} \left[\frac{e^{(t-1)x}}{t-1} \right]_{0}^{\infty}$$ [1]

For $-1 < t < 1$ this gives:

$$M_X(t) = \frac{1}{2} \left(\frac{1}{t+1} \right) - \frac{1}{2} \left(\frac{1}{t-1} \right)$$ [1]

Rearranging this gives:

$$M_X(t) = \frac{1}{1-t^2} = (1-t^2)^{-1}$$ [1]

(ii)(b)   **Mean and variance**

Differentiating the MGF from part (i) using the chain rule gives:

$$M_X'(t) = 2t(1-t^2)^{-2}$$ [½]

Hence:

$$E(X) = M_X'(0) = 0$$ [1]

Differentiating a second time using the product rule and the chain rule gives:

$$M_X''(t) = 2(1-t^2)^{-2} + 8t^2(1-t^2)^{-3}$$ [½]

Hence:

$$E(X^2) = M_X''(0) = 2 + 0 = 2$$

So the variance is given by:

$$var(X) = E(X^2) - E^2(X) = 2 - 0^2 = 2$$ [1]
                                                                         [Total 6]

Solution X1.12

Generating functions are covered in Chapter 2.

(i)   ***CGF of a gamma distribution***

From the definition of a moment generating function:

$$M_X(t) = E\left(e^{tX}\right) = \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x} \, dx \qquad [1]$$

We almost have a $Gamma(\alpha, \lambda - t)$ PDF, so putting in the appropriate constants:

$$M_X(t) = \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \int_0^\infty \frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} \, dx \qquad [1]$$

$$= \left(\frac{\lambda}{\lambda-t}\right)^\alpha \qquad\qquad \text{provided } t < \lambda \qquad [\tfrac{1}{2}]$$

This follows since $\displaystyle\int_0^\infty \frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} \, dx$ is the integral of a $Gamma(\alpha, \lambda-t)$ PDF over its

whole range, so its value is 1.

Dividing the numerator and denominator by λ gives:

$$M_X(t) = \left(\frac{1}{1-t/\lambda}\right)^\alpha = \left(1-\frac{t}{\lambda}\right)^{-\alpha} \qquad\qquad t < \lambda \qquad [\tfrac{1}{2}]$$

Hence, the cumulant generating function is:

$$C_X(t) = \ln M_X(t) = -\alpha \ln\left(1-\frac{t}{\lambda}\right) \qquad\qquad t < \lambda \qquad [1]$$

[Total 4]

Alternatively we could use a substitution of $y = (\lambda - t)x$ into the initial integral obtained above:

$$M_X(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x} \, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{y}{\lambda-t} \right)^{\alpha-1} e^{-y} \frac{1}{\lambda-t} \, dy$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\lambda-t} \right)^\alpha \int_0^\infty y^{\alpha-1} e^{-y} \, dy$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\lambda-t} \right)^\alpha \Gamma(\alpha) \qquad \text{by definition of } \Gamma(\alpha) \qquad [1]$$

This gives:

$$M_X(t) = \left(\frac{1}{1-t/\lambda} \right)^\alpha = \left(1 - \frac{t}{\lambda} \right)^{-\alpha} \qquad t < \lambda \qquad [1]$$

$$C_X(t) = \ln M_X(t) = -\alpha \ln\left(1 - \frac{t}{\lambda} \right) \qquad t < \lambda \qquad [1]$$

(ii)   *Coefficient of skewness*

The definition of the coefficient of skewness is:

$$\frac{\text{skew}(X)}{[\text{var}(X)]^{1.5}} \qquad [\frac{1}{2}]$$

So using the fact that $\text{var}(X) = C_X''(0)$ and $\text{skew}(X) = C_X'''(0)$, we obtain:

$$C_X'(t) = \frac{\alpha}{\lambda} \left(1 - \frac{t}{\lambda} \right)^{-1} \qquad [\frac{1}{2}]$$

$$C_X''(t) = \frac{\alpha}{\lambda^2} \left(1 - \frac{t}{\lambda} \right)^{-2} \quad \Rightarrow \quad \text{var}(X) = C_X''(0) = \frac{\alpha}{\lambda^2} \qquad [1]$$

$$C_X'''(t) = \frac{2\alpha}{\lambda^3} \left(1 - \frac{t}{\lambda} \right)^{-3} \quad \Rightarrow \quad \text{skew}(X) = C_X'''(0) = \frac{2\alpha}{\lambda^3} \qquad [1]$$

$$\Rightarrow \quad \text{Coefficient of skewness} = \frac{2\alpha / \lambda^3}{\left(\alpha / \lambda^2 \right)^{1.5}} = \frac{2\alpha}{\alpha^{1.5}} = \frac{2}{\sqrt{\alpha}} \qquad [1]$$

[Total 4]

Solution X1.13

Joint distributions are covered in Chapter 3 and conditional moments are covered in Chapter 4.

(i)(a)   *Show X and Y not independent*

If X and Y are independent, then:

$$P(X=x, Y=y) = P(X=x)P(Y=y) \qquad \text{for all } x, y$$

This is not this case. For example:

$$P(X=0, Y=0) = \frac{1}{2} \neq \frac{3}{4} \times \frac{9}{16} \tag{1}$$

Any suitable counterexample will be sufficient to obtain the mark, provided that the student explains why it shows dependence.

(i)(b)   *Show X and Y not uncorrelated*

If X and Y are uncorrelated, then:

$$\text{corr}(X,Y)=0 \;\Rightarrow\; \text{cov}(X,Y)=0 \;\Rightarrow\; E(XY)=E(X)E(Y)$$

We have:

$$E(X) = \sum_x xP(X=x) = \left(0 \times \frac{3}{4}\right) + \left(1 \times \frac{3}{16}\right) + \left(2 \times \frac{1}{16}\right) = \frac{5}{16} \tag{1}$$

$$E(Y) = \sum_y yP(Y=y) = \left(0 \times \frac{9}{16}\right) + \left(1 \times \frac{1}{8}\right) + \left(2 \times \frac{5}{16}\right) = \frac{3}{4} \tag{½}$$

$$E(XY) = \sum_x \sum_y xyP(X=x, Y=y)$$

$$= \left(0 \times 0 \times \frac{1}{2}\right) + (0 \times 1 \times 0) + \cdots + (2 \times 2 \times 0) = \left(1 \times 1 \times \frac{1}{8}\right) + \left(1 \times 2 \times \frac{1}{16}\right) = \frac{1}{4} \tag{1}$$

Since $E(XY) = \frac{1}{4} \neq \frac{5}{16} \times \frac{3}{4} = E(X)E(Y)$, they are *not* uncorrelated. [½]

[Total 4]

(ii)   *When is relationship true?*

This relationship is always true (provided all the quantities involved make sense). [1]

We've included this question to emphasise that you can always use this result. There is no requirement for X and Y to be independent, uncorrelated, normally distributed or whatever.

(iii)(a) **E[X + Y | X = 1]**

Since $P(X=1) = \frac{3}{16}$:

$$E[X+Y \mid X=1] = \sum_x \sum_y (x+y)P(X=x, Y=y \mid X=1)$$

$$= \sum_x \sum_y (x+y)\frac{P(X=1, Y=y)}{P(X=1)}$$

$$= 2 \times \frac{1/8}{3/16} + 3 \times \frac{1/16}{3/16} \qquad [1]$$

$$= 2 \times \frac{2}{3} + 3 \times \frac{1}{3} = 2\frac{1}{3} \qquad [1]$$

(iii)(b) **E[X | Y = 2]**

Since $P(Y=2) = \frac{5}{16}$:

$$E[X \mid Y=2] = \sum_x x P(X=x \mid Y=2)$$

$$= \sum_x x \frac{P(X=x, Y=2)}{P(Y=2)}$$

$$= 0 \times \frac{1/4}{5/16} + 1 \times \frac{1/16}{5/16} + 2 \times \frac{0}{5/16} \qquad [1]$$

$$= \left(0 \times \frac{4}{5}\right) + \left(1 \times \frac{1}{5}\right) + (2 \times 0) = \frac{1}{5} \qquad [1]$$

(iii)(c) **var[X | Y = 2]**

$$\mathrm{var}(X \mid Y=2) = E[X^2 \mid Y=2] - (E[X \mid Y=2])^2 \qquad [1]$$

But:

$$E[X^2 \mid Y=2] = \sum_x x^2 P(X=x \mid Y=2) = \left(0^2 \times \frac{4}{5}\right) + \left(1^2 \times \frac{1}{5}\right) = \frac{1}{5} \qquad [1]$$

So:

$$\mathrm{var}(X \mid Y=2) = \frac{1}{5} - \left(\frac{1}{5}\right)^2 = \frac{4}{25} \qquad [1]$$

[Total 7]

(iv) $E\left[E\left(Y^2 \mid X\right)\right]$

$E[Y^2 \mid X]$ depends on the value of X. So we need to work out $E[Y^2 \mid X = x]$ for each possible value of x:

$$E[Y^2 \mid X = 0] = \sum_y y^2 P(Y = y \mid X = 0) = \left(0^2 \times \tfrac{2}{3}\right) + \left(2^2 \times \tfrac{1}{3}\right) = 1\tfrac{1}{3}$$

$$E[Y^2 \mid X = 1] = \sum_y y^2 P(Y = y \mid X = 1) = \left(1^2 \times \tfrac{2}{3}\right) + \left(2^2 \times \tfrac{1}{3}\right) = 2$$

$$E[Y^2 \mid X = 2] = \sum_y y^2 P(Y = y \mid X = 2) = 0$$

Combining these, we have:

$$E[Y^2 \mid X] = \begin{cases} 1\tfrac{1}{3} & \text{if } X = 0 \\ 2 & \text{if } X = 1 \\ 0 & \text{if } X = 2 \end{cases} \qquad [1]$$

To calculate $E[E(Y^2 \mid X)]$, we need to apply the probabilities of the possible values of X:

$$E[E(Y^2 \mid X)] = \sum_x E[Y^2 \mid X = x] P(X = x) \qquad\qquad [1]$$

$$= \left(1\tfrac{1}{3} \times \tfrac{3}{4}\right) + \left(2 \times \tfrac{3}{16}\right) + \left(0 \times \tfrac{1}{16}\right) = 1\tfrac{3}{8} \qquad [1]$$

[Total 3]

(v) $E\left(Y^2\right)$

The marginal probabilities for Y are:

$$P(Y = 0) = \tfrac{1}{2} + \tfrac{1}{16} = \tfrac{9}{16}$$

$$P(Y = 1) = \tfrac{1}{8}$$

and $$P(Y = 2) = \tfrac{1}{4} + \tfrac{1}{16} = \tfrac{5}{16}$$

So: $$E[Y^2] = \sum_y y^2 P(Y = y) = \left(0^2 \times \tfrac{9}{16}\right) + \left(1^2 \times \tfrac{1}{8}\right) + \left(2^2 \times \tfrac{5}{16}\right) = 1\tfrac{3}{8} \qquad [1]$$

As expected, $E[E(Y^2 \mid X)] = E[Y^2]$. [1]

[Total 2]

Assignment X2 Solutions

Markers: This document sets out one approach to solving each of the questions (sometimes with alternatives). Please give credit for any other valid approaches.

Solution X2.1

The F distribution is covered in Chapter 6.

(i)   **$P(F_{9,24} < 3.256)$**

From page 174 of the *Tables*, we see that $P(F_{9,24} > 3.256) = 0.01$. Hence:

$$P(F_{9,24} < 3.256) = 1 - P(F_{9,24} > 3.256) = 1 - 0.01 = 0.99$$ [1]

(ii)   **$P(F_{3,5} < 0.18836)$**

Since 0.18836 is less than 1, we need to use the $\dfrac{1}{F_{m,n}}$ result:

$$P(F_{3,5} < 0.18836) = P\left(\frac{1}{F_{3,5}} > \frac{1}{0.18836}\right) = P(F_{5,3} > 5.309) = 0.10$$ [1]

(iii)   ***The value of a***

Since 95% of the distribution is greater than a , this tells us that it must be a lower critical point. Hence we need to use the $\dfrac{1}{F_{m,n}}$ result again:

$$P(F_{8,6} > a) = P\left(\frac{1}{F_{8,6}} < \frac{1}{a}\right) = P\left(F_{6,8} < \frac{1}{a}\right) = 0.95 \Rightarrow \quad P\left(F_{6,8} > \frac{1}{a}\right) = 0.05$$

$$\Rightarrow \quad \frac{1}{a} = 3.581$$

$$\Rightarrow \quad a = 0.279$$ [1]

[Total 3]

Solution X2.2

The t distribution is covered in Chapter 6.

Using $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ from page 22 of the *Tables*, we have:

$$\frac{\bar{X}-0}{S/\sqrt{9}} \sim t_8 \quad \Rightarrow \quad \frac{3\bar{X}}{S} \sim t_8 \tag{1}$$

The required probability can then be calculated as:

$$P(\bar{X} > S) = P\left(\frac{\bar{X}}{S} > 1\right) = P\left(\frac{3\bar{X}}{S} > 3\right) = P(t_8 > 3) \tag{1}$$

From the *Tables* on page 163, the required probability to 2 decimal places is 0.01. [1]

[Total 3]

Interpolation produces a value of 0.0089, but given to 2 decimal places this is 0.01.

Solution X2.3

We first need to calculate the difference between the waist measurements (before minus after):

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6
Difference, d	8	1	28	11	25	6

[½]

Calculating the mean and sample standard deviation of these values:

$$\bar{d} = \frac{79}{6} \tag{½}$$

$$s_d = \sqrt{\frac{8^2 + 1^2 + 28^2 + 11^2 + 25^2 + 6^2 - 6\left(\frac{79}{6}\right)^2}{5}} = \sqrt{\frac{1{,}631 - 6\left(\frac{79}{6}\right)^2}{5}} = \sqrt{118.167} = 10.87 \tag{1}$$

We can then calculate the confidence interval:

$$\bar{d} \pm t_{0.05,5} \frac{s_d}{\sqrt{n}} = \frac{79}{6} \pm 2.015 \frac{10.8704}{\sqrt{6}} = (4.22, 22.11) \tag{1}$$

[Total 3]

Solution X2.4

Likelihoods are covered in Chapter 7.

Now:

$$P(X = \text{an odd number}) = p + pq^2 + pq^4 + \cdots$$

This is an infinite geometric series with $a = p$ and $r = q^2$, so using $S_\infty = \dfrac{a}{1-r}$:

$$P(X = \text{an odd number}) = \frac{p}{1-q^2} = \frac{p}{(1-q)(1+q)} = \frac{1}{1+q} \qquad [1]$$

Similarly:

$$P(X = \text{an even number}) = pq + pq^3 + pq^5 + \cdots$$

This is an infinite geometric series with $a = pq$ and $r = q^2$, so using $S_\infty = \dfrac{a}{1-r}$:

$$P(X = \text{an even number}) = \frac{pq}{1-q^2} = \frac{pq}{(1-q)(1+q)} = \frac{q}{1+q} \qquad [1]$$

Alternatively, and more simply, students may use:

$$P(X = \text{an even number}) = 1 - P(X = \text{an odd number}) = 1 - \frac{1}{1+q} = \frac{q}{1+q}$$

So the likelihood of equal numbers of even and odd values of X is:

$$\binom{2n}{n}\left(\frac{1}{1+q}\right)^n\left(\frac{q}{1+q}\right)^n = \binom{2n}{n}\frac{q^n}{(1+q)^{2n}} \qquad [1]$$

[Total 3]

Students may also give the alternative answer of $\dbinom{2n}{n}\dfrac{p^{2n}q^n}{(1-q^2)^{2n}}$.

Solution X2.5

Confidence intervals for σ^2 are covered in Chapter 8.

Using $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, we can obtain a 90% confidence interval for σ^2 from:

$$0.9 = P\left(\chi^2_{n-1;0.95} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1;0.05} \right)$$

which gives:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1;0.05}} , \frac{(n-1)s^2}{\chi^2_{n-1;0.95}} \right) \tag*{[½]}$$

From this data set, we have:

$$s^2 = \frac{1}{15}\left[1,168 - 16 \times \left(\frac{128}{16} \right)^2 \right] = 9.6 \tag*{[½]}$$

This gives:

$$\left(\frac{15 \times 9.6}{25.00} , \frac{15 \times 9.6}{7.261} \right) = (5.76, 19.8) \tag*{[1]}$$

However, we require a 90% confidence interval for the standard deviation, so taking the square root:

$$(2.4, 4.45) \tag*{[1]}$$
$$\text{[Total 3]}$$

Solution X2.6

Confidence intervals for μ are covered in Chapter 8.

(i)   ***Confidence interval for mean (known variance)***

We know that $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has a $N(0,1)$ distribution. So a symmetrical 95% confidence interval will be given by:

$$0.95 = P(-1.96 < Z < 1.96) \quad \text{or} \quad \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} = \pm 1.96 \quad \text{or} \quad \overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}} \qquad \text{[1]}$$

For the data values given, $\overline{x} = 47.6$.

So the confidence interval is:

$$47.6 \pm 1.96 \times \frac{\sqrt{20}}{\sqrt{5}} = 47.6 \pm 3.92 = (43.68, 51.52) \qquad \text{[1]}$$

[Total 2]

(ii)   ***Confidence interval for mean (unknown variance)***

Since $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ has a t_{n-1} distribution, we will need to use the *t* tables. We want the value of the distribution for which $\alpha = 0.025$. From the *Tables* this value is 2.776.

So the confidence interval is $\overline{x} \pm 2.776\dfrac{s}{\sqrt{n}}$, where: [½]

$$s^2 = \frac{1}{4}\Big[11,746 - 5 \times 47.6^2 \Big] = 104.3 \qquad \text{[½]}$$

So the confidence interval is $47.6 \pm 12.68 = (34.92, 60.28)$.   [1]

[Total 2]

Solution X2.7

The Central Limit theorem and continuity corrections are covered in Chapter 5.

(i)   ***Probability using normal approximation***

Let X be the number of squares moved up the drain pipe each turn. Hence:

$$X = \begin{cases} -1 & \frac{1}{6} \\ 0 & \frac{5}{12} \\ +1 & \frac{1}{3} \\ +2 & \frac{1}{12} \end{cases}$$
[1]

These probabilities are derived from the information given in the table. For example, $P(X = -1)$ is the probability of obtaining a score of 2, 3 or 4 when two dice are rolled.

To use a normal approximation, we first need to calculate $E(X)$ and $var(X)$:

$$E(X) = \sum x P(X = x) = -\frac{1}{6} + 0 + \frac{1}{3} + \frac{1}{6} \ = \ \frac{1}{3}$$
[½]

$$E(X^2) = \sum x^2 P(X = x) = \frac{1}{6} + 0 + \frac{1}{3} + \frac{1}{3} \ = \ \frac{5}{6}$$

$$\Rightarrow \ \ var(X) = E(X^2) - E^2(X) = \frac{5}{6} - \frac{1}{9} \ = \ \frac{13}{18}$$
[½]

Using the Central Limit Theorem, we know that the approximate distribution of the total movement after 15 turns is:

$$\sum X \ \dot{\sim}\ N\left(15 \times \frac{1}{3}, 15 \times \frac{13}{18}\right) = N\left(5, 10\tfrac{5}{6}\right)$$
[1]

So to obtain $P\left(\sum X > 8\right)$, the probability that a spider has moved up more than 8 squares, we first use a continuity correction $P\left(\sum X > 8\right) \rightarrow P\left(\sum X > 8.5\right)$.
[1]

Standardising, by setting $z = \frac{8.5-5}{\sqrt{10\tfrac{5}{6}}} = 1.063$, gives:

$$P\left(\sum X > 8.5\right) = P(Z > 1.063) = 1 - P(Z < 1.063)$$

$$= 1 - 0.85619 = 0.14381$$
[1]

[Total 5]

(ii)   **Suitability of normal approximation**

The Central Limit Theorem requires *n* to be large. Fifteen turns is not large, therefore this will be a poor approximation. [1]

Solution X2.8

Maximum likelihood estimation is covered in Chapter 7.

(i)   **Log-likelihood**

The likelihood function is:

$$L(\lambda) = \text{constant} \times \left[P(0 \le X < 1)\right]^{30} \left[P(1 \le X < 2)\right]^{15} \left[P(2 \le X < \infty)\right]^{5}$$ [½]

where the constant arises from the different possible distinct arrangements of the 50 observations amongst the 3 groups (as we don't know exactly which 30 values are in the first group, *etc*). [½]

The value of the constant is $\frac{50!}{30!15!5!}$.

Using $F(x)$ from page 11 of the *Tables*:

$$P(0 \le X < 1) = F(1) = 1 - e^{-\lambda}$$

$$P(1 \le X < 2) = F(2) - F(1) = e^{-\lambda} - e^{-2\lambda}$$

$$P(X \ge 2) = 1 - F(2) = e^{-2\lambda}$$ [½]

Using these probabilities:

$$L(\lambda) = \text{constant} \times (1 - e^{-\lambda})^{30}(e^{-\lambda} - e^{-2\lambda})^{15}(e^{-2\lambda})^{5}$$ [½]

$$= \text{constant} \times (1 - e^{-\lambda})^{30}(e^{-\lambda})^{15}(1 - e^{-\lambda})^{15}(e^{-10\lambda})$$ [½]

$$= \text{constant} \times e^{-25\lambda}(1 - e^{-\lambda})^{45}$$

$$\Rightarrow \ln L(\lambda) = \text{constant} - 25\lambda + 45\ln(1 - e^{-\lambda})$$ [½]

[Total 3]

(ii)   **Maximum likelihood estimate**

Differentiating the log-likelihood with respect to λ gives:

$$\frac{d}{d\lambda}\ln L = -25 + \frac{45e^{-\lambda}}{1-e^{-\lambda}}$$ [1]

Setting this derivative equal to 0:

$$\frac{e^{-\lambda}}{1-e^{-\lambda}} = \frac{25}{45} \;\; \Rightarrow \;\; e^{-\lambda} = \frac{5}{14} \;\; \Rightarrow \;\; \hat{\lambda} = \ln\frac{14}{5} = 1.030$$ [1]

Checking the second derivative:

$$\frac{d^2}{d\lambda^2}\ln L = \frac{-45e^{-\lambda}\left(1-e^{-\lambda}\right) - 45e^{-\lambda}e^{-\lambda}}{\left(1-e^{-\lambda}\right)^2} = \frac{-45e^{-\lambda}}{\left(1-e^{-\lambda}\right)^2} < 0 \Rightarrow \max$$ [1]

[Total 3]

Solution X2.9

Maximum likelihood estimation and the method of moments are covered in Chapter 7.

(i)   **MLE**

The likelihood function based on a sample of n observations is given by:

$$L = \prod_{i=1}^{n} 2e^{-2(x_i-\theta)} = 2^n e^{-2(\sum x_i - n\theta)} \text{ , provided } x_1, \cdots, x_n \geq \theta$$ [1]
$$(ie \min_i x_i \geq \theta)$$

So $\hat{\theta}$, the MLE of θ, is the value of θ that maximises $2^n e^{-2(\sum x_i - n\theta)}$ subject to the condition that $\theta \leq \min x_i$ (otherwise the likelihood is zero).

When $\theta \leq \min x_i$, we have:

$$L = 2^n e^{-2(\sum x_i - n\theta)} \;\; \Rightarrow \;\; \ln L = n\ln 2 - 2(\sum x_i - n\theta) \;\; \Rightarrow \;\; \frac{d}{d\theta}\ln L = 2n > 0$$ [½]

ie L increases as θ increases.

So the MLE of θ is the highest value of θ subject to the condition that $\theta \leq \min x_i$, *ie* $\hat{\theta} = \min X_i$ is the maximum likelihood estimator of θ.   [1]

From this sample, the maximum likelihood estimate of θ is 1.56.   [½]
[Total 3]

(ii)   *Method of moments estimate*

Now $X - \theta$ has an *Exp*(2) distribution, so:

$$E[X] = E[X - \theta] + \theta = \frac{1}{2} + \theta \qquad\qquad [1]$$

Alternatively, from first principles:

$$E(X) = \int\limits_{\theta}^{\infty} 2xe^{-2(x-\theta)}\, dx$$

$$= \left[-xe^{-2(x-\theta)} \right]_{\theta}^{\infty} + \int\limits_{\theta}^{\infty} e^{-2(x-\theta)}\, dx \qquad\qquad \text{by parts}$$

$$= \theta + \left[-\frac{1}{2} e^{-2(x-\theta)} \right]_{\theta}^{\infty} = \theta + \frac{1}{2}$$

Equating $E(X)$ to the sample mean, $\bar{x}$, gives:

$$\tilde{\theta} + \frac{1}{2} = \bar{x} = \frac{1}{n}\sum x_i = 2.25 \qquad\qquad [½]$$

So the method of moments estimate of θ is $2.25 - 0.5 = 1.75$.   [½]

[Total 2]

(iii)   *Comment*

One of the observed values was less than the method of moments estimate of θ. So the method of moments gives an estimate of θ that is not 'possible' in this case.   [½]

This contrasts with the situation for maximum likelihood estimators, which, provided they exist, must, by definition, give feasible estimates.   [½]

[Total 1]

Solution X2.10

Approximations to the binomial distribution are covered in Chapter 5.

(i)   ***Approximate probability when n = 400***

We have $X \sim Bin(400, 0.01)$. Using a Poisson approximation:

$$X \sim Bin(400, 0.01) \doteqdot Poi(4)$$ [1]

We require $P(9 \leq X \leq 16)$. Using the cumulative Poisson tables on pages 176 and 180:

$$P(9 \leq X \leq 16) = P(X \leq 16) - P(X < 9) = P(X \leq 16) - P(X \leq 8)$$ [1]

$$= 1.00000 - 0.97864 = 0.02136$$ [1]

[Total 3]

A normal approximation is not valid here as $np = 4 < 5$. Students should only receive 1 mark for using this method and obtaining the answer of 0.0119.

(ii)   ***Approximate probability when n = 3,000***

We have $X \sim Bin(3000, 0.01)$. Using a normal approximation:

$$X \doteqdot N(3000 \times 0.01, 3000 \times 0.01 \times 0.99) = N(30, 29.7)$$ [1]

This approximation is valid since $np = 30 > 5$ and $n(1-p) = 2,970 > 5$.

Using a continuity correction we have $P(9 \leq X \leq 16) \approx P(8.5 < X < 16.5)$.   [1]

Evaluating this:

$$P(9 \leq X \leq 16) = P\left(\frac{8.5-30}{\sqrt{29.7}} < Z < \frac{16.5-30}{\sqrt{29.7}} \right)$$

$$= P(-3.945 < Z < -2.477)$$ [1]

$$= P(Z < -2.477) - P(Z < -3.945)$$

$$= [1 - P(Z < 2.477)] - [1 - P(Z < 3.945)]$$

$$= P(Z < 3.945) - P(Z < 2.477)$$

$$= 0.99996 - 0.99337 = 0.00659$$ [1]

[Total 4]

Using the complete decimal when interpolating would result in 0.00658.

Solution X2.11

Relationships between gamma distributions are covered in Chapter 5.

(i)   ***Sum of two independent gamma random variables***

Let $X_1 \sim Gamma(\alpha_1, \lambda)$, $X_2 \sim Gamma(\alpha_2, \lambda)$ and $Z = X_1 + X_2$. Then:

$$M_Z(t) = E\left(e^{tZ}\right) = E\left(e^{tX_1 + tX_2}\right) = E\left(e^{tX_1} e^{tX_2}\right)$$

$$= E\left(e^{tX_1}\right) E\left(e^{tX_2}\right) \qquad\qquad \text{by independence}$$

$$= M_{X_1}(t) M_{X_2}(t) \qquad\qquad\qquad\qquad\qquad\qquad [1]$$

Therefore:

$$M_Z(t) = \left(1 - \frac{t}{\lambda}\right)^{-\alpha_1} \left(1 - \frac{t}{\lambda}\right)^{-\alpha_2} = \left(1 - \frac{t}{\lambda}\right)^{-(\alpha_1 + \alpha_2)}$$

This is the MGF of a $Gamma(\alpha_1 + \alpha_2, \lambda)$ distribution. Hence, by the uniqueness property of MGFs, $Z \sim Gamma(\alpha_1 + \alpha_2, \lambda)$.   [1]

[Total 2]

(ii)   ***MGF***

Generalising the result of part (i), where $\alpha_1 = \cdots = \alpha_n = \alpha$, gives:

$$\sum X_i \sim Gamma\left(\sum \alpha_i, \lambda\right) \sim Gamma(n\alpha, \lambda) \qquad\qquad [1]$$

So the MGF of $\sum X_i$ is $\left(1 - \frac{t}{\lambda}\right)^{-n\alpha}$. Hence:

$$M_{2\lambda \sum X_i}(t) = E\left(e^{2\lambda t \sum X_i}\right) = M_{\sum X_i}(2\lambda t) = \left(1 - \frac{2\lambda t}{\lambda}\right)^{-n\alpha} = (1 - 2t)^{-n\alpha} \qquad [1]$$

This is the MGF of a $\chi^2_{2n\alpha}$ distribution. Hence, by the uniqueness property of MGFs, $2\lambda \sum X_i \sim \chi^2_{2n\alpha}$.   [1]

[Total 3]

(iii)   **Probability of sample mean exceeding 40**

Here $\sum X_i = 5\bar{X} \;\Rightarrow\; 10\lambda\bar{X} \sim \chi^2_{10\alpha} \;\Rightarrow\; \bar{X} \sim \chi^2_{20}$.   [1]

Using the percentage points for the χ^2 distribution on page 169 of the *Tables*:

$$P(\bar{X} > 40) = P(\chi^2_{20} > 40) = 0.005$$ [1]

[Total 2]

Solution X2.12

Confidence intervals for a Poisson distribution are covered in Chapter 8.

(i)   **Normal approximation confidence interval**

Since $X \sim Poi(\lambda)$, the distribution of X can be approximated by $N(\lambda, \lambda)$. So $\dfrac{\bar{X}-\lambda}{\sqrt{\hat{\lambda}/n}}$ has

approximately a $N(0, 1)$ distribution.   [1]

So $0.99 = P\left(-2.5758 < \dfrac{\bar{X}-\lambda}{\sqrt{\hat{\lambda}/n}} < 2.5758 \right)$.   [½]

Using $\hat{\lambda} = \bar{x} = 0.9$ as an approximation for λ in the denominator, and rearranging:

$$0.9 - 2.5758\sqrt{\tfrac{0.9}{10}} < \lambda < 0.9 + 2.5758\sqrt{\tfrac{0.9}{10}}$$ [½]

which gives an approximate confidence interval for λ of:

$$0.9 \pm 0.773 = (0.127, 1.673)$$ [1]

[Total 3]

(ii)   **Comment**

The approximation is not brilliant as:

•   the Poisson parameter is small (the approximation is better for large values)   [1]

•   the sample size of 10 is small (the approximation is better for large samples)   [1]

•   an estimate for λ is used in the variance.   [1]

[Maximum 2]

(iii)   ***Accurate confidence interval***

Since $\sum X_i \sim Poi(10\lambda)$, the equations required to obtain the accurate confidence interval are:

$$\sum_{x=0}^{9} \frac{e^{-10\lambda}(10\lambda)^x}{x!} = 0.005 \text{ to obtain the upper limit} \hspace{3cm} [1]$$

$$\sum_{x=0}^{8} \frac{e^{-10\lambda}(10\lambda)^x}{x!} = 0.995 \text{ to obtain the lower limit} \hspace{3cm} [1]$$

[Total 2]

Solution X2.13

Maximum likelihood estimation, bias and mean square error are covered in Chapter 7.

(i)   ***MLE of mean***

The likelihood is:

$$L(\mu) = P(X_1 = x_1) \times \ldots \times P(X_n = x_n) = \frac{\mu^{x_1}}{x_1!}e^{-\mu} \times \ldots \times \frac{\mu^{x_n}}{x_n!}e^{-\mu} = \frac{\mu^{\sum x_i}}{x_1! \ldots x_n!}e^{-n\mu} \hspace{1cm} [1]$$

Taking logs:

$$\ln L(\mu) = \sum x_i \ln \mu - \ln(x_1! \ldots x_n!) - n\mu \hspace{3cm} [\tfrac{1}{2}]$$

Differentiating and setting equal to zero:

$$\frac{d}{d\mu}\ln L(\mu) = \frac{\sum x_i}{\mu} - n \;\Rightarrow\; \frac{\sum x_i}{\hat{\mu}} = n \;\Rightarrow\; \hat{\mu} = \frac{\sum x_i}{n} = \bar{x} \hspace{2cm} [\tfrac{1}{2}]$$

Checking that this gives a maximum:

$$\frac{d^2}{d\mu^2}\ln L(\mu) = -\frac{\sum x_i}{\mu^2} \le 0 \;\Rightarrow\; \max \hspace{3cm} [\tfrac{1}{2}]$$

So the estimator is $\hat{\mu} = \bar{X}$. $\hspace{8cm}$ [½]

[Total 3]

(ii)   **Bias and MSE**

The bias of $\hat{\mu}$ is given by $E(\hat{\mu}) - \mu$. Now:

$$E(\hat{\mu}) = E(\overline{X}) = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\sum \mu = \frac{1}{n}n\mu = \mu$$

Hence the bias is zero.   [1]

Students must show the working to receive the mark.

The MSE of $\hat{\mu}$ is given by $\text{var}(\hat{\mu}) + bias^2(\hat{\mu})$. Now:

$$\text{var}(\hat{\mu}) = \text{var}(\overline{X}) = \text{var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\sum \text{var}(X_i) = \frac{1}{n^2}\sum \mu = \frac{1}{n^2}n\mu = \frac{\mu}{n}$$

So the MSE of $\hat{\mu}$ is $\frac{\mu}{n}$.   [1]

[Total 2]

Students must show the working to receive the mark. Half marks to be given if students give the answer σ^2/n (and fail to utilise the fact that the variance of $Poi(\mu)$ is μ).

(iii)   **Variance attains CRLB**

The CRLB is given by $-\dfrac{1}{E\left[\dfrac{d^2}{d\mu^2}\ln L(\mu)\right]}$. Using the second derivative from part (i):

$$E\left[\frac{d^2}{d\mu^2}\ln L(\mu)\right] = E\left[-\frac{\sum X_i}{\mu^2}\right] = -\frac{1}{\mu^2}E\left[\sum X_i\right] = -\frac{1}{\mu^2}\sum E[X_i]$$

$$= -\frac{1}{\mu^2}\sum \mu = -\frac{1}{\mu^2}n\mu = -\frac{n}{\mu}$$ [1]

So the CRLB is given by $\frac{\mu}{n}$. This is the same as the variance of $\hat{\mu}$. Hence the variance attains the CRLB.   [1]

[Total 2]

Solution X2.14

MLE and CRLB are covered in Chapter 7.

(i)   *Confidence interval for mean*

Here, $n = 10$, so using the result we are given in the question, we have $20\lambda\bar{X} \sim \chi^2_{20}$.

We want a confidence interval for the mean, so we need a confidence interval for $\dfrac{1}{\lambda}$. We have:

$$0.90 = P(10.85 < \chi^2_{20} < 31.41) \qquad [\frac{1}{2}]$$

$$= P(10.85 < 20\lambda\bar{X} < 31.41)$$

$$= P\left(\frac{10.85}{20\bar{X}} < \lambda < \frac{31.41}{20\bar{X}}\right) \qquad [\frac{1}{2}]$$

So a 90% confidence interval for λ is given by:

$$\left(\frac{10.85}{20 \times 680}, \frac{31.41}{20 \times 680}\right) = (0.000798, 0.00231) \qquad [1]$$

Therefore a 90% confidence interval for the mean, $\mu = 1/\lambda$, is given to 3 SF by:

$$(433, 1250) \qquad [1]$$
$$[\text{Total 3}]$$

(ii)   *Likelihood function and MLE of the mean*

Writing the PDF of the exponential distribution in terms of its mean, μ:

$$f(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}$$

So the likelihood function based on *n* observations is:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\mu} e^{-\frac{1}{\mu}x_i} = \frac{1}{\mu^n} e^{-\frac{1}{\mu}\sum x_i} \qquad [1]$$

Taking logs:

$$\ln L(\mu) = -n\ln\mu - \frac{1}{\mu}\sum x_i \qquad [\frac{1}{2}]$$

Differentiating with respect to μ:

$$\frac{d}{d\mu}\ln L(\mu) = \frac{-n}{\mu} + \frac{\sum x_i}{\mu^2} \qquad [\frac{1}{2}]$$

Setting the derivative equal to 0:

$$\frac{n}{\mu} = \frac{\sum X_i}{\mu^2}$$

[½]

This rearranges to give:

$$\mu = \frac{\sum X_i}{n} = \bar{X}$$

[½]

The second derivative of the log-likelihood is:

$$\frac{d^2}{d\mu^2} \ln L(\mu) = \frac{n}{\mu^2} - \frac{2\sum X_i}{\mu^3} = \frac{n\mu - 2n\bar{X}}{\mu^3}$$

[1]

Evaluating this at the point $\mu = \bar{X}$ gives:

$$\frac{d^2}{d\mu^2} \ln L(\mu) \bigg|_{\mu=\bar{X}} = \frac{n\bar{X} - 2n\bar{X}}{\bar{X}^3} = -\frac{n}{\bar{X}^2} < 0 \quad \Rightarrow \quad \text{max}$$

So the maximum likelihood estimator of μ is $\bar{X}$.

[1]

[Total 5]

(iii)(a)   **CRLB**

The formula for the *CRLB* of the mean μ is:

$$CRLB = -\frac{1}{E\left[\dfrac{d^2}{d\mu^2} \ln L(\mu)\right]}$$

From part (ii), we have:

$$\frac{d^2}{d\mu^2} \ln L(\mu) = \frac{n}{\mu^2} - \frac{2\sum X_i}{\mu^3}$$

So:

$$E\left[\frac{d^2}{d\mu^2} \ln L(\mu)\right] = E\left[\frac{n}{\mu^2} - \frac{2\sum X_i}{\mu^3}\right] = \frac{n}{\mu^2} - \frac{2}{\mu^3} E\left(\sum X_i\right) = \frac{n}{\mu^2} - \frac{2}{\mu^3} \sum E(X_i) = \frac{n}{\mu^2} - \frac{2}{\mu^3} n\mu$$

$$= \frac{n}{\mu^2} - \frac{2n}{\mu^2} = -\frac{n}{\mu^2}$$

[1½]

Hence, $CRLB(\mu) = \dfrac{\mu^2}{n}$.

[½]

(iii)(b)   *Estimated standard error*

The estimated standard error of $\overline{X}$ is:

$$\sqrt{\frac{\hat{\mu}^2}{n}} = \sqrt{\frac{680^2}{10}} = \sqrt{46,240} = 215.0 \qquad [1]$$

[Total 3]

(iv)(a)   *Confidence interval for mean*

The asymptotic distribution of $\hat{\mu} = \overline{X}$ is $N(\mu, CRLB)$. Therefore:

$$\frac{\overline{X} - \mu}{\sqrt{CRLB}} \approx N(0,1) \qquad [1]$$

and using this approximate distribution:

$$P\left(-1.6449 < \frac{\overline{X} - \mu}{\sqrt{CRLB}} < 1.6449\right) = 0.9 \qquad [1]$$

So, using the particular value for the CRLB obtained in (iii)(b), an approximate 90% confidence interval for μ is:

$$\overline{x} \pm 1.6449\sqrt{CRLB} = 680 \pm 1.6449 \times 215 = 680 \pm 354 = (326, 1034) \qquad [1]$$

Alternatively, using the expression for the CRLB involving μ (from (iii)(a)), and the result
$\frac{\overline{X} - \mu}{\sqrt{CRLB}} \approx N(0,1)$, *we could say:*

$$0.9 = P\left(-1.6449 < \frac{\overline{X} - \mu}{\mu/\sqrt{n}} < 1.6449\right)$$

$$= P\left(\mu\left(1 - \frac{1.6449}{\sqrt{n}}\right) < \overline{X} < \mu\left(1 + \frac{1.6449}{\sqrt{n}}\right)\right)$$

$$= P\left(\frac{1 - \frac{1.6449}{\sqrt{n}}}{\overline{X}} < \frac{1}{\mu} < \frac{1 + \frac{1.6449}{\sqrt{n}}}{\overline{X}}\right)$$

$$= P\left(\frac{\overline{X}}{1 + \frac{1.6449}{\sqrt{n}}} < \mu < \frac{\overline{X}}{1 - \frac{1.6449}{\sqrt{n}}}\right) \qquad [2]$$

So an approximate 90% confidence interval for μ is:

$$\left(\frac{\overline{x}}{1 + \frac{1.6449}{\sqrt{n}}}, \frac{\overline{x}}{1 - \frac{1.6449}{\sqrt{n}}}\right) = \left(\frac{680}{1 + \frac{1.6449}{\sqrt{10}}}, \frac{680}{1 - \frac{1.6449}{\sqrt{10}}}\right) = (447.3, 1417) \qquad [1]$$

(iv)(b)   *Compare confidence intervals*

The exact confidence interval of (433, 1250) from part (i) is very different from the approximate confidence intervals of (326, 1034) or (447, 1417) from part (iv)(a).

The normal approximation used in part (iv)(a) requires a large sample. We have a sample of only 10 values. Hence the approximation is not very good.   [1]
[Total 4]

Assignment X3 Solutions

Markers: This document sets out one approach to solving each of the questions (sometimes with alternatives). Please give credit for any other valid approaches.

Solution X3.1

Estimating regression parameters is covered in Chapter 11.

First calculate the values of s_{xx} and s_{xy}:

$$s_{xx} = \sum x^2 - n\bar{x}^2 = 126 - 4 \times 5^2 = 26$$

$$s_{xy} = \sum xy - n\bar{x}\bar{y} = 210.1 - 4 \times 5 \times 8.275 = 44.6 \qquad\qquad [\tfrac{1}{2}]$$

$$\Rightarrow \quad \hat{b} = \frac{s_{xy}}{s_{xx}} = \frac{44.6}{26} = 1.715 \qquad\qquad [\tfrac{1}{2}]$$

$$[\text{Total } 1]$$

Solution X3.2

Estimating regression parameters is covered in Chapter 11.

$$s_{yy} = \sum y^2 - n\bar{y}^2 = 350.49 - 4 \times 8.275^2 = 76.5875 \qquad\qquad [\tfrac{1}{2}]$$

The estimate for the variance parameter is given by:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(s_{yy} - \frac{s_{xy}^2}{s_{xx}} \right) = \frac{1}{2}\left[76.5875 - \frac{44.6^2}{26} \right]$$

$$= 0.0407 \qquad\qquad [\tfrac{1}{2}]$$

$$[\text{Total } 1]$$

Solution X3.3

Confidence intervals for b are covered in Chapter 11.

We know that $\dfrac{\hat{b}-b}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_2$.   [½]

So, using the tables of the t_2 distribution:

$$0.95 = P(-4.303 < t_2 < 4.303) = P\left(-4.303 < \frac{\hat{b}-b}{\sqrt{\hat{\sigma}^2/S_{xx}}} < 4.303 \right)$$

$$= P\left(\hat{b} - 4.303\sqrt{\hat{\sigma}^2/S_{xx}} < b < \hat{b} + 4.303\sqrt{\hat{\sigma}^2/S_{xx}} \right)$$ [½]

Substituting in the numerical values, $\hat{b} = 1.715$, $\hat{\sigma}^2 = 0.04067$ and $s_{xx} = 26$, we obtain
$1.715 \pm 0.170 = (1.54, 1.89)$ as the appropriate confidence interval.   [1]

[Total 2]

Solution X3.4

The correlation coefficient is covered in Chapter 10, the coefficient of determination is covered in Chapter 11.

(i)   ***Sample correlation coefficient***

The sample correlation coefficient is given by:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{44.6}{\sqrt{26 \times 76.5875}} = 0.9995$$ [1]

(ii)(a)   ***Confidence interval for correlation coefficient***

Using Fisher's approximation with the full value of r, the observed value of z_r is:

$$z_r = \tanh^{-1} r = 4.116619$$ [½]

We know that $Z_r \overset{.}{\sim} N(z_\rho, 1)$, where $z_\rho = \tanh^{-1}\rho$. So $\dfrac{Z_r - z_\rho}{\sqrt{1}} \overset{.}{\sim} N(0,1)$. Using the appropriate
percentage points from the *Tables*:   [½]

$$0.95 = P(-1.96 < Z_r - z_\rho < 1.96)$$ [½]

Rearranging, a 95% confidence interval for z_ρ is $z_r \pm 1.96$, *ie* (2.1566, 6.0766).   [½]

Therefore, a 95% confidence interval for ρ is $(0.974, 1.000)$.   [1]

Using $r = 0.9995$ *gives a confidence interval of* $(0.975, 1.000)$.

Alternatively, students may use the transformation $z_r = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$ *for full marks.*

(ii)(b)   *Coefficient of determination*

The coefficient of determination is given by:

$$R^2 = r^2 = 0.9995^2 = 99.9\%$$ [1]

The coefficient of determination gives the proportion of variability explained by the model. Since nearly all of the variability is explained, this tells us that the linear regression model is a very good fit.   [1]
[Total 5]

Solution X3.5

Tests for the binomial distribution are covered in Chapter 9.

The size of a test is the probability of rejecting H_0, given that H_0 is true. Here, assuming that $p = 0.4$, we have:

$$P(\text{reject } H_0) = P(X \geq 25)$$

Using a normal approximation to the binomial distribution, where $Bin(50, 0.4) \overset{\sim}{\cdot} N(20, 12)$, with a continuity correction, we have:

$$P[X \geq 25] \approx P[X > 24.5]$$ [1]

$$= P\left(Z > \frac{24.5 - 20}{\sqrt{12}}\right)$$

$$= P[Z > 1.299]$$ [1]

From the tables of the normal distribution:

$$1 - \Phi(1.299) = 1 - 0.90303 = 0.09697$$

So the approximate size of the test is 9.7%.   [1]
[Total 3]

Solution X3.6

Errors are covered in Chapter 9.

(i)   **Type I error**

A Type I error occurs if you reject H_0 when H_0 is true.   [1]

*Award no marks for definitions that include **probabilities.***

(ii)   **Type II error**

A Type II error occurs if you don't reject H_0 when H_0 is false.   [1]

*Award no marks for definitions that include **probabilities.***

(iii)   **Size**

The *size* of a test is the *probability* of rejecting H_0 when H_0 is true (*ie* the probability of a Type I error).   [1]

(iv)   **Power**

The *power* of a test is the *probability* of rejecting H_0 when H_0 is false (and may be a function of an unknown parameter).   [1]

Alternatively, it could be defined as $1 - P(\text{Type II error})$.

Solution X3.7

Tests for μ are covered in Chapter 9.

(i)   **Testing the mean**

We are testing: $H_0 : \mu = 9$ *vs* $H_1 : \mu \neq 9$ (σ^2 unknown)

Under H_0, the statistic $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ has a t_9 distribution.

Using the data given:

$$\overline{x} = 10.5$$

$$s^2 = \frac{1}{9}\{1{,}232.46 - 10 \times 10.5^2\} = 14.44$$ [1]

So the value of our test statistic is:

$$\frac{10.5 - 9}{\sqrt{14.44/10}} = 1.24827$$ [1]

Comparing this with the tables of the t_9 distribution, we find that we have a probability value in excess of $0.1 \times 2 = 20\%$, as the test here is two-sided.   [½]

Alternatively, we could say that, testing at the 5% significance level, the critical values are ± 2.262 and the value of the test statistic lies between the two critical values.

So we have insufficient evidence to reject H_0 at the 5% level, therefore it is reasonable to assume that the population mean is 9.   [½]

[Total 3]

(ii)   *Testing the variance*

We are now testing: $H_0 : \sigma^2 = 8$ *vs* $H_1 : \sigma^2 \neq 8$

Under H_0, the statistic $\dfrac{(n-1)S^2}{\sigma^2}$ has a χ_9^2 distribution. So here our test statistic is:

$$\frac{9 \times 14.44}{8} = 16.245$$ [1]

Comparing this value with the tables of the χ_9^2 distribution, we find that we have a probability value slightly in excess of 10% (since the test is two-sided).   [1]

Alternatively, we could say that, testing at the 5% significance level, the critical values are 2.700 and 19.02. So the value of the test statistic lies between the two critical values.

So we have insufficient evidence to reject H_0 at the 5% level, therefore it is reasonable to assume that the population variance is 8.   [1]

[Total 3]

Solution X3.8

Tests for a binomial distribution are covered in Chapter 9.

(i)   *Testing the proportion*

We are testing: $H_0 : p = 0.42$ *vs* $H_1 : p > 0.42$

If X is the number of people who supported the government then:

$$X \sim Bin(5000, p) \stackrel{.}{\sim} N(5000p, 5000pq)$$

Hence:

$$\frac{X - 5,000p}{\sqrt{5,000pq}} \stackrel{.}{\sim} N(0,1) \quad \text{or} \quad \frac{\hat{p} - p}{\sqrt{pq/5,000}} \stackrel{.}{\sim} N(0,1)$$ [1]

where $\hat{p} = X/5,000$.

Using a continuity correction, the value of our test statistic is:

$$\frac{2{,}184\frac{1}{2} - 2{,}100}{\sqrt{1{,}218}} = 2.421 \quad \text{or} \quad \frac{\frac{2{,}184.5}{5{,}000} - 0.42}{\sqrt{0.2436/5{,}000}} = 2.421 \qquad [1]$$

Comparing this with the normal distribution tables, we find that we have a probability value approximately equal to 0.8%. So we have sufficient evidence to reject H_0 at the 0.8% level, therefore it is reasonable to assume that the proportion of those who would vote for the current government is greater than 42%.   [1]

[Total 3]

Alternatively, we could say that the upper 5% point of $N(0,1)$ is 1.6449 and the upper 1% point is 2.3263. So we would reject the null hypothesis at both the 5% and 1% significance levels and hence reach the conclusion stated above.

(ii)   ***Testing the change in the proportions***

We will use X_1 and X_2 to represent the number of people supporting the government before and after the scandal. So:

$$X_1 \sim Bin(5000, p_1) \quad \text{and} \quad X_2 \sim Bin(3000, p_2)$$

We are testing:

$$H_0 : p_1 = p_2 \quad vs \quad H_1 : p_1 \neq p_2$$

We have:

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{5{,}000} + \dfrac{\hat{p}(1-\hat{p})}{3{,}000}}} \stackrel{.}{\sim} N(0,1) \qquad [1]$$

where $\hat{p} = \dfrac{2{,}185 + 1{,}191}{5{,}000 + 3{,}000} = \dfrac{3{,}376}{8{,}000} = 0.422$.

Also $\hat{p}_1 = \dfrac{2{,}185}{5{,}000} = 0.437$ and $\hat{p}_2 = \dfrac{1{,}191}{3{,}000} = 0.397$ which gives:

$$\frac{0.437 - 0.397}{\sqrt{\dfrac{0.422 \times 0.578}{5{,}000} + \dfrac{0.422 \times 0.578}{3{,}000}}} = 3.507 \qquad [1]$$

Comparing this with the normal distribution tables, we find that we have a probability value of $2 \times 0.022\% = 0.04\%$ (since it is a two-sided test). So we have sufficient evidence to reject H_0 at the 0.1% level, therefore it is reasonable to assume that the proportion of those who would vote for the current government is not the same as before the scandal.   [1]

[Total 3]

Alternatively, testing at the 5% significance level gives critical values of ± 1.96, and testing at the 1% significance level gives critical values of ± 2.5758. Since the test statistic is more extreme than the critical values, we reject the null hypothesis at both the 5% and 1% significance levels and reach the conclusion stated above. Note that technically, we would have to carry out a one-sided test to say it was less than before.

Solution X3.9

Correlation is covered in Chapter 10.

(i)   **Pearson's correlation coefficient**

For these data values:

$$s_{xx} = \sum x^2 - n\bar{x}^2 = 3{,}255.91 - \frac{144.1^2}{11} = 1{,}368.2$$

$$s_{yy} = \sum y^2 - n\bar{y}^2 = 7{,}481.26 - \frac{262.4^2}{11} = 1{,}221.8$$

$$s_{xy} = \sum xy - n\bar{x}\,\bar{y} = 2{,}495.43 - \frac{144.1 \times 262.4}{11} = -942.01 \qquad [\frac{1}{2}]$$

So the correlation coefficient is:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = -\frac{942.01}{\sqrt{1{,}368.2 \times 1{,}221.8}} = -0.7286 \qquad [\frac{1}{2}]$$

This value indicates a reasonably strong negative linear relationship.   [1]

[Total 2]

(ii)   *Spearman's rank correlation coefficient*

For these data values we have:

x	y	Rank x	Rank y	d	d^2
0	42.3	1	11	−10	100
1.1	30.7	2	9	−7	49
17.3	26.3	9	7	2	4
10.6	36.8	5	10	−5	25
25.1	8.9	10	1	9	81
5.2	25.1	4	5	−1	1
11.8	10.8	6	3	3	9
40	10	11	2	9	81
15.6	25.2	8	6	2	4
13.8	17.2	7	4	3	9
3.6	29.1	3	8	−5	25

Summing, we get:

$$\sum d^2 = 388 \tag{1}$$

Therefore the Spearman's rank correlation coefficient is:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 388}{11(11^2 - 1)} = -0.7636 \tag{1}$$

This value indicates a reasonably strong monotonically decreasing relationship (as we would have expected given the answer to (i)).   [1]

[Total 3]

(iii)   *Advantage of Spearman's rank correlation coefficient*

Since Spearman's rank correlation coefficient only considers ranks rather than the actual values, the value of the coefficient is less affected by outliers in the data than Pearson's correlation coefficient. Hence the Spearman's rank correlation coefficient is more robust.   [1]

(iv)   **Test**

We are testing:

> H_0 :   there is no association between age at death and the average number of cigarettes smoked per day

> H_1 :   there is some association between age at death and the average number of cigarettes smoked per day

We have $E[R_s] = 0$ and $\text{var}[R_s] = \dfrac{1}{n-1}$. The test statistic is:

$$\frac{r_s - 0}{\sqrt{\frac{1}{n-1}}} = \frac{-0.7636 - 0}{\sqrt{\frac{1}{10}}} = -2.415 \qquad [1]$$

This is a two-tailed test. Since $-2.415 < -1.96$, we have evidence at the 5% level to reject H_0 . It appears that there is some negative correlation. [1]

We should only use the normal approximation when we have a large sample size, so here we should be cautious about conclusions drawn from this approximate test. [1]

[Total 3]

(v)   **Kendall's rank correlation coefficient**

Writing the data in order of the x values, and showing the ranks for both x and y , we get:

x	y	Rank x	Rank y
0	42.3	1	11
1.1	30.7	2	9
3.6	29.1	3	8
5.2	25.1	4	5
10.6	36.8	5	10
11.8	10.8	6	3
13.8	17.2	7	4
15.6	25.2	8	6
17.3	26.3	9	7
25.1	8.9	10	1
40	10	11	2

[½]

We need to calculate the number of concordant and discordant pairs. From the Course Notes we have 'The concordant pairs (C) are the number of observations below the current one in the table that have a higher rank for the y and the discordant pairs (D) are the number of observations below which have a lower rank for the y', so:

x	y	Rank x	Rank y	C	D
0	42.3	1	11	0	10
1.1	30.7	2	9	1	8
3.6	29.1	3	8	1	7
5.2	25.1	4	5	3	4
10.6	36.8	5	10	0	6
11.8	10.8	6	3	3	2
13.8	17.2	7	4	2	2
15.6	25.2	8	6	1	2
17.3	26.3	9	7	0	2
25.1	8.9	10	1	1	0
40	10	11	2	n/a	n/a
			Sum	12	43

[1½]

For example for $x = 0$, the y rank is 11. In the 'Rank y' column there are no values higher than 11, hence C is 0. Similarly, in the 'Rank y' column all ten values are lower than 11, hence D is 10.

From this we can see that $n_c = 12$ and $n_d = 43$, so the Kendall's rank correlation coefficient is:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} = \frac{12 - 43}{11 \times 10/2} = -0.5636$$

[1]

[Total 3]

Alternatively, we can consider each pair individually. The table showing the concordant and discordant pairs is as follows:

(x,y)	0, 42.3	1.1, 30.7	17.3, 26.3	10.6, 36.8	25.1, 8.9	5.2, 25.1	11.8, 10.8	40, 10	15.6, 25.2	13.8, 17.2	3.6, 29.1
0,42.3		d	d	d	d	d	d	d	d	d	d
1.1,30.7			d	c	d	d	d	d	d	d	d
17.3,26.3				d	d	c	c	d	c	c	d
10.6,36.8					d	c	d	d	d	d	c
25.1,8.9						d	d	c	d	d	d
5.2,25.1							d	d	c	d	d
11.8,10.8								d	c	c	d
40,10									d	d	d
15.6,25.2										c	d
13.8,17.2											d
3.6,29.1											

From this table we can see that $n_c = 12$ and $n_d = 43$, so again the Kendall's rank correlation coefficient is:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} = \frac{12 - 43}{11 \times 10/2} = -0.5636$$

Solution X3.10

Goodness-of-fit tests and contingency tables are covered in Chapter 9.

(i)   ***Contingency table test***

We are testing:

H_0 :   there is no association between sex and colour-blindness (*ie* they are independent)

H_1 :   there is an association between sex and colour-blindness (*ie* they are not independent)

The expected values on the 2×2 contingency table are:

	male	female	
normal	950.5	950.5	1,901
colour-blind	49.5	49.5	99
	1,000	1,000	2,000

[1]

Since $\dfrac{\text{(row total)} \times \text{(column total)}}{\text{grand total}} = \dfrac{1,901 \times 1,000}{2,000} = 950.5$, *etc.*

The degrees of freedom are $(2-1) \times (2-1) = 1$.   [1]

Using $\chi^2 = \sum \dfrac{(O-E)^2}{E}$:

$$\chi_1^2 = \frac{(908-950.5)^2}{950.5} + \frac{(993-950.5)^2}{950.5} + \frac{(92-49.5)^2}{49.5} + \frac{(7-49.5)^2}{49.5} \qquad [1]$$

$$= 1.9003 + 1.9003 + 36.490 + 36.490$$

$$= 76.78 \qquad [1]$$

Since this exceeds even the 0.05% critical value of 12.12, we have overwhelming evidence at the 0.05% level to reject H_0. We therefore conclude that there **is** an association between sex and colour-blindness. [1]

[Total 5]

(ii)   *Goodness of fit test*

Using the maximum likelihood estimate of $\hat{q} = 0.0895$ in the formulae, we obtain the following numbers:

	male	female
normal	910.5	992.0
colour-blind	89.5	8.010

[1]

We are testing:

H_0 :   the model is a good fit.

H_1 :   the model is not a good fit.

Notice here that we are no longer dealing with a contingency table despite the presentation of the data. We are doing a χ^2 goodness-of-fit test.

The degrees of freedom are (no. of groups) $-1-$ (no. of parameters estimated) which is $4-1-1=2$ since q was estimated using the data. [1]

$$\chi_2^2 = \frac{(908-910.5)^2}{910.5} + \frac{(993-992)^2}{992} + \frac{(92-89.5)^2}{89.5} + \frac{(7-8.01)^2}{8.01} \qquad [1]$$

$$= 0.00686 + 0.00101 + 0.06983 + 0.12735$$

$$= 0.205 \qquad [1]$$

Since this is less than the 5% critical value of 5.991, we have insufficient evidence at the 5% level to reject H_0. We therefore conclude that the model *is* a good fit. [1]

[Total 5]

Solution X3.11

The confidence intervals used in this question are covered in Chapter 8.

(i)   *Confidence interval for mean*

Let the true mean age of wood found at this site be μ. Then:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$ [½]

Now, using the t_9 distribution:

$$-2.262 < \frac{\overline{X} - \mu}{S/\sqrt{n}} < 2.262$$ [½]

$$\overline{X} - 2.262 \frac{S}{\sqrt{n}} < \mu < \overline{X} + 2.262 \frac{S}{\sqrt{n}}$$

From our sample:

$$\overline{x} = \frac{46,960}{10} = 4,696$$ [½]
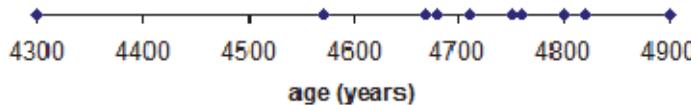
$$s^2 = \frac{1}{9}\{220,772,800 - 10 \times 4,696^2\} = 27,627$$ [½]

Therefore, a 95% confidence interval for μ is $4,696 \pm 119 = (4,577, \ 4,815)$.   [1]

[Total 3]

(ii)   *Dotplot*

The dotplot for the ages of the sample:

[1]

Given that our data set is small, the confidence interval in part (i) requires that the ages are normally distributed.   [½]

The plot seems to show that 4,300 is very different to the other values and so it may be an outlier. In which case the underlying distribution is not normal, and our confidence interval is not valid. However, more data is needed for us to be sure.   [½]

[Total 2]

(iii)   *Minimum sample size needed*

We require:

$$2 \times t_{n-1;0.025} \frac{s}{\sqrt{n}} \leq 200 \quad \Rightarrow \quad \frac{t_{n-1;0.025}}{\sqrt{n}} \leq 0.60164 \qquad [1]$$

Trial and improvement leads to $\dfrac{t_{13;0.025}}{\sqrt{14}} = 0.5773$ and $\dfrac{t_{12;0.025}}{\sqrt{13}} = 0.6043$. Therefore a sample

size of at least 14 is required.   [2]

[Total 3]

(iv)   *Confidence interval for difference between means*

We will use μ_X and μ_Y to denote the true mean age of wood at the first and second sites respectively. Then:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_P^2 \left(\dfrac{1}{n_X} + \dfrac{1}{n_Y} \right)}} \sim t_{n_X + n_Y - 2} \qquad [½]$$

Now, using the t_{16} distribution:

$$-2.120 < \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{s_P^2 \left(\dfrac{1}{10} + \dfrac{1}{8} \right)}} < 2.120 \qquad [½]$$

$$196 - 2.120 \sqrt{s_P^2 \left(\frac{1}{10} + \frac{1}{8} \right)} < (\mu_X - \mu_Y) < 196 + 2.120 \sqrt{s_P^2 \left(\frac{1}{10} + \frac{1}{8} \right)}$$

Now:

$$s_Y^2 = \frac{1}{7} \{162,280,000 - 8 \times 4,500^2\} = 40,000$$

$$s_P^2 = \frac{9 \times 27,627 + 7 \times 40,000}{10 + 8 - 2} = \frac{528,640}{16} = 33,040 \qquad [1]$$

Therefore a 95% confidence interval for $\mu_X - \mu_Y$ is $196 \pm 182.8 = (13.2, \ 379)$.   [1]

[Total 3]

(v)   *Confidence interval for ratio of variances*

Using $\dfrac{S_X^2 / S_Y^2}{\sigma_X^2 / \sigma_Y^2} \sim F_{n_X-1, n_Y-1}$, we have an $F_{9,7}$ distribution with upper and lower bounds of 3.677

and $1/3.293$.   [1]

$$\Rightarrow \quad \frac{1}{3.293} < \frac{27,627/40,000}{\sigma_X^2 / \sigma_Y^2} < 3.677$$

$$\Rightarrow \quad \frac{27,627/40,000}{3.677} < \frac{\sigma_X^2}{\sigma_Y^2} < 27,627/40,000 \times 3.293 \qquad [1]$$

This gives a confidence interval for $\dfrac{\sigma_X^2}{\sigma_Y^2}$ of $(0.188, 2.27)$.   [1]

Since this confidence interval contains 1, the assumption of equal variances used in the
confidence interval in part (iv) looks reasonable.   [1]

                                                                          [Total 4]
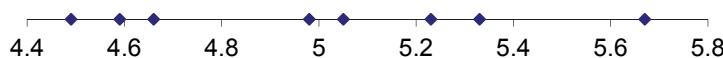
Solution X3.12

Two sample tests on means, variances and paired data are covered in Chapter 9.

(i)(a)(1) *Dotplot*

The dotplot is as follows:

Existing desiccant (A)

New desiccant (B)

                                                                                  [1]

The plots suggest that the new desiccant may extract more water. The spread of values is similar
for each desiccant.   [1]

(i)(a)(2) *Test equality of variances*

We use the F test to test for equality of variance. The hypotheses are:

$$H_0 : \sigma_A^2 = \sigma_B^2 \quad vs \quad H_1 : \sigma_A^2 \neq \sigma_B^2$$

Under H_0, the statistic $\dfrac{S_A^2}{S_B^2} \bigg/ \dfrac{\sigma_A^2}{\sigma_B^2}$ has an $F_{7,7}$ distribution.

Calculating the sample variances:

$$s_A^2 = \frac{1}{7}\{201.1574 - 8 \times 5^2\} = 0.16534 \qquad\qquad\qquad\qquad [\frac{1}{2}]$$

$$s_B^2 = \frac{1}{7}\{210.3659 - 8 \times 5.11375^2\} = 0.16606 \qquad\qquad\qquad [\frac{1}{2}]$$

So our test statistic is $\dfrac{0.16534}{0.16606} \bigg/ 1 = 0.9957$.   [1]

Since this lies between the critical values of 4.995 and 0.2002 we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to assume that the two population variances are equal. This is not surprising, since the spreads on our two data plots looked very similar.   [1]

Alternatively, we could have used $\dfrac{S_B^2}{S_A^2} \bigg/ \dfrac{\sigma_B^2}{\sigma_A^2}$ to get a ratio of 1.0043. This also has an $F_{7,7}$

distribution and we would draw the same conclusion as before.

(i)(b)   *Test if new desiccant extracts more moisture*

We now use a two-sample t test. The hypotheses are:

$$H_0: \mu_A = \mu_B \quad vs \quad H_1: \mu_B > \mu_A$$

This test is one-sided. The sample means are $\overline{A} = 5$, $\overline{B} = 5.11375$.

$$s_P^2 = \frac{7s_A^2 + 7s_B^2}{14} = \frac{2.3198}{14} = 0.1657 \qquad\qquad\qquad\qquad [1]$$

Our test statistic is:

$$T = \frac{(\overline{B} - \overline{A}) - (\mu_B - \mu_A)}{\sqrt{s_P^2\left(\dfrac{1}{8} + \dfrac{1}{8}\right)}} = \frac{0.11375}{\sqrt{0.041425}} = 0.559 \qquad\qquad [1]$$

which has a t_{14} distribution under H_0.

We compare this with the 5% point of the t_{14} distribution, which is 1.761. Our value lies well within the acceptance region. We have insufficient evidence to reject H_0 on the basis of this data. It does not appear that the new desiccant extracts any more moisture than the existing one.   [1]

[Total 8]

(ii)   **Paired t test**

We now perform a paired *t* test, looking at the differences. We use:

$$\frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \sim t_{n-1}$$ [1]

Let D represent the difference in moisture level extracted, so that $D = B - A$, and the values of D are:

0.16, -0.02, 0.17, 0.23, 0.24, -0.10, 0.13, 0.10   [1]

We must test whether these can be considered to come from a normal population with mean zero. We have the hypotheses:

$$H_0 : \mu_D = 0 \quad vs \quad H_1 : \mu_D > 0$$

Again, the test is one-sided.

Calculating the sample mean and variance:

$$\bar{D} = \bar{B} - \bar{A} = 0.11375$$

$$s_D^2 = \frac{1}{7}\left\{\sum D^2 - n\bar{D}^2\right\} = \frac{1}{7}\{0.2023 - 8 \times 0.11375^2\} = 0.01411$$ [1]

Under H_0, $\dfrac{\bar{D} - 0}{S_D / \sqrt{8}}$ has a t_7 distribution. So the observed value of our test statistic is:

$$\frac{0.11375}{\sqrt{0.01411 / 8}} = 2.7083$$ [1]

Comparing this with t_7 (critical value at 5% = 1.895), we have a result significant at the 5% level (and indeed at the 2½% level). This suggests that the new desiccant is more efficient at extracting moisture than the old one.   [1]
                                                                                                       [Total 5]

(iii)   **Comment**

The paired test shows that there was a significant difference between the two desiccators, whereas the two-sample test does not indicate any significant difference.   [1]

The small but significant difference between the two desiccants is masked in the two-sample test because the test statistic for the two-sample test is calculated using the pooled variance (which is 0.1657) rather than the sample variance of the differenced data (which is 0.01411). A smaller variance leads to a larger test statistic, which means we are more likely to reject the null hypothesis. In other words, the increased power of the paired test enables a significant difference to be identified.   [1]
                                                                                                       [Total 2]

Solution X3.13

The material tested in this question is covered in Chapter 11.

(i)(a)   *Derive least squares estimators*

Since $Y_i = a + bx_i + e_i \implies e_i = Y_i - a - bx_i$, the sum of squares is:

$$Q = \sum e_i^2 = \sum (Y_i - a - bx_i)^2$$

Differentiating with respect to a and b, we have:

$$\frac{\partial Q}{\partial a} = \sum 2(Y_i - a - bx_i) \times (-1)$$

$$\frac{\partial Q}{\partial b} = \sum 2(Y_i - a - bx_i) \times (-x_i)$$

[1]

Setting these two expressions to zero, we have:

$$\sum Y = na + b\sum x \qquad \dots \ (1)$$

$$\sum xY = a\sum x + b\sum x^2 \quad \dots \ (2)$$

Multiplying equation (1) by $\sum x$ and equation (2) by n, we obtain:

$$\sum x \sum Y = na\sum x + b\left(\sum x\right)^2 \qquad \dots \ (3)$$

$$n\sum xY = na\sum x + nb\sum x^2 \qquad \dots \ (4)$$

Subtracting equation (4) from equation (3), we obtain:

$$\sum x \sum Y - n\sum xY = b\left[\left(\sum x\right)^2 - n\sum x^2\right]$$

[1]

Rearranging this gives:

$$\hat{b} = \frac{\sum xY - \frac{\sum x \sum Y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \ \text{ or } \ \hat{b} = \frac{S_{xy}}{S_{xx}}$$

[1]

and this is our least squares estimator for b. The estimator for a is given by $\hat{a} = \bar{y} - \hat{b}\bar{x}$ (this can be obtained by rearranging equation (1)).

[1]

Award the marks for the alternative method using substitution.

(i)(b)   *Maximum likelihood estimators*

The answer would not have differed at all. For a normal distribution, maximum likelihood and least squares obtain the same estimates.   [1]

Since $Y_i = a + bx_i + e_i$, where $e_i \sim N(0, \sigma^2)$, gives $Y_i \sim N(a + bx_i, \sigma^2)$ the likelihood function using maximum likelihood estimation is:

$$L(a,b) = \prod \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{Y_i - a - bx_i}{\sigma}\right)^2\right\} \Rightarrow \ln(L) = \text{constant} - \frac{1}{2\sigma^2}\sum(Y_i - a - bx_i)^2$$

So maximising L is equivalent to minimising $\sum(Y_i - a - bx_i)^2$, which is identical to the criterion used above to find the least squares estimators. So the MLEs for a and b are equal to the least squares estimators in this case.   [1]

[Total 6]

(ii)   *Regression coefficients*

$$\hat{b} = \frac{s_{xy}}{s_{xx}} = \frac{826}{42} = 19.667 \tag*{[1]}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 105 - 19.667 \times 4.5 = 16.5 \tag*{[1]}$$

[Total 2]

(iii)   *Testing slope parameter*

We wish to test: $H_0 : b = 22$ *vs* $H_1 : b \neq 22$

Under H_0, the statistic $\dfrac{\hat{b} - b}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$, where $\hat{\sigma}^2$ is given by: [½]

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{6}\left(16{,}492 - \frac{826^2}{42}\right) = 41.2222 \tag*{[1]}$$

So the value of our test statistic here is:

$$\frac{19.667 - 22}{0.9907} = -2.355 \tag*{[1]}$$

This is a two-sided test. So, using values of the t_6 distribution, the probability value for this test statistic is about $0.03 \times 2 = 6\%$.   [½]

So we have insufficient evidence to reject H_0 at the 5% level, therefore it is reasonable to assume that the slope of the model is £22 per hour.   [1]

[Total 4]

(iv)(a) *Confidence interval for the average cost of a job lasting 4 hours*

Our estimate of the mean predicted cost is:

$$\hat{\mu} = \hat{a} + \hat{b} \times 4 = 16.5 + 19.667 \times 4 = 95.17 \qquad [1]$$

The standard error is given by:

$$\sqrt{\left(\frac{1}{8} + \frac{(4-4.5)^2}{42}\right)41.22} = \sqrt{5.3981} = 2.323 \qquad [1]$$

This gives a confidence interval, using the t_6 tables, of:

$$95.17 \pm 1.943 \times 2.323 = 95.17 \pm 4.51 \simeq (90.7, 99.7) \qquad [1]$$

(iv)(b) *Confidence interval for an individual job lasting 6 hours*

Our estimate of the individual predicted cost is:

$$\hat{\mu} = \hat{a} + \hat{b} \times 6 = 16.5 + 19.667 \times 6 = 134.5 \qquad [1]$$

The standard error is given by:

$$\sqrt{\left(1 + \frac{1}{8} + \frac{(6-4.5)^2}{42}\right)41.22} = \sqrt{48.583} = 6.970 \qquad [1]$$

This gives a confidence interval of:

$$134.5 \pm 1.943 \times 6.970 = 134.5 \pm 13.54 \simeq (121, 148) \qquad [1]$$

[Total 6]

(v) *Comment*

The confidence interval for the individual job is wider (£27) than the confidence interval for the average cost (£9). So there is greater uncertainty over an individual result than an average one.   [1]

Assignment X4 Solutions

Markers: This document sets out one approach to solving each of the questions (sometimes with alternatives). Please give credit for any other valid approaches.

Solution X4.1

This question applies the Core Reading in Chapter 13.

(i) *Conjugate prior for p*

The likelihood of observing $x_1, \ldots, x_n$ is given by:

$$L(p) = \frac{\Gamma(k+x_1)}{\Gamma(x_1+1)\Gamma(k)} p^k (1-p)^{x_1} \times \cdots \times \frac{\Gamma(k+x_n)}{\Gamma(x_n+1)\Gamma(k)} p^k (1-p)^{x_n}$$

$$\propto p^{nk}(1-p)^{\sum x_i} \hspace{4cm} [1]$$

If the prior distribution for p is $beta(\alpha, \beta)$, then its PDF is:

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \propto \quad p^{\alpha-1}(1-p)^{\beta-1} \hspace{2cm} [\tfrac{1}{2}]$$

So the PDF of the posterior distribution is given by:

$$f(p \mid \underline{x}) \propto \text{prior} \times \text{likelihood} = p^{\alpha-1}(1-p)^{\beta-1} \times p^{nk}(1-p)^{\sum x_i}$$

$$\propto p^{nk+\alpha-1}(1-p)^{\sum x_i+\beta-1} \hspace{4cm} [\tfrac{1}{2}]$$

This has the form of a $beta(\alpha^*, \beta^*)$ distribution, where $\alpha^* = nk+\alpha$ and $\beta^* = \sum x_i + \beta$. $[\tfrac{1}{2}]$

Both the posterior and the prior have beta distributions. Therefore the conjugate prior is indeed a beta distribution. $[\tfrac{1}{2}]$

[Total 3]

Alternatively students could adopt the following approach, where it is not necessary to do quite so much in terms of mathematical workings:

The likelihood of observing $x_1, \ldots x_n$ is given by:

$$L(p) = \frac{\Gamma(k+x_1)}{\Gamma(x_1+1)\Gamma(k)} p^k (1-p)^{x_1} \times \cdots \times \frac{\Gamma(k+x_n)}{\Gamma(x_n+1)\Gamma(k)} p^k (1-p)^{x_n}$$

$$\propto p^{nk}(1-p)^{\sum x_i} \hspace{4cm} [1]$$

The likelihood is of the form of a $beta\left(nk+1, \sum x_i + 1\right)$ PDF. $\hspace{2cm}$ [1]

The posterior PDF is proportional to the prior PDF multiplied by the likelihood. If the prior and the likelihood have beta forms, then the posterior must also have a beta distribution. [½]

Therefore the conjugate prior is indeed a beta distribution. [½]

[Total 3]

(ii) **Bayesian estimate for p under quadratic loss**

Substituting in the given values we have a posterior distribution of $beta(\alpha^*, \beta^*)$ where:

$$\alpha^* = nk + \alpha = 12 \times 2 + 3 = 27 \qquad \beta^* = \sum x_i + \beta = 8 + 4 = 12 \qquad \text{[½]}$$

The Bayesian estimate for p under quadratic loss is the mean of the posterior distribution:

$$\frac{\alpha^*}{\alpha^* + \beta^*} = \frac{27}{39} = 0.692 \qquad \text{[½]}$$

[Total 1]

Solution X4.2

This question tests the material in Chapter 13, on the derivation of the posterior distribution where the prior distribution is uninformative.

The likelihood function for μ is given by:

$$L(\mu) = \frac{1}{x_1 \sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x_1 - \mu}{\sigma}\right)^2} \times \cdots \times \frac{1}{x_n \sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x_n - \mu}{\sigma}\right)^2} \propto e^{-\frac{1}{2}\sum\left(\frac{\ln x_i - \mu}{\sigma}\right)^2} \qquad \text{[1]}$$

Since the prior distribution for μ is uninformative, we take the prior distribution to be constant. *Alternatively, we could say $f(\mu) = 1/k$ where $k \to \infty$.* [½]

So the PDF of the posterior distribution for μ is given by:

$$f(\mu \mid \underline{x}) \propto \text{prior x likelihood} = \text{constant} \times e^{-\frac{1}{2}\left(\frac{\ln x_i - \mu}{\sigma}\right)^2} \propto e^{-\frac{1}{2}\sum\left(\frac{\ln x_i - \mu}{\sigma}\right)^2} \qquad \text{[½]}$$

We now want to write this as the PDF of a normal distribution, with μ as the variable. So we want it to look like:

$$f(\mu \mid \underline{x}) \propto e^{-\frac{1}{2}\left(\frac{\mu - \mu_*}{\sigma_*}\right)^2} = e^{-\frac{1}{2\sigma_*^2}(\mu^2 - 2\mu_*\mu + \mu_*^2)} \qquad (*)$$

So, expanding the bracket in our posterior and summing the terms (ignoring any expressions that do not involve μ, since these can be absorbed into the constant term):

$$f(\mu\,|\,\underline{x}) \propto e^{-\frac{1}{2\sigma^2}\sum(\ln x_i - \mu)^2} = e^{-\frac{1}{2\sigma^2}\sum\left((\ln x_i)^2 - 2\mu\ln x_i + \mu^2\right)}$$

$$= e^{-\frac{1}{2\sigma^2}\left(\sum(\ln x_i)^2 - 2\mu\sum\ln x_i + n\mu^2\right)} \propto e^{-\frac{1}{2\sigma^2}\left(-2\mu\sum\ln x_i + n\mu^2\right)}$$

$$= e^{-\frac{n}{2\sigma^2}\left(\mu^2 - 2\mu\frac{\sum\ln x_i}{n}\right)} \qquad\qquad [1\tfrac{1}{2}]$$

Completing the square (the missing term is absorbed into the constant):

$$f(\mu\,|\,\underline{x}) \propto e^{-\frac{1}{2(\sigma^2/n)}\left(\mu - \frac{\sum\ln x_i}{n}\right)^2} \qquad\qquad [1]$$

Comparing this with equation $(*)$, we see that we have a normal distribution with parameters:

$$\mu_* = \frac{\sum\ln x_i}{n} \quad \text{and} \quad \sigma_*^2 = \frac{\sigma^2}{n} \qquad\qquad [\tfrac{1}{2}]$$

[Total 5]

Solution X4.3

This question tests the material from Chapter 15.

(i)   *EBCT Model 2 assumptions*

The assumptions are:

- The distribution of each X_j, $j = 1, \ldots, n$, depends on a parameter, θ, whose value does not change over time (and the same for all the X_j's) but is unknown

- Given θ, the X_j's are independent (but not necessarily identically distributed)

- $E(X_j\,|\,\theta)$ does not depend on j

- $P_j \operatorname{var}(X_j\,|\,\theta)$ does not depend on j.

[½ mark each, total 2]

(ii)(a) *Proof of unconditional mean*

Using the conditional expectation formula from page 16 of the *Tables*:

$$E(X_j) = E[E(X_j | \theta)] = E[m(\theta)]$$ [1]

(ii)(b) *Proof of unconditional variance*

Using the conditional variance formula from page 16 of the *Tables*:

$$\text{var}(X_j) = E[\text{var}(X_j | \theta)] + \text{var}[E(X_j | \theta)]$$

$$= E\left[\frac{s^2(\theta)}{P_j}\right] + \text{var}[m(\theta)]$$

$$= \frac{1}{P_j} E[s^2(\theta)] + \text{var}[m(\theta)]$$ [1]

[Total 2]

(iii) *EBCT Model 2 credibility premium*

Using the formula from page 30 of the *Tables*, the credibility factor for Risk 1 is given by:

$$Z_1 = \frac{\displaystyle\sum_{j=1}^{n} P_{1j}}{\displaystyle\sum_{j=1}^{n} P_{1j} + \frac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

So the estimated value of Z_1 is:

$$\frac{460}{460 + \frac{2.870}{0.1172}} = 0.9495$$ [1]

We have $\bar{x}_1 = \frac{980}{460} = 2.130$ and $\bar{x} = 2.275$. Hence the empirical Bayes credibility estimate for Risk 1 *per policy sold* is:

$$\text{cred estimate} = Z_1 \bar{x}_1 + (1 - Z_1) \bar{x}$$

$$= 0.9495 \times 2.130 + (1 - 0.9495) \times 2.275$$

$$= 2.138$$ [1]

Therefore, the credibility premium for year 5 is:

$$140 \times 2.138 = 299.3 \text{ , } ie \text{ £299,300}$$ [1]

[Total 3]

Solution X4.4

This question tests the material in Chapter 13.

(i) *Posterior distribution for λ*

The likelihood function for λ is:

$$f(\underline{x} \mid \lambda) = \frac{e^{-\lambda} \lambda^{x_1}}{x_1!} \times \cdots \times \frac{e^{-\lambda} \lambda^{x_n}}{x_n!} \quad \propto \quad e^{-n\lambda} \lambda^{\sum x_i} \qquad\qquad [1]$$

The prior distribution for λ is:

$$f(\lambda) = \frac{2^5}{\Gamma(5)} \lambda^4 e^{-2\lambda} \quad \propto \quad \lambda^4 e^{-2\lambda} \qquad\qquad [\tfrac{1}{2}]$$

So the PDF of the posterior distribution for λ is given by:

$$f(\lambda \mid \underline{x}) \propto \text{prior} \times \text{likelihood} = \lambda^4 e^{-2\lambda} \times e^{-n\lambda} \lambda^{\sum x_i}$$

$$\propto \lambda^{4 + \sum x_i} e^{-(n+2)\lambda} \qquad\qquad [\tfrac{1}{2}]$$

This has the form of the PDF of a *Gamma*(α^*, λ^*) distribution, where $\alpha^* = 5 + \sum x_i$ and $\lambda^* = n + 2$. From the data, we have $n = 10$ and $\sum x_i = 12$, so $\alpha^* = 17$ and $\lambda^* = 12$. $\qquad$ [1]

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ [Total 3]

(ii)(a) *Bayesian estimate for λ under squared error loss*

Using a squared error loss function, the Bayesian estimate for λ is the mean of the posterior distribution. So the Bayesian estimate is given by:

$$\frac{\alpha^*}{\lambda^*} = \frac{17}{12} = 1.4167 \qquad\qquad [1]$$

(ii)(b) *Bayesian estimate for λ under zero-one loss*

Using a zero-one loss function, the Bayesian estimate for λ is the mode of the posterior distribution. The PDF of the *Gamma*(17,12) distribution is given by:

$$f(\lambda \mid \underline{x}) = \frac{12^{17}}{\Gamma(17)} \lambda^{16} e^{-12\lambda}$$

Taking logs (to make it easier to differentiate):

$$\log f(\lambda \mid \underline{x}) = 17\log 12 + 16\log \lambda - 12\lambda - \log\left[\Gamma(17)\right]$$ [½]

Differentiating with respect to λ and setting the result to zero gives us the mode of the posterior distribution and the Bayesian estimate for λ:

$$\frac{d}{d\lambda}[\log f(\lambda \mid \underline{x})] = \frac{16}{\lambda} - 12 = 0 \quad \Rightarrow \quad \lambda = \frac{16}{12} = 1.333$$ [1]

Checking that we do indeed obtain a maximum:

$$\frac{d^2}{d\lambda^2}[\log f(\lambda \mid \underline{x})] = -\frac{16}{\lambda^2} < 0 \quad \Rightarrow \quad \max$$ [½]

(ii)(c) *Bayesian estimate for λ under absolute loss*

Using an absolute error loss function, the Bayesian estimate for λ is the median of the posterior distribution. For $X \sim Gamma(17,12)$, we require the value of m such that:

$$P(X < m) = 0.5$$ [½]

Multiplying through by $2\lambda^*$ and using $2\lambda^* X \sim \chi^2_{2\alpha^*}$ from Page 12 of the *Tables*:

$$P(2\lambda^* X < 2\lambda^* m) = 0.5$$ [½]

$$P(\chi^2_{34} < 24m) = 0.5$$ [½]

From page 169 of the *Tables* we see that $P(\chi^2_{34} < 33.34) = 0.5$. Hence:

$$24m = 33.34 \quad \Rightarrow \quad m = 1.389$$

So the Bayesian estimate for λ is 1.389. [½]

[Total 5]

Solution X4.5

This question applies the bookwork in Chapter 12 on showing whether or not a distribution is a member of the exponential family.

(i)(a)   ***Exponential family***

Rewriting the PDF we get:

$$f(y) = \frac{\alpha^\alpha}{\mu^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{\alpha}{\mu}y} = \exp\left\{\alpha\ln\alpha - \alpha\ln\mu - \ln\Gamma(\alpha) + (\alpha-1)\ln y - \frac{\alpha y}{\mu}\right\}$$

$$= \exp\left\{\alpha\left(-\frac{y}{\mu} - \ln\mu\right) + \alpha\ln\alpha - \ln\Gamma(\alpha) + (\alpha-1)\ln y\right\} \qquad [\tfrac{1}{2}]$$

$$= \exp\left\{\frac{-\frac{1}{\mu}y - \ln\mu}{1/\alpha} + \alpha\ln\alpha - \ln\Gamma(\alpha) + (\alpha-1)\ln y\right\} \qquad [\tfrac{1}{2}]$$

This has the form of a member of the exponential family. Comparing with the formula on Page 27 of the *Tables*, we see that:

The natural parameter is $\theta = -\dfrac{1}{\mu}$. $\qquad$ [½]

The dispersion (or scale) parameter is $\phi = \alpha$. $\qquad$ [½]

Where a distribution has two parameters (eg normal, gamma, binomial), we can determine ϕ by taking it to be the 'other' parameter (ie the parameter other than μ) in the pdf formulation. Here the two parameters are α and μ so ϕ is taken to be α . For a one parameter distribution (eg Poisson, exponential) we take ϕ to be 1.

Note also though that other possibilities for the dispersion parameter exist, eg $\phi = 1/\alpha$. Full marks should be awarded for these provided the other functions are expressed correctly, eg $a(\phi) = \phi$.

Hence:

$$a(\phi) = \frac{1}{\alpha} = \frac{1}{\phi} \qquad [\tfrac{1}{2}]$$

$$b(\theta) = \ln\mu = \ln\left(-\frac{1}{\theta}\right) = -\ln(-\theta) \qquad [1]$$

$$c(y,\phi) = \phi\ln\phi - \ln\Gamma(\phi) + (\phi-1)\ln y \qquad [\tfrac{1}{2}]$$

These functions need to be expressed in terms of θ and ϕ to obtain the marks.

(i)(b) *Mean and variance*

Using the formulae on Page 27 of the *Tables*, we get:

$$E(Y) = b'(\theta) = -\frac{1}{\theta} = \mu \qquad \text{(or } \frac{\alpha}{\lambda} \text{ since } \mu = \frac{\alpha}{\lambda}\text{)}$$ [1]

$$\text{var}(Y) = a(\phi)b''(\theta) = \frac{1}{\phi} \times \frac{1}{\theta^2} = \frac{1}{\phi\theta^2} = \frac{\mu^2}{\alpha} \quad \text{(or } \frac{\alpha}{\lambda^2} \text{ since } \mu = \frac{\alpha}{\lambda}\text{)}$$ [1]

Working is essential to ensure students have used these formulae and not just copied the mean and variance from Page 12 of the Tables.

(i)(c) *Variance function*

The variance function is given by:

$$V(\mu) = b''(\theta) = \frac{1}{\theta^2} = \mu^2$$ [1]

Students do need to express this in terms of the mean, μ, in order to obtain the mark.

Alternatively, students could calculate the variance function directly using the mean and variance of a gamma distribution given on Page 12 of the Tables.

$$\text{var}(Y) = a(\phi)V(\mu) \implies V(\mu) = \frac{\text{var}(Y)}{a(\phi)} = \frac{\alpha/\lambda^2}{1/\alpha} = \left(\frac{\alpha}{\lambda}\right)^2 = \mu^2$$

[Total 7]

(ii) *Negative binomial distribution*

We have:

$$P(X = x) = \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} p^k (1-p)^x$$

We need the PF to contain the mean, μ. Since:

$$\mu = \frac{k(1-p)}{p} \implies p = \frac{k}{k+\mu}$$

We get:

$$P(X = x) = \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} \left(\frac{k}{k+\mu}\right)^k \left(\frac{\mu}{k+\mu}\right)^x$$ [½]

Rearranging gives:

$$P(X = x) = \exp\left\{ x\ln\left(\frac{\mu}{k+\mu}\right) - k\ln(k+\mu) + \ln\left(\frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)}\right) + k\ln k \right\}$$ [½]

The first two terms correspond to $\dfrac{x\theta - b(\theta)}{a(\phi)}$, but since the natural parameter θ is a function of

μ only, we need to separate out the scale parameter, which will be $\phi = k$, from these two

terms.   [½]

However there is no way to separate k from either the $x\ln\left(\dfrac{\mu}{k+\mu}\right)$ term or the $\ln(k+\mu)$ term. So

the negative binomial distribution cannot be expressed in the exponential family form.   [½]

[Total 2]

Solution X4.6

This material is from Chapter 15.

(i)   ***Effect on Z***

The credibility premium is calculated as:

$$\text{credibility premium} = Z \times \overline{X}_i + (1 - Z) \times \overline{X}$$

where $Z = \dfrac{n}{n + \frac{E}{V}}$ and n is the number of years of past data. As n increases, Z increases so more

emphasis is put on the data from a particular risk (direct data). This makes sense, since the more
information we have from the relevant risk, the less emphasis we will wish to place on the
collateral data.   [1]

$E[s^2(\theta)]$ is the average of the variability *within* each of the different risks in the group. As

$E[s^2(\theta)]$ increases, Z decreases so more emphasis is put on the collateral data. This makes sense,
since the more variable each individual risk's experience is, the less reliable it is. So we would
expect to rely more on the collateral data.   [1]

$\text{var}[m(\theta)]$ is a measure of the variability *between* the means of the different risks in the group. As
$\text{var}[m(\theta)]$ increases, Z increases so more emphasis is put on the particular risk. This makes sense,
since the larger the variability between the different risks, the less relevant the other risks are in
assessing the premium of our particular risk. So we want to rely more on the direct data.   [1]

[Total 3]

(ii) Credibility premium

The means for each of the risks are:

$$\overline{x}_1 = 728 \qquad \overline{x}_2 = 907 \qquad \overline{x}_3 = 1{,}150 \qquad \overline{x}_4 = 1{,}033$$

The estimated value of $E[m(\theta)]$ is the overall mean:

$$\overline{x} = \tfrac{1}{4}(728 + 907 + 1{,}150 + 1{,}033) = 954.5 \tag*{[1]}$$

The estimated value of $E[s^2(\theta)]$ is the mean of the variances:

$$\frac{1}{4}\sum_{i=1}^{4}\frac{1}{2}\sum_{j=1}^{3}(x_{ij} - \overline{x}_i)^2 = \tfrac{1}{4}(11{,}172 + 10{,}147 + 10{,}900 + 11{,}997) = 11{,}054 \tag*{[1]}$$

The estimated value of $\operatorname{var}[m(\theta)]$ is:

$$\frac{1}{3}\sum_{i=1}^{4}(\overline{x}_i - \overline{x})^2 - \frac{1}{3}\left[\frac{1}{4}\sum_{i=1}^{4}\frac{1}{2}\sum_{j=1}^{3}(x_{ij} - \overline{x}_i)^2\right]$$

$$= \frac{1}{3}\left[(728 - 954.5)^2 + (907 - 954.5)^2 + (1{,}150 - 954.5)^2 + (1{,}033 - 954.5)^2\right] - \tfrac{1}{3}\times 11{,}054$$

$$= \tfrac{1}{3}\times 97{,}941 - \tfrac{1}{3}\times 11{,}054 = 28{,}962.3 \tag*{[1]}$$

So, the estimated credibility factor is:

$$Z = \frac{3}{3 + \dfrac{11{,}054}{28{,}962.3}} = 0.88714$$

Using the values calculated the credibility premium for Risk 3 is:

$$0.88714 \times 1{,}150 + 0.11286 \times 954.5 = 1{,}128 \tag*{[1]}$$

$$\text{[Total 4]}$$

(iii) Addition of a fifth risk

This risk has a much higher mean, so the variance of the means, $\operatorname{var}[m(\theta)]$, will increase. It has a similar variance to the other risks, so the mean variance of each risk, $E[s^2(\theta)]$, will remain similar. [1]

Hence the proportionately larger $\operatorname{var}[m(\theta)]$ will lead to a larger Z, because more emphasis will be put on the direct data. [1]

$$\text{[Total 2]}$$

Solution X4.7

This question covers material in Chapter 12.

(i) **Completed table**

Model	Parameterised form of the linear predictor	Number of parameters	Scaled deviance
SA	$\alpha + \beta x$	2	238.4
SA + PT	$\alpha_i' + \beta x$	11	206.7
SA + PT + SA • PT	$\alpha_i' + \beta_i' x$	20	178.3
SA * PT + NB	$\alpha_i' + \beta_i' x + B_j$	25	166.2
SA * PT * NB	$\alpha_{ij}'' + \beta_{ij}'' x$	120	58.9

[½ mark per entry = total 3]

Any other dummy variable is acceptable, eg $\gamma_i + \beta x$ or $T_i' + \beta x$ for SA + PT, as long as the subscripts match up.

Working:

SA + PT $(\alpha + \beta x) + T_i = \alpha_i' + \beta x$ $2 + (10 - 1) = 11$ parameters

SA + PT + SA • PT = SA * PT $= (\alpha + \beta x) * T_i = \alpha_i' + \beta_i' x$ $2 \times 10 = 20$ parameters

SA * PT + NB $\alpha_i' + \beta_i' x + B_j$ $20 + (6 - 1) = 25$ parameters

(ii) **Comparing models**

Comparing SA+PT with SA

The difference in the scaled deviances is $238.4 - 206.7 = 31.7$.

This is greater than 16.92, the upper 5% critical value of a $\chi_{11-2}^2 = \chi_9^2$ distribution.

So SA + PT **is** a significant improvement over the SA model.   [1]

*Alternatively, using the $\Delta(\text{deviance}) > 2 \times \Delta(\text{parameters})$ approximation, we get $31.7 > 2 \times 9$ so the SA + PT model **is** a significant improvement over the SA model.*

Comparing SA∗PT with SA+PT

The difference in the scaled deviances is $206.7 - 178.3 = 28.4$.

This is greater than 16.92, the upper 5% critical value of a $\chi^2_{20-11} = \chi^2_9$ distribution.

So SA $\ast$ PT **is** a significant improvement over the SA + PT model.   [1]

Alternatively, using the $\Delta(\text{deviance}) > 2 \times \Delta(\text{parameters})$ *approximation, we get* $28.4 > 2 \times 9$ *so the* SA $\ast$ PT *model **is** a significant improvement over the* SA + PT *model.*

Comparing SA∗PT+NB with SA∗PT

The difference in the scaled deviances is $178.3 - 166.2 = 12.1$.

This is greater than 11.07, the upper 5% critical value of a $\chi^2_{25-20} = \chi^2_5$ distribution.

So SA $\ast$ PT + NB **is** a significant improvement over the SA $\ast$ PT model.   [1]

Alternatively, using the $\Delta(\text{deviance}) > 2 \times \Delta(\text{parameters})$ *approximation, we get* $12.1 > 2 \times 5$ *so the* SA $\ast$ PT *model **is** a significant improvement over the* SA $\ast$ PT + NB *model.*

Comparing SA∗PT∗NB with SA∗PT+NB

The difference in the scaled deviances is $166.2 - 58.9 = 107.3$.

This is less than 118.7, the upper 5% critical value of a $\chi^2_{120-25} = \chi^2_{95}$ distribution.

We have to interpolate in the tables between χ^2_{90} *and* χ^2_{100} *to get the figure of 118.7.*

So SA $\ast$ PT $\ast$ NB is **not** a significant improvement over the SA $\ast$ PT+NB model.   [1]

Alternatively, using the $\Delta(\text{deviance}) > 2 \times \Delta(\text{parameters})$ *approximation, we get* $107.3 \not> 2 \times 95$ *so the* SA $\ast$ PT $\ast$ NB *model is **not** a significant improvement over the* SA $\ast$ PT+NB *model.*

So the analyst should choose the SA $\ast$ PT+NB model.   [1]

[Total 5]

(iii)   *Further information*

The analyst should also check:

- that the SA $\ast$ PT+NB model is a significant improvement when the order is different, *eg* add the NB factor *before* the PT factor   [½]

- other models involving these rating factors, *eg* SA $\ast$ NB+PT   [½]

- the residuals of the proposed model (to ensure that it is a good fit to the data)   [½]

- the significance of the parameters of the proposed model (to ensure that all the estimated parameters are significantly different from zero).   [½]

[Total 2]

Solution X4.8

This question covers material in Chapters 7 and 11.

(i)   **MLE**

$$\hat{\mu} = \overline{x} = \frac{1,140 + 1,200 + 1,170 + 1,190}{4} = 1,175 \text{ (per 100,000 lives)}$$ [1]

(ii)(a)   **Correlation coefficient**

We have:

$$\overline{t} = 160/4 = 40 \qquad \text{and} \qquad \overline{x} = 1,179/4 = 294.75$$ [½]

$$s_{tt} = \sum t^2 - n\overline{t}^2 = 6,900 - 4 \times 40^2 = 500$$

$$s_{xx} = \sum x^2 - n\overline{x}^2 = 613,379 - 4 \times 294.75^2 = 265,868.75$$

$$s_{tx} = \sum tx - n\overline{t}\,\overline{x} = 57,515 - 4 \times 40 \times 294.75 = 10,355$$ [1]

Hence, the correlation coefficient is given by:

$$r = \frac{s_{tx}}{\sqrt{s_{tt}s_{xx}}} = \frac{10,355}{\sqrt{500 \times 265,868.75}} = 0.89811$$ [½]

We have strong positive correlation. So there is a positive link between death and age, *ie* we expect to get more deaths at older ages.   [½]

(ii)(b)   **Linear regression**

Our estimates are:

$$\hat{\beta} = \frac{s_{tx}}{s_{tt}} = \frac{10,355}{500} = 20.71$$

$$\hat{\alpha} = \overline{x} - \hat{\beta}\overline{t} = 294.75 - 20.71 \times 40 = -533.65$$ [1]

So our fitted regression line is:

$$\hat{x} = -533.65 + 20.71t$$ [½]

[Total 4]

Solution X4.9

(i)   *Exponential family*

The PDF from page 11 of the *Tables* is:

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2\right\}$$

Rewriting this into the required form gives:

$$f(y_i) = \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2 - \ln\sigma - \frac{1}{2}\ln 2\pi\right\}$$

$$= \exp\left\{\frac{-\frac{1}{2}y_i^2 + \mu_i y_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \ln\sigma - \frac{1}{2}\ln 2\pi\right\}$$

$$= \exp\left\{\frac{\mu_i y_i - \frac{1}{2}\mu_i^2}{\sigma^2} + \left(-\frac{y_i^2}{2\sigma^2} - \ln\sigma - \frac{1}{2}\ln 2\pi\right)\right\} \qquad [1]$$

Comparing this with the expression on page 27 of the *Tables*, we see that:

$$\theta_i = \mu_i, \qquad b(\theta_i) = \frac{1}{2}\mu_i^2 = \frac{1}{2}\theta_i^2, \qquad \phi = \sigma, \qquad a(\phi) = \sigma^2 = \phi^2$$

$$c(y_i, \phi) = \left(-\frac{y_i^2}{2\phi^2} - \ln\phi - \frac{1}{2}\ln 2\pi\right) \qquad [1]$$

[Total 2]

(ii)   *Natural parameter and variance function*

The natural parameter, θ_i, is simply μ_i. [1]

The variance function, $V(\mu_i)$ is given by $V(\mu_i) = b''(\theta_i)$ where we are differentiating with respect to θ_i. From part (i) we have $b(\theta_i) = \frac{1}{2}\theta_i^2$. Hence:

$$b'(\theta_i) = \theta_i \qquad [\frac{1}{2}]$$

$$\Rightarrow \quad V(\mu_i) = b''(\theta_i) = 1 \qquad [\frac{1}{2}]$$

[Total 2]

(iii)   *Pearson and deviance residual*

The Pearson residual of observation y_i from $Y_i \sim N(\mu_i, \sigma^2)$ is:

$$\frac{y_i - \hat{\mu}_i}{\sqrt{\sigma^2}} = \frac{y_i - \hat{\mu}_i}{\sigma} \qquad\qquad [1]$$

The PDF of $Y_i \sim N(\mu_i, \sigma^2)$ is:

$$f(y_i) = \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2 - \ln\sigma - \frac{1}{2}\ln 2\pi \right\}$$

The likelihood is:

$$L(\mu_1, \ldots, \mu_n) = f(y_1) \times \cdots \times f(y_n) = \exp\left[\sum_{i=1}^{n}\left\{ -\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2 - \ln\sigma - \frac{1}{2}\ln 2\pi \right\} \right] \qquad [\frac{1}{2}]$$

So the log-likelihood is:

$$\ln L(\mu_1, \ldots, \mu_n) = \sum_{i=1}^{n}\left\{ -\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2 - \ln\sigma - \frac{1}{2}\ln 2\pi \right\} \qquad [\frac{1}{2}]$$

In the fitted model we use the fitted values $\hat{\mu}_i$. Hence:

$$\ln L_M = \sum_{i=1}^{n}\left\{ -\frac{1}{2}\left(\frac{y_i - \hat{\mu}_i}{\sigma}\right)^2 - \ln\sigma - \frac{1}{2}\ln 2\pi \right\} \qquad [\frac{1}{2}]$$

In the saturated model the fitted values, $\hat{\mu}_i$, are the observed values, y_i. So:

$$\ln L_S = \sum_{i=1}^{n}\left\{ -\frac{1}{2}\left(\frac{y_i - y_i}{\sigma}\right)^2 - \ln\sigma - \frac{1}{2}\ln 2\pi \right\} = \sum_{i=1}^{n}\left\{ -\ln\sigma - \frac{1}{2}\ln 2\pi \right\} \qquad [\frac{1}{2}]$$

Hence the scaled deviance is:

$$-2\left(\ln L_M - \ln L_S\right) = 2\sum_{i=1}^{n}\left\{ \frac{1}{2}\left(\frac{y_i - \hat{\mu}_i}{\sigma}\right)^2 \right\} = \sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{\sigma}\right)^2 \qquad [\frac{1}{2}]$$

Therefore the deviance residual for y_i is:

$$\text{sign}(y_i - \hat{\mu}_i) \times \sqrt{\left(\frac{y_i - \hat{\mu}_i}{\sigma}\right)^2} = \text{sign}(y_i - \hat{\mu}_i)\frac{|y_i - \hat{\mu}_i|}{\sigma} = \frac{y_i - \hat{\mu}_i}{\sigma} \qquad [\frac{1}{2}]$$

This is the same as the Pearson residual.   [Total 4]

Solution X4.10

This question tests material from Chapter 13.

(i)(a) *Posterior distribution*

The prior distribution for θ has the form:

$$\text{Prior}(\theta) \propto \theta^{\beta-1}(1-\theta)^{\beta-1}$$

The random variable X has a binomial distribution with parameters m and θ. So the likelihood function based on a random sample $x_1, \ldots, x_n$ is:

$$L(\theta) = \binom{m}{x_1}\theta^{x_1}(1-\theta)^{m-x_1} \times \binom{m}{x_2}\theta^{x_2}(1-\theta)^{m-x_2} \times \cdots \times \binom{m}{x_n}\theta^{x_n}(1-\theta)^{m-x_n}$$

$$\propto \theta^{\sum x_i}(1-\theta)^{mn-\sum x_i} \tag{1}$$

The posterior distribution for θ is proportional to the product of the likelihood function and the prior distribution:

$$\text{Posterior}(\theta) \propto \theta^{\beta-1}(1-\theta)^{\beta-1} \times \theta^{\sum x_i}(1-\theta)^{mn-\sum x_i}$$

$$= \theta^{\beta-1+\sum x_i}(1-\theta)^{\beta-1+mn-\sum x_i} \tag{1}$$

We see that this has the form of another beta distribution, this time with parameters $\beta + \sum x_i$ and $\beta + mn - \sum x_i$. $\tag{1}$

Note that we have here another conjugate prior distribution. The beta distribution is the conjugate prior in this particular case.

(i)(b) *Maximum likelihood estimate*

We have already worked out the likelihood function based on the sample data:

$$L(\theta) = C\theta^{\sum x_i}(1-\theta)^{mn-\sum x_i}$$

Taking logs:

$$\log L = K + \sum x_i \log\theta + (mn - \sum x_i)\log(1-\theta) \tag{½}$$

Differentiating with respect to θ:

$$\frac{d}{d\theta}\log L = \frac{\sum x_i}{\theta} - \frac{mn - \sum x_i}{1-\theta} \tag{1}$$

Setting this expression equal to zero and rearranging, we get the MLE for θ:

$$\hat{\theta} = \frac{\sum x_i}{mn}$$

[1]

We can check that this is a maximum by differentiating the log likelihood a second time:

$$\frac{d^2}{d\theta^2} \log L = -\frac{\sum x_i}{\theta^2} - \frac{mn - \sum x_i}{(1-\theta)^2}$$

[½]

Since $mn - \sum x_i$ must be positive, this expression is negative, and the estimate we have found is indeed a maximum.

(i)(c)   *Bayesian estimate*

The Bayesian estimate under quadratic loss is the mean of the posterior distribution. We know that the posterior distribution is another beta distribution with parameters $\beta + \sum x_i$ and $\beta + mn - \sum x_i$ (from part (i)(a)). So the Bayesian estimate for θ is (using the formula for the mean of the beta distribution given in the *Tables*):

$$\hat{\theta} = \frac{\beta + \sum x_i}{\beta + \sum x_i + \beta + mn - \sum x_i} = \frac{\beta + \sum x_i}{2\beta + mn}$$

[1]

We need to rewrite this as a credibility estimate:

$$Z g(\underset{\sim}{x}) + (1-Z)\mu$$

where μ is the mean of our prior distribution, so $\mu = ½$.

[1]

We can split this up as follows:

$$\frac{\beta + \sum x_i}{2\beta + mn} = \frac{\sum x_i}{2\beta + mn} + \frac{\beta}{2\beta + mn}$$

$$= \frac{mn}{2\beta + mn} \times \frac{\sum x_i}{mn} + \frac{2\beta}{2\beta + mn} \times \frac{1}{2}$$

$$= Z \times \frac{\sum x_i}{mn} + (1-Z) \times \mu$$

[1]

where $Z = \dfrac{mn}{2\beta + mn}$.

[1]

(i)(d)   *Increase in data*

As the number of data points increases, both m and n increase (assuming that by data points we mean both the number of policies in the portfolio and also the number of observations in the sample). This means that for a fixed value of β, Z will increase as the amount of data increases, tending ultimately to one. This is not surprising, since we would expect to put more emphasis on the sample data (*ie* give it greater credibility) when the amount of sample data is large.   [1]

[Total 11]

(ii)(a)   **β = 1**

In the first case we have:

$$\hat{\theta} = \frac{1+12}{2+60} = \frac{13}{62} = 0.20968$$ [1]

and: $Z = \frac{60}{2+60} = \frac{60}{62} = 0.96774$ [1]

(ii)(b)   **β = 4**

In the second case we have:

$$\hat{\theta} = \frac{4+12}{8+60} = \frac{16}{68} = 0.23529$$ [1]

and: $Z = \frac{60}{8+60} = \frac{60}{68} = 0.88235$ [1]

When $\beta = 1$, the prior variance is (using the formula for the variance of the beta distribution given in the *Tables*):

$$\frac{1}{2^2 \times 3} = 1/12 = 0.08333$$

When $\beta = 4$ the prior variance is:

$$\frac{16}{8^2 \times 9} = 0.02778$$

So when β is larger we have a smaller prior variance. This corresponds to the situation where we are more certain about the value of θ. Both prior distributions have a mean of ½, so we think that the value of θ might be somewhere near ½. However, if we choose a larger value for β we are saying that we are more certain that the value of θ is close to ½.   [1]

This means that we want to put more emphasis on the prior distribution in our analysis, and less emphasis on the data available from the sample (*ie* the MLE). This means that we need a smaller value for Z, and this is in fact the case. Z is smaller when $\beta = 4$ than when $\beta = 1$.   [1]

[Total 6]

Solution X4.11

This question is based on Chapter 12 on GLMs.

(i)(a)   *Log-likelihood*

The PDF is:

$$f(y_i) = \frac{1}{\mu_i} e^{-\frac{1}{\mu_i} y_i} \qquad y_i > 0$$

which can be written as:

$$f(y_i) = \exp\left\{ -\ln\mu_i - \frac{1}{\mu_i} y_i \right\}$$

Note that by putting all the terms inside the exponential, the maths is easier later on when we take natural logs.

So the likelihood is:

$$L(\mu_i) = \prod_{i=1}^{15} f(y_i) = \exp\sum_{i=1}^{15}\left\{ -\ln\mu_i - \frac{1}{\mu_i} y_i \right\}$$

Hence the log-likelihood is:

$$\ln L(\mu_i) = \sum_{i=1}^{15}\left\{ -\ln\mu_i - \frac{1}{\mu_i} y_i \right\} \qquad\qquad \text{eqn (1)} \qquad [½]$$

Now substituting for the $\dfrac{1}{\mu_i}$ we get:

$$\ln L(\alpha,\beta) = \sum_{i=1}^{10}\left\{ \ln\alpha - \alpha y_i \right\} + \sum_{i=11}^{15}\left\{ \ln(\alpha+\beta) - (\alpha+\beta)y_i \right\} \qquad \text{eqn (2)} \qquad [½]$$

(i)(b)   *MLEs*

Differentiating with respect to α:

$$\frac{\partial}{\partial\alpha}\ln L(\alpha,\beta) = \sum_{i=1}^{10}\left\{ \frac{1}{\alpha} - y_i \right\} + \sum_{i=11}^{15}\left\{ \frac{1}{\alpha+\beta} - y_i \right\} \; = \; \frac{10}{\alpha} + \frac{5}{\alpha+\beta} - \sum_{i=1}^{15} y_i \qquad [1]$$

Differentiating with respect to β :

$$\frac{\partial}{\partial \beta} \ln L(\alpha, \beta) = \sum_{i=11}^{15} \left\{ \frac{1}{\alpha + \beta} - y_i \right\} = \frac{5}{\alpha + \beta} - \sum_{i=11}^{15} y_i \qquad [1]$$

Setting these two derivatives equal to zero, we obtain the equations:

$$\frac{10}{\hat{\alpha}} + \frac{5}{\hat{\alpha} + \hat{\beta}} - \sum_{i=1}^{15} y_i = 0 \qquad \text{eqn (3)}$$

$$\frac{5}{\hat{\alpha} + \hat{\beta}} - \sum_{i=11}^{15} y_i = 0 \qquad \text{eqn (4)}$$

From (4), we have $\dfrac{5}{\hat{\alpha} + \hat{\beta}} = \displaystyle\sum_{i=11}^{15} y_i$. Substituting this into (3) gives:

$$\frac{10}{\hat{\alpha}} - \sum_{i=1}^{10} y_i = 0 \quad \Rightarrow \quad \hat{\alpha} = \frac{10}{\displaystyle\sum_{i=1}^{10} y_i} \qquad [1]$$

We can rearrange (4) to get:

$$\hat{\alpha} + \hat{\beta} = \frac{5}{\displaystyle\sum_{i=11}^{15} y_i}$$

Now using $\hat{\alpha} = \dfrac{10}{\displaystyle\sum_{i=1}^{10} y_i}$ we find that:

$$\hat{\beta} = \frac{5}{\displaystyle\sum_{i=11}^{15} y_i} - \frac{10}{\displaystyle\sum_{i=1}^{10} y_i} \qquad [1]$$

Markers please note that to show these are MLEs, we should really do some second order differentiation. However, this is complicated in a two-parameter case and would not be expected in the exam.

(i)(c)   *Scaled deviance*

The scaled deviance is given by $2\{\ln L_S - \ln L_M\}$ where $\ln L_S$ is the log-likelihood of the saturated model and $\ln L_M$ is the log-likelihood of the current model.

In the saturated model the expected values, μ_i, are equal to the actual observed values, y_i. Hence replacing μ_i's with y_i's in equation (1) we get the log-likelihood for the saturated model:

$$\ln L_S = \sum_{i=1}^{15}\left\{-\ln y_i - \frac{1}{y_i}y_i\right\} = \sum_{i=1}^{15}\left\{-\ln y_i - 1\right\} \tag{[1]}$$

The log-likelihood of our current model is equation (2) with our estimates for α and β:

$$\ln L_M = \sum_{i=1}^{10}\left\{\ln\hat{\alpha} - \hat{\alpha}y_i\right\} + \sum_{i=11}^{15}\left\{\ln(\hat{\alpha}+\hat{\beta}) - (\hat{\alpha}+\hat{\beta})y_i\right\} \tag{[½]}$$

Hence the scaled deviance is:

$$2\left\{\sum_{i=1}^{15}\left(-\ln y_i - 1\right) - \sum_{i=1}^{10}\left(\ln\hat{\alpha} - \hat{\alpha}y_i\right) - \sum_{i=11}^{15}\left(\ln(\hat{\alpha}+\hat{\beta}) - (\hat{\alpha}+\hat{\beta})y_i\right)\right\}$$

$$= 2\left\{\sum_{i=1}^{10}\left(-\ln y_i - 1 - \ln\hat{\alpha} + \hat{\alpha}y_i\right) + \sum_{i=11}^{15}\left(-\ln y_i - 1 - \ln(\hat{\alpha}+\hat{\beta}) + (\hat{\alpha}+\hat{\beta})y_i\right)\right\} \tag{[½]}$$

Using the given notation we see that $\hat{\alpha} = \frac{1}{\overline{y}_1}$ and $\hat{\beta} = \frac{1}{\overline{y}_2} - \frac{1}{\overline{y}_1}$. Hence $\hat{\alpha} + \hat{\beta} = \frac{1}{\overline{y}_2}$. Substituting these gives:

$$= 2\left\{\sum_{i=1}^{10}\left(-\ln y_i - 1 + \ln\overline{y}_1 + \frac{y_i}{\overline{y}_1}\right) + \sum_{i=11}^{15}\left(-\ln y_i - 1 + \ln\overline{y}_2 + \frac{y_i}{\overline{y}_2}\right)\right\} \tag{[1]}$$

as required.   [Total 8]

(ii)(a)   *Deviance residual*

The deviance residual for y_1 is given by:

$$\text{sign}(y_1 - \hat{\mu}_1)d_1 \tag{[½]}$$

where $\sum d_i^2$ gives the scaled deviance.

We have $\hat{\mu}_1 = \frac{1}{\hat{\alpha}} = \overline{y}_1 = 430$.   [½]

So $\text{sign}(y_1 - \hat{\mu}_1) = \text{sign}(-5)$, which is negative.   [½]

From Part (i)(c) we have:

$$d_1^2 = 2\left\{-\ln y_1 - 1 + \ln \overline{y}_1 + \frac{y_1}{\overline{y}_1}\right\} = 2\left\{-\ln 425 - 1 + \ln 430 + \frac{425}{430}\right\} = 0.000136 \qquad [\frac{1}{2}]$$

Hence, the deviance residual is:

$$-\sqrt{0.000136} = -0.0117 \qquad [\frac{1}{2}]$$

(ii)(b)   *Pearson residual*

The Pearson residual for y_1 is given by:

$$\frac{y_1 - \hat{\mu}_1}{\sqrt{\text{var}(\hat{\mu}_1)}} \qquad [\frac{1}{2}]$$

where $\text{var}(\hat{\mu}_1)$ is the variance of Y_1 with any μ_1's in the formula replaced by their estimate, $\hat{\mu}_1$.

Since the variance of $Y_1 \sim Exp\left(\frac{1}{\mu_1}\right)$ is $\text{var}(Y_1) = \mu_1^2$, we have $\text{var}(\hat{\mu}_1) = \hat{\mu}_1^2$. $\qquad [\frac{1}{2}]$

From (ii)(a), we had $\hat{\mu}_1 = 430$, hence the Pearson residual is:

$$\frac{425 - 430}{\sqrt{430^2}} = -0.0116 \qquad [\frac{1}{2}]$$

(ii)(c)   *Which residual is appropriate?*

The distribution of Pearson residuals is skewed for non-normal data, whereas the distribution of deviance residuals is symmetrical and approximately normal. Therefore the deviance residuals are appropriate for the exponential distribution. $\qquad [1]$

A histogram of the appropriate residuals should be symmetrical and approximately normally distributed if the model is a good fit to the data. $\qquad [\frac{1}{2}]$

Also a plot of the appropriate residuals against the variables and factors (our α's and β's in this question) should be patternless. $\qquad [\frac{1}{2}]$

$$[\text{Total } 6]$$

(iii)(a)   *Significant improvement?*

Using the fact that when subtracting scaled deviances of nested models we get a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters, our test statistic is:

$$0.135 - 0.0120 = 0.123 \qquad [\frac{1}{2}]$$

Since the original model has 2 parameters (α and β) and the simplified model has 1 parameter (α), this is a realisation of a χ_1^2 random variable. $\qquad [\frac{1}{2}]$

The upper 5% critical value for a χ_1^2 is 3.841. Since $0.123 < 3.841$ the original model with β is **not** a significant improvement over the simplified model (*ie* adding the extra β does not significantly reduce the scaled deviance and hence improve the fit).   [1]

Alternatively, we could use the approximation $\Delta(\text{deviance}) > 2 \times \Delta(\text{parameters})$*, since* $0.123 \ngtr 2 \times 1$ *the model with* β *is not a significant improvement.*

(iii)(b)   *Significance of the β parameter*

Using our formula for $\hat{\beta}$ from part (i)(b) we get:

$$\hat{\beta} = \frac{5}{\sum_{i=11}^{15} y_i} - \frac{10}{\sum_{i=1}^{10} y_i} = \frac{5}{2,600} - \frac{10}{4,300} = -0.000403$$

Since $|\hat{\beta}| = 0.000403 \ngtr 2 \times s.e.(\beta) = 2 \times 0.000769$ the β parameter is **not** considered significantly different from zero, and so should not be included in the model.   [1]

This ties up with the result in Part (iii)(a) that the original model with β's is not a significant improvement over the simplified model and so the β parameter should not be included.   [1]

[Total 4]

All study material produced by ActEd is copyright and is sold for the exclusive use of the purchaser. The copyright is owned by Institute and Faculty Education Limited, a subsidiary of the Institute and Faculty of Actuaries.

Unless prior authority is granted by ActEd, you may not hire out, lend, give out, sell, store or transmit electronically or photocopy any part of the study material.

You must take care of your study material to ensure that it is not used or copied by anybody else.

Legal action will be taken if these terms are infringed. In addition, we may seek to take disciplinary action through the profession or through your employer.

These conditions remain in force after you have finished using the course.