

CT3 – P C – 14

Combined Materials Pack

ActEd Study Materials: 2014 Examinations

Subject CT3

Contents

Study Guide for the 2014 exams

Course Notes

Question and Answer Bank

Series X Assignments*

***Note:** The Series X Assignment Solutions should also be supplied with this pack unless you chose not to receive them with your study material.

If you think that any pages are missing from this pack, please contact ActEd's admin team by email at ActEd@bpp.com.

How to use the Combined Materials Pack

Guidance on how and when to use the Combined Materials Pack is set out in the *Study Guide for the 2014 exams*.

Important: Copyright Agreement

This study material is copyright and is sold for the exclusive use of the purchaser. You may not hire out, lend, give out, sell, store or transmit electronically or photocopy any part of it. You must take care of your material to ensure that it is not used or copied by anybody else. By opening this pack you agree to these conditions.

ISBN 978-1-4727-6178-1



9 781472 761781

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.

2014 Study Guide

Subject CT3

Introduction



This Study Guide contains all the information that you will need before starting to study Subject CT3 for the 2014 exams. **Please read this Study Guide carefully before reading the Course Notes**, even if you have studied for some actuarial exams before.

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** and are available separately from the Publications shop of the Institute and Faculty of Actuaries. You will also need a “permitted” scientific calculator from the list published in the Student Handbook. Please check the list carefully, since it is reviewed each year. You will find the list of permitted calculators and a link to the Publications shop on the profession’s website at www.actuaries.org.uk.

Contents

Section 1	The Subject CT3 course structure	Page 2
Section 2	ActEd study support	Page 3
Section 3	How to study to pass the exams	Page 13
Section 4	Frequently asked questions	Page 17
Section 5	Core Reading and the Syllabus	Page 19
Section 6	Syllabus	Page 22

1 The Subject CT3 course structure

There are four parts to the Subject CT3 course. The parts cover related topics and 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. This table should help you plan your progress across the study session.

Part	Chapter	Title	No. of pages	Syllabus items	4 half days	2 full days	3 full days
1	1	Summarising data	34	(i)	1	1	1
	2	Probability	37	(ii)			
	3	Random variables	35	(iii)			
	4	Probability distributions	69	(iii), (v)			
	5	Generating functions	46	(iv)			
2	6	Joint distributions	58	(vi)	2	2	2
	7	Conditional expectation	21	(xiv)			
	8	The Central Limit Theorem	19	(vii)			
	9	Sampling and statistical inference	28	(viii)			
3	10	Point estimation	56	(ix)	3	2	3
	11	Confidence intervals	48	(x)			
4	12	Hypothesis testing	65	(xi)	4	2	3
	13	Correlation and regression	62	(xii)			
	14	Analysis of variance	47	(xiii)			

2 ***ActEd study support***

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.

ActEd offers 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 ActEd produces. Note that not all products are available for all subjects.

LEARNING	LEARNING & REVISION	REVISION	REVISION & REHEARSAL	REHEARSAL
Course Notes	Q&A Bank X Assignments Combined Materials Pack (CMP) X Assignment Marking Tutorials Online Classroom	Flashcards Sound Revision MyTest	Revision Notes ASET Revision Tutorials	Mock Exam A Additional Mock Pack (AMP) Mock / AMP Marking

The products and services available for Subject CT3 are described 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 short questions to test your understanding.

Each chapter includes the relevant syllabus objectives, a chapter summary and, where appropriate, a page of important formulae or definitions.

“Learning & revision” products

Question and Answer Bank

The Question and Answer Bank provides a comprehensive bank of questions (including some past exam questions) with full solutions and comments.

The Question and Answer Bank is divided into five parts. The first four parts include a range of short and long questions to test your understanding of the corresponding part of the Course Notes. Part five consists of 100 marks of exam style questions.

X Assignments

The four Series X Assignments (X1 to X4) cover the material in Parts 1 to 4 respectively. Assignments X1 and X2 are 80-mark tests and should take you two and a half hours to complete. Assignments X3 and X4 are 100-mark tests and should take you three hours to complete. The actual Subject CT3 examination will have a total of 100 marks.

Combined Materials Pack (CMP)

The Combined Materials Pack (CMP) comprises the Course Notes, the Question and Answer Bank 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.

CMP Upgrade

The CMP Upgrade lists all significant changes to the Core Reading and ActEd material so that you can manually amend last year's study material to make it suitable for study for this year.

The upgrade includes replacement pages and additional pages where appropriate. If a large proportion of the material has changed significantly, making it inappropriate to include all changes, the upgrade will still explain what has changed and if necessary recommend that students purchase a replacement CMP or Course Notes at a significantly reduced price.

The CMP Upgrade can be downloaded free of charge from our website at www.ActEd.co.uk.

X Assignment Marking

We are happy to mark your attempts at the X assignments. Marking is not included with the Assignments or the CMP and you need to order it separately. You can submit your scripts by email, fax or post. Your script may be marked electronically, in which case you will be able to download your marked script via a secure link on the internet. Otherwise your marked script will be returned to you in the post.

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 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, it was found that students who attempt more than half the assignments have significantly higher pass rates.

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 send out full solutions with the Series X Assignments. However, if you order Series Marking at the same time as you order the Series X 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 scripts are received by ActEd 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.

If you live outside the UK you must ensure that your last script reaches the ActEd office earlier than this to allow the extra time needed to return your marked script.

Tutorials

ActEd 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.

ActEd runs a range of different tutorials at various locations, including online. Full details are set out in ActEd's *Tuition Bulletin*, which is sent regularly to all students based in the UK, Eire and South Africa and is also available from the ActEd 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** (two or three days) spread over the session.
- **A Block Tutorial** (two or three consecutive days) held two to eight weeks before the exam.

Online Classroom

The Online Classroom acts as either a valuable add-on to a face-to-face tutorial or a great alternative to a tutorial, particularly if you're not based in the UK or near a tutorial venue.

At the heart of the Online Classroom in each subject is a comprehensive, easily-searched collection of over 100 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 the ActEd website at www.ActEd.co.uk.

“Revision” products

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 an 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

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.

Choice of revision product

Different students will have preferences for different revision products.

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

Flashcards

- 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.

“Revision & rehearsal” products

Revision Notes

ActEd’s 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 six A5 booklets – perfect for revising on the train or tube to work. Each booklet covers one main theme of 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
- detailed analysis of key past exam questions, and
- other useful revision aids.

ActEd Solutions with Exam Technique (ASET)

The ActEd Solutions with Exam Technique (ASET) contains ActEd’s solutions to the previous four years’ exam papers, *ie* eight papers, plus comment and explanation. In particular it will highlight 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.

A “Mini-ASET” will also be available in the summer session covering the April Exam only.

Revision Tutorials

Revision Tutorials are intensive one-day face-to-face tutorials or half-day online tutorials in the final run-up to the exam.

They give you the opportunity to practise interpreting and answering past exam questions and to raise any outstanding queries with an ActEd tutor. These courses are most suitable if you have previously attended Regular Tutorials or a Block Tutorial in the same subject.

Details of how to apply for ActEd’s tutorials are set out in our *Tuition Bulletin*, which is available from the ActEd website at **www.ActEd.co.uk**.

“Rehearsal” products

Mock Exam A

Mock Exam A is a 100-mark mock exam paper and is available for students as a realistic test of their exam preparation. It is based on Mock Exam A from last year but it has been updated to reflect any changes to the Syllabus and Core Reading.

Additional Mock Pack (AMP)

The Additional Mock Pack (AMP) consists of two further 100-mark mock exam papers – Mock Exam B and Mock Exam C. This is ideal for students who are retaking and have already sat Mock Exam A, or for those who just want some extra question practice.

Mock / AMP Marking

We are happy to mark your attempts at Mock Exam A or the mock exams included within the AMP. The same general principles apply as for the X Assignment Marking. In particular:

- Mock Exam Marking is available for Mock Exam A and it applies to a specified subject, session and student
- Marking Vouchers can be used for Mock Exam A or the mock exams contained within the AMP; please note that attempts at the AMP can **only** be marked using Marking Vouchers.

Recall that:

- marking is not included with the products themselves and you need to order it separately
- you can submit your scripts by email, fax or post
- your script may be marked and returned to you electronically, or marked and returned by post.

Queries and feedback

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 ActEd's 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 **CT3@bpp.com** (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.

If you find an error in the course, please check the corrections page of our website (www.ActEd.co.uk/Html/paper_corrections.htm) to see if the correction has already been dealt with. Otherwise please send details via email to **CT3@bpp.com** or send a fax to **01235 550085**.

Each year ActEd 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 **CT3@bpp.com** or fax them to **01235 550085**.

The ActEd 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 **examinations@actuaries.org.uk**.

3 **How to study to pass the exams**

The CT Subject exams

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.

Subjects CT2 and CT7 are more “wordy” than the other subjects, including an “essay style” question or two in Subject CT7.

As a high level of mathematics is required in the courses it is important that your mathematical skills are extremely good. If you are a little rusty you may wish to consider buying the Foundation ActEd Course (FAC) available from ActEd. This covers all of the mathematical techniques that are required for the CT Subjects, some of which are beyond A-Level (or Higher) standard. It is a reference document to which you can refer when you need help on a particular topic.

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 CT 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 Question and Answer Bank (which includes many past exam questions), assignments, mock exams, the Revision Notes and ASET.

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 this course to “bed down” in your mind. Most successful students will probably aim to complete the course at least a month before the exam, thereby leaving a sufficient amount of time for revision. By finishing the course 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 of the course.

A sample CT subject study plan is available on our website at:

www.ActEd.co.uk/Html/help_and_advice_study_plans.htm

It includes details of useful dates, including assignment deadlines and tutorial finalisation dates.

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 Questions 1.4, 1.7 and 1.12 from the Question and Answer Bank”*, 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 8am and 10am 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 on page 1 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. Our suggested solutions are at the end of each chapter. 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.

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 self-assessment questions again a few days after reading the chapter itself.

You may like to attempt some questions from the Question and Answer Bank when you have completed a part of the course. 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 Question and Answer Bank (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 questions as you can from the Question and Answer Bank and past exam papers.

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.

4 Frequently asked questions

Q: What knowledge of earlier subjects should I have?

A: No knowledge will be assumed from other CT series courses.

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 you feel that you need to brush up on your mathematical skills before starting the course, you may find it useful to study the Foundation ActEd Course (FAC) or read an appropriate textbook. The full Syllabus for FAC, a sample of the Course Notes and an Initial Assessment to test your mathematical skills can be found on our website at www.acted.co.uk.

FAC Online Classroom

Please note that if you have purchased a CT3 CMP then you will receive complimentary access to the FAC online classroom. Like our other online classrooms, this consists of pre-recorded tuition covering the main points from the course with examples together with a dedicated forum for queries staffed by tutors. To access the online classroom please visit:

<https://learn bpp.com>

You should have received an email with your access details. If you have lost this then enter your username (which is your email address used by ActEd) and click the “Forgotten your password?” to have a new password emailed to you.

Should you have any problems with accessing the online classroom then please do email our admin team at ActEd@bpp.com.

Q: *What should I do if I've never studied statistics before?*

A: Whilst the Core Reading *does* cover all that you need to know, it is our experience that students who have not studied any statistics before can find the pace too fast to gain a sufficient grasp. To prevent such students being disadvantaged, we have developed the *Stats Pack*, which covers the basics at a much slower pace with plenty of examples. Sample pages can be found on our website at www.acted.co.uk.

The Stats Pack is also available with an online classroom.

Q: *What calculators am I allowed to use in the exam?*

A: Please refer to the Profession's website for the latest advice.

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/Html/paper_corrections.htm

to see if the correction has already been dealt with. Otherwise please send details via email to **CT3@bpp.com** or send a fax to **01235 550085**.

5 Core Reading and the Syllabus

Core Reading

The Syllabus for Subject CT3, and the Core Reading that supplements it, has been written by the Institute and Faculty of Actuaries to state the requirements of the examiners. 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 6 of this Study Guide. We recommend that you use the Syllabus as an important part of your study. The purpose of Core Reading is to give the examiners, tutors and students a clear, shared understanding of the depth and breadth of treatment required by the Syllabus. In examinations students are expected to demonstrate their understanding of the concepts in Core Reading. Examiners have the Core Reading available when setting papers.

Core Reading deals with each syllabus objective. Core Reading covers what is needed to pass the exam but the tuition material that has been written by ActEd enhances it by giving examples and further explanation of key points. The Subject CT3 Course Notes include the Core Reading in full, integrated throughout the course. 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 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

Core Reading accreditation

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

The following book has been used as the basis for several units:

An introduction to the mathematics of finance. McCutcheon, J. J.; Scott, W. F. Heinemann, 1986. ISBN: 043491228X, by permission of the authors who are the holders of copyright of the book. All rights reserved.

Changes to the Syllabus and Core Reading

The Syllabus and Core Reading are updated as at 31 May each year. The exams in April and September 2014 will be based on the Syllabus and Core Reading as at 31 May 2013.

We recommend that you always use the up-to-date Core Reading to prepare for the exams.

The Institute and Faculty of Actuaries' Copyright

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 Institute and Faculty of Actuaries or through your employer.

These conditions remain in force after you have finished using the course.

Past exam papers

You can download some past exam papers and Examiners' Reports from the profession's website at www.actuaries.org.uk.

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.

However, some students may find it useful to obtain a different viewpoint on a particular topic covered in Subject CT3. The following list of further reading for Subject CT3 has been prepared by the Institute and Faculty of Actuaries. This list is not exhaustive and other useful material may be available.

Mathematical statistics. Freund, John E - 7th ed. - Prentice Hall International, 2004. 614 pages. ISBN: 978-0131246461.

Available from the Publications Unit.

Calculators

Please refer to **www.actuaries.org.uk** for the latest advice on which calculators are permitted in the exams.

6 Syllabus

The full Syllabus for Subject CT3 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 Probability and Mathematical Statistics subject is to provide a grounding in the aspects of statistics and in particular statistical modelling that are of relevance to actuarial work.

Links to other subjects

Subjects CT4 — Models and CT6 — Statistical Methods: use the statistical concepts and models covered in this subject. These are then developed further in other subjects in particular Subject ST1 — Health and Care Specialist Technical, Subject ST7 — General Insurance — Reserving and Capital Modelling Specialist Technical and Subject ST8 — General Insurance — Pricing Specialist Technical.

Objectives

On completion of the subject the trainee actuary will be able to:

- (i) Summarise the main features of a data set (exploratory data analysis).
(Chapter 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.
 2. Describe the level/location of a set of data using the mean, median, mode, as appropriate.
 3. Describe the spread/variability of a set of data using the standard deviation, range, interquartile range, as appropriate.
 4. Explain what is meant by symmetry and skewness for the distribution of a set of data.

- (ii) Explain the concepts of probability. (Chapter 2)
1. Explain what is meant by a set function, a sample space for an experiment, and an event.
 2. Define probability as a set function on a collection of events, stating basic axioms.
 3. Derive basic properties satisfied by the probability of occurrence of an event, and calculate probabilities of events in simple situations.
 4. Derive the addition rule for the probability of the union of two events, and use the rule to calculate probabilities.
 5. Define the conditional probability of one event given the occurrence of another event, and calculate such probabilities.
 6. Derive Bayes' Theorem for events, and use the result to calculate probabilities.
 7. Define independence for two events, and calculate probabilities in situations involving independence.
- (iii) Explain the concepts of random variable, probability distribution, distribution function, expected value, variance and higher moments, and calculate expected values and probabilities associated with the distributions of random variables. (Chapters 3 and 4)
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.
 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. 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.
 4. Evaluate probabilities (by calculation or by referring to tables as appropriate) associated with distributions.

5. Derive the distribution of a function of a random variable from the distribution of the random variable.
- (iv) Define a probability generating function, a moment generating function, a cumulant generating function and cumulants, derive them in simple cases, and use them to evaluate moments. (Chapter 5)
1. Define and determine the probability generating function of discrete, integer-valued random variables.
 2. Define and determine the moment generating function of random variables.
 3. Define the cumulant generating function and the cumulants, and determine them for random variables.
 4. Use generating functions to determine the moments and cumulants of random variables, by expansion as a series or by differentiation, as appropriate.
 5. Identify the applications for which a probability generating function, a moment generating function, a cumulant generating function and cumulants are used, and the reasons why they are used.
- (v) Define basic discrete and continuous distributions, be able to apply them and simulate them in simple cases. (Chapter 4)
1. Define and be familiar with the discrete distributions: geometric, binomial, negative binomial, hypergeometric, Poisson and uniform on a finite set.
 2. Define and be familiar with the continuous distributions: normal, lognormal, exponential, gamma, chi-square, t , F , beta and uniform on an interval.
 3. Define a Poisson process and note the connection between Poisson processes and the Poisson distribution, and that a Poisson process may be equivalently characterised as: (1) the distribution of waiting times between events, (2) the distribution of process increments and (3) the behaviour of the process over an infinitesimal time interval.
 4. Generate basic discrete and continuous random variables using simulation methods.

- (vi) Explain the concepts of independence, jointly distributed random variables and conditional distributions, and use generating functions to establish the distribution of linear combinations of independent random variables. (Chapter 6)
1. Explain what is meant by jointly distributed random variables, marginal distributions and conditional distributions.
 2. Define the probability function/density function of a marginal distribution and of a conditional distribution.
 3. Specify the conditions under which random variables are independent.
 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.
 5. Define the probability function/density function of the sum of two independent random variables as the convolution of two functions.
 6. Derive the mean and variance of linear combinations of random variables.
 7. Use generating functions to establish the distribution of linear combinations of independent random variables.
- (vii) State the central limit theorem, and apply it. (Chapter 8)
1. State the central limit theorem for a sequence of independent, identically distributed random variables.
 2. Apply the central limit theorem to establish normal approximations to other distributions, and to calculate probabilities.
 3. Explain and apply a continuity correction when using a normal approximation to a discrete distribution.
- (viii) Explain the concepts of random sampling, statistical inference and sampling distribution, and state and use basic sampling distributions. (Chapter 9)
1. Explain what is meant by a sample, a population and statistical inference.
 2. Define a random sample from a distribution of a random variable.
 3. Explain what is meant by a statistic and its sampling distribution.

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.
 5. State and use the basic sampling distributions for the sample mean and the sample variance for random samples from a normal distribution.
 6. State and use the distribution of the t -statistic for random samples from a normal distribution.
 7. State and use the F distribution for the ratio of two sample variances from independent samples taken from normal distributions.
- (ix) Describe the main methods of estimation and the main properties of estimators, and apply them. (Chapter 10)
1. Describe the method of moments for constructing estimators of population parameters and apply it.
 2. Describe the method of maximum likelihood for constructing estimators of population parameters and apply it.
 3. Define the terms: efficiency, bias, consistency and mean squared error.
 4. Define the property of unbiasedness of an estimator and use it.
 5. Define the mean square error of an estimator, and use it to compare estimators.
 6. Describe the asymptotic distribution of maximum likelihood estimators and use it.
- (x) Construct confidence intervals for unknown parameters. (Chapter 11)
1. Define in general terms a confidence interval for an unknown parameter of a distribution based on a random sample.
 2. Derive a confidence interval for an unknown parameter using a given sampling distribution.
 3. Calculate confidence intervals for the mean and the variance of a normal distribution.

4. Calculate confidence intervals for a binomial probability and a Poisson mean, including the use of the normal approximation in both cases.
 5. Calculate confidence intervals for two-sample situations involving the normal distribution, and the binomial and Poisson distributions using the normal approximation.
 6. Calculate confidence intervals for a difference between two means from paired data.
- (xi) Test hypotheses. (Chapter 12)
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.
 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. Use a χ^2 test to test the hypothesis that a random sample is from a particular distribution, including cases where parameters are unknown.
 4. Explain what is meant by a contingency (or two-way) table, and use a χ^2 test to test the independence of two classification criteria.
- (xii) Investigate linear relationships between variables using correlation analysis and regression analysis. (Chapter 13)
1. Draw scatterplots for bivariate data and comment on them.
 2. Define and calculate the correlation coefficient for bivariate data, explain its interpretation and perform statistical inference as appropriate.
 3. Explain what is meant by response and explanatory variables.
 4. State the usual simple regression model (with a single explanatory variable).
 5. Derive and calculate the least squares estimates of the slope and intercept parameters in a simple linear regression model.

6. Perform statistical inference on the slope parameter in simple linear regression.
 7. Calculate R^2 (coefficient of determination) and describe its use to measure the goodness of fit of a linear regression model.
 8. Use a fitted linear relationship to predict a mean response or an individual response with confidence limits.
 9. Use residuals to check the suitability and validity of a linear regression model.
 10. State the usual multiple linear regression model (with several explanatory variables).
- (xiii) Explain the concepts of analysis of variance and use them. (Chapter 14)
1. Describe the circumstances in which a one-way analysis of variance can be used.
 2. State the usual model for a one-way analysis of variance and explain what is meant by the term treatment effects.
 3. Perform a simple one-way analysis of variance.
- (xiv) Explain the concepts of conditional expectation and compound distribution, and apply them. (Chapter 7)
1. Define the conditional expectation of one random variable given the value of another random variable, and calculate such a quantity.
 2. Show how the mean and variance of a random variable can be obtained from expected values of conditional expected values, and apply this.
 3. Derive the moment generating function of the sum of a random number of independent, identically distributed random variables (a compound distribution), and use the result to calculate the mean and variance of such a distribution.

CT3 Index

$\text{corr}(X, Y)$	Ch6	p22-23
$\text{cov}(X, Y)$	Ch6	p20-22
$C_X(t)$	Ch5	p20-23
$E(X)$	Ch3	p11-13
$E(Y X)$	Ch7	p3-6
		Ch13 p3
$G_X(t)$	Ch5	p3-9, 14, 23
IQR	Ch1	p16-18
$M_X(t)$	Ch5	p10-19, 23
$o(h)$	Ch4	p29
$P(A B)$	Ch2	p16-17, 22
Q_1, Q_3	Ch1	p16-17
s^2	Ch1	p14
S^2	Ch9	p6, 8-9
$\text{skew}(X)$	Ch3	p18
$\text{var}(X)$	Ch3	p14-15
\bar{x}	Ch1	p10
\bar{X}	Ch8 Ch9	p2 p5, 8
Z	Ch4 Ch9	p21-23 p8
$\Gamma(\alpha)$	Ch4	p14
χ^2	Ch4 Ch9	p16-18 p8
μ	Ch3	p11-13
μ_3	Ch3	p18
ρ	Ch6	p22-23
σ^2	Ch3	p14-15
τ_i	Ch14	p6
\emptyset	Ch2	p4
\cap	Ch2	p4
\cup	Ch2	p4
\subset	Ch2	p3
\in	Ch2	p3
A', \bar{A}	Ch2	p4

Analysis of variance (ANOVA).....	Ch14	p3-15
Least significant difference analysis.....	Ch14	p22-25
Approximations		
Distribution of \bar{X}	Ch8 Ch9	p2 p8
Distribution of MLE, $\hat{\theta}$	Ch10	p26
Normal approx. of binomial.....	Ch8	p4-5
Normal approx. of chi-square	Ch8	p6
Normal approx. of gamma.....	Ch8	p6
Normal approx. of Poisson	Ch8	p5-6
Poisson approx. of binomial	Ch4	p11
Asymptotic		
Distribution of MLE	Ch10	p26
Distribution of \bar{X}	Ch8 Ch9	p2 p8
Bar Chart.....	Ch1	p4
Bayes' Theorem.....	Ch2	p22
Beta		
Distribution.....	Ch4	p19
Function	Ch4	p19
Bias	Ch10	p21-22
Box and whisker diagram	Ch1	p23
Boxplot	Ch1	p23
Central Limit Theorem	Ch8 Ch9	p2 p8
Chi-square.....	Ch4 Ch9	p16-18 p8
Coefficient		
Correlation		
sample, r	Ch13	p8
population, ρ	Ch5	p22-23
Determination	Ch13	p19
Skewness		
sample.....	Ch1	p22
population	Ch3	p18
Compound distributions.....	Ch7	p9-12
Conditional		
Expectation	Ch7	p3-6
Probabilities	Ch2	p16-17, 22
Probability function	Ch6	p9-10
Probability density function.....	Ch6	p11
Variance.....	Ch6	p7
Confidence intervals		
Mean		
μ	Ch11	p9
paired data mean difference, μ_D	Ch11	p26-27
treatment, μ_i	Ch14	p19
difference between 2 means, $\mu_1 - \mu_2$	Ch11	p21-22
difference between 2 treatment means.....	Ch14	p20-21

Confidence intervals (<i>ctd</i>)			
MLEs, $\hat{\theta}$	Ch10	p26	
Poisson parameter			
λ	Ch11	p18-20	
difference between 2 parameters, $\lambda_1 - \lambda_2$	Ch11	p25	
Proportion			
p	Ch11	p13-16	
difference between 2 proportions, $p_1 - p_2$	Ch11	p24	
Slope parameter, β	Ch13	p22-23	
Response			
individual	Ch13	p25-26	
mean	Ch13	p24, 26	
Variance			
σ^2	Ch11	p11	
linear regression residual, σ^2	Ch13	p21	
ANOVA residual, σ^2	Ch14	p13	
ratio of 2 variances, σ_1^2/σ_2^2	Ch11	p23	
Contingency table	Ch12	p33-35	
Continuity correction	Ch8	p7-9	
	Ch12	p13, 15	
Continuous			
Data	Ch1	p2	
Distributions	Ch4	p13-26	
Random variables	Ch3	p4-7	
Convolutions	Ch6	p25-26	
	Ch7	p9	
Correlation	Ch6	p22-23	
	Ch13	p8-12	
Covariance	Ch6	p20-22	
Cramér-Rao Lower Bound (CRLB)	Ch10	p26	
Cumulant			
Generating Function (CGF)	Ch5	p20-22	
r th, κ_r	Ch5	p22	
Cumulative distribution function (CDF)			
for discrete random variables	Ch3	p6-7	
for continuous random variables	Ch3	p9-10	
Cumulative frequency			
Diagram	Ch1	p9	
Table	Ch1	p9	
Data			
Categorical			
attribute	Ch1	p3	
nominal	Ch1	p3	
ordinal	Ch1	p3	
Numerical			
discrete	Ch1	p2	
continuous	Ch1	p2	
Sample	Ch1	p2	

Discrete

Data.....	Ch1	p2
Distributions	Ch4	p3-12
Random variables	Ch3	p4-7
Distribution function (DF)		
for discrete random variables.....	Ch3	p6-7
for continuous random variables.....	Ch3	p9-10

Distributions

Discrete

Bernoulli	Ch4	p4
	Ch6	p30
Binomial	Ch4	p5-6
	Ch5	p6, 15
	Ch6	p30
Geometric	Ch4	p6-7
	Ch6	p31
Hypergeometric	Ch4	p9
Negative binomial.....	Ch4	p8-9
	Ch5	p7, 15
	Ch6	p31
Poisson.....	Ch4	p10-12
	Ch5	p8, 15
	Ch6	p32
Uniform	Ch4	p3
	Ch5	p6

Distributions

Continuous

Beta.....	Ch4	p19
Chi-square.....	Ch4	p16-18
	Ch5	p17
	Ch6	p35
	Ch9	p8
Exponential	Ch4	p15-16
	Ch5	p17
	Ch6	p34
<i>F</i> -distribution	Ch4	p26
	Ch9	p14
Gamma	Ch4	p14-18
	Ch5	p16
	Ch6	p34
Lognormal	Ch4	p24
Normal.....	Ch4	p20-23
	Ch5	p18-19
	Ch6	p35-36
	Ch9	p8, 10
<i>t</i> -distribution	Ch4	p25
	Ch9	p11-12
Uniform	Ch4	p13
Dotplot	Ch1	p8
Efficiency.....	Ch10	p23
Element	Ch2	p3
Errors	Ch12	p5

Estimate	Ch10	p4
Estimation		
Method of moments	Ch10	p3-8, 31
Method of maximum likelihood (MLE)	Ch10	p9-20, 31
Method of least squares	Ch13	p14
	Ch14	p8
Estimator	Ch10	p4
Distribution of.....	Ch10	p26
Properties of.....	Ch10	p21-25
Event	Ch2	p3
	Ch3	p4
Mutually Exclusive	Ch2	p4, 11
Independent.....	Ch2	p18
Expectation		
of a distribution, $E(X)$	Ch3	p11-13
using GFs to find.....	Ch5	p9, 11, 21
of compound distributions, $E(S)$	Ch7	p10
of a function of a distribution, $E[g(X)]$	Ch3	p12
of a function of two random variables $E[g(X,Y)]$	Ch6	p15-17
of a linear function, $E(aX + b)$	Ch3	p15-16
of linear combination, $E[c_1X_1 + \dots + c_nX_n]$	Ch6	p27
of a sum, $E[ag(X) + bh(Y)]$	Ch6	p18-19
of a product, $E[g(X)h(Y)]$	Ch6	p18-19
Conditional.....	Ch7	p3-6
Fisher's transformation	Ch13	p11
Frequency distribution	Ch1	p4
Grouped	Ch1	p5
Gamma		
Distribution	Ch4	p14-18
Function	Ch4	p14
Geometric series.....	Ch5	p26
Generating functions		
Cumulant (CGF)	Ch5	p20-23
Moment (MGF).....	Ch5	p10-19, 23
	Ch6	p33-36
	Ch7	p11
Probability (PGF).....	Ch5	p3-9, 14, 23
	Ch6	p29-32
Histogram.....	Ch1	p6
Hypothesis		
Alternative, H_1	Ch12	p3
Composite	Ch12	p3
Null, H_0	Ch12	p3
Simple	Ch12	p3
Test		see Tests

Independent		
Events	Ch2	p18
Functions of random variables.....	Ch6	p13
Random variables	Ch6	p11-14, 22
Interpolation.....	Ch4	p21
Stats Pack		
Interquartile range.....	Ch1	p16
Ch3		p18
Invariance property of MLEs.....	Ch10	p11
Least significant difference analysis of treatment means.....	Ch14	p21-24
Least squares estimation		
ANOVA model.....	Ch14	p8
Linear regression model.....	Ch13	p14
Level of a test.....	Ch12	p5
Likelihood.....	Ch10	p9
Method of maximum (MLE)	Ch10	p9-20, 31
Ratio test.....	Ch12	p6-7
Linear regression analysis		
Full model.....	Ch13	p21
Multiple model.....	Ch13	p30
Simple model.....	Ch13	p13-16
Lineplot	Ch1	p8
Location, measures of	Ch1	p10-12
Maximum likelihood, method of	Ch10	p9-20, 31
Mean		
Sample, \bar{x}	Ch1	p10
Ch8		p2
Ch9		p5, 8
Population, μ	Ch3	p11
see also expectation		
Mean square error (MSE)	Ch10	p23-25
Median		
Sample	Ch1	p11-12, 23
Population	Ch3	p19
Ch4		p15
Memoryless property	Ch4	p7, 16
Mode		
Sample	Ch1	p12
Population	Q&A	1.9
X		1.4
Moments (<i>see also mean, expectation, variance and skewness</i>)		
Generating function (MGF)	Ch5	p10-19, 23
Method of	Ch10	p3-8, 31
Population	Ch3	p17-18
using GFs to find	Ch5	p8, 11-12
of compound distributions	Ch7	p10
Sample	Ch1	p15
Neyman-Pearson Lemma	Ch12	p6

One-way analysis of variance (ANOVA)	see analysis of variance
Outliers	Ch1 p23
Partition	
Sample space.....	Ch2 p21
Variance in ANOVA model.....	Ch14 p11
Variance in linear regression model.....	Ch13 p11
Pivotal	
method of constructing confidence intervals	Ch11 p4-6
quantity	Ch11 p4
Point estimation	see estimation
Poisson	
Distribution	Ch4 p10-12
Process	Ch4 p12, 27-35
Waiting time between events	Ch4 p15, 34-35
Power	Ch12 p5
Probabilities	
Addition rule.....	Ch2 p12
Conditional.....	Ch2 p16-17, 22
Law of total.....	Ch2 p21
Multiplication rule	Ch2 p18
Probability density function (PDF)	Ch3 p8
Conditional.....	Ch6 p11
Joint (bivariate)	Ch6 p5-6
Marginal.....	Ch6 p8-9
Probability function (PF)	Ch3 p5
Conditional.....	Ch6 p9-10
Joint (bivariate)	Ch6 p3-4
Marginal.....	Ch6 p7-8
Process	
Poisson.....	Ch4 p12, 27-35
Waiting time	Ch4 p15, 34-35
p-value	Ch12 p7-8
Quartiles	
Lower, Q_1	Ch1 p16-17, 23
	Ch3 p19
Upper, Q_3	Ch1 p16-17, 23
	Ch3 p19
Random sample, X	Ch9 p3
Random variables.....	Ch3 p3
Continuous	Ch3 p8-10
Discrete	Ch3 p4-7
Distributions of	see distributions
Functions of	Ch3 p20-24
Simulation.....	Ch4 p36-40
Range	Ch1 p16
Residuals	
ANOVA model	Ch14 p17-18
Linear regression model.....	Ch13 p27-28

Sample

Correlation coefficient, r	Ch13	p8
Data.....	Ch1	p2
Mean, \bar{x}	Ch1	p10
	Ch8	p2
	Ch9	p5, 8
Median	Ch1	p11, 23
Mode.....	Ch1	p12
Moments	Ch1	p15
Random, X	Ch9	p3
Size, for confidence intervals.....	Ch11	p7-8
Space.....	Ch2	p3
	Ch3	p3
Standard deviation, s	Ch1	p13-14
Variance, s^2	Ch1	p13-14
	Ch9	p6, 8-9

Scatterplot

Sets	Ch2	p3
Complement, A' , \bar{A}	Ch2	p4
Intersection of, \cap	Ch2	p4
Mutually exclusive.....	Ch2	p4, 11
Null, \emptyset	Ch2	p4
Subset.....	Ch2	p3
	Ch3	p4
Union of, \cup	Ch2	p4

Simulation.....

Size of a test.....

Skewness

Coefficient

Sample	Ch1	p22
Population.....	Ch3	p18

Formula

Sample	Ch1	p22
Population $\mu_3 = \text{skew}(X)$	Ch3	p18
	Ch5	p21

Types

Spread, measures of

Standard deviation

Population, σ	Ch3	p14
Sample, s	Ch1	p13-14

Standard error

Statistic Ch9.....	Ch9	p5
	Ch12	p4

Stem and leaf diagrams.....

Subset	Ch2	p3
	Ch3	p4

Symmetry.....

Test			
ANOVA	Ch14	p13	
Best	Ch12	p6-7	
χ^2 goodness of fit.....	Ch12	p29	
Correlation coefficient, ρ	Ch13	p11	
Mean			
μ	Ch12	p9	
paired data mean difference, μ_D	Ch12	p25	
difference between 2 means, $\mu_1 - \mu_2$	Ch12	p17	
One-sided	Ch12	p4	
Poisson parameter			
λ	Ch12	p15	
difference between 2 parameters $\lambda_1 - \lambda_2$	Ch12	p23	
Proportion			
p	Ch12	p13	
difference between 2 proportions, $p_1 - p_2$	Ch12	p21	
Slope parameter, β	Ch13	p22-23	
Two-sided	Ch12	p4	
Variance			
σ^2	Ch12	p12	
ratio of 2 variances, σ_1^2 / σ_2^2	Ch12	p20	
Unbiasedness.....	Ch10	p21-22	
Uncorrelated.....	Ch6	p23	
Uniqueness property of GFs	Ch5	p4, 13	
Variance			
Sample, s^2	Ch1	p13-14	
	Ch9	p6, 8-9	
Pooled sample, s_p^2	Ch11	p22	
	Ch12	p17	
Population, $\sigma^2 = \text{var}(X)$	Ch3	p14-15	
	Ch5	p9, 21	
of a compound distribution	Ch7	p10	
of a linear function, $\text{var}(aX + b)$	Ch3	p15-16	
of linear combination, $\text{var}[c_1X_1 + \dots + c_nX_n]$	Ch6	p27	
of a sum, $\text{var}(X + Y)$	Ch6	p23-24	
of a difference, $\text{var}(X - Y)$	Ch6	p24	
Conditional	Ch7	p7	
Venn Diagrams	Ch2	p5-7, 13-14	
Waiting times	Ch4	p15, 34-35	

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.

Chapter 1

Summarising data



Syllabus objectives

- (i)
 - 1. *Summarise a set of data using a table or frequency distribution, and display it graphically using a line plot, a bar chart, histogram, stem and leaf plot, or other appropriate elementary device.*
 - 2. *Describe the level/location of a set of data using the mean, median, mode, as appropriate.*
 - 3. *Describe the spread/variability of a set of data using the standard deviation, range, interquartile range, as appropriate.*
 - 4. *Explain what is meant by symmetry and skewness for the distribution of a set of data.*

0 *Introduction*

This chapter deals with descriptive statistics, that is, the methodology for describing or summarising a set of data using tables, diagrams and numerical measures.

Presenting the data in a descriptive form is usually the first stage in any statistical analysis, as it allows us to spot any patterns in the data. The numerical measures mentioned are the “average” of the data (*ie* mean, median and mode) and the “spread” of the data (*ie* range, interquartile range (IQR) and variance).

If you have done any statistics before then this chapter will be very straightforward. For students who have not met this material before and would like more examples and careful explanations of where these results actually come from, the Stats Pack has been developed to help. See http://www.acted.co.uk/Html/paper_stats_pack.htm or contact StatsPack@bpp.com for further details.

1 Tabular and graphical methods

1.1 Types of data

Batch data are a set of related observations, such as the current inflation rates of EU countries.

Sample data are a set of observations selected from a population and designed to be representative of that population, such as the sums assured for a sample of 100 policies selected from a company's whole-life business.

The “sum assured” for a life insurance policy is the amount of benefit that is paid when the policyholder dies or the policy matures. “Whole-life business” refers to a particular type of policy. These terms are not important here.

The objectives of an analysis involving batch data will usually be to extract the important features by summarising the data. For an analysis involving sample data the objectives will be the same plus the main objective of making inferences about the population.

In other words, we can use sample data to determine certain properties of the underlying population from which the sample was taken. This might be useful for example, in setting up a model which could then be used to predict future behaviour, eg estimating the number of claims that will be made on a certain type of policy in a given time period.

Data involve the values of a variable and there are several types of variable:

NUMERICAL

Numerical data can be classified into two types: discrete and continuous. The distinction between discrete and continuous data is that discrete data can only take one of a set of particular values, whereas continuous data can take any value within a specified range (or the possible values are so close together that they can be considered to occupy a continuous range).

Discrete data arise from counting, eg numbers of actuaries, numbers of claims.

It is also possible to have discrete data that take negative or fractional values. To take a non-actuarial example, data obtained from measuring the spins of subatomic particles, which can take any “half-integer” value ($\dots, -1\frac{1}{2}, -1, -\frac{1}{2}, 0, +\frac{1}{2}, +1, +1\frac{1}{2}, \dots$), would also be discrete.

Continuous data arise from measuring, eg height, amount, age.

In actuarial work an “amount” often refers to an amount of money, eg an employee’s annual salary. “Age” here refers to a person’s “exact age”, not age last birthday. For example your age might be 21.85 years.

CATEGORICAL

Attribute (or dichotomous) data have only two categories, eg yes/no, male/female, claim/no claim.

Nominal data have several unordered categories, eg type of policy, nature of claim.

Ordinal data have several ordered categories, eg questionnaire responses such as “strongly in favour / ... / strongly against”.



Question 1.1

Answer the following dating agency questionnaire and state what type of data is required in each question:

1. *How old are you? (Give your age last birthday.)*
2. *How tall are you? (State as accurately as you can.)*
3. *What sex are you?*
4. *What colour are your eyes?*
5. *Do you smoke?*
6. *How would you rate your looks? (10 =Drop-dead gorgeous, 1= Seen better days)*

Most of our work in this course will relate to numerical data.

The distinction between discrete and continuous data is important because we will be using different statistical models to deal with each type.

1.2 Frequency distributions

The data from a discrete distribution can be summarised using a frequency distribution, that is, by counting the number of 0's, 1's, 2's, etc. For example, the number of children in a sample of 80 families might be summarised as follows:

Number of children under 16, x	Number of families in sample, f
0	8
1	12
2	28
3	19
4	7
5	4
6	1
7	1
8 or more	0

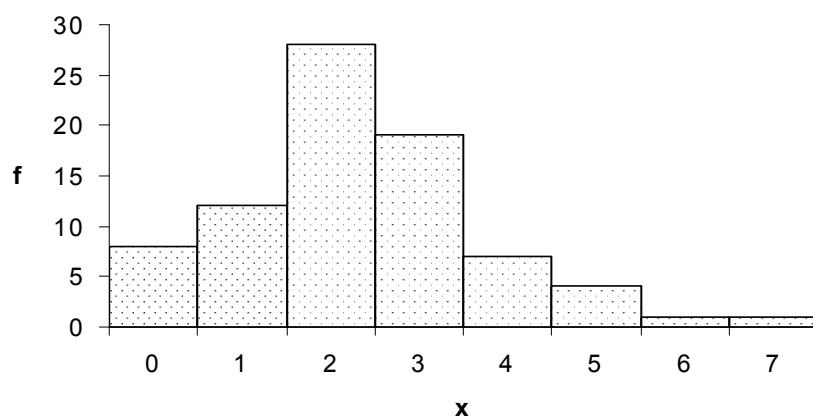


Question 1.2

How would you calculate how many children there were altogether?

To represent the data graphically, a bar chart is used.

Bar chart of number of children in families



(The above graph is part of Core Reading.)

1.3 Histograms and grouped frequency distributions

The next example involves a typical continuous variable and will introduce the idea of a *grouped frequency distribution* and a *histogram*. The data involve cash amounts given to the nearest £1. Cash amounts are actually discrete, being measured in whole numbers of pence, but here the sums are so large that they can be considered as continuous. In practice, all variables are discrete due to the fact that they will be rounded to a certain accuracy and nothing can be measured to infinitesimal accuracy.

For example, a sample of 100 claims for damage due to water leakage on an insurance company's household contents policies might be as follows:

243	306	271	396	287	399	466	269	295	330
425	324	228	113	226	176	320	230	404	487
127	74	523	164	366	343	330	436	141	388
293	464	200	392	265	403	372	259	426	262
221	355	324	374	347	261	278	113	135	291
176	342	443	239	302	483	231	292	373	346
293	236	223	371	287	400	314	468	337	308
359	352	273	267	277	184	286	214	351	270
330	238	248	419	330	319	440	427	343	414
291	299	265	318	415	372	238	323	411	494

These data might be summarised in the following grouped frequency distribution:

Group	Frequency
50–99	1
100–149	5
150–199	4
200–249	14
250–299	22
300–349	20
350–399	14
400–449	13
450–499	6
500–549	1

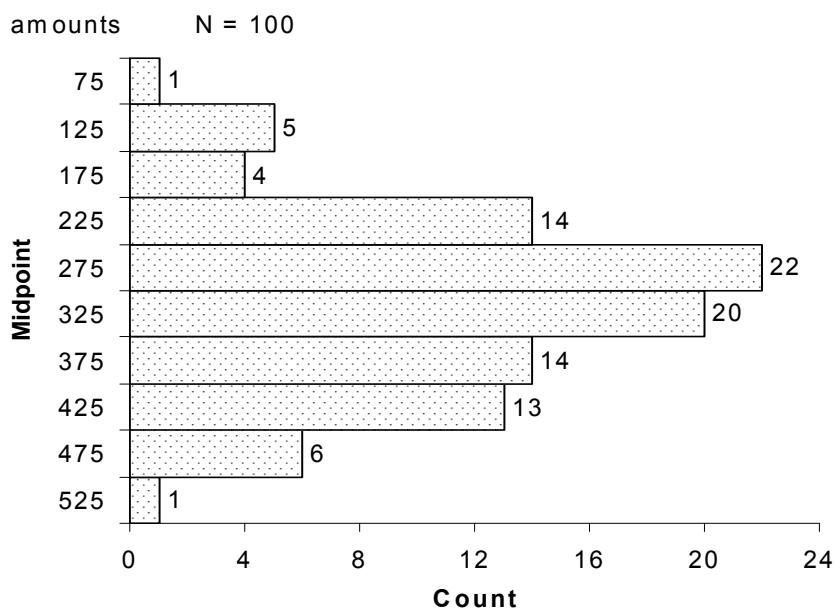
This is a grouped frequency distribution with equal group sizes.

A histogram is similar to a bar chart, but with a continuous scale.

In other words, a bar chart is called a histogram when it is used to present continuous numerical data. So the responses to any of the six questions on the dating agency questionnaire could be represented as a bar chart, but only Question 2 could be shown as a histogram.

Histograms are usually presented with vertical rectangles, but the one given below is unusual as it is in horizontal form.

In this histogram the continuous scale has been broken down into categories by dividing the data values into £50 bands. The bar labelled “125”, for example, represents the 5 data values (113, 127, 141, 113 and 135) that were in the range 100 to 149 (inclusive).



The midpoints on this diagram have been rounded to the nearest whole number for convenience. The actual values should be $\frac{49.5 + 99.5}{2} = 74.5, \frac{99.5 + 149.5}{2} = 124.5, \dots$

The above grouped frequency distribution and histogram have equal group widths. In some situations it may be convenient to have one or two wider groups at the extremes of the distributions. For such cases it should be noted that it is the areas of the rectangles that are proportional to the frequencies not their heights.

This is an important point to note and you may wonder why this is the case. Think of a histogram as a “fairer” bar chart: if a group has twice the width, it should only be half the height for a *given* frequency.

**Question 1.3**

If we had had a single group containing all values greater than or equal to 350, how would you have drawn the bar representing this group?

**Question 1.4**

Write down a formula to calculate the correct height for any bar in a histogram.

1.4 Stem and leaf diagrams

An alternative to the histogram is the stem and leaf display. It gives a visual representation similar to the histogram but does not lose the detail of the individual data points in the grouping. Here is a stem and leaf display for the water leakage data claim amounts:

0	7
1	11344
1	6888
2	012233334444
2	56667777778899999999
3	0001112222233334444
3	55556677777799
4	000001122333444
4	677899
5	2

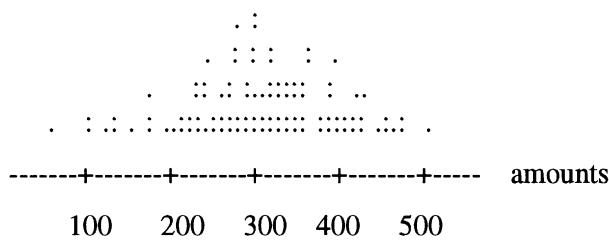
The stems are on the left with units of £100 and the leaves are on the right with units of £10. So the individual data points can be represented, although they are rounded to the nearest £10. It is usual to put a key on the diagram to show this.

These diagrams are useful to observe the general shape of the distribution of data and can be used to calculate values such as the median or interquartile range.

1.5 Lineplots

For smaller data sets another alternative diagram is the *dotplot* or *lineplot* in which the data points are plotted as “dots” or “crosses” along a line with a scale.

Here is a computer generated dotplot for the water leakage claim amount data. This computer package hasn't lined things up perfectly. The notches and numbers on the horizontal scale both need to be moved slightly to the left.



Here is a lineplot for the first row of 10 amounts from the claims data:



If there are repeats of any data points (eg two values of 300) in a lineplot, the crosses are placed above each other as in a dotplot diagram.

The dotplot/lineplot is used frequently throughout the CT3 course:

- To check the normality of a data set. This is required for some confidence intervals (Chapter 11), hypothesis tests (Chapter 12), analysis of variance (Chapter 14) and as a test for the appropriateness of the linear regression model (Chapter 13).
- To check whether two data sets have a common variance. This is required for the two-sample *t*-test (Chapter 12) and for analysis of variance (Chapter 14).

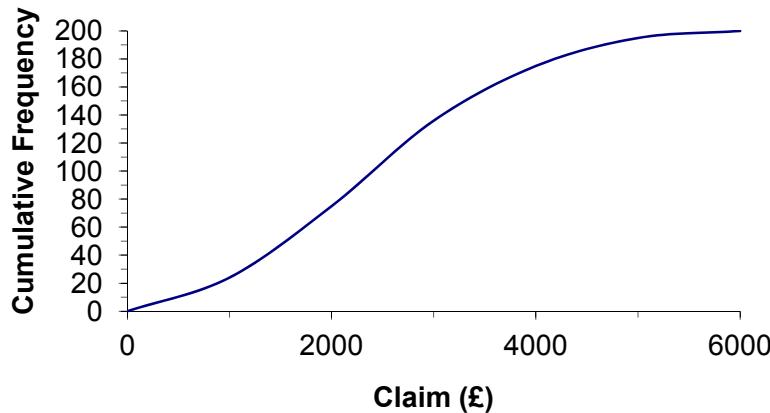
1.6 Cumulative frequency tables

Cumulative frequencies are obtained by accumulating the frequencies to give the total number of observations up to and including the value or group in question. For grouped data it is natural to relate the cumulative frequency to the upper boundaries of the groups.

The following table shows the cumulative frequencies for 200 motor insurance claims received by an office in a month.

Claim size	Cumulative frequency
up to £1,000	24
up to £2,000	75
up to £3,000	136
up to £4,000	175
up to £5,000	195
up to £6,000	200

A cumulative frequency diagram can be drawn from these data as follows:



Notice on the diagram that the cumulative frequency is plotted against the *highest* claim size in the group.

The cumulative frequency table or diagram is commonly used to find the median or interquartile range.

2 Measures of location

There are a number of different quantities, which can be used to estimate the central point of a sample. These are called measures of central tendency, or measures of location.

The quantities described in this section are just different ways of calculating the “average” value for the data set.

2.1 The sample mean

By far the most common measure for describing the location of a set of data is the mean.

For a set of observations denoted by x_1, x_2, \dots, x_n or $x_i, i = 1, 2, \dots, n$ the mean is

defined by $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (read as “x bar”).

For a frequency distribution with possible values x_1, x_2, \dots, x_k with corresponding frequencies f_1, f_2, \dots, f_k , where $\sum f_i = n$, the mean is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_i .$$

For example, for the family size distribution data the mean number of children in the sample is, from the frequency distribution:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{186}{80} = 2.325$$



Question 1.5

Confirm the above figure obtained for the mean number of children using the family size data in Section 1.2.

For grouped data the midpoint of each group would normally be used in the frequency distribution to determine the mean.



Question 1.6

Calculate \bar{x} for the water leakage data from the frequency distribution table in Section 1.3 using 74.5, 124.5, etc as midpoints for the bands.

2.2 The median

Another useful measure of location is the median. Consider placing the n observations in order of magnitude. The median is a value, which splits the data set into two equal halves, so that half the observations are less than the median and half are greater than the median.

This definition says “*a* value”, rather than “*the* value”, because there will often be a range of possible values satisfying this condition.

If n is odd, then the median is the middle observation. If n is even, then the median is the midpoint of the middle two observations. This can be conveniently expressed as the $(n + 1)/2$ th observation.

One of the potential advantages of the median for certain data sets is that it is robust or resistant to the effects of extreme observations.

Extreme observations can cause problems because they can have a disproportionate effect on the calculated value of some of the quantities we have been looking at. The word “robust” is used with this special meaning in statistics. There is a whole area of statistical theory called “robust statistics”, which deals with this issue.

For the family size distribution data, the 40th and 41st observations are both 2. So the median number of children is 2.



Question 1.7

The ages of seven policyholders in a portfolio of insurance policies are as follows:

39 34 26 41 70 34 28

- (i) Find the median age of the policyholders in this portfolio.
- (ii) Another policyholder aged 41 years is added to the portfolio. Find the median age of policyholders in the portfolio.
- (iii) Why would the mean be a poor measure of central tendency for these data?



Question 1.8

Calculate the median for the water leakage data given that, when the data values are arranged in ascending order, the 50th and 51st values are £314 and £318.

However, we would not be able to calculate exactly the median of the water leakage data from the grouped frequency distribution. This is because we would not be able to tell where the actual values are. What we do is assume that the data is uniformly spread over each class and use interpolation.

Some statisticians then argue that, under these assumptions, the median is better chosen to be at the midpoint, *ie* $\frac{1}{2}n$. Hence, the Core Reading hints at this when it says:

The median is that value corresponding to a cumulative frequency of 50% and it can be read from a plot of the cumulative frequency distribution.



Question 1.9

Estimate the median for the motor insurance claims from the cumulative frequency table in Section 1.6.

The examiners will give credit for either approach to a grouped frequency table.

2.3 *The mode*

A third measure of location is the mode. It is defined as the value which occurs with the greatest frequency or the most typical value. Its use in practice is limited but there are occasions when, for example, a company is interested in the most typical policyholder.

For the family size distribution data the largest frequency is 28, so the modal number of children in the sample is 2.

3 Measures of spread

The location of a data set is usually the main feature of interest. Another feature of interest is the spread (or variability or dispersion or scatter): how widely spread the data are about the mean (or other measure of location).

3.1 The sample standard deviation

The most commonly used measure of spread is the *standard deviation*. Essentially it is a measure of how far on average the observations are from the mean.

For a data set x_i , $i = 1, 2, \dots, n$ with mean \bar{x} , consider an individual observation x_i . Then $(x_i - \bar{x})$ is the distance of x_i from the mean, also called the deviation of x_i from the mean.

Here we are thinking of x_i and \bar{x} as points plotted on a line. In mathematical language, $x_i - \bar{x}$ (which can be a positive or a negative number) is the “displacement” and $|x_i - \bar{x}|$ is the “distance” between the points, which is always a positive number. (The vertical lines indicate the “absolute value”.)

The sum of the squares of these deviations is:

$$\sum(x_i - \bar{x})^2$$

Because each of the terms in this sum is a squared quantity (and hence a positive number) the total must be a positive number.

The variance is this total divided by $n - 1$.

You might think it would be more logical to divide by n , rather than $n - 1$. We will see later in Chapter 10 that dividing by $n - 1$ makes the sample variance an *unbiased estimator* of the population variance. That is to say, on average, the sample variance gives us the correct value for the real variance of the whole population.

You might also think it would be more logical to use $\sum|x_i - \bar{x}|$ instead of $\sum(x_i - \bar{x})^2$ in this calculation. The reason this is not done is simply because the absolute function is not a nice function to deal with mathematically – in particular, the graph of $|x|$ has a nasty kink when $x = 0$, whereas the graph of x^2 is always a nice “smooth” curve.

The standard deviation is the positive square root of the variance. The symbol s^2 is used to denote the variance:

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

The standard deviation and variance can be calculated more easily using the alternative formula:

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - n\bar{x}^2 \right]$$

For the family size distribution data the variance of the number of children for the sample of families is given by:

$$s^2 = \frac{1}{n-1} \left[\sum f_i x_i^2 - n\bar{x}^2 \right] = \frac{1}{79} \left[592 - 80 \times \left(\frac{186}{80} \right)^2 \right] = 2.02$$

And the standard deviation is $s = \sqrt{2.02} = 1.42$.

Note that where the original data has “units”, eg if the values were amounts measured in Euros, the variance (being based on a sum of squares) would be measured in “square Euros”. But the standard deviation (being the square root of the variance) would be measured in the original units – Euros in this case. For this reason, standard deviations are often more “friendly” to work with in numerical comparisons.



Question 1.10

Show that the two formulae $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$ and $s^2 = \frac{1}{n-1} \left[\sum x_i^2 - n\bar{x}^2 \right]$ are mathematically equivalent.



Question 1.11

Given that for the water leakage data:

$$n = 100 \quad \sum x = 31,353 \quad \sum x^2 = 10,687,041$$

Calculate the:

- (i) sample mean
- (ii) sample standard deviation.

3.2 Moments

The mean and variance are special cases of a set of summary measures called the **moments** of a set of data.

In general the k th-order moment about the value α is defined by:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \alpha)^k$$

Here α is any fixed number.

So the mean is the first-order moment about the origin, and the variance is essentially the second-order moment about the mean with a divisor of $n-1$ rather than n .

Again, we are thinking of points plotted along a line, so that the “origin” just refers to the zero position, ie $\alpha = 0$. When α equals the mean of the distribution, we get a *central moment*.

3.3 The range

The **range** is a very simple measure of spread defined, as its name suggests, by the difference between the largest and smallest observations in the data set.

$$R = \max_i(x_i) - \min_i(x_i)$$

For the family size distribution data, $R = 7 - 0 = 7$.

The range is a poor measure of the spread of the data as it relies on the extreme values, which aren't necessarily representative of the data as a whole.



Question 1.12

What is the range of the water leakage data?

3.4 The interquartile range

The **interquartile range (IQR)** is another measure of spread which is like the range but which is not affected by the data extremes.

First we must define the quartiles of a set of data:

Just as the median divides a set of data into two halves, the quartiles divide a set of data into four quarters. They are denoted by Q_1 , Q_2 and Q_3 .

Thus a quarter of the data have values less than Q_1 , a quarter have values between Q_1 and Q_2 , a quarter have values between Q_2 and Q_3 , and the last quarter have values greater than Q_3 . Note that Q_2 is just the median, while Q_1 is called the lower quartile and Q_3 the upper quartile.

When dealing with grouped data, the quartiles can be read from the cumulative frequency diagram or calculated using interpolation.

When dealing with discrete ungrouped data, we determine the appropriate value in a similar way to that used to calculate the median.

Recall that the median was specified as the $(n + 1)/2$ th observation. In a similar way Q_1 can be defined to be the $(n + 2)/4$ th observation counting from below and Q_3 as the same counting from above, with relevant interpolation if needed.

The lower quartile, the median and the upper quartile are also referred to as the 25th, 50th and 75th percentiles. The p th percentile corresponds to the $\left(\frac{p}{100} \times n + \frac{1}{2}\right)$ th data value.

The interquartile range is defined as $Q_3 - Q_1$.

So, for the family size distribution data, the interquartile range is calculated as follows:

Since $n = 80$, Q_1 is based on the $(80 + 2) / 4 = "20\frac{1}{2}"$ th observation.

From the frequency distribution, the 20th observation is 1 and the 21st is 2 and so $Q_1 = 1.5$. From the other end, the 20th and 21st observations are both 3 and so $Q_3 = 3$.

So the IQR = 3 – 1.5 = 1.5 .

Note: An alternative to the definitions of Q_1 and Q_3 given above is:

Q_1 is the $\frac{n+1}{4}$ th observation counting from below;

Q_3 is the $\frac{n+1}{4}$ th observation counting from above.

Using this alternative, the IQR = 3 – 1.25 = 1.75 .

Either pair of definitions is acceptable.

You may think it odd that there is more than one definition for the quartiles. Five definitions are given here <http://mathworld.wolfram.com/Quartile.html> and these don't include the first Core Reading definition nor the formulae used by Excel! There is much disagreement between statisticians over how the quartiles are "best" defined. The disagreements are over whether the median is included in each half of the data set, whether the formulae have good statistical properties and whether the formulae can be adapted to quintiles, etc. A great explanation can be found at <http://mathforum.org/library/drmath/view/60969.html>.

For our purposes either of these definitions can be used (even though they can give different answers for the IQR) and both will be given full credit in the exam.

**Question 1.13**

Calculate the interquartile range (using the first definition) for the water leakage data given that, when the data values are arranged in ascending order:

- the 25th and 26th values are 259 and 261
- the 75th and 76th values are 373 and 374.

Again, we would not be able to calculate exactly the IQR of the water leakage data from the grouped frequency distribution as we would not be able to tell where the actual values are. Again, we assume that the data is uniformly spread over each class and use interpolation and so some statisticians then argue that, under these assumptions, the quartiles are better chosen to be $\frac{1}{4}n$ and $\frac{3}{4}n$.

**Question 1.14**

Estimate the interquartile range for the motor insurance claims from the cumulative frequency table in Section 1.6.

4 Symmetry and skewness

The next feature of interest is the shape of the distribution of a data set, that is, whether it is symmetric or skewed to one side or the other.

The approximate shape of a distribution can be determined by looking at a histogram, stem and leaf display or dotplot.

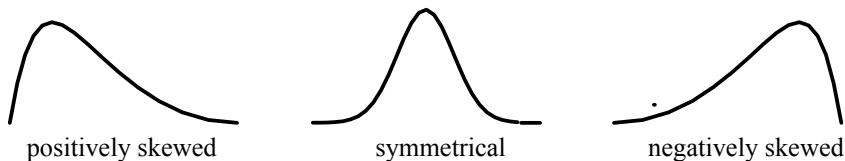
Here are histograms for three data sets each of 200 observations:

- the first is fairly symmetrical
- the second is positively skewed
- the third is negatively skewed.

(The histograms are shown on the next page.)

It should be noted that for sample data, a small data set of 10 observations is unlikely to show up the skewness of a population unless its skewness is very severe, whereas a large data set of 200 observations will reflect the shape of the population quite well.

In the more usual format, a positively skewed, a symmetrical and a negatively skewed distribution would look like this:

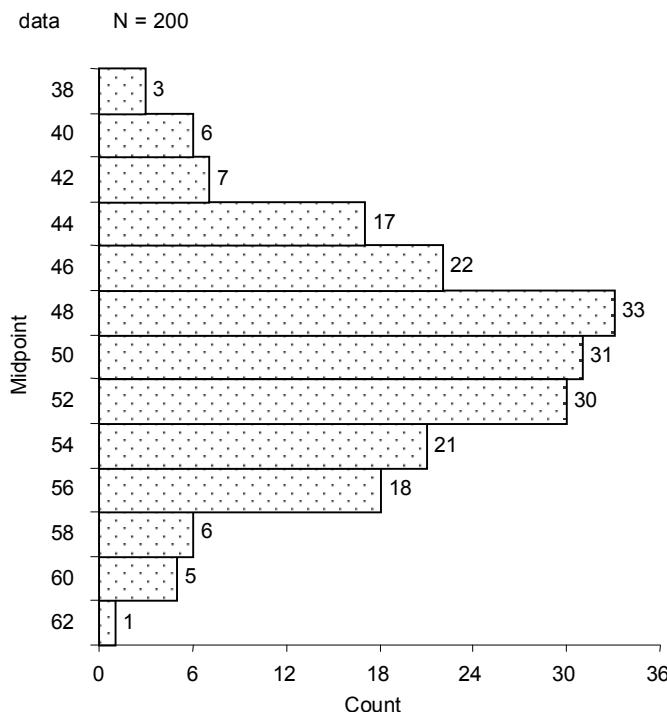


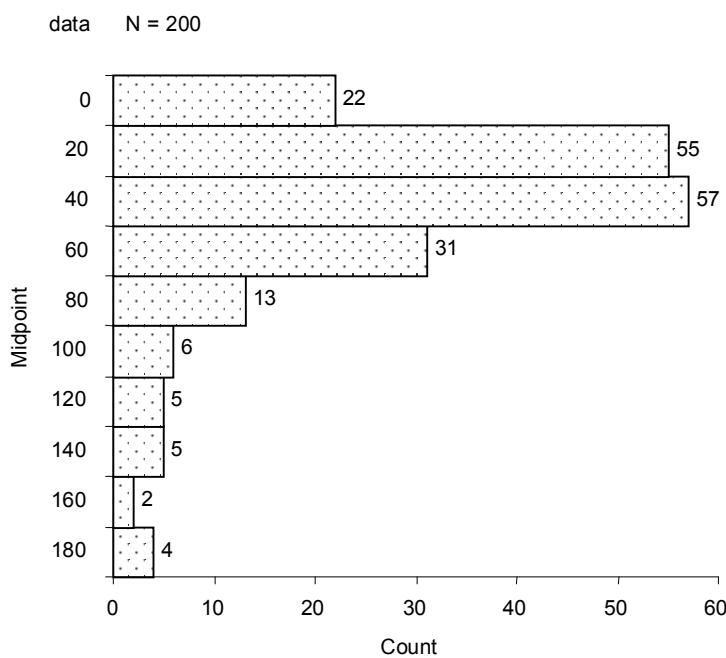
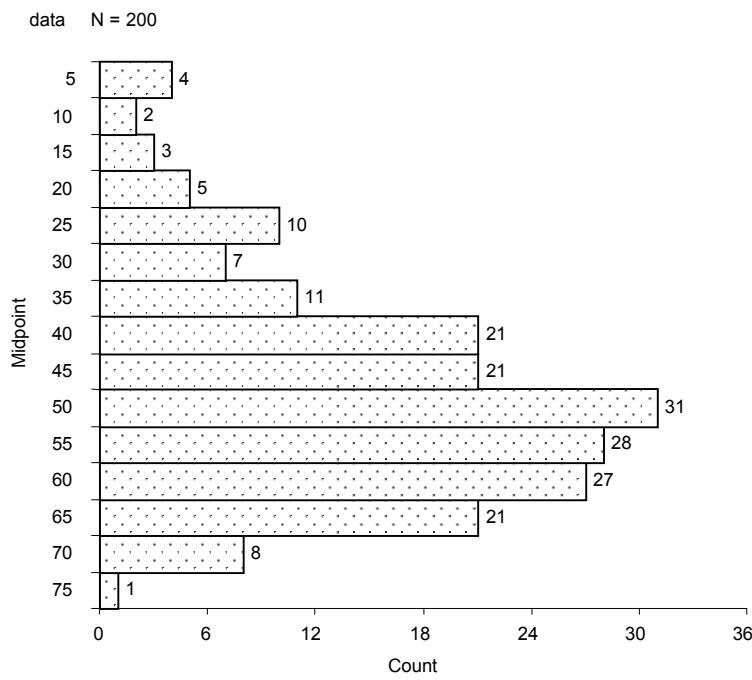
The reason for describing these as positively or negatively skewed is that the skewness can be measured numerically by calculating the third central moment. For a positively (negatively) skewed distribution, this works out to a positive (negative) answer. For a symmetrical distribution it equals zero.

You may wonder why the skewness is called positive when the ‘hump’ is on the left. This is because there are more data points to the right (on the positive side) of the ‘hump’ than the left and vice versa for negative skewness.

Positively skew distributions are most common in actuarial work because we are often dealing with quantities, such as claim amounts, which must be positive but have no upper limit. Hence, most of the probability distributions we will meet in Subjects CT3 and CT6 are positively skewed (*eg* the Poisson, exponential, gamma and lognormal distributions).

Symmetric



Positively skew**Negatively skew**

4.1 Measuring skewness for data

One particular measure of skewness is based on the third moment about the mean:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

The cubic power in this formula gives a positive or negative value depending on which side of the mean the value x_i is. Consequently, positively skewed distributions with a long tail on the right give a positive value, and negatively skewed distribution with a long tail on the left give a negative value.

The coefficient of skewness is a scaled version of this moment obtained by dividing it by the second moment about the mean raised to the power 3/2.

Recall that the second moment was almost our sample variance formula (but dividing by n instead of $n-1$). Denoting this by s_n^2 , we have:

$$\text{coeff of skew} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1.5}} = \frac{\text{skewness}}{\left(s_n^2 \right)^{1.5}} = \frac{\text{skewness}}{s_n^3}$$

We can see that we are dividing by a cubic measure. Since the skewness is a cubic measure, we obtain a dimensionless measure when we divide by another cubic measure.

Note that this does not depend on the units of the data and is such that its sign reflects the skewness with a value of zero corresponding to a symmetric set of data.



Question 1.15

Given that for the water leakage data:

$$n = 100 \quad \sum (x_i - \bar{x})^2 = 856,934.91 \quad \sum (x_i - \bar{x})^3 = -11,949,848.3946$$

Calculate the:

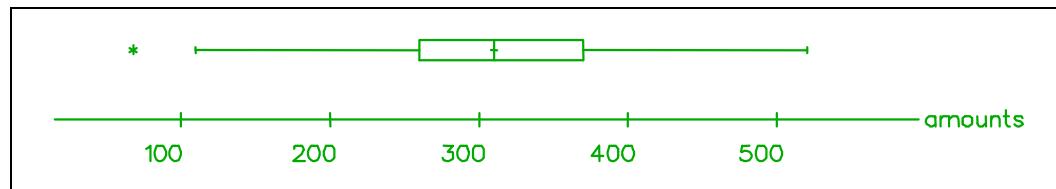
4.2 Boxplots

The quartiles together with the extremes provide a useful way of displaying a data set in summary form, called a *boxplot*. It consists of a box or rectangle with ends at Q_1 and Q_3 divided with a line at the median Q_2 . Then lines are drawn from Q_1 and Q_3 out to the extremes.

These diagrams are also called “box and whisker” diagrams (the whiskers referring to the lines protruding at each end).

Here is a boxplot of the water leakage claim amounts data. This boxplot was generated by a computer package which labels potential “outliers”, here by an asterisk. *Outliers* are observations which are in some way detached from the main bulk of the observations.

Outliers are just untrustworthy (“rogue”) values present in the data. In actuarial work outliers may sometimes be values that have been recorded incorrectly, eg a pensioner’s age given as 27 instead of 72, or an amount of money typed with an extra or missing 0 on the end. However, they may also be a genuine part of the data.



The median can be read off as a measure of location, the interquartile range (and the range) as measures of spread, and the relative symmetry can also be observed.



Question 1.16

Sketch a possible boxplot for each of the following data sets:

- (i) the annual salaries of the employees of a medium-sized company
- (ii) the scores obtained by a class of students in an easy exam

The precise numerical values are not important, but your sketches should show clearly the likely “shape” of the distributions.

5 Exam Question



Past Exam Question (Subject C1, April 1998, Q10)

In the claims department of an insurance office various quantities are computed at the end of each day's business. On Monday, 10 claims are received for a particular class of policy. The mean claim amount is calculated to be £426 and the standard deviation to be £112.

On Tuesday it is found that one of Monday's claims for £545, was classified wrongly and it is removed from the set of 10 claims. Calculate the resulting mean and standard deviation of the reduced set of 9 claims.



Chapter 1 Summary

Numerical data can be discrete or continuous.

Categorical data can be dichotomous (attribute), nominal or ordinal.

Data can be presented either in tabular form (using a frequency table, a cumulative frequency table or a stem and leaf diagram) or in graphical form (using a lineplot, a dotplot, a boxplot, a bar chart or a histogram).

The *location* of a data set can be summarised using the mean, the median or the mode.

The *spread* of a data set can be summarised using the standard deviation, the range or the interquartile range.

The variance measures the spread squared.

Third moments can be used to summarise the *skewness* (*i.e.* the degree of asymmetry) of a data set.



Chapter 1 Formulae

Measures of location

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{or} \quad \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$M = \left(\frac{1}{2}n + \frac{1}{2}\right)^{th}$ value or $\frac{1}{2}n$ from a grouped frequency table

Measures of spread

$$R = \max_i(x_i) - \min_i(x_i)$$

$$IQR = Q_3 - Q_1, \quad Q_3 = \left(\frac{3}{4}n + \frac{1}{2}\right)^{th} \text{ value} \quad Q_1 = \left(\frac{1}{4}n + \frac{1}{2}\right)^{th} \text{ value}$$

$$\text{alternatively} \quad Q_3 = \left(\frac{3}{4}n + \frac{3}{4}\right)^{th} \text{ value} \quad Q_1 = \left(\frac{1}{4}n + \frac{1}{4}\right)^{th} \text{ value}$$

or $\frac{3}{4}n$ and $\frac{1}{4}n$ from a grouped frequency table

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]$$

$$\text{or} \quad \frac{1}{\sum f_i - 1} \sum_{i=1}^n f_i (x_i - \bar{x})^2 = \frac{1}{\sum f_i - 1} \left[\sum f_i x_i^2 - n\bar{x}^2 \right]$$

Measures of skewness

$$\text{skewness} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

$$\text{coeff of skew} = \frac{\text{skewness}}{s_n^3} \quad \text{where} \quad s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Sample moments

$$k\text{th moment} = \frac{1}{n} \sum_{i=1}^n x_i^k$$

$$k\text{th moment about } \alpha = \frac{1}{n} \sum_{i=1}^n (x_i - \alpha)^k$$

$$k\text{th central moment} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

Chapter 1 Solutions

Solution 1.1

1. Discrete
2. Continuous
3. Dichotomous
4. Nominal
5. Dichotomous
6. Ordinal

Solution 1.2

The total number of children is $n = \sum f_i x_i = 8 \times 0 + 12 \times 1 + 28 \times 2 + \dots + 1 \times 7 = 186$.

Solution 1.3

The total area of the last four groups must remain the same. So these groups would be replaced by a single “thick” group with a count of $\frac{34}{4} = 8.5$.

Solution 1.4

Since the area (= height \times width) is proportional to the frequency:

$$\text{height} = \frac{\text{frequency}}{\text{width}}$$

Solution 1.5

The mean is calculated as follows:

$$\begin{aligned}\bar{x} &= \frac{1}{80}(8 \times 0 + 12 \times 1 + 28 \times 2 + 19 \times 3 + 7 \times 4 + 4 \times 5 + 1 \times 6 + 1 \times 7) \\ &= \frac{1}{80} \times 186 = 2.325\end{aligned}$$

Solution 1.6

The mean is:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1 \times 74.5 + 5 \times 124.5 + \dots + 1 \times 524.5}{1 + 5 + \dots + 1} = \frac{31,200}{100} = £312$$

This figure is an approximation because we have used the midpoints and have assumed that the values in each group are uniformly distributed. For comparison, the value of \bar{x} calculated from the original data values is £313.53.

Solution 1.7

- (i) First placing the ages in increasing order of magnitude:

26 28 34 34 39 41 70

The median is the $(7 + 1)/2 = 4$ th observation, which is 34.

- (ii) The ages in increasing order of magnitude are now:

26 28 34 34 39 41 41 70

The median is the $(8 + 1)/2 = 4.5$ th observation, which is $(34 + 39)/2 = 36.5$.

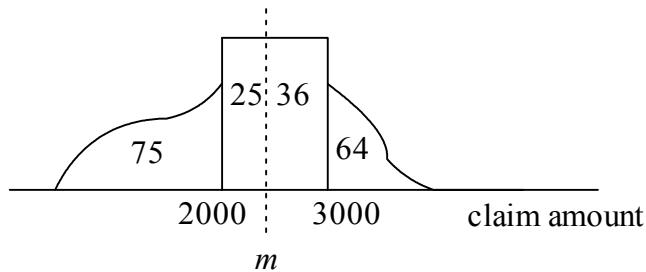
- (iii) The mean for the original seven policyholders is 38.86. This figure has been affected by the extreme value of 70, whereas the median is *robust* (resistant to the effects of extreme observations).

Solution 1.8

The median is the $(100+1)/2 = 50.5$ th observation, which is $(314+318)/2 = £316$.

Solution 1.9

The median value splits the data set into two equal halves, so that half of the observations (100) lie below the median and half lie above the median.



We can see that 75 observations are equal to or below £2,000 and that 64 observations are above £3,000, so the median lies in the group $2,000 < X \leq 3,000$. If we assume the values are distributed uniformly and apply linear interpolation, the median, m , is given by:

$$\frac{m - 2,000}{3,000 - 2,000} = \frac{100 - 75}{136 - 75} \Rightarrow m = 2,000 + \left(\frac{25}{61} \right) \times 1,000 = £2,410$$

We've rounded this to the nearest £ because our calculation is not likely to be very accurate. (We have assumed the values are distributed uniformly within (very few) groups).

Alternatively, using the $\frac{1}{2}(n+1)$ definition for the median, we would be 25.5 observations into the 2000-3000 group. This would give a median of:

$$m = 2,000 + \left(\frac{25.5}{61} \right) \times 1,000 = £2,418$$

Solution 1.10

We need to show that $\sum(x_i - \bar{x})^2$ is equivalent to $\sum x_i^2 - n\bar{x}^2$.

Expanding the first expression and taking any constant factors outside the sums gives:

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

Since $\sum_{i=1}^n x_i = n\bar{x}$ and $\sum_{i=1}^n 1 = n$, we get:

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

So the two methods of calculating s^2 are equivalent.

Solution 1.11

Now we have:

$$\sum x = 31,353 \quad \sum x^2 = 10,687,041$$

$$(i) \quad \bar{x} = \frac{31,353}{100} = £313.53$$

$$(ii) \quad s^2 = \frac{1}{99} [10,687,041 - 100 \times 313.53^2] = 8655.91$$

So the sample standard deviation is:

$$s = \sqrt{8655.91} = £93.04$$

Solution 1.12

After a bit of searching in the original data table, we find that the smallest and largest values are 74 and 523, giving a range of £449.

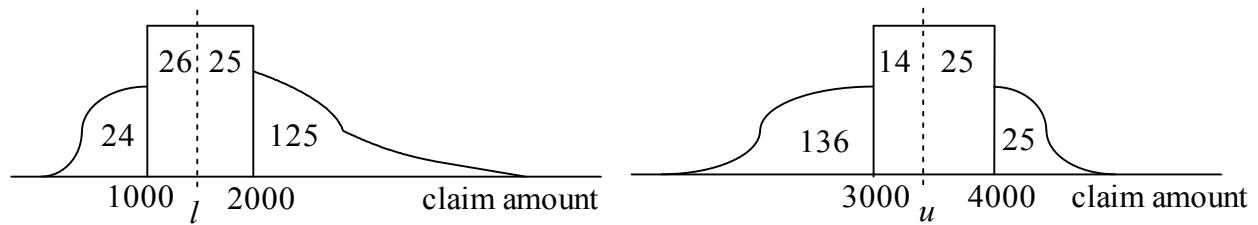
Solution 1.13

The quartiles correspond to the $25\frac{1}{2}$ th and $75\frac{1}{2}$ th values. So the IQR is:

$$IQR = 373\frac{1}{2} - 260 = £113.50$$

Solution 1.14

The quartiles split the data set into four equal quarters, so that a quarter of the observations (50) lie below the lower quartile (l) and a quarter lie above the upper quartile (u).



If we assume the values are distributed uniformly and apply linear interpolation, the lower quartile, l , is given by:

$$l = 1,000 + \left(\frac{26}{51} \right) \times 1,000 = £1,510$$

and the upper quartile, u , is given by:

$$u = 3,000 + \left(\frac{14}{39} \right) \times 1,000 = £3,359$$

The estimated interquartile range is therefore:

$$IQR = £3,359 - £1,510 = £1,849$$

Alternatively, using the $\frac{1}{4}n + \frac{1}{2}$ and $\frac{3}{4}n + \frac{1}{2}$ definitions for the quartiles, we get:

$$l = 1,000 + \left(\frac{26.5}{51} \right) \times 1,000 = £1,520 \quad u = 3,000 + \left(\frac{14.5}{39} \right) \times 1,000 = £3,372$$

$$\Rightarrow IQR = £3,372 - £1,520 = £1,852$$

Alternatively, using the $\frac{1}{4}n + \frac{1}{4}$ and $\frac{3}{4}n + \frac{3}{4}$ definitions for the quartiles, we get:

$$l = 1,000 + \left(\frac{26.25}{51} \right) \times 1,000 = £1,515 \quad u = 3,000 + \left(\frac{14.75}{39} \right) \times 1,000 = £3,378$$

$$\Rightarrow IQR = £3,378 - £1,515 = £1,863$$

Solution 1.15

- (a) The skewness is given by:

$$\frac{\sum (x_i - \bar{x})^3}{n} = -\frac{11,949,848.3946}{100} = -119,498.483946$$

The negative value tells us the water leakage claims are negatively skewed.

- (b) The second central moment is:

$$\frac{\sum (x_i - \bar{x})^2}{n} = \frac{856,934.91}{100} = 8,569.3491$$

Hence the coefficient of skewness is given by:

$$-\frac{119,498.483946}{8,569.3491^{1.5}} = -0.15064$$

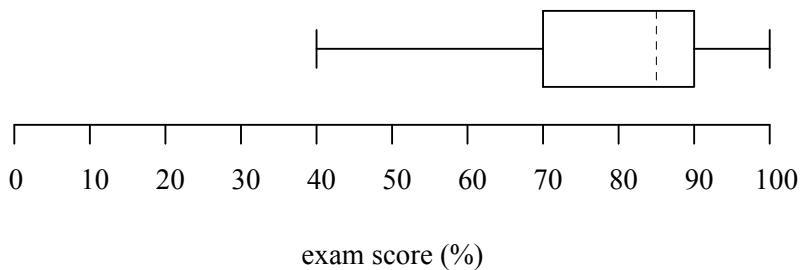
Now the skewness is “standardised” we can see that it is only slightly negatively skewed (as it is close to a value of zero which means the distribution is symmetrical).

Solution 1.16(i) ***Annual salaries***

This data set is likely to have a positively skew distribution. A boxplot might look like this:

(ii) ***Exam scores***

This data set is likely to have a negatively skew distribution. A boxplot might look like this:

**Past Exam Question (Subject C1, April 1998, Q10)**

This means that we can calculate the total of the claims:

$$\sum x = 426 \times 10 = 4,260$$

Removing the unwanted claim means that $n = 9$ and $\sum x = 4,260 - 545 = 3,715$.

This gives a new mean of $\frac{3,715}{9} = £412.78$.

$$112^2 = \frac{1}{9} (\sum x^2 - 10 \times 426^2) \Rightarrow \sum x^2 = (9 \times 112^2 + 10 \times 426^2) = 1,927,656$$

Removing the unwanted claim gives $\sum x^2 = 1,927,656 - 545^2 = 1,630,631$.

The new variance is:

$$S^2 = \frac{1}{8} \left(1,630,631 - 9 \times \left(\frac{3715}{9} \right)^2 \right) = 12,145.19$$

i.e the standard deviation is £110.21.

Chapter 2

Probability



Syllabus objectives

- (ii) 1. Explain what is meant by a set function, a sample space for an experiment, and an event.
2. Define probability as a set function on a collection of events, stating basic axioms.
3. Derive basic properties satisfied by the probability of occurrence of an event, and calculate probabilities of events in simple situations.
4. Derive the addition rule for the probability of the union of two events, and use the rule to calculate probabilities.
5. Define the conditional probability of one event given the occurrence of another event, and calculate such probabilities.
6. Derive Bayes' Theorem for events, and use the result to calculate probabilities.
7. Define independence for two events, and calculate probabilities in situations involving independence.

0 **Introduction**

Probability is a *numerical* way of describing how likely something is to happen. This chapter looks at calculating probabilities using set theory. We also look at various rules for calculating “or”, “and” and “conditional” probabilities, including Bayes’ Theorem.

Much of this material will be familiar to anyone who has studied statistics before. For students who have not met probability (and in particular probability tree diagrams) before and would like more examples and careful explanations, the Stats Pack has been recently developed to help. See http://www.acted.co.uk/Html/paper_stats_pack.htm or contact StatsPack@bpp.com for further details. Alternatively, refer to an A-level (or equivalent) statistics textbook.

Please be aware that permutations and combinations are assumed knowledge for CT3. Therefore probability questions in CT3 can (and do) involve their use. Again, if you are unfamiliar with these please refer to Stats Pack or an A-level (or equivalent) textbook.

1 Sets

1.1 Basic terminology

A set is defined as a collection of objects and each individual object is called an element of that set.

For example, the days of the week form a *set*:

$$D = \{\text{Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}\}$$

Tuesday is an *element* of the set D . We write this as:

$$\text{Tuesday} \in D$$

However, we are interested in using sets for probabilities:

Supposing an experiment is defined as any sort of operation whose outcome cannot be predicted in advance with certainty, the sample space S for such an experiment is the set of all possible outcomes that might be observed.

For example, supposing the experiment is one roll of a normal six-sided dice, the sample space would be defined as $S = \{1, 2, 3, 4, 5, 6\}$, ie all the possible numbers that can be rolled.



Question 2.1

Write down the sample space for an experiment where the scores of two dice are added.

An event A is defined as a subset of the sample space S , which contains any element of S . A is only a subset of S (written as $A \subset S$) if every element that belongs to A also belongs to S .

For example, $C = \{\text{Tuesday, Thursday}\}$ is a subset of the days of the week set D given above. We write $C \subset D$. It might be helpful to think of the subset symbol as a bit like a “less than” symbol.

In the example above, we could define throwing an even number as an event, therefore $A = \{2, 4, 6\}$. The event is said to have occurred if any one of its elements is the outcome observed, for example if a 2 was rolled.

Think of an event as anything we might wish to observe from our experiment.

The null set \emptyset is the set with no elements.

Therefore: $\emptyset = \{\}$

An event B is said to be complementary to another A in a sample space, S , if B contains all the elements of S that are not in A . The complement of A is normally written as A' or \bar{A} .

Considering the example of rolling a dice, where $S = \{1, 2, 3, 4, 5, 6\}$. If $A = \{2, 4, 6\}$, then $A' = \{1, 3, 5\}$.

We can think of the complement as “not”, ie all the elements *not* in A .

1.2 Set operations

The union of two sets, A and B , written as $A \cup B$, is the set that consists of all the elements that belong to A or B or both, for example if $A = \{1, 2\}$ and $B = \{1, 3\}$ then $A \cup B = \{1, 2, 3\}$.

We can think of the union as “or”, ie all the elements in A or B (or both).

The intersection of two sets, A and B , written as $A \cap B$ is the set that consists of all elements that belong to both A and B so in the example above, $A \cap B = \{1\}$.

We can think of the intersection as “and”, ie all the elements in A and B .

If there are no elements common to both sets, they are known as *mutually exclusive*, for example, if $A = \{1, 2\}$ and $B = \{3, 4\}$ then $A \cap B = \emptyset$, the null set.



Question 2.2

Write down the sets $A \cap B$ and $A \cup B$ where:

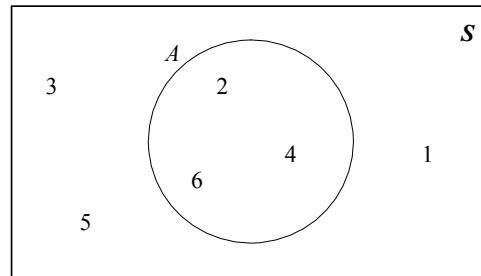
$$A = \{2, 8, 11, 16, 121\}$$

$$B = \{\text{the square numbers less than } 144\}$$

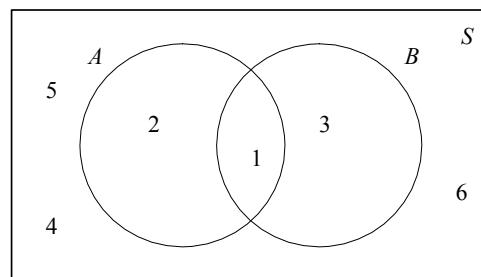
2 Venn diagrams

2.1 Basic diagrams

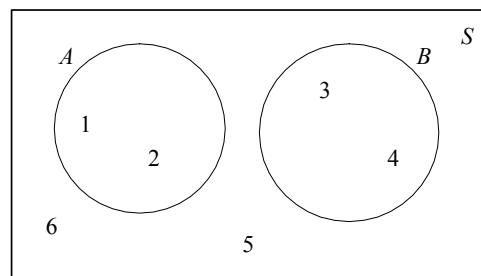
A convenient way to represent sets is by drawing Venn diagrams.



The rectangle shows the complete sample space for our dice roll, $S = \{1, 2, 3, 4, 5, 6\}$. The circle shows the event “rolling an even number”, $A = \{2, 4, 6\}$. Note how A is contained within S , because $A \subset S$.



The Venn diagram above for a dice roll shows the sets $A = \{1, 2\}$ and $B = \{1, 3\}$. Since 1 is in both sets it is placed in the overlap (the intersection) between A and B .



The Venn diagram above for a dice roll shows the sets $A = \{1, 2\}$ and $B = \{3, 4\}$. These sets are mutually exclusive – they have no elements in common, so they are drawn with no overlap.

**Question 2.3**

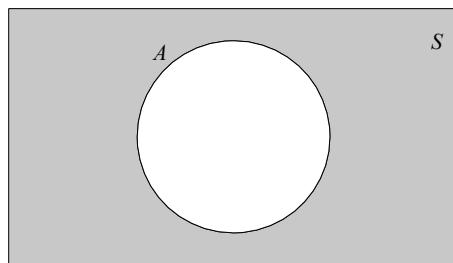
Draw a separate Venn diagram for each of the following sample spaces:

(i) $A = \{1\}$, $B = \{1, 2, 3\}$ from $S = \{1, 2, 3, 4, 5, 6\}$

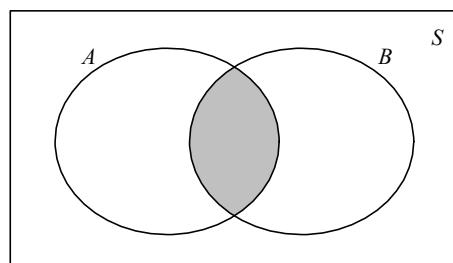
(ii) $A = \{1, 4, 7\}$, $B = \{3, 7, 8, 9\}$, $C = \{4, 7, 8, 9\}$ from $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

2.2 Set operations on Venn diagrams

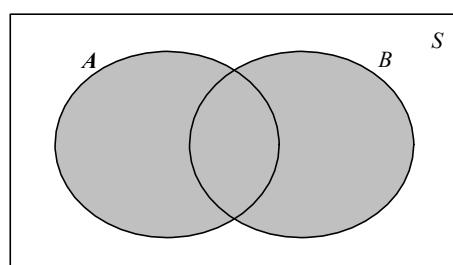
We can use shading to show the appropriate part represented by an operation:



The diagram above shows A' (the complement of A), ie everything that is *not* in A .



The diagram above shows $A \cap B$ (the intersection of A and B), ie everything in A and B .

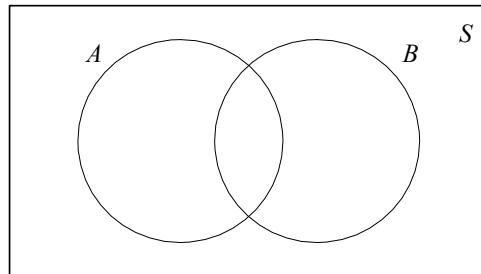


The diagram above shows $A \cup B$ (the union), ie everything in A or B (or both).



Question 2.4

For the Venn diagram below, use shading to identify the following regions:



- (i) $A' \cap B$ (ii) $A' \cap B'$ (iii) $(A \cap B)'$ (iv) $(A \cup B)'$

2.3 Using Venn diagrams to solve problems

Instead of writing the elements on the diagram we could simply write the total number of elements in each region (or the probability of each region).

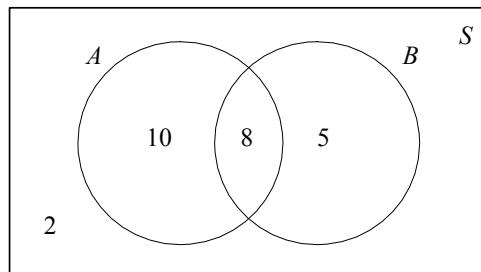


Example

In a group of 25 people, 18 have a mortgage, 13 own some shares and 2 people have neither a mortgage nor any shares. How many people have both?

Solution

Adding up the number of people who have a mortgage and own shares, we should get 23, since there are two people who have neither. The actual figure is 31, but this includes the people who have a mortgage and own shares twice. Therefore there must be 8 people in the intersection.



There are 8 people with both a mortgage and some shares.

**Question 2.5**

In a group of students 60% have passed CT1, 45% have passed CT3 and 25% have passed both. What percentage of students have passed neither?

3 **Probabilities**

In the previous section, we used sets to define events. A probability is a *numerical* way of describing how likely (or not) an event is to happen.

If each of the elements in the sample space are equally likely, then we can define the probability of event A as:

$$P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } S}$$



Example

Find the probability of rolling an even number on an ordinary dice.

Solution

We have a sample space of $S = \{1, 2, 3, 4, 5, 6\}$.

Defining “throwing an even number” as event A , we have $A = \{2, 4, 6\}$.

So the probability of throwing an even number is given by:

$$P(A) = \frac{3}{6} = \frac{1}{2}$$



Question 2.6

One card is picked from an ordinary pack of 52 playing cards. What is the probability of obtaining:

- (i) a diamond
- (ii) an ace
- (iii) the ace of diamonds
- (iv) a jack, queen or king.

3.1 Basic probability axioms

We now look at the basic rules of this probability definition:

The three basic probability axioms can be summarised as follows:

1. $P(S) = 1$

The relative frequency of an event that is certain to occur must be 1. The sample space, S contains all possible outcomes and therefore the probability of S must be 1.

For example, when we are rolling a dice $S = \{1, 2, 3, 4, 5, 6\}$. When $A = S$, ie the event of getting any one of the numbers 1 to 6 on a dice, the probability is:

$$P(S) = \frac{6}{6} = 1$$

The probability of getting any one of the numbers 1 to 6 on a dice is certain! So this result is saying that we give *certain* events a probability of 1.

It follows that for event A from sample space S that, $P(A') = 1 - P(A)$.

Suppose the probability of an individual dying is 0.2, then the probability of the individual *not* dying is $1 - 0.2 = 0.8$.

2. $P(A) \geq 0$ for all $A \subset S$

The relative frequency of occurrence of any event must not be negative, that is, probabilities can never be negative.

If we define event A as “rolling a 7 on a dice”, then the probability of this is:

$$P(A) = \frac{0}{6} = 0$$

So the probability of an *impossible* event is 0.

So rules 1 and 2 together are telling us that probabilities lie between 0 (impossible) and 1 (certain).

$$3. \quad P(A \cup B) = P(A) + P(B) \quad \text{if } A \cap B = \emptyset$$

If two events cannot occur simultaneously, because they are mutually exclusive, the probability of an event defined by their union is equal to the sum of the probabilities of the two events. This property is known as **additivity**.

Mutually exclusive events are events that cannot occur simultaneously, ie $A \cap B = \emptyset$. Taking our dice rolling example again, if $A = \{1, 4\}$ and $B = \{2, 3, 5\}$ then A and B are mutually exclusive.

We have $P(A) = \frac{2}{6}$, $P(B) = \frac{3}{6}$ and $A \cup B = \{1, 2, 3, 4, 5\}$. So:

$$P(A \cup B) = \frac{5}{6} = \frac{2}{6} + \frac{3}{6} = P(A) + P(B)$$

The reason this rule works is that there is no “overlap” between the elements in A and B .

Suppose we now consider non-mutually exclusive events when rolling a dice. If $A = \{1, 2\}$ and $B = \{1, 3\}$, we have $P(A) = \frac{2}{6}$, $P(B) = \frac{2}{6}$ and $A \cup B = \{1, 2, 3\}$. So:

$$P(A \cup B) = \frac{3}{6} \neq \frac{2}{6} + \frac{2}{6} = P(A) + P(B)$$

The rule has broken down because the element 1 is in *both* events A and B . This element has been counted twice on the RHS of the equation, once in $P(A)$ and once in $P(B)$.



Question 2.7

One card is picked from an ordinary pack of 52 playing cards. Consider the following events $A = \{\text{pick a 7}\}$, $B = \{\text{pick an ace}\}$ and $C = \{\text{pick a club}\}$.

Show that Axiom 3, namely $P(A \cup B) = P(A) + P(B)$:

- (i) works for events A and B
- (ii) doesn't work for events A and C and explain why.

3.2 The addition rule

Axiom 3 is often known as the special addition rule. For a more general case, where two sets are not necessarily mutually exclusive, the rule can be extended as follows:

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

Axiom 3 is a special case of this rule when A and B are mutually exclusive. In which case $A \cap B = \emptyset$, so $P(A \cap B) = 0$, and hence $P(A \cup B) = P(A) + P(B)$.

How can we practically show that this rule is true? Recall that, for the non-mutually exclusive events $A = \{1, 2\}$ and $B = \{1, 3\}$, when rolling a dice, we had $P(A) = \frac{2}{6}$, $P(B) = \frac{2}{6}$ and $A \cup B = \{1, 2, 3\}$. So:

$$P(A \cup B) = \frac{3}{6} \neq \frac{2}{6} + \frac{2}{6} = P(A) + P(B)$$

The rule has broken down because the element $\{1\}$ is in *both* events A and B (*i.e.* $A \cap B = \{1\}$). This element has been counted twice on the RHS of the equation, once in $P(A)$ and once in $P(B)$. So, to fix this error, we need to remove the $\{1\}$, *i.e.* we need to remove $A \cap B$. Hence:

$$P(A \cup B) = \frac{3}{6} = \frac{2}{6} + \frac{2}{6} - \frac{1}{6} = P(A) + P(B) - P(A \cap B)$$

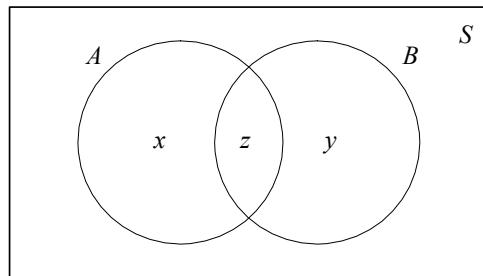
We can also prove this using Venn diagrams. Core Reading now demonstrates this rule more formally.

This can be shown as follows:

Consider the mutually exclusive events:

$A \cap B'$ elements of A which are not in B with probability x	
$A' \cap B$ elements of B which are not in A	with probability y
$A \cap B$ elements in both A and B	with probability z

On a Venn diagram we have:



It can be seen that $P(A) = x + z$, and $P(B) = y + z$

$$\begin{aligned}
 P(A \cup B) &\text{ can also be expressed as } P(A' \cap B) + P(A \cap B') + P(A \cap B) \\
 &= x + y + z \\
 &= (x + z) + (y + z) - z \\
 &= P(A) + P(B) - P(A \cap B)
 \end{aligned}$$

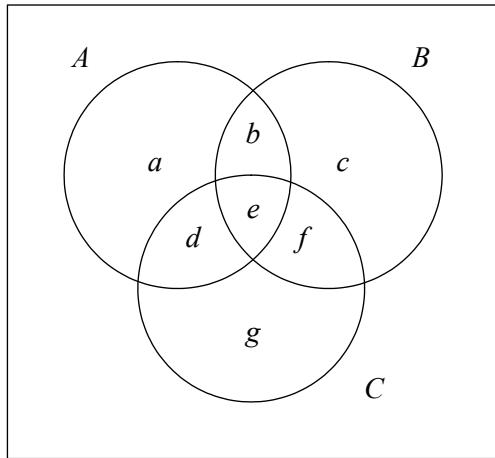


Question 2.8

A contestant on a game show is asked two questions. The probability that she gets the first question correct is 0.3 and the probability that she gets the second question correct is 0.4. Given that the probability that she gets both questions correct is 0.1, calculate the probability that:

- (i) she gets either the first, the second or both questions right
- (ii) she gets both questions wrong.

The addition rule can be extended to three events A , B and C . Consider the Venn diagram:



Using the letters a to g to stand for the appropriate probabilities, we can see that:

$$P(A \cup B \cup C) = a + b + c + d + e + f + g$$

and:

$$P(A) = a + b + d + e \quad P(B) = b + c + e + f \quad P(C) = d + e + f + g$$



Question 2.9

Use the letters a to g to write down expressions for:

- | | |
|---------------------|---------------------------|
| (i) $P(A \cap B)$ | (ii) $P(A \cap C)$ |
| (iii) $P(B \cap C)$ | (iv) $P(A \cap B \cap C)$ |

We can then show that:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C) \end{aligned}$$



Question 2.10

Use your results from Question 2.9 to prove the above result.



Example

Students at a performing arts college can choose to study one or more classes of acting, dance or singing. The probability that a student is studying acting is 0.5, dance 0.65, singing 0.55, acting or dancing 0.8, acting and singing 0.25, dancing and singing 0.25. Find the probability that a student studies all three classes.

Solution

We are given:

$$P(A) = 0.5 \quad P(D) = 0.65 \quad P(S) = 0.55$$

$$P(A \cup D) = 0.8 \quad P(A \cap S) = 0.25 \quad P(D \cap S) = 0.25$$

Since students must study at least one of these classes, we also have:

$$P(A \cup D \cup S) = 1$$

We require $P(A \cap D \cap S)$, but to use our three-event addition rule we also require $P(A \cap D)$. We obtain this from:

$$P(A \cup D) = P(A) + P(D) - P(A \cap D)$$

Hence:

$$1 = 0.5 + 0.65 + 0.55 - 0.35 - 0.25 - 0.25 + P(A \cap D \cap S)$$

$$\Rightarrow P(A \cap D \cap S) = 0.15$$



Question 2.11

Customers at a restaurant may order any combination of chips, salad or onion rings. The probability that a customer chooses onion rings is 0.3, salad 0.4, chips and salad 0.15, chips and onion rings 0.15, salad or onion rings 0.55, all three 0.05, none 0.2. Calculate the probability a customer chooses:

- (i) chips (ii) chips only

4 Conditional probabilities

Consider the two events, A and B . We might wish to know the probability that event A occurred, given the occurrence of the event B . This is known as a conditional probability and is denoted thus:

$$P(A | B)$$

$P(A | B)$ is read as “the probability of event A occurring *given* that event B has already occurred” or “probability of A *given* B ” for short. This is called a conditional probability as the probability depends (*i.e.* is conditional) on event B .

The conditional probability of A occurring given B can be expressed as:

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

The above formula can be explained as representing the occasions that event A occurs with B relative to the occasions that B occurs (with or without A).



Example

Consider picking a card from an ordinary pack of playing cards. If we have the events:

$$A = \{\text{pick a spade}\} \quad B = \{\text{pick an 8}\}$$

calculate the probability of picking a spade *given* that we have picked an 8, *i.e.* calculate $P(A | B)$.

Solution

Since there are only four 8's in the pack and only one of them is a spade, we conclude that $P(A | B) = \frac{1}{4}$.

Checking this intuitive answer using the formula:

$$A \cap B = \{\text{8 of spades}\} \Rightarrow P(A \cap B) = \frac{1}{52}$$

$$A' \cap B = \{\text{8 of clubs, 8 of diamonds, 8 of hearts}\} \Rightarrow P(A' \cap B) = \frac{3}{52}$$

$$\Rightarrow P(A | B) = \frac{P(A \cap B)}{P(A \cap B) + P(A' \cap B)} = \frac{\frac{1}{52}}{\frac{1}{52} + \frac{3}{52}} = \frac{1}{4}$$

$(A \cap B)$ and $(A' \cap B)$ are mutually exclusive. Noting that $(A \cap B) \cup (A' \cap B) = B$, this can be rearranged thus:

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

This is the version of the formula that is commonly used in practice. We have simplified the denominator using the addition rule for mutually exclusive events:

$$P[(A \cap B) \cup (A' \cap B)] = P(A \cap B) + P(A' \cap B) = P(B)$$



Example

Consider our previous example of picking a card from an ordinary pack of cards:

$$A = \{\text{pick a spade}\} \quad B = \{\text{pick an 8}\}$$

Calculate the probability of picking a spade *given* that we have picked an 8, *i.e.* calculate $P(A|B)$.

Solution

We obtain the same answer as before – but our calculation is much simpler:

$$A \cap B = \{\text{8 of spades}\} \Rightarrow P(A \cap B) = \frac{1}{52}$$

$$B = \{\text{pick an 8}\} \Rightarrow P(B) = \frac{4}{52}$$

$$\Rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{4}{52}} = \frac{1}{4}$$



Question 2.12

In a group of 24 actuaries, 20 have worked for Life Office A and 12 have worked for Life Office B. Everyone in the group has worked for at least one of the two companies. What is the probability that an actuary picked at random has worked for:

- (i) Life Office A and Life Office B
- (ii) Life Office A given that they have worked for Life Office B?

4.1 The multiplication rule

Rearranging our formula for conditional probabilities, we obtain:

$$P(A \cap B) = P(B)P(A | B)$$

Recall that $A \cap B$ can be interpreted as A and B . We can use this rule to calculate the probability of events A and B both happening.

4.2 Independent events

Events A and B are said to be *independent* if whether or not event B has occurred gives us no information on whether event A has occurred. This can be expressed algebraically as follows:

$$A \text{ and } B \text{ are independent if } P(A) = P(A | B) = P(A | B')$$

Given that:

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

Then if A and B are independent:

$$P(A \cap B) = P(A)P(B)$$

This is a special case of the multiplication rule when events A and B are independent.



Example

Two dice are thrown. Find the probability of rolling a 5 on both dice.

Solution

$$A = \{\text{roll a 5 on the 1st dice}\} \Rightarrow P(A) = \frac{1}{6}$$

$$B = \{\text{roll a 5 on the 2nd dice}\} \Rightarrow P(B) = \frac{1}{6}$$

Since these events are independent:

$$P(A \cap B) = P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

4.3 Tree diagrams

A tree diagram is a convenient representation of probabilities:

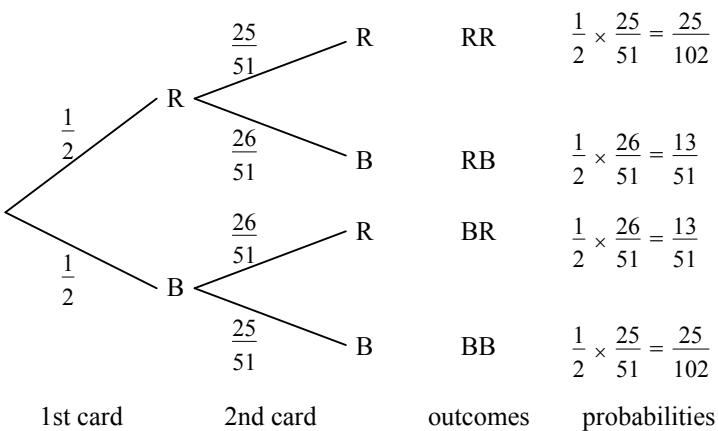


Example

A card is picked from an ordinary pack of 52 playing cards, *without replacement*, and then another one is picked. What is the probability of picking:

- (i) two red cards
- (ii) one of each colour?

Solution



A tree diagram is used to show the possible outcomes. In this case we can either get red or a black card. So the number of branches is the number of different possibilities, and the numbers on the branches are the probabilities of those particular events happening. Notice that the probabilities on any particular set of branches add up to one.

To calculate probabilities, go along the branches of the tree from left to right to get to the end, and multiply together any probabilities that you have passed – using the *multiplication rule*. If there is more than one route through the tree to give the answer you require, then sum the answers from the different routes – using the *addition rule* for mutually exclusive events.

$$(i) P(RR) = \frac{1}{2} \times \frac{25}{51} = \frac{25}{102}$$

$$(ii) P(1 R \text{ and } 1 B) = P(RB) + P(BR) = \left(\frac{1}{2} \times \frac{26}{51}\right) + \left(\frac{1}{2} \times \frac{26}{51}\right) = \frac{13}{51} + \frac{13}{51} = \frac{26}{51}$$

**Question 2.13**

A box of chocolates contains 8 milk chocolates and 4 plain chocolates. A chocoholic eats three chocolates. Calculate the probability that:

- (i) all three are milk chocolates
- (ii) exactly one is a plain chocolate.

Using probability tree diagrams makes it much easier to calculate conditional probabilities:

**Question 2.14**

In a restaurant, 45% of the customers are female. 74% of females choose from the *à la carte* menu, whilst only 37% of males do. The rest choose from the set menu. What is the probability that:

- (i) a customer orders from the set menu
- (ii) a customer ordering from the *à la carte* menu is female?

4.4 Law of total probability

Consider the set space, S , being divided into a partition of n mutually exclusive events, E_i where $i = 1, 2, 3, \dots, n$, then:

$$E_i \cap E_j = \emptyset, \text{ and}$$

$$E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n = S$$

For example, a partition of our dice sample space $S = \{1, 2, 3, 4, 5, 6\}$ could be:

$$E_1 = \{1\}, E_2 = \{2, 3\}, E_3 = \{4, 5, 6\}$$

There are many other possibilities, as long as the E_i 's are mutually exclusive (*ie* there is no overlap between them) and together make up the whole sample space (*ie* they are *exhaustive*) it is a partition.

And for any $A \subset S$:

$$A = (A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3) \cup \dots \cup (A \cap E_n)$$

For example, if A is the event “roll an even number” we have $A = \{2, 4, 6\}$. This can be made up of all the intersections with our partition from above:

$$A \cap E_1 = \emptyset$$

$$A \cap E_2 = \{2\}$$

$$A \cap E_3 = \{4, 6\}$$

Hence:

$$A = \{2, 4, 6\} = \emptyset \cup \{2\} \cup \{4, 6\}$$

Therefore:

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3) + \dots + P(A \cap E_n)$$

$$= \sum_{j=1}^n P(A \cap E_j)$$

This result is known as the law of total probability.

4.5 Bayes' Theorem

This theorem is named after the Reverend T Bayes and is used extensively in Bayesian methods of statistical inference, which will be covered in more detail in Subject CT6, Statistical Methods.

The result is as follows:

Let $E_1, E_2, E_3, \dots, E_n$ be a partition of S and define event $A \subset S$

Using the conditional probability:

$$P(E_i | A) = \frac{P(E_i \cap A)}{P(A)} \quad (1)$$

also, the relationship:

$$P(E_i \cap A) \equiv P(A \cap E_i) = P(E_i)P(A | E_i)$$

(ie using $P(E_i \cap A) \equiv P(A \cap E_i)$ and then the multiplication rule for $P(A \cap E_i)$)

and the law of total probability:

$$P(A) = \sum_{j=1}^n P(A \cap E_j)$$

then, by substituting for $P(E_i \cap A)$ and $P(A)$ in equation (1),

the result is:

$$P(E_i | A) = \frac{P(E_i)P(A | E_i)}{\sum_{j=1}^n P(E_j)P(A | E_j)}, \quad i = 1, 2, 3, \dots, n$$

Essentially Bayes' formula allows us to "turnaround" conditional probabilities, ie calculate $P(E_i | A)$ given only information about $P(A | E_i)$.

The values $P(E_j)$ are known as prior probabilities, the event A is some event which is known to have occurred and the conditional probability $P(E_i | A)$ is known as the posterior probability.

This formula, with slightly different notation, is given on page 5 of the *Tables*.



Example

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% Birmingham and 20% Edinburgh. The probabilities of a train arriving late in Manchester, Edinburgh or Birmingham are 30%, 20% and 25% respectively.

If a late train is picked at random from the group under consideration, what is the probability that it terminated in Manchester?

Solution

We want $P(\text{Manchester} \mid \text{Late})$.

If M is the event “A train chosen at random terminated in Manchester” (and E and B have corresponding definitions), and L is the event “A train chosen at random runs late”, then:

$$\begin{aligned} P(M \mid L) &= \frac{P(M)P(L \mid M)}{P(M)P(L \mid M) + P(E)P(L \mid E) + P(B)P(L \mid B)} \\ &= \frac{0.6 \times 0.3}{(0.6 \times 0.3) + (0.2 \times 0.2) + (0.2 \times 0.25)} \\ &= 66.7\% \end{aligned}$$



Question 2.15

A person has three routes to get to work. The probability that he arrives on time using routes A, B and C are 60%, 62% and 70% respectively. If he is equally likely to choose any of the routes, and arrives at work on time, what is the probability that he chose route B?

5 Exam questions



Past Exam Question (Subject C1, April 1994, Q1)

In a certain constituency, 30% of the voters are “blue collar” workers, of whom 46% voted Conservative at the last United Kingdom election. Of the remaining voters, 36% voted Conservative.

Consider a voter selected at random from those who voted Conservative in this constituency. What is the probability that this voter is a “blue collar” worker?



Past Exam Question (Subject C2, September 1997, Q5)

A coin is selected at random from a pair of coins and tossed. Coin 1 is a double-headed coin (*i.e* a head on both sides). Coin 2 is a standard unbiased coin.

The result of the toss is a head. What is the probability that it was coin 1 which was tossed?



Chapter 2 Summary

A *set* is a collection of objects, called *elements*. A is a *subset* of B , written $A \subset B$, if all the elements in A are contained in B . The *complement* of A , written A' , is the set of all the elements *not* in A . The empty set is denoted \emptyset .

The *union* of A and B , written $A \cup B$, is the set of all elements in A or B or both. The *intersection* of A and B , written $A \cap B$, is the set of all elements in A and B .

Venn diagrams are used to represent sets and the relationships between them.

A *sample space*, S , is the set of all the possible outcomes from an experiment. An *event* is anything we might wish to observe from our experiment.

Probabilities are a numerical way of describing how likely an event is to happen. A formula for equally likely elements is given overleaf. Probabilities lie between 0 (impossible) and 1 (certain).

We can use the *addition rule* and the *multiplication rule* (see overleaf) to calculate probabilities. Tree diagrams are a helpful way of working out probabilities.

The conditional probabilities of A occurring given that B has already occurred is written $P(A|B)$. The formula is given overleaf.

Events A and B are *mutually exclusive* if $A \cap B = \emptyset$. Events A and B *independent* if $P(A|B) = P(A)$.

E_1, \dots, E_n is a *partition* of S if the E_i 's are mutually exclusive and together make up the whole set S .

Bayes' Theorem (see overleaf) allows us to "turnaround" conditional probabilities, *i.e.* calculate $P(E_i | A)$ given only information about $P(A | E_i)$.



Chapter 2 Formulae

Probabilities

For equally likely elements:

$$P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } S}$$

Addition rule

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

For mutually exclusive events $A \cap B = \emptyset$

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

Multiplication rule

$$P(A \cap B) = P(A)P(B | A) \quad \text{or} \quad P(B)P(A | B)$$

For independent events $P(A | B) = P(A)$ and $P(B | A) = P(B)$, so:

$$P(A \cap B) = P(A)P(B)$$

Conditional probabilities

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

Bayes' Theorem

For a partition E_i , $i = 1, 2, \dots, n$

$$P(E_i | A) = \frac{P(E_i)P(A | E_i)}{\sum_{j=1}^n P(E_j)P(A | E_j)}, \quad i = 1, 2, 3, \dots, n$$

Chapter 2 Solutions

Solution 2.1

When throwing two dice, we can get a total from 2 to 12. Hence:

$$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Note that (unlike the sample space for a single dice) each element in this sample space is *not* equally likely. For example, a total of 7 occurs more than, say, a total of 8.

Solution 2.2

We have:

$$A = \{2, 8, 11, 16, 121\}$$

$$B = \{1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121\}$$

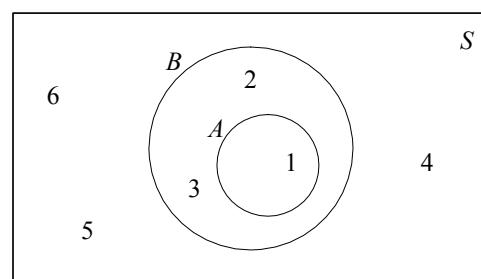
Hence:

$$A \cap B = \{16, 121\}$$

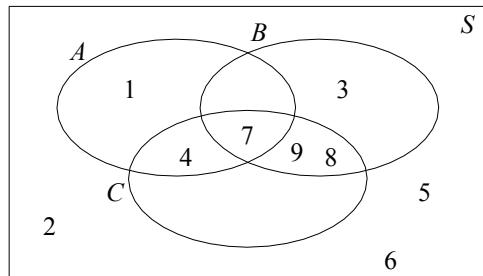
$$A \cup B = \{1, 2, 4, 8, 9, 11, 16, 25, 36, 49, 64, 81, 100, 121\}$$

Solution 2.3

- (i) Since all the elements in A are contained in B , we have $A \subset B$. This is shown on a Venn diagram as follows:

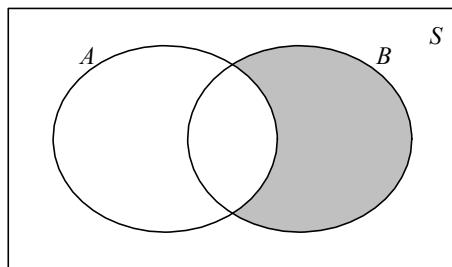


- (ii) We have three sets. Taking care to ensure each number is in the correct overlaps between sets, we obtain:

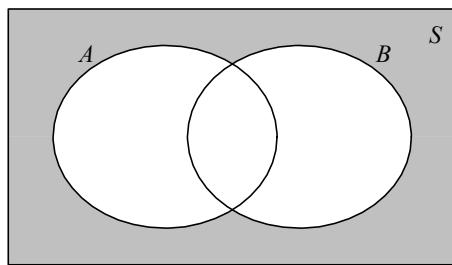


Solution 2.4

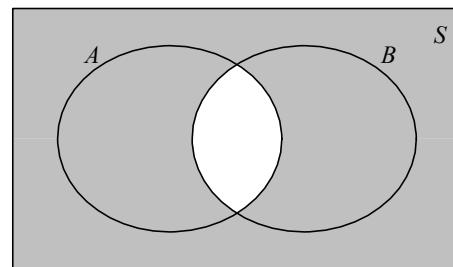
- (i) $A' \cap B$ is everything not in set A **and** in set B :



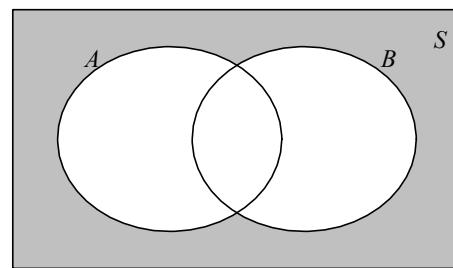
- (ii) $A' \cap B'$ is everything not in set A **and** not in set B :



- (iii) $(A \cap B)'$ is everything not in set A **and** set B (*i.e.* everything outside of the intersection of A and B).

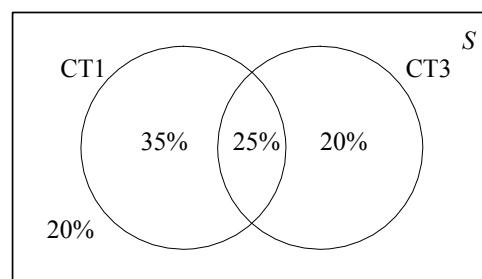


- (iv) $(A \cup B)'$ is everything not in set A **or** set B (*i.e.* everything outside of the union of A and B).



Solution 2.5

Since 25% have passed both CT1 and CT3, there must be $60\% - 25\% = 35\%$ who have only passed CT1. There must be $45\% - 25\% = 20\%$ who have only passed CT3. This leaves $100\% - 35\% - 25\% - 20\% = 20\%$ who have passed neither exam:



Solution 2.6

(i) $\frac{13}{52} = \frac{1}{4}$

(ii) $\frac{4}{52} = \frac{1}{13}$

(iii) $\frac{1}{52}$

(iv) $\frac{12}{52} = \frac{3}{13}$

Solution 2.7(i) We have $P(A) = \frac{4}{52}$, $P(B) = \frac{4}{52}$ and:

$$A \cup B = \{7 \text{ hearts}, 7 \text{ clubs}, 7 \text{ diamonds}, 7 \text{ spades}, \\ \text{Ace hearts, Ace clubs, Ace diamonds, Ace spades}\}$$

So:

$$P(A \cup B) = \frac{8}{52} = \frac{4}{52} + \frac{4}{52} = P(A) + P(B)$$

This rule works if that there is no “overlap” between the elements in A and B .(ii) We have $P(A) = \frac{4}{52}$, $P(C) = \frac{13}{52}$ and:

$$A \cup C = \{7 \text{ hearts}, 7 \text{ diamonds}, 7 \text{ spades}, \text{Ace clubs, 2 clubs, 3 clubs,} \\ 4 \text{ clubs, 5 clubs, 6 clubs, 7 clubs, 8 clubs, 9 clubs, 10 clubs,} \\ \text{Jack clubs, Queen clubs, King clubs}\}$$

So:

$$P(A \cup C) = \frac{16}{52} \neq \frac{4}{52} + \frac{13}{52} = P(A) + P(B)$$

This is because the 7 of clubs is in both A and C and so is counted twice on the RHS of the equation.

Solution 2.8

If $A = \{\text{get 1st question correct}\}$ and $B = \{\text{get 2nd question correct}\}$ then we are told:

$$P(A) = 0.3, P(B) = 0.4 \text{ and } P(A \cap B) = 0.1$$

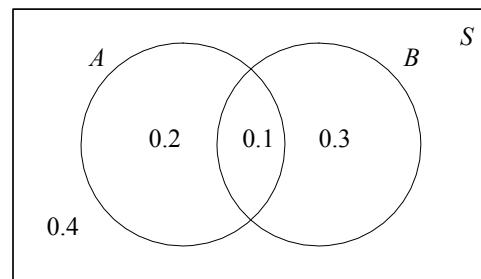
- (i) We want $P(A \cup B)$, so using the addition rule:

$$P(A \cup B) = 0.3 + 0.4 - 0.1 = 0.6$$

- (ii) We want the probability that both questions are wrong, that is the complement of getting at least one question right. Hence the probability is:

$$= 1 - 0.6 = 0.4$$

Alternatively, we could have used a Venn diagram as follows:



Since, $P(A) = 0.3$ and $P(A \cap B) = 0.1$ there must be $0.3 - 0.1 = 0.2$ just in set A alone. Similarly we get $0.4 - 0.1 = 0.3$ in set B alone. It is then easy to see the total probability in both sets is $0.2 + 0.1 + 0.3 = 0.6$ and the probability that neither question is right, ie the area outside both sets, is 0.4.

Solution 2.9

$$(i) \quad P(A \cap B) = b + e$$

$$(ii) \quad P(A \cap C) = d + e$$

$$(iii) \quad P(B \cap C) = e + f$$

$$(iv) \quad P(A \cap B \cap C) = e$$

Solution 2.10

Replacing the terms on the RHS of the equation we obtain:

$$\begin{aligned} & P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \\ &= (a + b + d + e) + (b + c + e + f) + (d + e + f + g) - (b + e) - (d + e) \\ &\quad - (e + f) + e \\ &= a + b + c + d + e + f + g \\ &= P(A \cup B \cup C) \end{aligned}$$

Solution 2.11

We are given:

$$P(O) = 0.3 \quad P(S) = 0.4$$

$$P(C \cap S) = 0.15 \quad P(C \cap O) = 0.15 \quad P(S \cup O) = 0.55 \quad P(C \cap S \cap O) = 0.05$$

Since we also have $P(\text{none}) = 0.2$, this gives us:

$$P(C \cup O \cup S) = 1 - 0.2 = 0.8$$

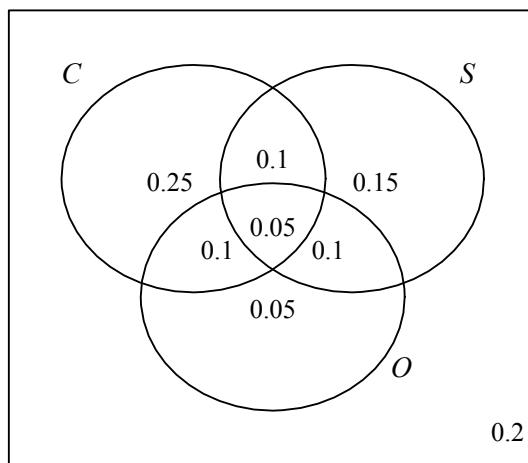
- (i) We require $P(C)$, but to use our three-event addition rule we also require $P(S \cap O)$. We get this from:

$$\begin{aligned} P(S \cup O) &= P(S) + P(O) - P(S \cap O) \\ \Rightarrow 0.55 &= 0.4 + 0.3 - P(S \cap O) \Rightarrow P(S \cap O) = 0.15 \end{aligned}$$

Hence:

$$0.8 = P(C) + 0.4 + 0.3 - 0.15 - 0.15 - 0.15 + 0.05 \\ \Rightarrow P(C) = 0.5$$

- (ii) We require $P(C \cap S' \cap O')$. This is not so easy to get by the three-event addition rule (as too many new events need to be calculated). However, it's easy to get if we place all the known probabilities into a Venn diagram:



We can see that $P(C \text{ only}) = 0.25$.

Solution 2.12

If A is the event “has worked for company A” and B is the event “has worked for company B”, then we are told that:

$$P(A) = \frac{20}{24}, P(B) = \frac{12}{24} \text{ and } P(A \cup B) = 1$$

- (i) Using the addition rule, we obtain:

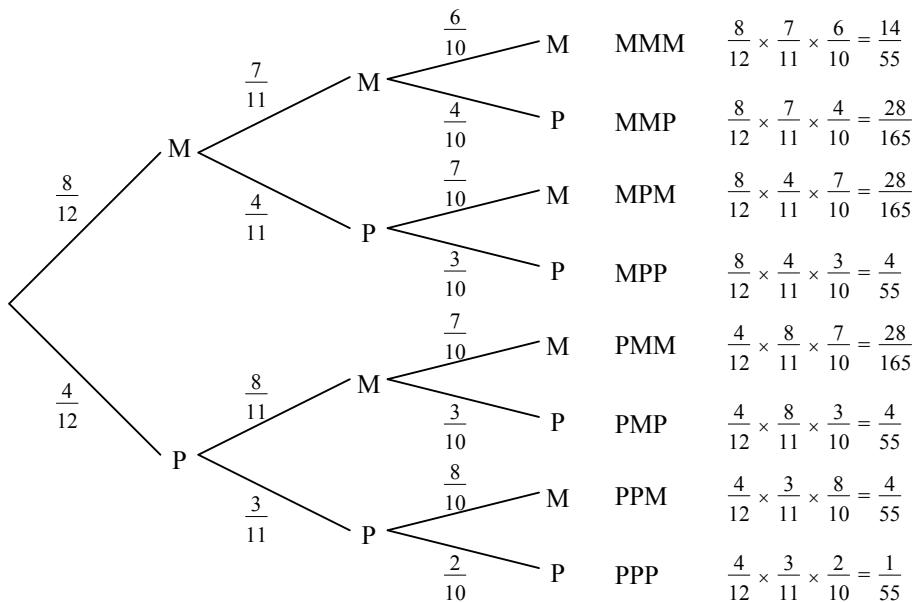
$$P(A \cap B) = \frac{20}{24} + \frac{12}{24} - 1 = \frac{8}{24}$$

(ii) Using the conditional probability formula, we obtain:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{8}{24}}{\frac{12}{24}} = \frac{2}{3}$$

Solution 2.13

If M is the event “eats a milk chocolate” and P is the event “eats a plain chocolate”, then the tree diagram is as follows:



(i) $P(3 \text{ milk chochs}) = P(MMM) = \frac{14}{55}$

(ii) $P(1 \text{ plain choc}) = P(MMP) + P(MPM) + P(PMM)$

$$= \frac{28}{165} + \frac{28}{165} + \frac{28}{165} = \frac{28}{55}$$

We could also have solved this more quickly using combinations:

$${}^3C_1 \times \left(\frac{8}{12} \times \frac{7}{11} \times \frac{4}{10} \right) = \frac{28}{55}$$

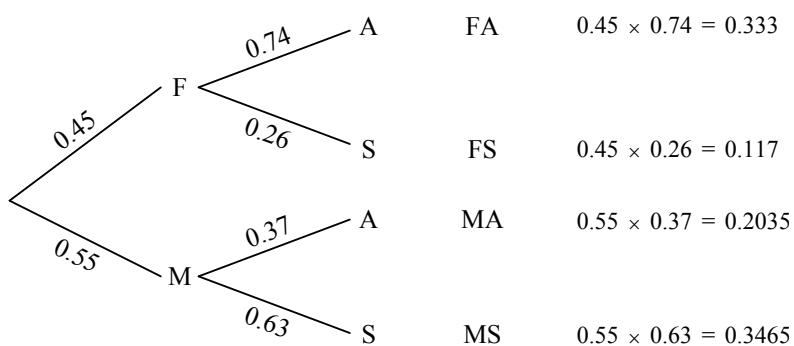
where 3C_1 is the number of ways of obtaining one “success” (in this case plain chocolate) and two “failures” (in this case milk chocolate) out of three trials, and nC_r or $\binom{n}{r}$ is calculated as $\frac{n!}{r!(n-r)!}$.

Please note that permutations and combinations **are assumed knowledge for CT3** and so can be (and have been) asked in the examination.

Tree diagrams and combinations are covered in further detail in the Stats Pack.

Solution 2.14

If F is the event “is female”, M is the event “is male”, A is the event “chooses from “à la carte”, and S is the event “chooses from the set menu”, then the tree diagram is as follows:



$$\begin{aligned}
 \text{(i)} \quad P(S) &= P(FS) + P(MS) \\
 &= 0.117 + 0.3465 \\
 &= 0.4635
 \end{aligned}$$

- (ii) This is a conditional probability as we are told that the customer *has chosen* from the à la carte menu – so this is already *given*.

$$P(F | A) = \frac{P(F \cap A)}{P(A)} = \frac{0.333}{0.333 + 0.2035} = 0.621$$

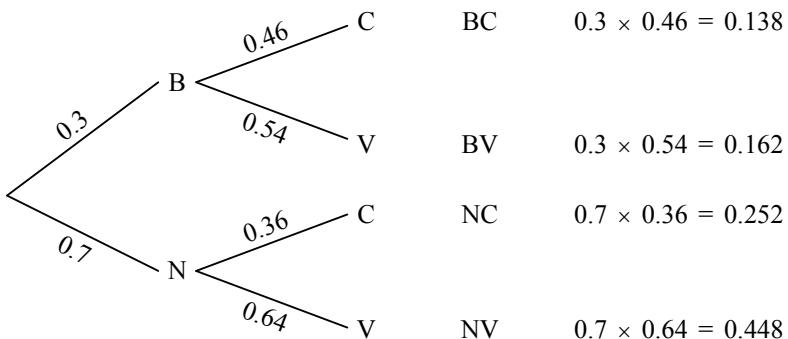
Solution 2.15

If A is the event “chooses route A”, B is the event “chooses route B”, C is the event “chooses route C”, and T is the event “arrives on time to work”, then:

$$\begin{aligned} P(B | T) &= \frac{P(B)P(T | B)}{P(A)P(T | A) + P(B)P(T | B) + P(C)P(T | C)} \\ &= \frac{\frac{1}{3} \times 0.62}{\frac{1}{3} \times 0.6 + \frac{1}{3} \times 0.62 + \frac{1}{3} \times 0.7} \\ &= 32.3\% \end{aligned}$$

Past Exam Question (Subject C1, April 1994, Q1)

Using B for “blue collar”, N for “not blue collar”, C for “vote Conservative” and V for “vote something else” we obtain:



We want:

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

So we need to calculate:

$$\begin{aligned} P(C) &= P(BC) + P(NC) \\ &= 0.138 + 0.252 \\ &= 0.39 \end{aligned}$$

Hence:

$$P(B | C) = \frac{0.138}{0.39} = 0.354$$

Past Exam Question (Subject C2, September 1997, Q5)

We require $P(\text{coin 1} | \text{head})$. Using Bayes' Theorem:

$$\begin{aligned} P(\text{coin 1} | \text{head}) &= \frac{P(\text{head} | \text{coin 1})P(\text{coin 1})}{P(\text{head} | \text{coin 1})P(\text{coin 1}) + P(\text{head} | \text{coin 2})P(\text{coin 2})} \\ &= \frac{1 \times \frac{1}{2}}{1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}} \\ &= \frac{2}{3} \end{aligned}$$

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.

Chapter 3

Random variables



Syllabus objectives

- (iii) 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.*
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. *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.*
4. *Evaluate probabilities (by calculation or by referring to tables as appropriate) associated with distributions.*
5. *Derive the distribution of a function of a random variable from the distribution of the random variable.*

0 Introduction

In this chapter we will introduce the concept of a random variable and study the properties of probability distributions.

This chapter builds upon the work of Chapter 1 in that we will be calculating the mean, variance and skewness. However, we will now do this for a *population* and not a sample. A population is not what you might think from say, geography. For example, if we looked at all the babies being born in the world we would not necessarily find equal numbers of male and female. They are merely a realisation (*i.e.* a sample) of the true underlying distribution. We model this underlying distribution as a probability distribution.

If you have studied statistics to A-Level standard or equivalent you should find this chapter straightforward. For students who have not met random variables before and would like more examples and careful explanations, Stats Pack has been developed to help. See http://www.acted.co.uk/Html/paper_stats_pack.htm for further details.

Calculating probabilities for continuous random variables involves integration. Please be aware that exam questions can (and do) make use of integration by parts and integration by substitution. If you are rusty on these, you would be wise to sort them out as soon as possible! Refer to an A-Level (or equivalent) textbook or purchase FAC from ActEd.

1 Discrete random variables

1.1 Random variables

A sample space is the set of all possible outcomes of an experiment.

A random variable is a rule for associating a number with each element in a sample space.

So, if w is an element of the sample space S (ie w is one of the possible outcomes of the experiment concerned) and the number x is associated with this outcome, then $X(w) = x$.

The “number” referred to here may be a whole number but it need not necessarily be a whole number. A good way of thinking about this is as a “numerical value”.

It is X that is the random variable. It is conventional to denote the random variable by a capital letter and the possible values it can take by a small letter.



Example 3.1

Suppose there are 8 balls in a bag. The random variable X is the weight, in kg, of a ball selected at random. Balls 1, 2 and 3 weigh 0.1kg, balls 4 and 5 weigh 0.15kg and balls 6, 7 and 8 weigh 0.2kg. Using the notation above, write down this information.

Solution

The x values are the weights and the w values are the numbers of the balls, which gives us:

$$X(1) = 0.1, X(2) = 0.1, X(3) = 0.1,$$

$$X(4) = 0.15, X(5) = 0.15,$$

$$X(6) = 0.2, X(7) = 0.2, X(8) = 0.2$$

The sample space consists of the individual balls that can be picked out.

1.2 Discrete random variables

If the range of the function (the set of all possible values for x) is a finite set:

eg $\{1, 2, 3, 4, 5, 6\}$

or $\{x: x = a + b, a = 1, 2, 3, 4, 5, 6, b = 1, 2, 3, 4, 5, 6\}$

or a countably infinite set:

eg $\{0, 1, 2, 3, \dots\}$

or $\{x: x = y^2, y = \dots -3, -2, -1, 0, 1, 2, 3, \dots\}$

then X is a discrete random variable.

Random variables can alternatively be “continuous”. We will meet continuous random variables in Section 2.

1.3 Probabilities

Probabilities are defined on events (subsets of S). So what is meant by “ $P(X = x)$ ”?

Suppose there are 8 possible outcomes, and that outcomes w_1 , w_2 , and w_3 are all associated with the number x_1 , outcomes w_4 and w_5 are both associated with x_2 , and outcomes w_6 , w_7 , and w_8 are all associated with x_3 . Then the meaning will become clear:

By $P(X = x_1)$ is meant $P(E_1)$ where $E_1 = \{w_1, w_2, w_3\}$

By $P(X = x_2)$ is meant $P(E_2)$ where $E_2 = \{w_4, w_5\}$

By $P(X = x_3)$ is meant $P(E_3)$ where $E_3 = \{w_6, w_7, w_8\}$


Example 3.2

Referring back to Example 3.1, describe what is meant by $P(X = 0.1)$.

Solution

$P(X = 0.1)$ means the probability that either ball 1, ball 2 or ball 3 is selected.

1.4 Probability functions

The function $f_X(x) = P(X = x)$ for each x in the range of X is the probability function (PF) of X – it specifies how the total probability of 1 is divided up amongst the possible values of X and so gives the probability distribution of X .

Probability functions are also known as “probability distribution functions”.

Note the requirements for a function to qualify as the probability function of a discrete random variable:

$$f_X(x) \geq 0 \quad \text{for all } x \text{ within the range of } X$$

$$\sum_x f_X(x) = 1$$


Example 3.3

Write down the probability function for X as defined in Example 3.1.

Solution

$$f_X(0.1) = P(X = 0.1) = \frac{3}{8}$$

$$f_X(0.15) = P(X = 0.15) = \frac{2}{8} = \frac{1}{4}$$

$$f_X(0.2) = P(X = 0.2) = \frac{3}{8}$$

**Question 3.1**

Write down the probability function for N , the number of “tails” obtained when three fair coins are tossed.

1.5 Cumulative distribution functions

The cumulative distribution or cumulative distribution function (CDF) of X is also very important:

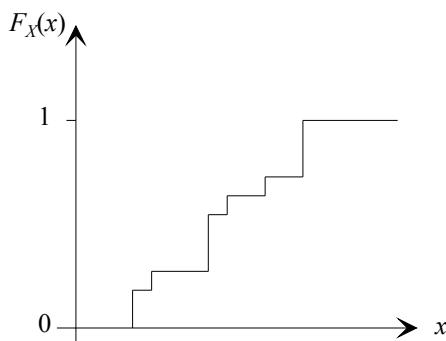
$$F_X(x) = P(X \leq x)$$

gives the probability that X assumes a value that does not exceed x .

Note that CDF is sometimes abbreviated to just DF (or distribution function). This is the notation used in the *Tables*.

The function $F_X(x)$ is defined for all real values of x . The graph of $F_X(x)$ against x starts at a height of 0 then increases by jumps as values of x are reached for which $P(X = x)$ is positive. Once all possible values are included $F_X(x)$ takes its maximum value of 1. $F_X(x)$ is called a step function.

The graph of such a step function could look like this:





Example 3.4

What is the CDF of X in Example 3.1?

Solution

$$F_X(0) = P(X \leq 0) = 0$$

$$F_X(0.1) = P(X \leq 0.1) = \frac{3}{8}$$

$$F_X(0.15) = P(X \leq 0.15) = \frac{3}{8} + \frac{2}{8} = \frac{5}{8}$$

$$F_X(0.2) = P(X \leq 0.2) = \frac{5}{8} + \frac{3}{8} = 1$$

This is generally written as:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0.1 \\ \frac{3}{8} & \text{for } 0.1 \leq x < 0.15 \\ \frac{5}{8} & \text{for } 0.15 \leq x < 0.2 \\ 1 & \text{for } 0.2 \leq x \end{cases}$$



Question 3.2

A random sample of four policyholders is taken from a group of eight, comprising three men and five women. Determine the probability function of X , the number of female policyholders in the sample, and calculate $F_X(2)$.

2 Continuous random variables

2.1 Definition

The range of a continuous random variable is an interval (or a collection of intervals) on the real line:

eg $\{x: x > 0\}$ or $\{x: -\infty < x < \infty\}$ or $\{x: 0 < x < 1\}$



Question 3.3

How does this differ from the range of a discrete random variable?

2.2 Probability density function

The probability associated with an interval of values, (a, b) say, is represented as $P(a < X < b)$ or $P(a \leq X \leq b)$ – these have the same value – and is the area under the curve of the probability density function (PDF) from a to b . So probabilities can be evaluated by integrating the PDF $f_X(x)$. This relationship defines the PDF.

Thus:

$$P(a < X < b) = \int_a^b f_X(x) dx$$

The conditions for a function to serve as a PDF are as follows:

$$f_X(x) \geq 0 \quad -\infty \leq x \leq \infty$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

Here the limits are given as $\pm \infty$, but in real life you would use the values of x that are applicable in the context. For example, if you were given that $a < x < b$, your limits would be a and b .

You should have noticed that these conditions are equivalent to those of the probability function for a discrete random variable, where summation is replaced by integration for the continuous case.

2.3 Cumulative distribution function

The cumulative distribution function (CDF) is defined to be the function:

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

For a continuous random variable, $F_X(x)$ is a continuous, non-decreasing function, defined for all real values of x .

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

Note that because we're using x in the upper limit, we cannot also use it inside the integral. We have changed the notation inside the integral from $f_X(x)$ to $f_X(t)$ and we're integrating this from $t = -\infty$ to $t = x$ to obtain the required probability.



Example 3.5

The continuous random variable W has the PDF $f_W(w) = 12w^2(1-w)$, for $0 < w < 1$. Calculate $P(W < \frac{1}{2})$ and determine an expression for $F_W(w)$.

Solution

The probability is found by integrating the PDF over the relevant range of values, which here are 0 to $\frac{1}{2}$.

$$\begin{aligned} P(W < \frac{1}{2}) &= \int_0^{\frac{1}{2}} 12w^2(1-w) dw \\ &= \left[4w^3 - 3w^4 \right]_0^{\frac{1}{2}} \\ &= 4 \times \frac{1}{8} - 3 \times \frac{1}{16} = \frac{5}{16} \end{aligned}$$

$F_W(w)$, or the CDF is also found by integration, this time with an upper limit of w :

$$F_W(w) = \int_0^w 12t^2(1-t) dt = \left[4t^3 - 3t^4 \right]_0^w = w^3(4-3w)$$

**Question 3.4**

A continuous random variable has the probability density function $f_X(x) = ke^{-2x}$, for $x > 0$. Calculate k , and $P(X < 5.27)$.

3 Expected values

Expected values are numerical summaries of important characteristics of the distributions of random variables.

3.1 Mean

$E[X]$ gives a “typical” value of X in that it is a measure of the average/centre/location/level of the distribution of X . It is called the mean of the distribution of X , or just the mean of X , and is usually denoted μ .

In other words, expectation and mean are the same thing. Very often in statistics we use μ for mean.

$E[X]$ is calculated by summing (discrete case) or integrating (continuous case) the product:

value \times probability of assuming that value

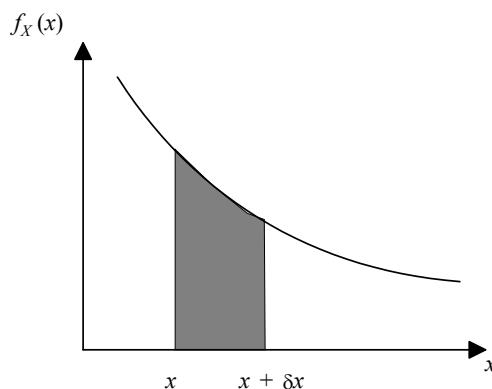
over all values which X can assume.

Thus:

$$E[X] = \sum_x xf_x(x) \text{ for the discrete case}$$

and:

$$E[X] = \int_{-\infty}^{\infty} xf_X(x) dx \text{ for the continuous case}$$



Note that the $f_X(x) dx$ component of the integral is just the probability of getting a value of X in the interval $(x, x + \delta x)$.

**Example 3.6**

Calculate the mean of N defined in Question 3.1, ie the number of tails obtained when three coins are tossed.

Solution

The mean is calculated as the value of N multiplied by the probability of getting that value, and then summing over all possible values of N .

$$E(N) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5$$

Comment:

It is important to remember two things here. Firstly the value obtained need not be a whole number. The average number of children in a family is 2.2 even though you cannot get 0.2 of a child. Also since our distribution was symmetrical, we know that the mean is in the middle, hence a mean of 1.5 here.

**Question 3.5**

Calculate the expected value of the number of female policyholders in Question 3.2.

**Example 3.7**

Calculate $E(W)$ for the random variable given in Example 3.5 with PDF:

$$f_W(w) = 12w^2(1-w) \quad 0 < w < 1$$

Solution

The mean is calculated as.

$$\int_0^1 w f_W(w) dw = \int_0^1 12w^3(1-w) dw = \int_0^1 12w^3 - 12w^4 dw = \left[3w^4 - \frac{12}{5}w^5 \right]_0^1 = 0.6$$



Question 3.6

Calculate $E(X)$ for the random variable given in Question 3.4 with PDF:

$$f_X(x) = ke^{-2x} \quad x > 0$$

For functions of random variables, the same method of summing or integrating the product of the value of the function and the probability of assuming that value can be used.

In other words, to calculate the expected value of a function, say $g(x)$, we replace the value x in the above formula by $g(x)$.

Thus:

$$E[g(x)] = \sum_x g(x)f_x(x) \text{ for the discrete case}$$

and:

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f_x(x) dx \text{ for the continuous case}$$

Note: Expected values do not necessarily exist – if the sum or integral required is not well-defined or not finite then the expected value concerned does not exist.



Question 3.7

A random variable X has PDF:

$$f_X(x) = \frac{1}{9}x^2 \quad 0 < x < 3$$

Calculate $E(2X + 1)$.

3.2 Variance and standard deviation

The variance σ^2 is a measure of the spread/dispersion/variability of the distribution. Specifically it is a measure of the spread of the distribution about its mean.

Formally, $\text{var}[X] = E[\{X - E[X]\}^2]$ is the expected value (or mean) of the squared deviation of X from its mean. The standard deviation, σ , is the positive square root of this – hence the term sometimes used “root mean squared deviation”.

Variance can also be abbreviated to $\text{var}(X)$. Both notations will be used in this course to enable you to be familiar with both versions.

Simplifying: $\text{var}[X] = E[X^2] - \mu^2$

We will prove that this formula is indeed a simplification of $E[\{X - E[X]\}^2]$ in Section 3.3. Some authors write $[E(X)]^2$ as $E^2[X]$.

Note: The units in which the variance is expressed are squared units – the units of the standard deviation (SD) are the same as those of the variable itself. If X is a measurement in cm, then σ^2 is in cm^2 , and σ is in cm.



Example 3.8

Calculate the variance of the number of tails obtained when three coins are tossed. You may wish to refer back to Question 3.1 and Example 3.6.

Solution

The mean was found in Example 3.6 to be 1.5. We need to calculate $E(N^2)$, which is:

$$E(N^2) = 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8} = 3$$

This gives $\text{var}(N) = 3 - 1.5^2 = 0.75$.



Example 3.9

Calculate $\text{var}(W)$ for the random variable given in Example 3.5 with PDF:

$$f_W(w) = 12w^2(1-w) \quad 0 < w < 1$$

Solution

The mean was found to be 0.6.

We need $E[W^2]$, which is given by:

$$E[W^2] = \int_0^1 12w^4(1-w) dw = \int_0^1 12w^4 - 12w^5 dw = \left[\frac{12}{5}w^5 - 2w^6 \right]_0^1 = 0.4$$

The variance is $0.4 - 0.6^2 = 0.04$.



Question 3.8

Calculate $\text{var}(X)$ for the random variable given in Question 3.4:

$$f_X(x) = 2e^{-2x} \quad x > 0$$

3.3 Linear functions of X

Now consider changing the origin and the scale of X .

Let $Y = aX + b$. Let $E[X] = \mu$.

$$E[Y] = E[aX + b] = a\mu + b$$

$$\text{So } Y - E[Y] = aX + b - [a\mu + b] = a[X - \mu].$$

$$\text{Hence } \text{var}[Y] = E[(Y - E[Y])^2] = a^2 E[(X - \mu)^2] = a^2 \text{var}[X].$$

These are important results. The result for the expected value can be thought of simply as “whatever you multiply the random variable by or add to it, you do the same to the mean”.

However, the addition of a constant to a random variable does not alter the variance. This should make sense since the variance is a measure of spread and the spread is not altered when the same constant is added to all values. When you multiply the random variable by a constant you multiply the standard deviation by the same value, so the variance is multiplied by that constant squared. If this doesn't seem obvious you could try this on a few simple examples.



Question 3.9

If a random variable X has a mean of 3 and standard deviation of 2, calculate:

(i) $E[2X - 4]$

(ii) $\text{var}[3X + 2]$

(iii) $\text{var}[3 - 4X]$

(iv) $E\left[\frac{3X + 2}{4}\right]$

Note that if X has mean μ and standard deviation σ , then $Z = \frac{X - \mu}{\sigma}$, the standardised version of X , has mean 0 and standard deviation 1.

This relationship will prove to be very useful when we look at the Normal distribution later in this chapter.

If we now want to extend the result for expectations, we can say:

In general:

$$E\left[\sum_i c_i g_i(X)\right] = \sum_i c_i E[g_i(X)]$$



Example 3.10

Prove that $\text{var}(X) = E[X^2] - \mu^2$.

Solution

Let $E(X) = \mu$. From the definition of variance, we have $\text{var}(X) = E[(X - \mu)^2]$.

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

This is the proof of the formula for variance that was given earlier.

3.4 Moments

Moments are just particular expected values that summarise features of a distribution.

$E[(X - c)^k]$ is the k th moment or k th-order moment of X about c . Moments about the mean are called central moments.

A non-central moment refers to a moment about zero. Notice the similarity between this result and the sample moments defined in Chapter 1, Section 3.2.

$E[X]$ ($= \mu$), is a first-order moment, which provides information on the average value.

$E[X]$ and $E[X^2]$, moments “up to second order” provide information on spread – this information is used in calculating the standard deviation.

$E[X]$, $E[X^2]$, and $E[X^3]$, moments “up to third order”, provide information on one aspect of the shape of a distribution, namely skewness.

The third central moment, μ_3 , can be expanded as follows:

from
$$(X - \mu)^3 = X^3 - 3\mu X^2 + 3\mu^2 X - \mu^3$$

comes
$$\mu_3 = E[(X - \mu)^3] = E[X^3] - 3\mu E[X^2] + 2\mu^3$$

using the rules for the use of the expectation operator $E[]$ noted earlier.

You can use the unexpanded version in questions *ie* calculating $E[(X - \mu)^3]$ from first principles. If there are only a few values of X in the question, this may be quicker than expanding $(X - \mu)^3$.



Question 3.10

Show how this formula for the third central moment is derived.

The sign of μ_3 gives the direction of the skew. Positive skewness corresponds to $\mu_3 > 0$ while symmetry corresponds to $\mu_3 = 0$.

And, of course, negative skewness corresponds to $\mu_3 < 0$.

The usual measure of the skewness of a distribution or random variable based on moments is the coefficient of skewness given by $\frac{\mu_3}{\sigma^3}$. Note that it is a dimensionless measure (there are no units, since the units of measurement of the numerator and denominator are the same).

This is an important result to remember as exam questions often ask for the coefficient of skewness.



Question 3.11

What can you say about the first and second central moments, μ_1 and μ_2 , of a random variable X ?

3.5 Other quantities

Just as with data, there are other quantities that summarise features of a distribution. For example, an alternative to the mean is the median, defined for any random variable X as m , such that:

$$P(X < m) \leq 0.5 \leq P(X \leq m)$$

Recall in Chapter 1 the sample median was the middle data value. The population median is simply the value in the middle of the distribution.

In particular, if X is continuous the median m is defined as the solution of:

$$P(X < m) = \int_{-\infty}^m f_X(x) dx = 0.5$$

Similarly quartiles can also be defined in an obvious way.

The lower and the upper quartiles of a random variable X are the values q_1 and q_3 such that:

$$P(X < q_1) = 0.25 \quad P(X < q_3) = 0.75$$

And so the interquartile range, IQR , (an alternative to the standard deviation as a measure of spread) is given by:

$$IQR = q_3 - q_1$$



Question 3.12

For the random variable, X , given in Question 3.4 with PDF:

$$f_X(x) = 2e^{-2x} \quad x > 0$$

calculate the:

- (a) median
- (b) interquartile range.

4 Functions of a random variable

When deriving the probability distribution of Y , where Y is a function of X , $Y = u(X)$ say, from the known distribution of X , the procedure is straightforward for discrete variables but for continuous variables there are several methods available. Methods are illustrated in this section based first on the distribution function $F_x(x)$, and second directly on the PDF $f_x(x)$. Often both methods can be used, although one may be easier and quicker to carry out than the other.

F and f will denote the CDF and PF/PDF respectively of Y ; F_x and f_x will denote the CDF and PF/PDF respectively of X .

4.1 Discrete random variables

All that has to be done is to find the value y of Y which corresponds to the value (or values) x of X – the probabilities in the given distribution of X are unaffected.

For example suppose that X has the following distribution:

X	0	1	2
$P(X = x)$	0.3	0.5	0.2

Let $Y = 2X + 1$. Then it will have the following distribution:

Y	1	3	5
$P(Y = y)$	0.3	0.5	0.2



Question 3.13

The random variable X has probability function:

$$P(Z = z) = \binom{n}{z} \mu^z (1 - \mu)^{n-z} \quad z = 0, 1, \dots, n$$

where n is a positive integer and $0 < \mu < 1$.

Determine the probability function of $Y = Z/n$.

4.2 Continuous random variables

$F(Y) = P(Y \leq y)$ [$= P(Y < y)$] is required. The event “ $Y < y$ ” is equivalent to the event “ $u(X) < y$ ” (as $Y = u(X)$), so this latter event is expressed in terms of the values which X must take. Then $P[u(X) < y]$ can be found from F_x . We rearrange the formula to get $P[X < u^{-1}(y)]$ and then since $F_X(x) = P[X < x]$, we just need to find $F_X(u^{-1}(y))$. This can be done either by using the F_X formula from the *Tables* or by integration.

Having found $F(y)$, obtain the PDF $f(y)$, if required, by differentiation. $F(y)$ and/or $f(y)$ may be recognisable as the distribution function/density function of a “standard” distribution, eg uniform, gamma, normal. These standard distributions can be found in Chapter 4. Alternatively, their PDFs are listed in the *Tables*.

For example, if $y = u(x)$ is such that a unique inverse $x = w(y) = u^{-1}(y)$ exists and $u(x)$ is an increasing function, then:

$$F(y) = P(Y < y) = P[u(X) < y] = P[X < w(y)] = F_x[w(y)]$$

and hence differentiating using the chain rule $f(y) = f_x[w(y)] \frac{dw(y)}{dy}$

In the case that $u(x)$ is decreasing:

$$F(y) = P(Y < y) = P[u(X) < y] = P[X > w(y)] = 1 - F_x[w(y)]$$

and $f(y) = -f_x[w(y)] \frac{dw(y)}{dy}$.

As $\frac{dw(y)}{dy}$ is negative in this case, both cases are summed up in the result:

$$f(y) = f_x[w(y)] \left| \frac{dw(y)}{dy} \right|$$



Example 3.11

The random variable T has an exponential distribution (see page 11 *Tables*) with PDF:

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0$$

By considering the distribution function, determine the distribution of $U = e^{-\lambda T}$.

Solution

Consider the distribution function of U :

$$F_U(u) = P(U \leq u) = P(e^{-\lambda T} \leq u) = P(-\lambda T \leq \ln u) = P\left(T \geq -\frac{\ln u}{\lambda}\right)$$

Expressing this as an integral:

$$\int_{-\frac{1}{\lambda} \ln u}^{\infty} \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t} \right]_{-\frac{1}{\lambda} \ln u}^{\infty} = 0 - (-u) = u$$

Alternatively, we could use the distribution function of the exponential distribution given on page 11 of the Tables:

$$P\left(T \geq -\frac{\ln u}{\lambda}\right) = 1 - P\left(T \leq -\frac{\ln u}{\lambda}\right) = 1 - F_T\left(-\frac{\ln u}{\lambda}\right) = 1 - \left(1 - e^{-\lambda \times -\frac{\ln u}{\lambda}}\right) = u$$

Having determined $F_U(u)$, we need to differentiate to obtain $f_U(u)$:

$$f_U(u) = F'_U(u) = 1.$$

Since T can take values in the range $0 < T < \infty$, U can take values in the range $0 < U < 1$. So U has a $U(0,1)$ distribution.

Alternatively, jumping straight to the formula $f_U(u) = f_T[w(u)] \times \left| \frac{d}{du} w(u) \right|$ we obtain:

$$u = e^{-\lambda t} \Rightarrow t = w(u) = -\frac{1}{\lambda} \ln u \Rightarrow \frac{d}{du} w(u) = -\frac{1}{\lambda u}$$

$$\Rightarrow f_U(u) = \lambda e^{-\lambda \times -\frac{1}{\lambda} \ln u} \times \left| -\frac{1}{\lambda u} \right| = e^{\ln u} \times \frac{1}{u} = u \times \frac{1}{u} = 1$$

**Question 3.14**

- (i) Determine the cumulative distribution function for the random variable having the PDF:

$$f(x) = 2\beta x e^{-\beta x^2}, \quad x > 0$$

where β is a positive constant.

- (ii) Hence, derive the PDF of $Y = X^2$ if X has the distribution above.

The method used above relies on calculating probabilities using a distribution function or integration. However, some distributions have complicated PDFs so that the integral needed to calculate the distribution functions can not be evaluated analytically. For example:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

For these an alternative method is required which is listed in the appendix to this chapter.

5 Exam question



Past Exam Question (Subject C1, April 1994, Q11)

Let X have a uniform distribution on $(0,1)$, so the probability density function of X is

$$f(x) = 1, \quad 0 < x < 1$$

and let θ be a positive constant.

- (i) Derive the probability density function of $Y = -\theta \ln X$.
- (ii) Determine $E(Y)$.

6 Appendix

This section contains an alternative method of calculating the PDF of a function of a random variable where the PDF is not easy to integrate.

We do this by considering the probability integral. We have $\int_{x_{\min}}^{x_{\max}} f_X(x) dx = 1$, and we

can then make the substitution $y = u(x)$ into the integral to get an integral of the form

$\int_{y_{\min}}^{y_{\max}} f_Y(y) dy = 1$. $f_Y(y)$ will then represent the PDF of Y , with the appropriate limits

on y being y_{\max} and y_{\min} .



Example 3.12

The random variable T has PDF:

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0$$

By considering the probability integral, derive the distribution of $U = e^{-\lambda T}$.

Solution

The probability integral for T is $\int_0^{\infty} \lambda e^{-\lambda t} dt = 1$.

If we substitute $u = e^{-\lambda t}$ we obtain:

$$1 = \int_0^{\infty} \lambda e^{-\lambda t} dt = \int_1^0 -du = \int_0^1 1 du$$

So U has PDF $f_U(u) = 1$, and takes values in the range $0 < u < 1$, ie it has a $U(0,1)$ distribution.

This page has been left blank so that you can keep the chapter summaries together for revision purposes.



Chapter 3 Summary

Random variables are used to model features of a population using probabilities.

A discrete random variable has a probability function (PF), $P(X = x)$. This defines how the probability is split between the values the variable can take. The PF satisfies:

$$\sum_x P(X = x) = 1 \quad \text{and} \quad P(X = x) \geq 0$$

A continuous random variable has a probability density function (PDF), $f_X(x)$. The PDF satisfies:

$$\int_x f_X(x) dx = 1 \quad \text{and} \quad f_X(x) \geq 0$$

We can use the PDF to obtain probabilities as follows:

$$P(a < X < b) = \int_a^b f_X(x) dx .$$

The cumulative distribution function (CDF), for both discrete and continuous random variables is given by:

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

For continuous random variables $F'_X(x) = f_X(x)$.

Using formulae given overleaf, we can calculate the:

- mean μ
- variance σ^2
- skewness μ_3

and other central and non-central moments of a random variable.



Chapter 3 Formulae

Population mean (or expectation)

$$\mu = E(X) = \sum_x xP(X=x) \quad \text{or} \quad \int_{-\infty}^{\infty} xf_X(x) dx$$

$$E[g(X)] = \sum_x g(x)P(X=x) \quad \text{or} \quad \int_{-\infty}^{\infty} g(x)f_X(x) dx$$

Population variance

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

Population skewness

$$\mu_3 = \text{skew}(X) = E[(X - \mu)^3] = E[X^3] - 3\mu E[X^2] + 2\mu^3$$

$$\text{coefficient of skewness} = \frac{\mu_3}{\sigma^3}$$

Population moments

$$k\text{th moment} = E[X^k]$$

$$k\text{th moment about } c = E[(X - c)^k]$$

$$k\text{th central moment} = E[(X - \mu)^k]$$

Population median and IQR

m such that $P(X < m) = 0.5$

$IQR = q_3 - q_1$ where $P(X < q_1) = 0.25$ and $P(X < q_3) = 0.75$

Linear functions of X

$$E[aX + b] = aE[X] + b$$

$$\text{var}[aX + b] = a^2 \text{ var}[X]$$

Functions of a random variable

$$Y = u(X) \Rightarrow f_Y(y) = f_X[u^{-1}(y)] \left| \frac{du^{-1}(y)}{dy} \right|$$

Chapter 3 Solutions

Solution 3.1

There are eight possible outcomes, which are equally likely, namely TTT, TTH, THT, HTT, THH, HTH, HHT, HHH.

This gives us:

$$f_N(0) = P(N = 0) = \frac{1}{8}$$

$$f_N(1) = P(N = 1) = \frac{3}{8}$$

$$f_N(2) = P(N = 2) = \frac{3}{8}$$

$$f_N(3) = P(N = 3) = \frac{1}{8}$$

Solution 3.2

X can take the values 1 to 4. To find the probabilities we need to use combinations.

Total number of groups of 4 that could be picked = ${}^8C_4 = 70$.

We have:

$$P(X = 1) = \frac{{}^3C_3 {}^5C_1}{{}^8C_4} = \frac{1}{14}$$

$$P(X = 2) = \frac{{}^3C_2 {}^5C_2}{{}^8C_4} = \frac{3}{7}$$

$$P(X = 3) = \frac{{}^3C_1 {}^5C_3}{{}^8C_4} = \frac{3}{7}$$

$$P(X = 4) = \frac{{}^3C_0 {}^5C_4}{{}^8C_4} = \frac{1}{14}$$

Recall that ${}^nC_r = \frac{n!}{(n-r)!r!}$ where $n! = n \times (n-1) \times (n-2) \times \dots \times 1$.

Combinations are assumed knowledge for the actuarial exams and if you are unfamiliar with them, we would advise you to refer to Stats Pack or an A-level (or equivalent) textbook.

So the probability function of X is given by:

X	1	2	3	4
$P(X = x)$	$\frac{1}{14}$	$\frac{3}{7}$	$\frac{3}{7}$	$\frac{1}{14}$

Notice that the probabilities add to 1.

We can then calculate $F_X(2) = P(X \leq 2) = 0.5$.

Solution 3.3

The range of a discrete random variable is a finite (or countably infinite) set of numbers.

Solution 3.4

If X is a random variable, then $\int f_X(x) dx = 1$.

Here:

$$\int_0^\infty k e^{-2x} dx = 1 \Rightarrow \left[-\frac{k}{2} e^{-2x} \right]_0^\infty = 1 \Rightarrow \frac{k}{2} = 1 \Rightarrow k = 2$$

$$P(X < 5.27) = \int_0^{5.27} k e^{-2x} dx = \left[-\frac{k}{2} e^{-2x} \right]_0^{5.27}$$

But since $k = 2$:

$$P(X < 5.27) = -e^{-10.54} + 1 = 0.99997$$

Solution 3.5

Calculating this from the formula gives:

$$E[X] = 1 \times \frac{1}{14} + 2 \times \frac{3}{7} + 3 \times \frac{3}{7} + 4 \times \frac{1}{14} = 2.5$$

Alternatively, we could write this down immediately since it is a symmetrical distribution.

Solution 3.6

$$E[X] = \int_0^{\infty} 2xe^{-2x} dx$$

Using integration by parts, where $u = x$, we obtain:

$$E[X] = \int_0^{\infty} 2xe^{-2x} dx = \left[-xe^{-2x} \right]_0^{\infty} + \int_0^{\infty} e^{-2x} dx = \left[-xe^{-2x} \right]_0^{\infty} + \left[-\frac{1}{2}e^{-2x} \right]_0^{\infty} = \frac{1}{2}$$

Solution 3.7

Using $E[g(X)] = \int g(x)f(x) dx$, we obtain:

$$\begin{aligned} E(2X + 1) &= \int (2x + 1)f(x) dx \\ &= \int_0^3 (2x + 1) \frac{1}{9}x^2 dx = \int_0^3 \frac{2}{9}x^3 + \frac{1}{9}x^2 dx = \left[\frac{2}{36}x^4 + \frac{1}{27}x^3 \right]_0^3 = 5.5 \end{aligned}$$

Solution 3.8

We need to determine $E[X^2]$ first. Using integration by parts with $u = x^2$, we obtain:

$$E[X^2] = \int_0^\infty 2x^2 e^{-2x} dx = \left[-x^2 e^{-2x} \right]_0^\infty + \int_0^\infty 2xe^{-2x} dx$$

The first term in this expression is zero.

The last integral is just $E[X]$, and using the fact that $E[X] = \frac{1}{2}$ from Solution 3.6, we have:

$$\text{var}[X] = E[X] - E^2[X] = 0.25$$

Solution 3.9

$$(i) \quad E[2X - 4] = 2E[X] - 4 = 2$$

$$(ii) \quad \text{var}(3X + 2) = 9 \text{ var}[X] = 36$$

$$(iii) \quad \text{var}(3 - 4X) = 16 \text{ var}[X] = 64$$

$$(iv) \quad E\left[\frac{3X + 2}{4}\right] = \frac{3}{4}E[X] + \frac{1}{2} = 2.75$$

Solution 3.10

$$\begin{aligned} \mu_3 &= E[(X - \mu)^3] = E[X^3 - 3\mu X^2 + 3\mu^2 X - \mu^3] \\ &= E[X^3] - E[3\mu X^2] + E[3\mu^2 X] - E[\mu^3] \\ &= E[X^3] - 3\mu E[X^2] + 3\mu^2 E[X] - E[\mu^3] \\ &= E[X^3] - 3\mu E[X^2] + 3\mu^3 - \mu^3 \\ &= E[X^3] - 3\mu E[X^2] + 2\mu^3 \end{aligned}$$

Solution 3.11

$$\mu_1 = E[X - \mu] = E[X] - \mu = 0$$

So μ_1 is always zero.

$$\mu_2 = E[(X - \mu)^2] = E[X^2] - \mu^2 = \text{var}(X)$$

So μ_2 is just the variance.

Solution 3.12

- (a) The median, m , is the value such that:

$$P(X < m) = \int_0^m 2e^{-2x} dx = 0.5$$

Integrating gives:

$$\left[-e^{-2x} \right]_0^m = 1 - e^{-2m} = 0.5 \Rightarrow m = -\frac{1}{2} \ln 0.5 = 0.3466$$

- (b) Similarly the lower and upper quartiles are the values q_1 and q_3 such that:

$$P(X < q_1) = \int_0^{q_1} 2e^{-2x} dx = \left[-e^{-2x} \right]_0^{q_1} = 1 - e^{-2q_1} = 0.25$$

$$P(X < q_3) = \int_0^{q_3} 2e^{-2x} dx = \left[-e^{-2x} \right]_0^{q_3} = 1 - e^{-2q_3} = 0.75$$

Hence:

$$q_1 = -\frac{1}{2} \ln 0.75 = 0.1438$$

$$q_3 = -\frac{1}{2} \ln 0.25 = 0.6931$$

$$IQR = 0.6931 - 0.1438 = 0.5493$$

Solution 3.13

The probability function of Y is given by:

$$P(Y = y) = P\left(\frac{Z}{n} = y\right) = P(Z = ny)$$

So using the PF of Z given in the question:

$$P(Y = y) = P(Z = ny) = \binom{n}{ny} \mu^{ny} (1 - \mu)^{n-ny}$$

The original PF was for $z = 0, 1, \dots, n$ whereas now it is for $y = 0, \frac{1}{n}, \dots, 1$.

Solution 3.14

The cumulative distribution function is:

$$F(x) = \int_0^x 2\beta t e^{-\beta t^2} dt = \left[-e^{-\beta t^2} \right]_0^x = 1 - e^{-\beta x^2}$$

So the distribution function of Y is:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = P(X \leq \sqrt{y})$$

since X can only take positive values. This is equal to $F_X(\sqrt{y}) = 1 - e^{-\beta y}$.

Therefore:

$$f_Y(y) = F'_Y(y) = \beta e^{-\beta y}$$

In Chapter 4, we will see that this is the PDF of an exponential distribution with parameter β .

Past Exam Question (Subject C1, April 1994, Q11)

- (i) Consider the distribution function of Y :

$$F_Y(y) = P(Y \leq y) = P(-\theta \ln X \leq y) = P\left(\ln X \geq -\frac{y}{\theta}\right) = P\left(X \geq e^{-\frac{y}{\theta}}\right)$$

Expressing this as an integral:

$$\int_{e^{-\frac{y}{\theta}}}^1 1 dx = [x]_{e^{-\frac{y}{\theta}}}^1 = 1 - e^{-\frac{y}{\theta}}$$

Alternatively, we could use the uniform distribution function given on page 13 of the Tables.

We then need to differentiate to obtain $f_Y(y)$:

$$f_Y(y) = F'_Y(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}}.$$

Since X can take values in the range $0 \leq X \leq 1$, Y can take values in the range $0 \leq Y < \infty$.

Alternatively, using the formula $f_Y(y) = f_X[w(y)] \times \left| \frac{d}{dy} w(y) \right|$ we obtain:

$$\begin{aligned} y = -\theta \ln x &\Rightarrow x = e^{-\frac{y}{\theta}} \Rightarrow \frac{dx}{dy} = -\frac{1}{\theta} e^{-\frac{y}{\theta}} \\ &\Rightarrow f_Y(y) = 1 \times \left| -\frac{1}{\theta} e^{-\frac{y}{\theta}} \right| = \frac{1}{\theta} e^{-\frac{y}{\theta}} \end{aligned}$$

- (ii) We have $f(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}}$, $0 \leq y < \infty$, so the mean is given by:

$$E(Y) = \int y f(y) dy = \int_0^\infty \frac{1}{\theta} y e^{-\frac{y}{\theta}} dy$$

Using integration by parts this gives:

$$E(Y) = \left[-ye^{-\frac{y}{\theta}} \right]_0^\infty + \int_0^\infty e^{-\frac{y}{\theta}} dy = 0 + \left[-\theta e^{-\frac{y}{\theta}} \right]_0^\infty = 0 - (-\theta) = \theta$$

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.

Chapter 4

Probability distributions



Syllabus objectives

- (iii) 4. Evaluate probabilities (by calculation or by referring to tables as appropriate) associated with distributions.
- (v) 1. Define and be familiar with the discrete distributions: geometric, binomial, negative binomial, hypergeometric and uniform on a finite set.
- 2. Define and be familiar with the continuous distributions: normal, lognormal, exponential, gamma, chi-square, t, F, beta and uniform on an interval.
- 3. Define a Poisson process and note the connection between Poisson processes and the Poisson distribution, and that a Poisson process may be equivalently characterised as: (1) the distribution of waiting times between events, (2) the distribution of process increments and (3) the behaviour of the process over an infinitesimal time interval.
- 4. Generate basic discrete and continuous random variables using simulation methods.

0 Introduction

Chapter 3 looked at random variables. In this chapter, we will consider some important distributions.

All of the relevant formulae for the values you need for these distributions are contained in the Formulae and Tables for Examinations (usually simply denoted *Tables*). Hence you should concentrate on being able to *apply* them to calculate the mean, variance, coefficient of skewness and probabilities.

Also, in order to find many of the probabilities you will require statistical tables. These are available from the Profession and you should purchase a copy as soon as possible as they are essential to your Subject CT3 course. Some of the tables have been reproduced at the end of this chapter so that you are able to progress through the course. However, these are *not* the same format as the *Tables*, which are the ones that will be available to you in the examination. Therefore, it would be wise to get hold of a set promptly so that you are thoroughly familiar with 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.

Whilst the Core Reading *does* cover all that you need to know, it is our experience that students who have not studied any statistics before can find the pace and brevity of this chapter an obstacle to gaining a sufficient grasp. To prevent such students being disadvantaged, we have developed the Stats Pack, which covers the basics at a slower pace with plenty of examples. See http://www.acted.co.uk/Html/paper_stats_pack.htm or contact StatsPack@bpp.com for further details.

1 **Important discrete distributions**

In this section we will look at the standard discrete distributions that we will use for modelling and their properties.

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 in this section 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 $S = \{1, 2, 3, \dots, k\}$.

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

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

$$\text{Distribution: } P(X = x) = \frac{1}{k} \quad (x = 1, 2, 3, \dots, k)$$

Moments:

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

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

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

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

**Question 4.1**

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

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: $P(\{s\}) = \theta$, $P(\{f\}) = 1 - \theta$

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

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

Moments: $\mu = \theta$

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

A Bernoulli variable is also called an “indicator” variable – its value can be used to indicate whether or not some specified event, A say, occurs. Set $X = 1$ if A occurs, 0 if A does not occur. If $P(A) = \theta$ 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 if s/he dies.

For example, if X is the number of sixes obtained when a fair die is thrown, $\theta = \frac{1}{6}$, $(1 - \theta) = \frac{5}{6}$ and $P(X = 0) = \frac{5}{6}$ and $P(X = 1) = \frac{1}{6}$.

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 $P(\{s\}) = \theta$.

Such a sequence is called a “sequence of n independent, identical, Bernoulli (θ) trials” or, for short, a “sequence of n Bernoulli (θ) trials”.

A quick way of saying independent and identically distributed is IID. You 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.

$$\text{Distribution: } P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad x = 0, 1, 2, \dots, n; \quad 0 < \theta < 1$$

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

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

The fact that a $Bin(n, \theta)$ distribution is the sum of n independent and identical Bernoulli (θ) trials is important and will be used in Chapter 6 to prove some important results.

Moments: $\mu = n\theta$

$$\sigma^2 = n\theta(1-\theta)$$

Very often, when using the binomial distribution, θ will be written as p , the probability of success, and rather than use $1-\theta$, we would use q .

For example, if 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 ${}^{10}C_1 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^9 = 0.323$. Notice that there are 10 $({}^{10}C_1)$ ways of obtaining exactly one “six”, ie the “six” could be on the first throw, the second throw, or the tenth throw.



Question 4.2

What is 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%?

1.4 Geometric distribution

Consider again a sequence of independent, identical Bernoulli trials with $P(\{s\}) = \theta$. The variable of interest now is the number of trials that have 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 $P(X = x) = \theta(1-\theta)^{x-1}$, $x = 1, 2, 3, \dots$
 $(0 < \theta < 1)$

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

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

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 4.3

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

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

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, you will need to remember that $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

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

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

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 4.4

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

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

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



Question 4.5

What is the variance for this formulation?

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, X might be the number of phone calls required to make the fifth sale.

$$\text{Distribution: } P(X = x) = \binom{x-1}{k-1} \theta^k (1-\theta)^{x-k} \quad x = k, k+1, \dots; \quad 0 < \theta < 1$$

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

This distribution satisfies the recurrence relationship:

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

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

In this model k must be a positive integer.

$$\text{Moments: } \mu = \frac{k}{\theta}$$

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

Note: The mean and variance are just k times those for the geometric (θ) 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 4.6

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

Another formulation of the negative binomial distribution is sometimes used.

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

$$\text{Then } P(Y = y) = \binom{k + y - 1}{y} \theta^k (1 - \theta)^y, \quad y = 0, 1, 2, 3, \dots, \text{ with mean } \mu = \frac{k(1 - \theta)}{\theta}.$$

$Y = X - k$, 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 in Section 2.2 of this chapter).

Note that for this formulation k does not need to be an integer, we simply require $k > 0$.

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.

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

The mean is given by $\mu = \frac{nk}{N}$, which parallels the “ $\mu = n\theta$ ” result for the binomial distribution – the initial proportion of successes here being $\frac{k}{N}$.

The binomial, with $\theta = \frac{k}{N}$, provides a good approximation to the hypergeometric in many situations.

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

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



Question 4.7

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, what is the probability that none of the group selected have the qualification?

Calculate the answer:

- (i) exactly
- (ii) using the binomial approximation.

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 (*i.e* 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.

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, θ) distributions as $n \rightarrow \infty$ and $\theta \rightarrow 0$ together, such that the mean $n\theta$ is held constant at the value λ . The limit leads to the distribution of the Poisson variable, with parameter λ .

Here $\lambda = n\theta$.

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

This distribution satisfies 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:

$$n\theta(1 - \theta) = n\left(\frac{\lambda}{n}\right)\left(1 - \frac{\lambda}{n}\right) = \lambda\left(1 - \frac{\lambda}{n}\right) \rightarrow \lambda \text{ as } n \rightarrow \infty$$

This suggests that X has variance λ . This is in fact also the case.

So $\mu = \sigma^2 = \lambda$.



Question 4.8

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



Question 4.9

If the rate at which goals are scored in a game of football is, on average, three every match, calculate the probability that more than 5 goals are scored in a match.

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

We will look at other approximations in Chapter 8.



Question 4.10

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, calculate the probability of exactly 2 such deaths occurring in a given year.

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.



Question 4.11

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.

2 Important continuous distributions

2.1 Uniform distribution

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

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

$X \sim U(\alpha, \beta)$ is often written as shorthand for “the random variable X has a continuous uniform distribution between α and β ”.

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

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



Question 4.12

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

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 4.13

If $Y \sim U(50,150)$, what are $P(Y > 74)$ and $P(50 < Y < 126)$?

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 $Gamma(\alpha, \lambda)$.

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

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$$

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 $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

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

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: $\mu = \frac{\alpha}{\lambda}$, $\sigma^2 = \frac{\alpha}{\lambda^2}$



Question 4.14 (messy)

Prove the formulae given for the mean and variance.



Question 4.15

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

Special case 1: exponential distribution

Gamma with $\alpha = 1$.

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

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

$$\text{Moments: } \mu = \frac{1}{\lambda}, \quad \sigma^2 = \frac{1}{\lambda^2}$$

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



Question 4.16

What is the median of the $Exp(\lambda)$ distribution? (Don't forget that the median is the value of m such that $P(X \leq m) = \frac{1}{2}$.)

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 .

$$\begin{aligned} P(T > t) &= P(0 \text{ events in time } t) \\ &= P(X = 0) \quad \text{where } X \sim \text{Poisson}(\lambda t), \text{ so} \\ &= e^{-\lambda t} \end{aligned}$$

$$P(T < t) = 1 - e^{-\lambda t}$$

$$f_T(t) = \lambda e^{-\lambda t}$$



Question 4.17

Claims to a general insurance company's 24 hour call centre occur at a rate of 3 per hour. What is the probability that the next call arrives after more than $\frac{1}{2}$ hour?

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. So for the exponential distribution we can also show that:

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

This is like saying that the probability that we wait a further 10 minutes given that we have already waited 20 minutes is just the probability of waiting 10 minutes.



Question 4.18

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

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 k th event in a Poisson process with rate λ .

The fact that a $\text{Gamma}(\alpha, \lambda)$ distribution is the sum of α independent and identical $\text{Exp}(\lambda)$ trials is important and will be used in Chapter 6 to prove some important results.

Special case 2: chi-square (χ^2) distribution with parameter “degrees of freedom” v

Gamma with $\alpha = \frac{v}{2}$ where v is a positive integer, and $\lambda = \frac{1}{2}$.

So the PDF of a χ^2 distribution is:

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

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

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

Since integrating the PDF is a bit messy extensive probability tables for the chi-square distribution can be found on pages 164-166 of the *Tables*. The probability tables are not given in the appendix of this chapter, however the percentage points tables are.

Another result that we will use in later work is:

If $W \sim \text{Gamma}(\alpha, \lambda)$, then $2\lambda W$ has a $\chi^2_{2\alpha}$ distribution (ie a chi-square distribution with 2α degrees of freedom).

This 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 parameters, in the *Tables*. If you haven't yet obtained a copy of the *Tables*, recall that the chi-square tables are reproduced at the end of this chapter.

We can prove this result using moment generating functions which you will meet in Chapter 5.



Example 4.1

If the random variable X has a χ^2 distribution with five degrees of freedom, calculate:

- (a) $P(X > 6.5)$
- (b) $P(X < 11.8)$.

Solution

Using the χ^2 probabilities given in the appendix or 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.

The theory behind interpolation is covered more fully in the Foundation ActEd Course. Alternatively, consult a pure mathematics textbook.

We are now going to repeat Question 4.15 using the χ^2 result:



Question 4.19

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

2.3 Beta distribution

This is another versatile family of distributions with 2 positive parameters. The range of the variable is $\{x: 0 < x < 1\}$.

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

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$

The relationship between beta functions and gamma functions is:

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

The PDF of a beta distribution is defined by:

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

Moments:

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

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

The beta distribution is a very 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 fractions, such as the fraction of a batch that is defective or the percentage of claims that are over £1,000.



Question 4.20

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 .

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, θ)

It is also used to approximate the Poisson distribution. Both of these approximations are covered in Chapter 8.

- (iii) it provides a model for the sampling distributions of various statistics – see Chapter 9
- (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 Chapter 12.

- (v) it is a “building block” for many other distributions.

The distribution has 2 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 Chapter 5.

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

We can prove this result using moment generating functions which we will meet in Chapter 5.

The x -values $\mu, \mu + \sigma, \mu + 2\sigma, \mu + 3\sigma$ 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 the same way – transform to standard normal via $z = \frac{X - \mu}{\sigma}$ and look up in the tables.

The table that you will need to use is on pages 160-161 of the *Tables*. If you haven't yet obtained a copy of the *Tables*, recall that the standard normal tables are given at the end of this chapter.

The probabilities in the table are “left hand” probabilities, in other words they give $P(Z < z)$. You will need to use interpolation to get more precise probabilities than the tabulated values, using the method we will use in parts (ii), (iii) and (iv) of Example 4.2.

A brief outline of using interpolation on the normal tables is given in Appendix A. For further details please consult the Stats Pack, FAC or a pure maths text book.

Since Z is symmetrical about zero, it follows that:

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

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



Example 4.2

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

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

The following answers use interpolation on the tables:

- (ii) $P(X > 30) = P(Z > 0.833) = 1 - P(Z < 0.833) = 1 - 0.79758 = 0.20242$
- (iii) $P(X < 20) = P(Z < -0.833) = 1 - P(Z < 0.833) = 1 - 0.79758 = 0.20242$
- (iv) We need to simplify the expression involving the absolute value:

$$\begin{aligned} P(|X - 25| < 4) &= P(-4 < X - 25 < 4) \\ &= P(21 < X < 29) \\ &= P(X < 29) - P(X < 21) \\ &= P(Z < 0.667) - P(Z < -0.667) \\ &= P(Z < 0.667) - [1 - P(Z < 0.667)] \\ &= 0.49522 \end{aligned}$$



Question 4.21

If $X \sim N(14,20)$, calculate:

- (i) $P(X < 14)$
- (ii) $P(X > 20)$
- (iii) $P(X < 9)$
- (iv) r such that $P(X > r) = 0.41294$

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 follow:

95% and 99% intervals:

$$P(Z < 1.96) = 0.97500 \text{ so } P(0 < Z < 1.96) = 0.97500 - 0.5 = 0.47500$$

$$\therefore P(-1.96 < Z < 1.96) = 2 \times 0.47500 = 0.95$$

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 “3- σ limits”. (The range of a large set of observations from a normal distribution is usually about 6 or 7 standard deviations).

Questions on all of these distributions can involve any of the features from this chapter. For example:



Question 4.22

Determine the third non-central moment of the normal distribution with mean 10 and variance 25.

Note: A handy result is that if Z has a standard normal distribution then Z^2 has a χ_1^2 distribution. This can be used to find results easily for $E(Z^2)$ and $\text{var}(Z^2)$.

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 can write $X \sim \log N(\mu, \sigma^2)$.

The PDF of the lognormal distribution is defined by:

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

Notice that the lower limit for x is 0 and not $-\infty$, as it was for the normal distribution. This is because $\log x$ is not defined for values of x below zero.



Question 4.23

If $W \sim \log N(5,6)$, what is $P(W > 3,000)$?

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 4.24

If the mean of the lognormal distribution is 9.97 and the variance is 635.61, what are the parameters μ and σ^2 ?

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 CT1, Chapter 15, to calculate the probabilities associated with accumulating funds.

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.

The PDF of the *t*-distribution is defined by:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } -\infty < x < \infty$$

Calculating probabilities by integrating this PDF is no small task! Fortunately, we will only be expected to look up probabilities using page 163 in the *Tables*. A limited version is reproduced at the end of this chapter.



Question 4.25

Use the *t*-tables to calculate:

- (i) $P(t_{15} < 1.341)$
- (ii) a such that $P(t_8 > a) = 0.01$
- (iii) $P(t_{24} < -0.5314)$

This 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 9, 11 and 12.

2.7 F-distribution

If two independent random variables, X and Y have χ^2 distributions with parameter 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 .

The PDF of the F distribution is defined by:

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \cdot x^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2}x\right)^{-\frac{1}{2}(n_1+n_2)}$$

for $x > 0$ and $f(x) = 0$ elsewhere.

Once again, it is simply too nightmarish to calculate probabilities by integrating this PDF. We find probabilities by using the F -tables given on pages 170-174 of the *Tables*. A limited version is reproduced at the end of this chapter.

The F -distribution is *not* symmetrical. Given that only upper tail probabilities are given in the *Tables*, we will need to 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 9.



Question 4.26

Use the F -tables to calculate:

- (i) $P(F_{5,12} < 3.106)$
- (ii) a such that $P(F_{7,4} > a) = 0.01$

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 9, 11, 12 and 14.

3 **The Poisson process**

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

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

This was 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* is where we let the time vary. So instead of looking at the number of events occurring per unit time, we now look at the number of events occurring in time t .



Question 4.27

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.

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

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



Question 4.28

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.

3.1 Deriving the Poisson process

In Section 1.7, we stated that the Poisson distribution could be modelled from events occurring singly 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 instead.

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

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**

ie we start counting at zero.

(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 3 months *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 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) $P(N(t+h) = r | N(t) = r) = 1 - \lambda h + o(h)$

$$P(N(t+h) = r+1 | N(t) = r) = \lambda h + o(h) \quad (4.1)$$

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

(Note that a function $f(h)$ is described as $o(h)$ if $\lim_{h \rightarrow 0} \frac{f(h)}{h} = 0$)

- (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 . (4.2)

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. Note that 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.



Example 4.3

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) or claims 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 .

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:

$$p_n(t) = \exp\{-\lambda t\} \frac{(\lambda t)^n}{n!} \quad (4.3)$$

This will be proved by deriving and solving a “differential-difference” equation.

Recall from Chapter 2, that for a partition B_1, \dots, B_k , the probability of any event A is:

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

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:

$$\begin{aligned} P(n \text{ at time } t + h) &= P(n \text{ at time } t + h | n \text{ at time } t)P(n \text{ at time } t) \\ &\quad + P(n \text{ at time } t + h | n - 1 \text{ at time } t)P(n - 1 \text{ at time } t) \\ &\quad + \dots \end{aligned}$$

Hence using (4.1) and our $p_n(t)$ notation, we obtain:

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

Thus:

$$p_n(t + h) - p_n(t) = \lambda h[p_{n-1}(t) - p_n(t)] + o(h) \quad (4.4)$$

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

Now recall the formal definition of differentiation:

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

Now divide (4.4) by h , and let h go to zero from above to get the differential-difference equation:

$$\begin{aligned} \lim_{h \rightarrow 0^+} \frac{p_n(t+h) - p_n(t)}{h} &= \lim_{h \rightarrow 0^+} \frac{\lambda h[p_{n-1}(t) - p_n(t)] + o(h)}{h} \\ \frac{d}{dt} p_n(t) &= \lambda[p_{n-1}(t) - p_n(t)] \end{aligned} \quad (4.5)$$

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

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

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

$$\begin{aligned} P(0 \text{ at time } t+h) &= P(0 \text{ at time } t+h | 0 \text{ at time } t)P(0 \text{ at time } t) \\ p_0(t+h) &= p_0(t)[1 - \lambda h + o(h)] \end{aligned}$$

So:

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

and therefore:

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

Hence:

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

$$\frac{d}{dt} p_0(t) = -\lambda p_0(t) \quad (4.6)$$

Now all we have to do is show that $p_n(t) = \exp\{-\lambda t\} \frac{(\lambda t)^n}{n!}$ satisfies the differential equations (4.5) and (4.6).

**Question 4.29**

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

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

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

We can, however, prove this result more rigorously, but:

The completion of this proof requires familiarity with probability generating functions (see Chapter 5, Section 1).

So skip onto Section 3.2 for now and come back and look at this after you have studied Chapter 5.

Solve for $p_n(t)$ by introducing the probability generating function $G(s,t)$ defined by:

$$G(s,t) = \sum_{n=0}^{\infty} s^n p_n(t)$$

In Chapter 5 notation $G_{N(t)}(s) = E[s^{N(t)}]$.

so that:

$$\frac{d}{dt} G(s,t) = \sum_{n=0}^{\infty} s^n \frac{d}{dt} p_n(t)$$

Note that we use s as the dummy variable for the generating function in order to avoid confusion with t denoting time.

We are assuming here that s is a fixed number.

Now multiply (4.5) by s^n and sum over all values of n to get:

$$\sum_{n=1}^{\infty} s^n \frac{d}{dt} p_n(t) = \lambda \sum_{n=1}^{\infty} s^n p_{n-1}(t) - \lambda \sum_{n=1}^{\infty} s^n p_n(t)$$

Now add (4.6) to the above identity to get:

$$\sum_{n=0}^{\infty} s^n \frac{d}{dt} p_n(t) = \lambda \sum_{n=1}^{\infty} s^n p_{n-1}(t) - \lambda \sum_{n=0}^{\infty} s^n p_n(t)$$

which can be written as:

$$\frac{d}{dt} G(s,t) = \lambda s G(s,t) - \lambda G(s,t)$$

Here we have used the definition of $G(s,t)$ directly in the first and last sums. If we take a factor of s out of the middle sum it leaves $\sum_{i=1}^n s^{n-1} p_{n-1}(t)$, which is an equivalent way of writing $G(s,t)$.

Or, equivalently:

$$\frac{1}{G(s,t)} \frac{d}{dt} G(s,t) = \lambda(s-1) \quad (4.7)$$

Since the left hand side of (4.7) is the same as the derivative with respect to t of $\log G(s,t)$, (4.7) can be integrated to find that:

$$\log G(s,t) = \lambda t(s-1) + c(s)$$

where $c(s)$ is some function of s . $c(s)$ can be identified by noting that when $t = 0$, $p_0(t) = 1$ and $P_n(t) = 0$ for $n = 1, 2, 3, \dots$

This is because we are certain that there have been no claims at time zero.

Hence $G(s,0) = 1$ and $\log G(s,0) = 0 = c(s)$.

Normally when integrating we would put in a constant. Here, as we are integrating with respect to t , the “constant” can in fact be a function of the other variable s (but not of t).

Thus:

$$G(s, t) = \exp\{\lambda t(s - 1)\}$$

which is the probability generating function for the Poisson distribution with parameter λt . Since there is a one-to-one relationship between probability generating functions and distribution functions, the distribution of $N(t)$ is Poisson with parameter λt .

Using the uniqueness property of PGFs from Chapter 5.

This section of work has just shown that, for fixed t , $N(t) \sim Poi(\lambda t)$.

3.2 Deriving the waiting time process

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

This study of the Poisson process concludes by considering the distribution of the time to the first event and the times between events.

Time to the first event

Let the random variable T_1 denote the time of the first event. Then, for a fixed value of t , if no events have occurred by time t , $T_1 > t$. Hence:

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

This last step follows from the formula for the probability function of a $Poisson(\lambda t)$ distribution with $x = 0$.

And:

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

so that T_1 has an exponential distribution with parameter λ .

This is because the RHS matches the formula for the distribution function of an exponential distribution.

Times between events

For $i = 2, 3, \dots$, let the random variable T_i denote the time between the $(i-1)$ -th and the i -th events. Then:

$$\begin{aligned} P(T_{n+1} > t \mid \sum_{i=1}^n T_i = r) &= P(\sum_{i=1}^{n+1} T_i > t + r \mid \sum_{i=1}^n T_i = r) \\ &= P(N(t+r) = n \mid N(r) = n) \\ &= P(N(t+r) - N(r) = 0 \mid N(r) = n) \end{aligned}$$

By condition (4.2) (we now use the independence of the number of events in different time periods):

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

Finally:

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

since the number of events in a time interval of length t does not depend on when that time interval starts (condition (4.1)). Thus inter-event times also have an exponential distribution with parameter λ .

Note that 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. This is referred to as the memoryless property of the exponential distribution.



Question 4.30

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.

4 Random number simulation

Random variables for discrete and/or continuous distributions can be generated using a basic simulation technique as described below:

First a set of random numbers distributed uniformly between 0 and 1 is generated. This can be done using a simple computer package.

For example, the Ran# button on a calculator, or the RND() function in EXCEL both produce pseudorandom numbers between 0 and 1. They can't produce *true* random numbers as they use computer algorithms to obtain them!

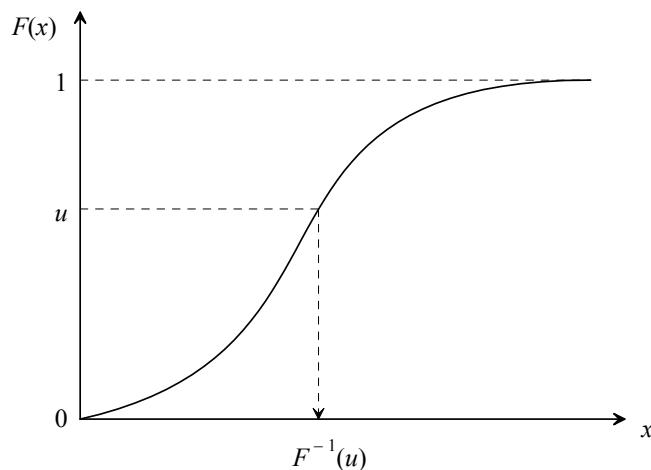
However, in the exam we'll use the list of "pseudorandom values from $U(0,1)$ " given on page 190 of the *Tables*.

The random numbers are then converted into values following the specified probability distribution using the inverse transform method as follows:

4.1 Continuous distributions

Supposing the random number generated was denoted U and the desired output was a random variable X with distribution $f(x)$. This can be obtained from U using the cumulative distribution, $F(x)$.

We have a random value, u , between 0 and 1. Recall that the cumulative distribution, $F(x)$, varies from 0 to 1:



By setting $u = F(x)$ 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.

Let U be the probability that X takes on a value less than or equal to x :

i.e. $U = P(X \leq x)$ denoted as $F(x)$

Hence x can be derived as:

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

This method obviously 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.

Taking the exponential distribution as an example, from Section 2.2 above, for the exponential distribution with parameter λ :

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

Hence:

$$x = -\frac{1}{\lambda} \ln(1 - U)$$



Example 4.4

Simulate three random values from an $Exp(0.1)$ distribution using the random 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:

$$\begin{aligned} x &= -10 \ln(1 - 0.113) = 1.20 \\ x &= -10 \ln(1 - 0.608) = 9.36 \\ x &= -10 \ln(1 - 0.003) = 0.03 \end{aligned}$$

**Question 4.31**

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

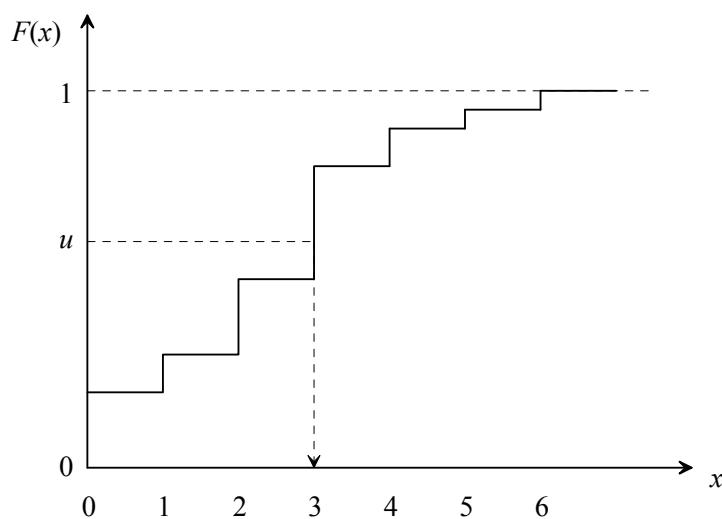
0.07 0.628 0.461

4.2 Discrete distributions

Discrete random variables don't have a distribution function to invert. The distribution function, $F(x)$, is just the sum of the probabilities so far:

$$\text{eg} \quad F(5) = P(X \leq 5) = P(X = 0) + P(X = 1) + \cdots + P(X = 5)$$

Given a random value, u , from $U(0,1)$ we can still read off the x value from the distribution graph as follows:



From the graph, we can see that 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 .

For discrete probability distributions a cumulative distribution function is calculated based on the probability function $P[X = x]$ as follows:

Let U be the probability that X takes on a value less than or equal to x , then:

$X = x_j$ if:

$$P[X = x_1] + P[X = x_2] + \cdots + P[X = x_{j-1}] \leq U < P[X = x_1] + P[X = x_2] + \cdots + P[X = x_j]$$

$$ie \quad F(x_{j-1}) \leq U < F(x_j)$$

Using the Poisson distribution, parameter μ as an example:

$$\sum_{x=0}^{x=j-1} e^{-\mu} \cdot \frac{\mu^x}{x!} \leq U < \sum_{x=0}^{x=j} e^{-\mu} \cdot \frac{\mu^x}{x!}$$



Example 4.5

Simulate two random values from a $Poi(2)$ distribution using the random values 0.721 and 0.128 from $U(0,1)$.

Solution

$$P(X = x) = e^{-\mu} \frac{\mu^x}{x!}, \quad x = 0, 1, 2, \dots$$

$$P(X = 0) = e^{-2} = 0.1353 \quad \Rightarrow \quad F(0) = 0.1353$$

$$P(X = 1) = 2e^{-2} = 0.2707 \quad \Rightarrow \quad F(1) = 0.4060$$

$$P(X = 2) = \frac{2^2}{2!} e^{-2} = 0.2707 \quad \Rightarrow \quad F(2) = 0.6767$$

$$P(X = 3) = \frac{2^3}{3!} e^{-2} = 0.1804 \quad \Rightarrow \quad F(3) = 0.8571, \quad etc.$$

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 have used the cumulative Poisson tables on page 175 of the Tables instead of calculating the values.

**Question 4.32**

Generate three random values from a $\text{Bin}(4, 0.6)$ using the following random values from $U(0,1)$:

0.588 0.222 0.906

5 Exam questions

Exam questions on this chapter can cover any of the distributions. The following are short questions on two distributions just to assess whether you are able to calculate probabilities.



Past Exam Question (Subject C1, April 1994, Q7)

For a certain type of insurance business, the number of claims per policy in a year has a Poisson distribution with mean 0.4

Consider a policy, which you know, has given rise to at least one claim in the last year. The probability that this policy has in fact given rise to exactly two claims in the least year is:

- | | | | |
|--------------|--------------|--------------|--------------|
| A 0.054 | B 0.163 | C 0.330 | D 0.992 |
|--------------|--------------|--------------|--------------|



Past Exam Question (Subject C1, April 1998, Q6)

The number of accidents in a factory is represented by a Poisson distribution averaging 2 accidents per 5 days. Evaluate the probability that the time from one accident to the next exceeds 3 days.

- | | | | |
|--------------|--------------|--------------|--------------|
| A 0.301 | B 0.449 | C 0.551 | D 0.699 |
|--------------|--------------|--------------|--------------|

This page has been left blank so that you can pull the appendices out for ease of reference.

6 Appendix – Statistical tables

Pages 43 to 49 contain some of the tables required for this course. They are reproduced for students who have yet to obtain a copy of “Formulae and Tables for Examinations of the Faculty and Institute of Actuaries”.

It should be noted that these are *not* the same format as the Actuarial Tables, which are the ones that will be available to you in the examination. Therefore, it would be wise to get hold of a set promptly so that you are thoroughly familiar with them. They are available directly from the Profession.

6.1 Use of standard normal tables (Appendix A)

Appendix A gives the cumulative probabilities $\Phi(z) = P(Z \leq z)$ for the standard normal distribution $N(0,1)$. The table gives probabilities for z values with 2 decimal places, however it is expected that students use *at least* 3 decimal places and interpolate.

To obtain $\Phi(1.268) = P(Z < 1.268)$, we read off the values of z either side: $\Phi(1.26) = P(Z < 1.26) = 0.89617$ and $\Phi(1.27) = P(Z < 1.27) = 0.89796$. We now use interpolation:

$$\Phi(1.268) \approx 0.89617 + \frac{1.268 - 1.26}{1.27 - 1.26} (0.89796 - 0.89617) = 0.89760$$

Further examples are given in 0.

6.2 Use of chi-square tables (Appendix B)

Appendix B gives the cumulative probabilities $P(\chi_v^2 < x)$ for a χ_v^2 distribution. The column title gives the *degrees of freedom*, v , of the chi-square distribution.

To find the probability that $P(\chi_{12}^2 < 8)$, we look down the 12 degrees of freedom column until we reach the row for 8. The probability is 0.2149.

Again for values between the given rows we are expected to use interpolation.

6.3 Using the tables of the t-distribution (Appendix C)

Appendix C gives the percentage points. That is, the percentage of the distribution that lies *above* the values given in the table. The number on the left of each row is the *degrees of freedom*. The table only gives the upper tail percentage points. This is because the *t*-distribution is symmetrical about zero.

To find the value a such that $P(t_9 > a) = 0.05$, we look along the 9 degrees of freedom row until we reach the 5% column. The value is 1.833.

To find the value a such that $P(t_{22} < a) = 0.01$, we use the fact the t -distribution is symmetrical. So $P(t_{22} < a) = P(t_{22} > -a) = 0.01$, hence $-a = 2.508 \Rightarrow a = -2.508$.

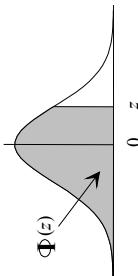
6.4 Using the tables of the F-distribution (Appendices D-G)

Appendices D, E, F and G give the upper 10%, 5%, 2½% and 1% points of the distribution respectively, ie the percentage of the distribution that lies *above* the values given in the table. For an $F_{a,b}$ distribution the a is the *degrees of freedom* along the top row, whereas the b is the *degrees of freedom* down the left-hand column. As only the upper tail is given, we will use the $\frac{1}{F_{m,n}}$ result to obtain lower tail values.

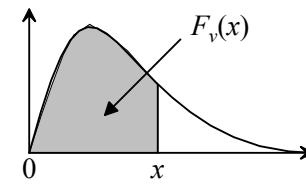
To find the value b such that $P(F_{6,14} > b) = 0.10$, we refer to the 10% F table. We then look in the 6th column and 14th row. The value of b is 2.243.

For a lower-tail percentage, eg find the value b such that $P(F_{11,5} < b) = 0.05$, we use $P(F_{11,5} < b) = P\left(F_{5,11} > \frac{1}{b}\right) = 0.05$. Referring to the 5% F table, we look in the 5th column and 11th row to get 3.204. Hence $\frac{1}{b} = 3.204 \Rightarrow b = 0.3121$.

Appendix A: Probabilities for the standard normal distribution



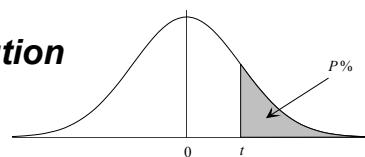
$\Phi(z)$	z	$\Phi(z)$																	
0.00	0.50000	0.40	0.65542	0.80	0.78814	1.20	0.88493	1.60	0.94520	2.00	0.97725	2.40	0.99180	2.80	0.99744	3.20	0.99931	3.60	0.99984
0.01	0.50399	0.41	0.65910	0.81	0.79103	1.21	0.88686	1.61	0.94630	2.01	0.97778	2.41	0.99202	2.81	0.99752	3.21	0.99934	3.61	0.99985
0.02	0.50798	0.42	0.66276	0.82	0.79389	1.22	0.88877	1.62	0.94738	2.02	0.97831	2.42	0.99224	2.82	0.99760	3.22	0.99936	3.62	0.99985
0.03	0.51197	0.43	0.66640	0.83	0.79673	1.23	0.89065	1.63	0.94845	2.03	0.97882	2.43	0.99245	2.83	0.99767	3.23	0.99938	3.63	0.99986
0.04	0.51595	0.44	0.67003	0.84	0.79955	1.24	0.89251	1.64	0.94950	2.04	0.97932	2.44	0.99266	2.84	0.99774	3.24	0.99940	3.64	0.99986
0.05	0.51994	0.45	0.67364	0.85	0.80234	1.25	0.89435	1.65	0.95053	2.05	0.97982	2.45	0.99286	2.85	0.99781	3.25	0.99942	3.65	0.99987
0.06	0.52392	0.46	0.67724	0.86	0.80511	1.26	0.89617	1.66	0.95154	2.06	0.98030	2.46	0.99305	2.86	0.99788	3.26	0.99944	3.66	0.99987
0.07	0.52790	0.47	0.68082	0.87	0.80785	1.27	0.89796	1.67	0.95254	2.07	0.98077	2.47	0.99324	2.87	0.99795	3.27	0.99946	3.67	0.99988
0.08	0.53188	0.48	0.68439	0.88	0.81057	1.28	0.89973	1.68	0.95352	2.08	0.98124	2.48	0.99343	2.88	0.99801	3.28	0.99948	3.68	0.99988
0.09	0.53586	0.49	0.68793	0.89	0.81327	1.29	0.90147	1.69	0.95449	2.09	0.98169	2.49	0.99361	2.89	0.99807	3.29	0.99950	3.69	0.99989
0.10	0.53983	0.50	0.69146	0.90	0.81594	1.30	0.90320	1.70	0.95543	2.10	0.98214	2.50	0.99379	2.90	0.99813	3.30	0.99952	3.70	0.99989
0.11	0.54380	0.51	0.69497	0.91	0.81859	1.31	0.90490	1.71	0.95637	2.11	0.98257	2.51	0.99396	2.91	0.99819	3.31	0.99953	3.71	0.99990
0.12	0.54776	0.52	0.68847	0.92	0.82121	1.32	0.90658	1.72	0.95728	2.12	0.98300	2.52	0.99413	2.92	0.99825	3.32	0.99955	3.72	0.99990
0.13	0.55172	0.53	0.70194	0.93	0.82381	1.33	0.90824	1.73	0.95818	2.13	0.98341	2.53	0.99430	2.93	0.99831	3.33	0.99957	3.73	0.99990
0.14	0.55567	0.54	0.70540	0.94	0.82639	1.34	0.90988	1.74	0.95907	2.14	0.98382	2.54	0.99446	2.94	0.99836	3.34	0.99958	3.74	0.99991
0.15	0.55962	0.55	0.70884	0.95	0.82894	1.35	0.91149	1.75	0.95994	2.15	0.98422	2.55	0.99461	2.95	0.99841	3.35	0.99960	3.75	0.99991
0.16	0.56356	0.56	0.71226	0.96	0.83147	1.36	0.91308	1.76	0.96080	2.16	0.98461	2.56	0.99477	2.96	0.99846	3.36	0.99961	3.76	0.99992
0.17	0.56749	0.57	0.71566	0.97	0.83398	1.37	0.91466	1.77	0.96164	2.17	0.98500	2.57	0.99492	2.97	0.99851	3.37	0.99962	3.77	0.99992
0.18	0.57142	0.58	0.71904	0.98	0.83646	1.38	0.91621	1.78	0.96246	2.18	0.98537	2.58	0.99506	2.98	0.99856	3.38	0.99964	3.78	0.99992
0.19	0.57535	0.59	0.72240	0.99	0.83891	1.39	0.91774	1.79	0.96327	2.19	0.98574	2.59	0.99520	2.99	0.99861	3.39	0.99965	3.79	0.99992
0.20	0.57926	0.60	0.72575	1.00	0.84134	1.40	0.91924	1.80	0.96407	2.20	0.98610	2.60	0.99534	3.00	0.99865	3.40	0.99966	3.80	0.99993
0.21	0.58317	0.61	0.72907	1.01	0.84375	1.41	0.92073	1.81	0.96485	2.21	0.98645	2.61	0.99547	3.01	0.99869	3.41	0.99968	3.81	0.99993
0.22	0.58706	0.62	0.73237	1.02	0.84614	1.42	0.92220	1.82	0.96562	2.22	0.98679	2.62	0.99560	3.02	0.99874	3.42	0.99969	3.82	0.99993
0.23	0.59095	0.63	0.73565	1.03	0.84849	1.43	0.92364	1.83	0.96638	2.23	0.98713	2.63	0.99573	3.03	0.99878	3.43	0.99970	3.83	0.99994
0.24	0.59483	0.64	0.73891	1.04	0.85083	1.44	0.92507	1.84	0.96712	2.24	0.98745	2.64	0.99585	3.04	0.99882	3.44	0.99971	3.84	0.99994
0.25	0.59871	0.65	0.74215	1.05	0.85314	1.45	0.92647	1.85	0.96784	2.25	0.98778	2.65	0.99598	3.05	0.99886	3.45	0.99972	3.85	0.99994
0.26	0.60257	0.66	0.74537	1.06	0.85543	1.46	0.92785	1.86	0.96856	2.26	0.98809	2.66	0.99609	3.06	0.99889	3.46	0.99973	3.86	0.99994
0.27	0.60642	0.67	0.74857	1.07	0.85769	1.47	0.92922	1.87	0.96926	2.27	0.98840	2.67	0.99621	3.07	0.99893	3.47	0.99974	3.87	0.99995
0.28	0.61026	0.68	0.75175	1.08	0.85993	1.48	0.93056	1.88	0.96995	2.28	0.98870	2.68	0.99632	3.08	0.99896	3.48	0.99975	3.88	0.99995
0.29	0.61409	0.69	0.75490	1.09	0.86214	1.49	0.93189	1.89	0.97062	2.29	0.98899	2.69	0.99643	3.09	0.99900	3.49	0.99976	3.89	0.99995
0.30	0.61791	0.70	0.75804	1.10	0.86433	1.50	0.93319	1.90	0.97128	2.30	0.98928	2.70	0.99653	3.10	0.99903	3.50	0.99977	3.90	0.99995
0.31	0.62172	0.71	0.76115	1.11	0.86650	1.51	0.93448	1.91	0.97193	2.31	0.98956	2.71	0.99664	3.11	0.99906	3.51	0.99978	3.91	0.99995
0.32	0.62552	0.72	0.76424	1.12	0.86864	1.52	0.93574	1.92	0.97257	2.32	0.98983	2.72	0.99674	3.12	0.99910	3.52	0.99978	3.92	0.99996
0.33	0.62930	0.73	0.76730	1.13	0.87076	1.53	0.93699	1.93	0.97320	2.33	0.99010	2.73	0.99683	3.13	0.99913	3.53	0.99979	3.93	0.99996
0.34	0.63307	0.74	0.77035	1.14	0.87286	1.54	0.93822	1.94	0.97381	2.34	0.99036	2.74	0.99693	3.14	0.99916	3.54	0.99980	3.94	0.99996
0.35	0.63683	0.75	0.77337	1.15	0.87493	1.55	0.93943	1.95	0.97441	2.35	0.99061	2.75	0.99702	3.15	0.99918	3.55	0.99981	3.95	0.99996
0.36	0.64058	0.76	0.77637	1.16	0.87698	1.56	0.94062	1.96	0.97500	2.36	0.99086	2.76	0.99711	3.16	0.99921	3.56	0.99981	3.96	0.99996
0.37	0.64431	0.77	0.77935	1.17	0.87900	1.57	0.94179	1.97	0.97558	2.37	0.99111	2.77	0.99720	3.17	0.99924	3.57	0.99982	3.97	0.99996
0.38	0.64803	0.78	0.78230	1.18	0.88100	1.58	0.94295	1.98	0.97615	2.38	0.99134	2.78	0.99728	3.18	0.99926	3.58	0.99983	3.98	0.99997
0.39	0.65173	0.79	0.78524	1.19	0.88298	1.59	0.94408	1.99	0.97670	2.39	0.99158	2.79	0.99736	3.19	0.99929	3.59	0.99983	3.99	0.99997
0.40	0.65542	0.80	0.78814	1.20	0.88493	1.60	0.94520	2.00	0.97725	2.40	0.99180	2.80	0.99744	3.20	0.99931	3.60	0.99984	4.00	0.99997



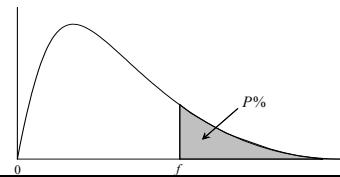
Appendix B: Probabilities for the χ^2 distribution

	1	2	3	4	5	6	7	8	9	10	11	12
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.5	0.5205	0.2212	0.0811	0.0265	0.0079	0.0022	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
1	0.6827	0.3935	0.1987	0.0902	0.0374	0.0144	0.0052	0.0018	0.0006	0.0002	0.0001	0.0000
1.5	0.7793	0.5276	0.3177	0.1734	0.0869	0.0405	0.0177	0.0073	0.0029	0.0011	0.0004	0.0001
2	0.8427	0.6321	0.4276	0.2642	0.1509	0.0803	0.0402	0.0190	0.0085	0.0037	0.0015	0.0006
2.5	0.8862	0.7135	0.5247	0.3554	0.2235	0.1315	0.0729	0.0383	0.0191	0.0091	0.0042	0.0018
3	0.9167	0.7769	0.6084	0.4422	0.3000	0.1912	0.1150	0.0656	0.0357	0.0186	0.0093	0.0045
3.5	0.9386	0.8262	0.6792	0.5221	0.3766	0.2560	0.1648	0.1008	0.0589	0.0329	0.0177	0.0091
4	0.9545	0.8647	0.7385	0.5940	0.4506	0.3233	0.2202	0.1429	0.0886	0.0527	0.0301	0.0166
4.5	0.9661	0.8946	0.7877	0.6575	0.5201	0.3907	0.2793	0.1906	0.1245	0.0780	0.0471	0.0274
5	0.9747	0.9179	0.8282	0.7127	0.5841	0.4562	0.3400	0.2424	0.1657	0.1088	0.0688	0.0420
5.5	0.9810	0.9361	0.8614	0.7603	0.6421	0.5185	0.4008	0.2970	0.2113	0.1446	0.0954	0.0608
6	0.9857	0.9502	0.8884	0.8009	0.6938	0.5768	0.4603	0.3528	0.2601	0.1847	0.1266	0.0839
6.5	0.9892	0.9612	0.9103	0.8352	0.7394	0.6304	0.5173	0.4086	0.3110	0.2283	0.1620	0.1112
7	0.9918	0.9698	0.9281	0.8641	0.7794	0.6792	0.5711	0.4634	0.3629	0.2746	0.2009	0.1424
7.5	0.9938	0.9765	0.9424	0.8883	0.8140	0.7229	0.6213	0.5162	0.4148	0.3225	0.2427	0.1771
8	0.9953	0.9817	0.9540	0.9084	0.8438	0.7619	0.6674	0.5665	0.4659	0.3712	0.2867	0.2149
8.5	0.9964	0.9857	0.9633	0.9251	0.8693	0.7963	0.7094	0.6138	0.5154	0.4199	0.3321	0.2551
9	0.9973	0.9889	0.9707	0.9389	0.8909	0.8264	0.7473	0.6577	0.5627	0.4679	0.3781	0.2971
9.5	0.9979	0.9913	0.9767	0.9503	0.9093	0.8527	0.7813	0.6981	0.6075	0.5146	0.4242	0.3403
10	0.9984	0.9933	0.9814	0.9596	0.9248	0.8753	0.8114	0.7350	0.6495	0.5595	0.4696	0.3840
10.5	0.9988	0.9948	0.9852	0.9672	0.9378	0.8949	0.8380	0.7683	0.6885	0.6022	0.5140	0.4278
11	0.9991	0.9959	0.9883	0.9734	0.9486	0.9116	0.8614	0.7983	0.7243	0.6425	0.5567	0.4711
11.5	0.9993	0.9968	0.9907	0.9785	0.9577	0.9259	0.8818	0.8251	0.7570	0.6801	0.5976	0.5134
12	0.9995	0.9975	0.9926	0.9826	0.9652	0.9380	0.8994	0.8488	0.7867	0.7149	0.6364	0.5543
12.5	0.9996	0.9981	0.9941	0.9860	0.9715	0.9483	0.9147	0.8697	0.8134	0.7470	0.6727	0.5936
13	0.9997	0.9985	0.9954	0.9887	0.9766	0.9570	0.9279	0.8882	0.8374	0.7763	0.7067	0.6310
13.5	0.9998	0.9988	0.9963	0.9909	0.9809	0.9643	0.9392	0.9042	0.8587	0.8030	0.7381	0.6662
14	0.9998	0.9991	0.9971	0.9927	0.9844	0.9704	0.9488	0.9182	0.8777	0.8270	0.7670	0.6993
14.5	0.9999	0.9993	0.9977	0.9941	0.9873	0.9755	0.9570	0.9304	0.8944	0.8486	0.7935	0.7301
15	0.9999	0.9994	0.9982	0.9953	0.9896	0.9797	0.9640	0.9409	0.9091	0.8679	0.8175	0.7586
15.5	0.9999	0.9996	0.9986	0.9962	0.9916	0.9833	0.9699	0.9499	0.9219	0.8851	0.8393	0.7848
16	0.9999	0.9997	0.9989	0.9970	0.9932	0.9862	0.9749	0.9576	0.9331	0.9004	0.8589	0.8088
16.5	1.0000	0.9997	0.9991	0.9976	0.9944	0.9887	0.9791	0.9642	0.9429	0.9138	0.8764	0.8306
17	1.0000	0.9998	0.9993	0.9981	0.9955	0.9907	0.9826	0.9699	0.9513	0.9256	0.8921	0.8504
17.5	1.0000	0.9998	0.9994	0.9985	0.9964	0.9924	0.9856	0.9747	0.9586	0.9360	0.9061	0.8683
18	1.0000	0.9999	0.9996	0.9988	0.9971	0.9938	0.9880	0.9788	0.9648	0.9450	0.9184	0.8843
18.5	1.0000	0.9999	0.9997	0.9990	0.9976	0.9949	0.9901	0.9822	0.9702	0.9529	0.9293	0.8987
19	1.0000	0.9999	0.9997	0.9992	0.9981	0.9958	0.9918	0.9851	0.9748	0.9597	0.9389	0.9115
19.5	1.0000	0.9999	0.9998	0.9994	0.9984	0.9966	0.9932	0.9876	0.9787	0.9656	0.9473	0.9228
20	1.0000	1.0000	0.9998	0.9995	0.9988	0.9972	0.9944	0.9897	0.9821	0.9707	0.9547	0.9329
21	1.0000	1.0000	0.9999	0.9997	0.9992	0.9982	0.9962	0.9929	0.9873	0.9789	0.9666	0.9496
22	1.0000	1.0000	0.9999	0.9998	0.9995	0.9988	0.9975	0.9951	0.9911	0.9849	0.9756	0.9625
23	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9983	0.9966	0.9938	0.9893	0.9823	0.9723
24	1.0000	1.0000	1.0000	0.9999	0.9998	0.9995	0.9989	0.9977	0.9957	0.9924	0.9873	0.9797
25	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9992	0.9984	0.9970	0.9947	0.9909	0.9852
26	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9995	0.9989	0.9980	0.9963	0.9935	0.9893
27	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9993	0.9986	0.9974	0.9954	0.9923
28	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9995	0.9990	0.9982	0.9968	0.9945
29	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9994	0.9988	0.9977	0.9961
30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9991	0.9984	0.9972

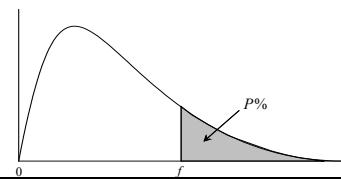
Appendix C: Percentage points for the t-distribution



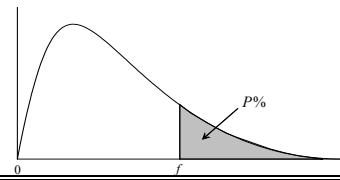
	40%	30%	25%	20%	15%	10%	5%	2.50%	1%	0.50%	0.10%	0.05%
1	0.3249	0.7265	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	0.2887	0.6172	0.8165	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.2767	0.5844	0.7649	0.9785	1.250	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.2707	0.5686	0.7407	0.9410	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.2672	0.5594	0.7267	0.9195	1.156	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	0.2648	0.5534	0.7176	0.9057	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.2632	0.5491	0.7111	0.8960	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.2619	0.5459	0.7064	0.8889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.2610	0.5435	0.7027	0.8834	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.2602	0.5415	0.6998	0.8791	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.2596	0.5399	0.6974	0.8755	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.2590	0.5386	0.6955	0.8726	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.2586	0.5375	0.6938	0.8702	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.2582	0.5366	0.6924	0.8681	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.2579	0.5357	0.6912	0.8662	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.2576	0.5350	0.6901	0.8647	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.2573	0.5344	0.6892	0.8633	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.2571	0.5338	0.6884	0.8620	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.2569	0.5333	0.6876	0.8610	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.2567	0.5329	0.6870	0.8600	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.2566	0.5325	0.6864	0.8591	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.2564	0.5321	0.6858	0.8583	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.2563	0.5317	0.6853	0.8575	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.2562	0.5314	0.6848	0.8569	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.2561	0.5312	0.6844	0.8562	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.2560	0.5309	0.6840	0.8557	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.2559	0.5306	0.6837	0.8551	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	0.2558	0.5304	0.6834	0.8546	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.2557	0.5302	0.6830	0.8542	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.660
30	0.2556	0.5300	0.6828	0.8538	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
32	0.2555	0.5297	0.6822	0.8530	1.054	1.309	1.694	2.037	2.449	2.738	3.365	3.622
34	0.2553	0.5294	0.6818	0.8523	1.052	1.307	1.691	2.032	2.441	2.728	3.348	3.601
36	0.2552	0.5291	0.6814	0.8517	1.052	1.306	1.688	2.028	2.434	2.719	3.333	3.582
38	0.2551	0.5288	0.6810	0.8512	1.051	1.304	1.686	2.024	2.429	2.712	3.319	3.566
40	0.2550	0.5286	0.6807	0.8507	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.2547	0.5278	0.6794	0.8489	1.047	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.2545	0.5272	0.6786	0.8477	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.2539	0.5258	0.6765	0.8446	1.041	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.2533	0.5244	0.6745	0.8416	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.290

Appendix D: 10% points for the F-distribution

	1	2	3	4	5	6	7	8	9	10	12	24
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	62.00
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.408	9.450
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230	5.216	5.176
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920	3.896	3.831
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297	3.268	3.191
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937	2.905	2.818
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703	2.668	2.575
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.502	2.404
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416	2.379	2.277
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323	2.284	2.178
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248	2.209	2.100
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188	2.147	2.036
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138	2.097	1.983
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095	2.054	1.938
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059	2.017	1.899
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028	1.985	1.866
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001	1.958	1.836
18	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977	1.933	1.810
19	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956	1.912	1.787
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937	1.892	1.767
21	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920	1.875	1.748
22	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904	1.859	1.731
23	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890	1.845	1.716
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877	1.832	1.702
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866	1.820	1.689
26	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855	1.809	1.677
27	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845	1.799	1.666
28	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836	1.790	1.656
29	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827	1.781	1.647
30	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819	1.773	1.638
32	2.869	2.477	2.263	2.129	2.036	1.967	1.913	1.870	1.835	1.805	1.758	1.622
34	2.859	2.466	2.252	2.118	2.024	1.955	1.901	1.858	1.822	1.793	1.745	1.608
36	2.850	2.456	2.243	2.108	2.014	1.945	1.891	1.847	1.811	1.781	1.734	1.595
38	2.842	2.448	2.234	2.099	2.005	1.935	1.881	1.838	1.802	1.772	1.724	1.584
40	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763	1.715	1.574
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707	1.657	1.511
120	2.748	2.347	2.130	1.992	1.896	1.824	1.767	1.722	1.684	1.652	1.601	1.447

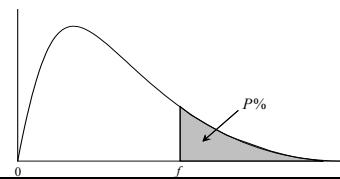
Appendix E: 5% points for the F-distribution


	1	2	3	4	5	6	7	8	9	10	12	24
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	249.1
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.45
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.745	8.638
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.774
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.527
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.841
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.410
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.284	3.115
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	2.900
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.737
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.609
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.505
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.420
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.349
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.288
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.235
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.190
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.150
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.114
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.082
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.250	2.054
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.028
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.204	2.005
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.183	1.984
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	1.964
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.148	1.946
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.132	1.930
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.118	1.915
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.104	1.901
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	1.887
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142	2.070	1.864
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123	2.050	1.843
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106	2.033	1.824
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091	2.017	1.808
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.003	1.793
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.917	1.700
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.834	1.608

Appendix F: 2½% points for the F-distribution

	1	2	3	4	5	6	7	8	9	10	12	24
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.6	963.3	968.6	976.7	997.3
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.46
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.12
4	12.22	10.65	9.979	9.604	9.364	9.197	9.074	8.980	8.905	8.844	8.751	8.511
5	10.01	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619	6.525	6.278
6	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461	5.366	5.117
7	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761	4.666	4.415
8	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295	4.200	3.947
9	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964	3.868	3.614
10	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717	3.621	3.365
11	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588	3.526	3.430	3.173
12	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374	3.277	3.019
13	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388	3.312	3.250	3.153	2.893
14	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285	3.209	3.147	3.050	2.789
15	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060	2.963	2.701
16	6.115	4.687	4.077	3.729	3.502	3.341	3.219	3.125	3.049	2.986	2.889	2.625
17	6.042	4.619	4.011	3.665	3.438	3.277	3.156	3.061	2.985	2.922	2.825	2.560
18	5.978	4.560	3.954	3.608	3.382	3.221	3.100	3.005	2.929	2.866	2.769	2.503
19	5.922	4.508	3.903	3.559	3.333	3.172	3.051	2.956	2.880	2.817	2.720	2.452
20	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837	2.774	2.676	2.408
21	5.827	4.420	3.819	3.475	3.250	3.090	2.969	2.874	2.798	2.735	2.637	2.368
22	5.786	4.383	3.783	3.440	3.215	3.055	2.934	2.839	2.763	2.700	2.602	2.332
23	5.750	4.349	3.750	3.408	3.183	3.023	2.902	2.808	2.731	2.668	2.570	2.299
24	5.717	4.319	3.721	3.379	3.155	2.995	2.874	2.779	2.703	2.640	2.541	2.269
25	5.686	4.291	3.694	3.353	3.129	2.969	2.848	2.753	2.677	2.613	2.515	2.242
26	5.659	4.265	3.670	3.329	3.105	2.945	2.824	2.729	2.653	2.590	2.491	2.217
27	5.633	4.242	3.647	3.307	3.083	2.923	2.802	2.707	2.631	2.568	2.469	2.195
28	5.610	4.221	3.626	3.286	3.063	2.903	2.782	2.687	2.611	2.547	2.448	2.174
29	5.588	4.201	3.607	3.267	3.044	2.884	2.763	2.669	2.592	2.529	2.430	2.154
30	5.568	4.182	3.589	3.250	3.026	2.867	2.746	2.651	2.575	2.511	2.412	2.136
32	5.531	4.149	3.557	3.218	2.995	2.836	2.715	2.620	2.543	2.480	2.381	2.103
34	5.499	4.120	3.529	3.191	2.968	2.808	2.688	2.593	2.516	2.453	2.353	2.075
36	5.471	4.094	3.505	3.167	2.944	2.785	2.664	2.569	2.492	2.429	2.329	2.049
38	5.446	4.071	3.483	3.145	2.923	2.763	2.643	2.548	2.471	2.407	2.307	2.027
40	5.424	4.051	3.463	3.126	2.904	2.744	2.624	2.529	2.452	2.388	2.288	2.007
60	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334	2.270	2.169	1.882
120	5.152	3.805	3.227	2.894	2.674	2.515	2.395	2.299	2.222	2.157	2.055	1.760

Appendix G: 1% points for the F-distribution



	1	2	3	4	5	6	7	8	9	10	12	24
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6234
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.42	99.46
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.60
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	13.93
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.466
6	13.75	10.92	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.313
7	12.25	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.074
8	11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.279
9	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.729
10	10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.327
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.021
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	3.780
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.587
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.427
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.294
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.181
17	8.400	6.112	5.185	4.669	4.336	4.101	3.927	3.791	3.682	3.593	3.455	3.083
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	2.999
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	2.925
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	2.859
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	2.801
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.749
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.702
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.659
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.620
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.585
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.552
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.522
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.495
30	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173	3.067	2.979	2.843	2.469
32	7.499	5.336	4.459	3.969	3.652	3.427	3.258	3.127	3.021	2.934	2.798	2.423
34	7.444	5.289	4.416	3.927	3.611	3.386	3.218	3.087	2.981	2.894	2.758	2.383
36	7.396	5.248	4.377	3.890	3.574	3.351	3.183	3.052	2.946	2.859	2.723	2.347
38	7.353	5.211	4.343	3.858	3.542	3.319	3.152	3.021	2.915	2.828	2.692	2.316
40	7.314	5.178	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.288
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.115
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.336	1.950

This page has been left blank so that you can keep the chapter summaries together for revision purposes.



Chapter 4 Summary

Standard discrete distributions covered in this course are the discrete uniform, Bernoulli, binomial, geometric, negative binomial, hypergeometric and Poisson.

Waiting times between events in a $Poi(\lambda)$ distribution have a $Exp(\lambda)$ distribution.

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 overleaf.

The t -distribution with k degrees of freedom is defined as:

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi_k^2/k}}$$

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 :

$$N(t) \sim Poi(\lambda t)$$

It can be derived by considering events occurring in a small time interval h . The waiting times between events in a Poisson process have an exponential distribution.

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 $F(x_{j-1}) < u \leq F(x_j)$

	Distribution	PF or PDF	Mean	Variance
Discrete Distributions	Discrete uniform	$\frac{1}{k}$	$\frac{k+1}{2}$	$\frac{k^2-1}{12}$
	Bernouilli	$\theta^x(1-\theta)^{1-x}$	θ	$\theta(1-\theta)$
	Binomial	$\binom{n}{x} \theta^x (1-\theta)^{n-x}$	$n\theta$	$n\theta(1-\theta)$
	Geometric	$\theta(1-\theta)^{x-1}$	$\frac{1}{\theta}$	$\frac{1-\theta}{\theta^2}$
	Negative binomial	$\binom{x-1}{k-1} \theta^k (1-\theta)^{x-k}$	$\frac{k}{\theta}$	$\frac{k(1-\theta)}{\theta^2}$
	Poisson	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ
Continuous Distributions	Continuous uniform	$\frac{1}{\beta-\alpha}$	$\frac{1}{2}(\alpha+\beta)$	$\frac{1}{12}(\beta-\alpha)^2$
	Gamma	$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
	Exponential	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
	Chi-square	$\frac{\left(\frac{1}{2}\right)^{\frac{v}{2}}}{\Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-\frac{1}{2}x}$	v	$2v$
	Beta	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
	Lognormal	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2}$	$e^{\mu+\frac{1}{2}\sigma^2}$	$e^{2\mu+\sigma^2} (e^{\sigma^2} - 1)$
	Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2

Chapter 4 Solutions

Solution 4.1

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

$$\begin{aligned}\sigma^2 &= E(X^2) - E^2(X) = \frac{(k+1)(2k+1)}{6} - \frac{(k+1)^2}{4} \\ &= \frac{2(2k^2 + 3k + 1) - 3(k^2 + 2k + 1)}{12} \\ &= \frac{k^2 - 1}{12}\end{aligned}$$

Solution 4.2

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.

Solution 4.3

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

Solution 4.4

Due to the “memoryless” property, the children she has so far are irrelevant when it comes to working out the probability of two girls. So the probability is $\left(\frac{1}{2}\right)^2 = 0.25$.

Solution 4.5

Since $Y = X - 1$:

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

using $\text{var}(aX + b) = a^2 \text{var}(X)$ and the result from the first formulation.

Solution 4.6

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

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

Solution 4.7

(i) The exact probability is:

$$\frac{28}{58} \times \frac{27}{57} \times \cdots \times \frac{24}{54} = 0.02145$$

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

$$P(X = 0) \doteq \left(\frac{28}{58}\right)^5 = 0.02622$$

Solution 4.8

This table shows the probabilities for a Poisson distribution (up to $X = 4$):

X	0	1	2	3	4
Probability	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2} e^{-\lambda}$	$\frac{\lambda^3}{3!} e^{-\lambda}$	$\frac{\lambda^4}{4!} e^{-\lambda}$

The mean is:

$$\begin{aligned}
 E(X) &= \lambda e^{-\lambda} + 2 \frac{\lambda^2}{2} e^{-\lambda} + 3 \frac{\lambda^3}{3!} e^{-\lambda} + 4 \frac{\lambda^4}{4!} e^{-\lambda} + \dots \\
 &= \lambda e^{-\lambda} + \lambda^2 e^{-\lambda} + \frac{\lambda^3}{2!} e^{-\lambda} + \frac{\lambda^4}{3!} e^{-\lambda} + \dots \\
 &= \lambda e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \right)
 \end{aligned}$$

Since $e^\lambda = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots$, 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 $E[X(X-1)]$:

$$\begin{aligned}
 E[X(X-1)] &= 2 \frac{\lambda^2}{2} e^{-\lambda} + 6 \frac{\lambda^3}{3!} e^{-\lambda} + 12 \frac{\lambda^4}{4!} e^{-\lambda} + \dots \\
 &= \lambda^2 e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \dots \right) \\
 &= \lambda^2 e^{-\lambda} e^\lambda \\
 &= \lambda^2
 \end{aligned}$$

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

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

Solution 4.9

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 $P(X > 5) = 1 - 0.916 = 0.084$.

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

Solution 4.10

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 $\theta = 1 \times 10^{-8}$. We can approximate this by using a Poisson distribution with:

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

Hence:

$$P(X = 2) = \frac{0.55^2}{2!} e^{-0.55} = 0.08726$$

Solution 4.11

The number of claims in a month has a $Poi(2)$ distribution; therefore the number of claims in a year has a $Poi(24)$ distribution.

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

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

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

Solution 4.12

If we return to the original definition of variance:

$$\begin{aligned} \text{var}[X] &= E[(X - \mu)^2] = \int_{\alpha}^{\beta} \left(x - \frac{1}{2}(\alpha + \beta)\right)^2 \frac{1}{\beta - \alpha} dx \\ &= \left[\frac{\left(x - \frac{1}{2}(\alpha + \beta)\right)^3}{3(\beta - \alpha)} \right]_{\alpha}^{\beta} \\ &= \frac{\left(\beta - \frac{1}{2}(\alpha + \beta)\right)^3}{3(\beta - \alpha)} - \frac{\left(\alpha - \frac{1}{2}(\alpha + \beta)\right)^3}{3(\beta - \alpha)} \\ &= \frac{\left(\frac{1}{2}(\beta - \alpha)\right)^3}{3(\beta - \alpha)} - \frac{\left(-\frac{1}{2}(\beta - \alpha)\right)^3}{3(\beta - \alpha)} \\ &= \frac{1}{24}(\beta - \alpha)^2 + \frac{1}{24}(\beta - \alpha)^2 \\ &= \frac{1}{12}(\beta - \alpha)^2 \end{aligned}$$

Alternatively, this can be proved using $\text{var}(X) = E(X^2) - E^2(X)$, but this way relies on algebraic long division, which is not everybody's strong point!

Solution 4.13

The PDF is given by $\frac{1}{150-50} = \frac{1}{100}$. 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.

Solution 4.14

Remembering that the formulae for mean and variance are $E(X) = \int xf(x) dx$, and $\text{var}(X) = \int x^2 f(x) dx - E^2(X)$, using appropriate limits, we have:

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

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

$$\begin{aligned} E(X) &= \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha \times -\frac{1}{\lambda} e^{-\lambda x} \right]_0^\infty - \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} \alpha x^{\alpha-1} \times -\frac{1}{\lambda} e^{-\lambda x} dx \\ &= \frac{\alpha}{\lambda} \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \end{aligned}$$

But the integral is just 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:

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

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

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

These results can be proved far more easily using MGFs from Chapter 5.

Solution 4.15

Using integration by parts:

$$\begin{aligned} P(X > 4) &= \int_4^\infty \frac{1.5^2}{\Gamma(2)} x e^{-1.5x} dx \\ &= 2.25 \left\{ \left[-\frac{x}{1.5} e^{-1.5x} \right]_4^\infty + \int_4^\infty \frac{1}{1.5} e^{-1.5x} dx \right\} \\ &= 2.25 \left\{ \frac{4}{1.5} e^{-6} + \left[-\frac{1}{1.5^2} e^{-1.5x} \right]_4^\infty \right\} \\ &= 2.25 \left\{ \frac{4}{1.5} e^{-6} + \frac{1}{1.5^2} e^{-6} \right\} = 0.0174 \end{aligned}$$

Solution 4.16

This gives:

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

But since $\ln 0.5 = -\ln 2$, we can say $m = \frac{\ln 2}{\lambda}$.

Solution 4.17

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 $\lambda = 3$. Hence:

$$P(T > \frac{1}{2}) = \int_{\frac{1}{2}}^{\infty} 3e^{-3x} dx = [-e^{-3x}]_{\frac{1}{2}}^{\infty} = 0 - (-e^{-1\frac{1}{2}}) = 0.223$$

Solution 4.18

$$\begin{aligned} P(X > x + n \mid X > n) &= \frac{P(X > x + n, X > n)}{P(X > n)} \\ &= \frac{P(X > x + n)}{P(X > n)} \\ &= \frac{e^{-\lambda(x+n)}}{e^{-\lambda n}} \\ &= e^{-\lambda x} \\ &= P(X > x) \end{aligned}$$

Solution 4.19

Since $X \sim \text{Gamma}(2, 1.5)$, we know that $3X \sim \chi^2_4$. So:

$$P(X > 4) = P(3X > 12) = P(\chi^2_4 > 12) = 1 - 0.9826 = 0.0174$$

using the χ^2 probability tables given on page 165 of the *Tables* or in Appendix B to this chapter.

This gives us the same answer as Question 4.15 but without the integration by parts!

Alternatively, we could use interpolation from the χ^2 percentage points tables given on pages 168-169 of the Tables. This gives 0.019.

*The difference in these answers has occurred because linear interpolation in the upper tail of the chi square distribution is inaccurate. Therefore you would be wise to only use interpolation (if needed) on the **probability** tables, rather than the percentage points table, as these values are much closer together.*

Solution 4.20

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

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

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

Solution 4.21

(i) Since 14 is the mean, the probability is 0.5.

$$(ii) P(X > 20) = P\left(Z > \frac{20-14}{\sqrt{20}}\right) = P(Z > 1.342) = 1 - 0.91020 = 0.0898$$

$$(iii) P(X < 9) = P\left(Z < \frac{9-14}{\sqrt{20}}\right) = P(Z < -1.118) = 1 - 0.86821 = 0.13179$$

$$(iv) P(X > r) = P\left(Z > \frac{r-14}{\sqrt{20}}\right) = 0.41294, \text{ which gives:}$$

$$P\left(Z < \frac{r-14}{\sqrt{20}}\right) = 0.58706 \Rightarrow \frac{r-14}{\sqrt{20}} = 0.22 \Rightarrow r = 14.98$$

Solution 4.22

The third non-central moment is $E[X^3]$. Using the formula for the skewness, from Chapter 3:

$$E[(X - \mu)^3] = E[X^3] - 3\mu E[X^2] + 2\mu^3$$

and we also know that the skewness of the normal distribution is zero, so:

$$0 = E[X^3] - 3 \times 10 \times (25 + 10^2) + 2 \times 10^3 \Rightarrow E[X^3] = 1,750$$

Note that we have worked out $E[X^2]$ here by turning round the relationship $\text{var}(X) = E[X^2] - E^2[X]$.

Solution 4.23

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

$$P(W > 3,000) = P(\ln W > 8.006) = P(Z > 1.227) = 1 - 0.89009 = 0.10991$$

Solution 4.24

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

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 $\mu = 1.3$.

Solution 4.25

From the *Tables*:

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

(ii) $a = 2.896$

(iii) By symmetry:

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

Solution 4.26

From the *Tables*:

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

(ii) $a = 14.98$

Solution 4.27

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).

Solution 4.28

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:

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

Solution 4.29

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

$$\begin{aligned} \frac{d}{dt} p_n(t) &= \frac{d}{dt} \left\{ e^{-\lambda t} \frac{(\lambda t)^n}{n!} \right\} \\ &= -\lambda e^{-\lambda t} \frac{(\lambda t)^n}{n!} + e^{-\lambda t} \frac{n \lambda^n t^{n-1}}{n!} \\ &= -\lambda e^{-\lambda t} \frac{(\lambda t)^n}{n!} + \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \\ &= \lambda [p_{n-1}(t) - p_n(t)] \end{aligned}$$

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

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

Solution 4.30

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

$$\begin{aligned} P(X < 2) &= P(X = 0) + P(X = 1) \\ &= e^{-5} + 5e^{-5} \\ &= 0.0404 \end{aligned}$$

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, X , 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.

Solution 4.31

The distribution function for a $U(-1, 4)$ is:

$$F(x) = \frac{x+1}{5}$$

We could calculate this directly:

$$F(x) = P(X \leq x) = \int_{-1}^x \frac{1}{5} dt = \left[\frac{1}{5} t \right]_{-1}^x = \frac{x+1}{5}$$

Alternatively, we could just use the formula $F(x) = \frac{x-a}{b-a}$ on page 13 of the Tables.

We now set our random value, u , equal to this and rearrange:

$$u = \frac{x+1}{5} \Rightarrow x = 5u - 1$$

Substituting, we obtain:

$$x = 5 \times 0.07 - 1 = -0.65$$

$$x = 5 \times 0.628 - 1 = 2.14$$

$$x = 5 \times 0.461 - 1 = 1.305$$

Solution 4.32

The probability function for a $\text{Bin}(4, 0.6)$ is:

$$P(X = x) = \binom{4}{x} 0.6^x 0.4^{4-x}, \quad x = 0, 1, 2, 3, 4$$

Calculating the probabilities and the cumulative distribution function:

$$P(X = 0) = 0.4^4 = 0.0256 \quad \Rightarrow \quad F(0) = 0.0256$$

$$P(X = 1) = 4 \times 0.6 \times 0.4^3 = 0.1536 \quad \Rightarrow \quad F(1) = 0.1792$$

$$P(X = 2) = 6 \times 0.6^2 \times 0.4^2 = 0.3456 \quad \Rightarrow \quad F(2) = 0.5248$$

$$P(X = 3) = 4 \times 0.6^3 \times 0.4 = 0.3456 \quad \Rightarrow \quad F(3) = 0.8704$$

$$P(X = 4) = 0.6^4 = 0.1296 \quad \Rightarrow \quad F(4) = 1$$

Since $F(2) < 0.588 < F(3)$, our first simulated value is 3.

Since $F(1) < 0.222 < F(2)$, our first simulated value is 2.

Since $F(3) < 0.906 < F(4)$, our first simulated value is 4.

Alternatively, it is much quicker to use the cumulative binomial probabilities given on page 186 of the Tables.

Past Exam Question (Subject C1, April 1994 Q7)

Answer B.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 0.6703 = 0.3297$$

$$P(X = 2) = 0.0536$$

Using $P(A|B) = \frac{P(A \cap B)}{P(B)}$, we get:

$$P(X = 2 | X \geq 1) = \frac{P(X \geq 1 \cap X = 2)}{P(X \geq 1)} = \frac{P(X = 2)}{P(X \geq 1)} = \frac{0.0536}{0.3297} = 0.163$$

Past Exam Question (Subject C1, April 1998, Q6)

Answer A.

Working in days, the number of accidents, X , has a $Poi(0.4)$ distribution, so the waiting time (in days), T , has an $Exp(0.4)$ distribution. Hence:

$$P(T > 3) = \int_3^{\infty} 0.4e^{-0.4t} dt = \left[-e^{-0.4t} \right]_3^{\infty} = 0 - (e^{-1.2}) = 0.301$$

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

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.

Chapter 5

Generating functions



Syllabus objectives

- (iv) 1. Define and determine the probability generating function of discrete, integer-valued random variables.
2. Define and determine the moment generating functions of random variables.
3. Define the cumulant generating function and the cumulants and determine them for random variables.
4. Use generating functions to determine the moments and cumulants of random variables, by expansion as a series or by differentiation, as appropriate.
5. Identify the applications for which a probability generating function, 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 Chapters 3 and 4.
- (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 Chapter 6.
- (c) Determine the properties of compound distributions. This will be covered in Chapter 7 and then extended in Subject CT6.

In this chapter we will introduce three types of generating functions: probability generating functions (PGFs), moment generating functions (MGFs) and cumulant generating functions (CGFs). We will use them to derive formulae for the moments of statistical distributions. The names give the game away: PGFs are used to generate probabilities, 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 hung up on this chapter, as they want to know “where these definitions come from”. Basically, they were invented to make calculations of means and variances easier. Remember some of those horrendous proofs you did in Chapter 4 for the mean and variance of the binomial, negative binomial and gamma? The good news is you won’t have to use those any more. The bad news is you’ve already done them!!

The definitions require that you have a working knowledge of the exponential series and of geometric series. Geometric series are briefly recapped in the appendix to this chapter. For further study please refer to a pure maths textbook or use FAC.

The syllabus clearly says “define and determine”, so make sure you know the definitions of PGFs, MGFs and CGFs and can find them (where they exist) for all the distributions met in Chapter 4. In addition, the syllabus requires us to “determine the moments”, so ensure you can calculate $E(X)$ and $\text{var}(X)$ for each of these distributions.

1 Probability generating functions

1.1 General formula

A probability generating function (PGF) can be used to generate a set of probabilities, namely the probabilities associated with the values 0, 1, 2, 3, ... assumed by a counting variable which assumes non-negative integer values.

In other words, we will only define PGFs for counting variables, *ie* discrete random variables such as the binomial, Poisson, geometric and negative binomial distributions.

Let X be a counting variable which assumes values 0, 1, 2, 3, ... with probabilities $p_0, p_1, p_2, p_3, \dots$ respectively. A generating function, $G_X(t)$, is constructed for the probabilities as follows:

$$G_X(t) = p_0 + p_1t + p_2t^2 + p_3t^3 + \dots \quad (*)$$

The variable t here is a “dummy variable”. The letter used is not important, and you will also see letters such as s and r used instead.

From this it can be seen that $G_X(1) = 1$ and $G_X(0) = P(X = 0)$.

The first result follows because $G_X(1) = p_0 + p_1 + p_2 + \dots$, *ie* it is the sum of all the probabilities.



Example

Derive the PGF of a random variable, X , where $P(X = 0) = 0.5$, $P(X = 1) = 0.3$ and $P(X = 3) = 0.2$.

Solution

$$G_X(t) = 0.5 + 0.3t + 0.2t^3$$

Another way of writing (*) using series notation is $G_X(t) = \sum_{x=0}^{\infty} t^x P(X = x)$.

In Chapter 3 we defined the expectation of a function, $g(x)$, for a discrete random variable to be $E[g(X)] = \sum_x g(x)P(X=x)$. So, $E[t^X] = \sum_x t^x P(X=x)$.

Note that (*) is just the expression for the expected value of the function t^X – this gives the definition of a PGF:

Definition

The probability generating function, $G_X(t)$, of a counting variable X is given by:

$$G_X(t) = E[t^X]$$

for all values of t for which the expectation exists.

In mathematical work, saying that something “exists” means that it works out to a definite finite value, *ie* it doesn’t work out to infinity or an indeterminate value.

Note: $G_X(t)$ does exist at least for $|t| \leq 1$, and PGFs are unique, in the sense that two variables have the same PGF if and only if they have the same probability distribution. This last fact enables us to recognise a PGF and say that the variable concerned must have a particular distribution.

Since the coefficients of t^k in the expansion of $G_X(t)$ are probabilities, which are positive numbers less than 1, then if $|t| \leq 1$, $G_X(t)$ must be finite.



Question 5.1

Show that the PGF for the score from one throw of an unbiased die can be written as

$$G(t) = \frac{t}{6}(1-t^6)(1-t)^{-1}.$$



Question 5.2

Derive the PGF of the random variable X with probability function:

$$P(X=x) = \frac{3}{4^x} \quad x = 1, 2, 3, \dots$$

Now, since:

$$G_X(t) = E(t^X) = \sum_x t^x P(X = x) = p_0 + p_1 t + p_2 t^2 + p_3 t^3 + \dots$$

We see that the probabilities $p_k = P(X = k)$ are the numbers “in front of” the corresponding powers of t . **Thus, p_k = the coefficient of t^k in $G_X(t)$.**

This is why it is called a probability generating function. We can use this idea to find probabilities of a distribution if we know its PGF:



Question 5.3

The PGF of a discrete random variable, Y , is given by:

$$G_Y(t) = 1 + (t - 1)(0.2t + 0.6).$$

Determine $P(Y = 0)$, $P(Y = 1)$ and $P(Y = 2)$.

We will often have to use a series expansion in order to obtain the powers of t :



Question 5.4

The PGF of the discrete random variable, N , is $G_N(t) = \left(\frac{p}{1-qt}\right)^k$, where $p + q = 1$.

By writing it as $G_N(t) = p^k (1 - qt)^{-k}$, use a series expansion to find $P(N = 3)$ in terms of k .

1.2 Important examples

The PGFs for some of the distributions introduced earlier are found as follows.

Uniform

$$P(X = x) = \frac{1}{k}, \quad x = 1, 2, 3, \dots, k$$

$$G_X(t) = E[t^X] = \frac{1}{k}(t + t^2 + \dots + t^k) = \frac{t}{k} \frac{(1-t^k)}{(1-t)} \quad \text{for } t \neq 1$$

This is the simplest example of a discrete uniform distribution, but there are others that take values over a different range of integers.



Question 5.5

A random variable U can take any of the values $\{a, a+1, a+2, \dots, b-1, b\}$, where $a < b$, each with equal probability. Derive a formula for the PGF of U .

Binomial (n, θ) (including Bernoulli, for which $n=1$)

$$P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

$$G_X(t) = \sum_{x=0}^n \binom{n}{x} (\theta t)^x (1-\theta)^{n-x} = [\theta t + (1-\theta)]^n$$

You will often see p and q used instead of θ and $1-\theta$ as in the *Tables* on page 6.



Question 5.6

Derive from first principles the probability generating function of a Bernoulli random variable with parameter θ .

An alternative method of obtaining the PGF (and MGF) of a binomial random variable is given in Chapter 6, Section 4.2 using the fact that $Bin(n, \theta)$ is the sum of n independent $Bernoulli(\theta)$ random variables.

Type 1 Negative binomial (k, θ) (incl. geometric, for which $k = 1$)

For convenience let $P(\text{failure}) = 1 - \theta = \phi$

$$P(X = x) = \binom{x-1}{k-1} \theta^k \phi^{x-k} \quad x = k, k+1, k+2, \dots$$

$$\begin{aligned} G_X(t) &= \sum_{x=k}^{\infty} \binom{x-1}{k-1} t^x \theta^k \phi^{x-k} \\ &= (\theta t)^k \sum_{x=k}^{\infty} \binom{x-1}{k-1} (\phi t)^{x-k} \end{aligned}$$

When this sum is expanded, we get $1 + k\phi t + \frac{1}{2}k(k+1)(\phi t)^2 + \dots$, which, as we saw in, Question 5.4 is just $(1 - \phi t)^{-k}$. Therefore the PGF is:

$$G_X(t) = \left(\frac{\theta t}{(1 - \phi t)} \right)^k$$

Note: The summation is valid for $|\phi t| < 1$, ie for $|t| < \frac{1}{\phi}$.

An alternative method of obtaining the PGF (and MGF) of a negative binomial random variable, without the need for summation, is given in Chapter 6, Section 4.2. It uses the fact that a negative binomial with parameters (k, θ) is the sum of k independent geometric random variables with parameter θ .

Hypergeometric

PGF not used.

Poisson (λ)

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

$$G_X(t) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda t)^x}{x!} = e^{\lambda(t-1)}$$

The last step uses the exponential series expansion $e^{\lambda t} = \sum_{x=0}^{\infty} \frac{(\lambda t)^x}{x!}$. The exponential series is given on page 2 of the *Tables*.

1.3 Evaluating moments

The PGF $G_X(t)$ can be used quite easily to find low-order moments.

Expand the function t^X as follows (using a “Taylor expansion” about the point $t = 1$):

$$t^X = 1 + X(t-1) + \frac{X(X-1)(t-1)^2}{2!} + \frac{X(X-1)(X-2)(t-1)^3}{3!} + \dots$$

This is the same as writing t^X as $[1 + (t-1)]^X$ and using a binomial expansion.

Now taking expectations of both sides:

$$\begin{aligned} G_X(t) &= E[t^X] \\ &= 1 + (t-1)E[X] + \frac{(t-1)^2}{2!} E[X(X-1)] + \frac{(t-1)^3}{3!} E[X(X-1)(X-2)] + \dots \end{aligned}$$

Differentiation of this expansion with respect to t , and setting $t = 1$, produces the mean. Further differentiation gives formulae that can be used to determine other low-order moments.

The reason we will only determine other “low-order moments” is that the algebra becomes rather messy! It is *much* easier to obtain moments using MGFs or CGFs. However, examiners are not always that generous! So make sure you can remember the awkward formulae given below or learn the derivation.

Differentiating once and substituting $t = 1$ into the resulting expression:

$$G'_X(t) = E[X] + (t-1)E[X(X-1)] + \frac{(t-1)^2}{2!}E[X(X-1)(X-2)] + \dots$$

$$\Rightarrow G'_X(1) = E[X]$$

Differentiating again and substituting $t = 1$ into the resulting expression:

$$G''_X(t) = E[X(X-1)] + (t-1)E[X(X-1)(X-2)] + \dots$$

$$\Rightarrow G''_X(1) = E[X(X-1)]$$

$$= E[X^2] - E[X]$$

$$\Rightarrow E[X^2] = G''_X(1) + E[X] = G''_X(1) + G'_X(1)$$

Therefore, since $\text{var}[X] = E[X^2] - E^2[X]$, we obtain:

$$\text{var}[X] = G''_X(1) + G'_X(1) - [G'_X(1)]^2$$



Question 5.7

Calculate the mean and variance of a random variable, X , with PGF given by:

$$G_X(t) = \frac{3}{4-t}$$

An alternative and frankly simpler method to derive the mean and variance is to differentiate the definition of the PGF and substitute $t = 1$ in as before:

$$G_X(t) = \sum t^x P(X = x) = E(t^X)$$

$$\Rightarrow G'_X(t) = \sum xt^{x-1}P(X = x) = E(Xt^{X-1})$$

$$\Rightarrow G'_X(1) = \sum xP(X = x) = E(X)$$

$$G''_X(t) = \sum x(x-1)t^{x-2}P(X = x) = E[X(X-1)t^{X-2}]$$

$$\Rightarrow G''_X(1) = E[X(X-1)] \quad \text{etc}$$

2 Moment generating functions

2.1 General formula

A moment generating function (MGF) can be used to generate moments, namely the moments (about the origin) of the distribution of a random variable (discrete or continuous), ie $E(X), E(X^2), E(X^3), \dots$

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:

$$M_X(t) = E[e^{tX}]$$

for all values of t for which the expectation exists.

Note that MGFs can be defined for both discrete and continuous random variables.



Question 5.8

What is $M_X(0)$?

Now, in Chapter 3 we 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:

$$M_X(t) = E[e^{tX}] = \sum_x e^{tx}P(X=x) \quad \text{or} \quad \int_x e^{tx}f_X(x) dx .$$

Notice the similarity between the discrete MGF and the PGF. So obtaining the MGF for a discrete random variable will be very similar to finding the PGF.

**Question 5.9**

Derive the MGF of the random variable X with probability function:

$$P(X = x) = \frac{3}{4^x} \quad x = 1, 2, 3, \dots$$

Now to find the MGF of a continuous distribution:

**Question 5.10**

Find the MGF of the random variable X with probability density function:

$$f(x) = \frac{1}{2}(1-x) \quad -1 \leq x \leq 1$$

In a moment we'll look at how to obtain the MGFs of the standard distributions given in Chapter 4, but first let's find out how we can use MGFs to find moments:

Finding moments

In order to obtain moments, *ie* $E(X), E(X^2), E(X^3), \dots$, from the MGF we can either differentiate the MGF or use a series expansion of the MGF:

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.

Note: $M'_X(t) = E[Xe^{tX}]$, since $M_X(t) = E[e^{tX}]$ and $\frac{d}{dt}e^{tX} = Xe^{tX}$.

$$\Rightarrow M'_X(0) = E[Xe^0] = E[X]$$

Similarly:

$$M''_X(t) = E[X^2 e^{tX}] \Rightarrow M''_X(0) = E[X^2]$$

$$M'''_X(t) = E[X^3 e^{tX}] \Rightarrow M'''_X(0) = E[X^3]$$

etc



Question 5.11

Find the mean and variance of a random variable, X , with MGF given by:

$$M_X(t) = \left(1 - \frac{t}{5}\right)^{-1} \quad t < 5$$

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 messy.

Expanding the exponential function and taking expected values throughout (a procedure which is justifiable for the distributions here) gives:

$$\begin{aligned} M_X(t) &= E(e^{tX}) = E\left(1 + tX + \frac{t^2}{2!} X^2 + \frac{t^3}{3!} X^3 + \dots\right) \\ &= 1 + tE[X] + \frac{t^2}{2!} E[X^2] + \frac{t^3}{3!} E[X^3] + \dots \end{aligned}$$

from which it is seen that the r th moment of the distribution about the origin, $E[X^r]$, is obtainable as the coefficient of $\frac{t^r}{r!}$ in the power series expansion of the MGF.

To use this method to find moments, we will first need to obtain a series expansion of the MGF, just like we did for PGFs. We then equate the coefficients of the powers of t with the above expression:



Question 5.12

Use a series expansion to find $E(X)$, $E(X^2)$ and $E(X^3)$ of a random variable, X , with MGF given by:

$$M_X(t) = \left(1 - \frac{t}{5}\right)^{-1} \quad t < 5$$

Note: if we differentiate the series expansion for the MGF with respect to t and then substituted $t = 0$ this would again give $M'_X(0) = E(X)$, $M''_X(0) = E(X^2)$, ... as we got before:

$$\begin{aligned} M_X(t) &= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \dots \\ \Rightarrow M'_X(t) &= E(X) + tE(X^2) + \frac{t^2}{2!}E(X^3) + \dots \Rightarrow M'_X(0) = E(X) \\ \Rightarrow M''_X(t) &= E(X^2) + tE(X^3) + \dots \Rightarrow M''_X(0) = E(X^2) \quad \text{etc} \end{aligned}$$

Use 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.

So, given a distribution we can calculate moments. Therefore if we are given a set of moments, we can identify the distribution that they came from. This is demonstrated in the example below.

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 Chapter 6.



Question 5.13

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 .

**Example**

Identify the continuous distribution for which $E[X^k] = \frac{k!}{\lambda^k}$ where $k = 1, 2, 3, \dots$, and $\lambda > 0$.

Solution

The moment generating function of a random variable X is:

$$M_X(t) = 1 + tE[X] + \frac{t^2}{2!} E[X^2] + \frac{t^3}{3!} E[X^3] + \dots$$

Substituting in the values of the moments given:

$$M_X(t) = 1 + \frac{1}{\lambda}t + \frac{2!}{\lambda^2} \frac{t^2}{2!} + \frac{3!}{\lambda^3} \frac{t^3}{3!} + \dots = 1 + \frac{t}{\lambda} + \frac{t^2}{\lambda^2} + \frac{t^3}{\lambda^3} + \dots$$

This is $(1-t/\lambda)^{-1}$. By comparing this to standard MGFs we can see that the unknown distribution is the exponential distribution with parameter λ .

Recall from Question 5.2 that the PGF for the random variable, X , with probability function $P(X = x) = \frac{3}{4^x}$, $x = 1, 2, \dots$ was given by $\frac{3t}{4-t}$. The MGF of X was found in

0 to be $\frac{3e^t}{4-e^t}$. Effectively all we have done is to replace the t in the PGF by e^t in order to obtain the MGF.

For a counting variable which has a PGF $G_X(t)$ [$= E(t^X)$], then its MGF $M_X(t)$ can be found by substituting e^t for t in $G_X(t)$, ie $M_X(t) = G_X(e^t)$.

**Question 5.14**

Show that $M_X(t) = G_X(e^t)$ and $G_X(t) = M_X(\ln t)$.

Since we obtained the PGFs of the binomial, negative binomial and Poisson distributions earlier, we can now use this result to find their respective MGFs.

So:

Binomial (n, θ)

Recall that $G_X(t) = [\theta t + (1 - \theta)]^n$, hence:

$$\begin{aligned} M_X(t) &= (\theta e^t + \phi)^n \text{ where } \phi = 1 - \theta \\ &= [1 + \theta(e^t - 1)]^n \end{aligned}$$

This is written as $(q + pe^t)^n$ in the *Tables*.

Negative binomial (k, θ)

Recall that $G_X(t) = \left(\frac{\theta t}{(1 - \phi t)} \right)^k$, hence:

$$M_X(t) = \left[\frac{\theta e^t}{(1 - \phi e^t)} \right]^k$$

Poisson (λ)

Recall that $G_X(t) = e^{\lambda(t-1)}$, hence:

$$M_X(t) = \exp\{\lambda(e^t - 1)\}$$

Whilst using this handy shortcut to obtain the MGFs is useful, if an exam question asks you to derive the MGF of, say, a Poisson distribution, you would be expected to obtain the PGF first and *then* use $M_X(t) = G_X(e^t)$.

Alternatively, you could just derive the MGFs from first principles in a very similar way to that used for PGFs. Do practise this! Again, do note that there is an easier method of deriving the MGFs of the binomial and negative binomial given in Chapter 6, Section 4.2.

**Question 5.15**

Derive from first principles the moment generating function of a random variable X , where $P(X = x) = \theta(1 - \theta)^{x-1} \quad x = 1, 2, 3, \dots$.

2.2 Important examples – continuous variables

We will now look at how to calculate the MGF of some standard continuous distributions.

**Question 5.16**

Derive from first principles the moment generating function of a $U(a, b)$ distribution.

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 $y = (\lambda - t)x$, if $y = (\lambda - t)x$, then

$\frac{dy}{dx} = \lambda - t$, so the integral becomes:

$$M_X(t) = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{y}{\lambda - t} \right)^{\alpha-1} e^{-y} \frac{1}{\lambda - t} dy = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\lambda - t} \right)^\alpha y^{\alpha-1} e^{-y} dy$$

$$\text{So } \left(\frac{1}{\lambda} \right)^\alpha \Gamma(\alpha) M_X(t) = \int_0^\infty \left(\frac{1}{\lambda - t} \right)^\alpha y^{\alpha-1} e^{-y} dy = \left(\frac{1}{\lambda - t} \right)^\alpha \Gamma(\alpha).$$

In the last step we've used the definition of the gamma function. This is given on page 5 of the *Tables*.

$$\text{So } M_X(t) = \lambda^\alpha (\lambda - t)^{-\alpha} = \left(\frac{\lambda}{\lambda - t} \right)^\alpha = \left(\frac{1}{1 - t/\lambda} \right)^\alpha = \left(1 - \frac{t}{\lambda} \right)^{-\alpha}.$$

This formula only holds when $t < \lambda$ and is given on page 12 of the *Tables*.



Question 5.17

What is $E(e^{tX})$ for the gamma distribution when $t \geq \lambda$?

We can use the MGF to find the moments of the gamma distribution.

From $M_X(t) = \lambda^\alpha (\lambda - t)^{-\alpha}$:

$$M'_X(t) = \alpha \lambda^\alpha (\lambda - t)^{-\alpha-1} \text{ so } E[X] = M'_X(0) = \frac{\alpha}{\lambda}$$

$$M''_X(t) = \alpha(\alpha+1) \lambda^\alpha (\lambda - t)^{-\alpha-2} \text{ so } E[X^2] = M''_X(0) = \frac{\alpha(\alpha+1)}{\lambda^2}$$

$$\text{Hence, } \mu = \frac{\alpha}{\lambda}, \sigma^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda} \right)^2 = \frac{\alpha}{\lambda^2}.$$

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 will be $\theta = \frac{1}{\lambda}$.

Note: The MGF of the chi-square ν distribution is given by $M_X(t) = (1 - 2t)^{-\nu/2}$.



Question 5.18

Show that this is true.

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 here.

The result is:

$$M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$$



Question 5.19

Prove this result.

We can also check the moments of the normal distribution.

$$\text{Expanding, } M_X(t) = 1 + \left(\mu t + \frac{1}{2} \sigma^2 t^2 \right) + \frac{(\mu t + \frac{1}{2} \sigma^2 t^2)^2}{2!} + \dots$$

$$\begin{aligned} E[X] &= \text{coefficient of } t \\ &= \mu \end{aligned}$$

(confirming that the parameter μ does indeed represent the mean).

$E[X^2] = \text{coefficient of } \frac{t^2}{2!} = \sigma^2 + \mu^2$ 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 have differentiated 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 Z has MGF:

$$M_Z(t) = \exp(\frac{1}{2}t^2) = 1 + \frac{1}{2}t^2 + \frac{\left(\frac{1}{2}t^2\right)^2}{2!} + \dots$$

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 $E[(X - \mu)^3] = 0$, $E[(X - \mu)^4] = 3\sigma^4$.

Remember that $E[(X - \mu)^3]$ gives the skewness. We stated earlier that the normal distribution was symmetrical, hence we would expect μ_3 to be zero. This has now been proved.



Question 5.20

In this last result, we have used the fact that if we standardise a normal random variable X by calculating $Z = \frac{X - \mu}{\sigma}$, then Z has a standard normal distribution. Use moment generating functions to show that this is true.

The MGFs do not exist in closed form for the Beta and lognormal distributions. Hence, they are excluded from this section.

3 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:

$$C_X(t) = \ln M_X(t)$$

You can treat this as a definition of the CGF.



Question 5.21

The MGF of a binomial distribution is given by:

$$M(t) = (q + pe^t)^n$$

State the CGF of the binomial distribution.

and so $M_X(t) = e^{C_X(t)}$.

As a result if $C_X(t)$ is known it is easy to determine $M_X(t)$.

Finding moments

If we differentiate, we obtain:

$$C'_X(t) = \frac{M'_X(t)}{M_X(t)}$$

and:

$$C''_X(t) = \frac{M''_X(t)M_X(t) - (M'_X(t))^2}{M_X^2(t)}$$

Now $M_X(0) = 1$ so:

$$C'_X(0) = \frac{M'_X(0)}{M_X(0)} = \frac{E[X]}{1}$$

and:

$$\begin{aligned} C''_X(0) &= \frac{M''_X(0)M_X(0) - (M'_X(0))^2}{M_X^2(0)} \\ &= \frac{E[X^2](1) - (E[X])^2}{1^2} \\ &= \text{var}[X] \end{aligned}$$

Thus the first two derivatives of $C_X(t)$ evaluated at $t = 0$ give the mean and variance of X directly.

In fact the third derivative (evaluated at $t = 0$) gives the skewness as well.



Example

State the CGF of X where $X \sim \text{Gamma}(\alpha, \lambda)$. Hence prove that $E(X) = \frac{\alpha}{\lambda}$ and $\text{var}(X) = \frac{\alpha}{\lambda^2}$.

Solution

$$M_X(t) = \frac{\lambda^\alpha}{(\lambda - t)^\alpha} = \left(1 - \frac{t}{\lambda}\right)^{-\alpha} \Rightarrow C_X(t) = -\alpha \ln\left(1 - \frac{t}{\lambda}\right) \quad t < \lambda$$

Differentiating with respect to t , and substituting $t = 0$, we obtain:

$$C'_X(t) = -\alpha \times \frac{-\frac{1}{\lambda}}{\left(1 - \frac{t}{\lambda}\right)} = \frac{\alpha}{\lambda} \left(1 - \frac{t}{\lambda}\right)^{-1} \Rightarrow E(X) = C'_X(0) = \frac{\alpha}{\lambda}$$

$$C''_X(t) = -\frac{\alpha}{\lambda} \left(1 - \frac{t}{\lambda}\right)^{-2} \times -\frac{1}{\lambda} = \frac{\alpha}{\lambda^2} \left(1 - \frac{t}{\lambda}\right)^{-2} \Rightarrow \text{var}(X) = C''_X(0) = \frac{\alpha}{\lambda^2}$$



Question 5.22

Obtain the cumulant generating function of the normal distribution, and hence determine the mean, variance and coefficient of skewness of the normal distribution.

Definition:

The coefficient of $\frac{t^r}{r!}$ in the Maclaurin series of $C_X(t) = \ln M_X(t)$ is called the *r*th cumulant and is denoted by κ_r .

Alternatively, and more simply, we can obtain the cumulants simply by differentiating the CGF with respect to t and substituting $t = 0$ into the resulting expressions:

$$\kappa_1 = C'_X(0)$$

$$\kappa_2 = C''_X(0)$$

$$\kappa_3 = C'''_X(0) \quad etc$$

Cumulants are similar to moments. Note that the first three cumulants are the mean, the variance and the skewness.



Question 5.23

By using the CGF of a Poisson distribution, obtain the 2nd, 3rd and 4th cumulants.

You can see that the CGF is particularly useful where the MGF is in the form of an exponential function – it makes the differentiation a lot easier!

4 Linear functions

Suppose X has PGF $G_X(t)$ and the distribution of a linear function $Y = a + bX$ is of interest. The PGF of Y , $G_Y(t)$ say, can be obtained from that of X as follows:

$$G_Y(t) = E[t^Y] = E[t^{a+bX}] = t^a E[t^{bX}] = t^a G_X(t^b).$$



Question 5.24

If X has a *Poisson*(5) distribution, find the PGF of $Y = 2X + 3$.

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:

$$M_Y(t) = E[e^{tY}] = E[e^{t(a+bX)}] = e^{at} E[e^{btX}] = e^{at} M_X(bt)$$

We are now in a position to prove a result quoted in Chapter 4, Section 2.2:



Question 5.25

Use MGFs to show that if X has a *Gamma* (α, λ) distribution, then $2\lambda X$ has a $\chi^2_{2\alpha}$ distribution. Hence, if X is *Gamma* (20,0.4), estimate the probability that $X > 75$.

This is an important result as it is the only practical way to calculate probabilities of a gamma distribution in an exam.



Question 5.26

If $Y = a + bX$, find and simplify an expression for $C_Y(t)$ in terms of $C_X(t)$.

5 ***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 Chapter 6, Section 3.

A linear combination of the random variables X_1, \dots, X_n is:

$$c_1X_1 + \dots + c_nX_n$$

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 they are independent, then $X_1 + X_2$ has a $Poi(\mu_1 + \mu_2)$ distribution.

Moment generating functions can also be used to calculate moments for and specify compound distributions. This will be covered in detail in Chapter 7, Sections 4 and 5.

A compound distribution is a distribution that consists of a random number of random variables. For example:

$$S = X_1 + \dots + X_N$$

where N is a discrete random variable giving the number of the random variables X_i summed.

We can use MGFs (or CGFs) to obtain the moments of such a distribution.

6 Exam question



Past Exam Question (Subject C1, April 1996, Q10)

- (i) Determine the moment generating function of the two parameter exponential random variable X , defined by the probability density function:

$$f(x) = \lambda e^{-\lambda(x-\alpha)}, \quad x \geq \alpha \quad \text{where } \lambda, \alpha > 0.$$

- (ii) Hence, or otherwise, determine the mean and variance of the random variable X .

7 Appendix – geometric series

Definition

A series is the summation of terms in a sequence. A geometric series is a series where the next term is obtained by multiplying the previous term by a fixed number.

$$\begin{array}{ll} \text{eg} & \text{Series 1} \\ & 1 + 3 + 9 + 27 + 81 + \dots \\ & \text{Series 2} \\ & 6 + 3 + 1.5 + 0.75 + 0.375 + \dots \end{array}$$

A general expression for a geometric series is:

$$a + ar + ar^2 + ar^3 + \dots$$

where a is the first term and r is the common ratio (the fixed number we multiply by each time).

For Series 1 we have $a = 1$ and $r = 3$. For Series 2 we have $a = 6$ and $r = 0.5$.

Sum of the first n terms

The sum of the first n terms of a geometric series is given by:

$$S_n = \frac{a(1 - r^n)}{1 - r}$$

The sum of the first 10 terms for Series 1 is $\frac{1 \times (1 - 3^{10})}{1 - 3} = 29,524$. For Series 2 it is $\frac{6(1 - 0.5^{10})}{1 - 0.5} = 11.98828$.

Sum to infinity

The terms of Series 1 get bigger and so the summation will head to infinity. On the other hand the terms of series 2 get smaller and the summation will converge to a limit.

In general if $-1 < r < 1$ then the sum to infinity is given by:

$$S_\infty = \frac{a}{1 - r}$$

The sum to infinity of series 2 is $\frac{6}{1 - 0.5} = 12$.

8 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 5.
2. Attempt some of the questions in Part 1 of the **Question and Answer Bank**. 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 a recent survey suggest that attempting the assignments and having them marked improves your chances of passing the exam. Students have said:

“The marker gave really concise, constructive feedback, clearly making me aware of gaps in my knowledge and what needs to be done in order to achieve success in the exams. At the same time, the marker's comments were encouraging; ensuring that I stay motivated to do better in the next assignments. Thank you very much!”

Face-to-face 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. Here are a few comments made by past students:

“Face-to-face interaction I find is the best way to learn ... all the tutors I have had to date, have been enthusiastic and have helped by giving tips etc.”

Online Classroom – Alternatively / additionally, you might consider the Online Classroom to give you access to ActEd's expert tuition and additional support.

“I do like to order both online and face-to-face tutorials. There is so much to get through with face-to-face tutorials, online allows you to go through the material at your own pace and at the point that you are properly revising/learning specific areas of the course.”

You can find lots more information, including demos, on our website at www.ActEd.co.uk.

Buy online at www.ActEd.co.uk/estore

This page has been left blank so that you can keep the chapter summaries together for revision purposes.



Chapter 5 Summary

Generating functions are used to make it easier to find moments of distributions.

The probability generating function (PGF) of a counting random variable is defined to be:

$$G_X(t) = E[t^X]$$

The series expansion for PGFs is:

$$G_X(t) = P(X = 0) + tP(X = 1) + t^2P(X = 2) + t^3P(X = 3) + \dots$$

The moment generating function (MGF) of a random variable is defined to be:

$$M_X(t) = E[e^{tX}]$$

The series expansion for MGFs:

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \dots$$

The cumulant generating function (CGF) of a random variable is defined to be:

$$C_X(t) = \ln M_X(t)$$

Moments of a random variable can be found from its PGF, MGF or CGF using the formulae listed overleaf.

The *uniqueness property* means that if two variables have the same PGF, MGF or CGF then they have the same distribution.

If $Y = a + bX$, then:

$$G_Y(t) = t^a G_X(t^b), M_Y(t) = e^{at} M_X(bt) \text{ and } C_Y(t) = at + C_X(bt)$$



Chapter 5 Formulae

Probability Generating Functions

$$G_X(t) = E(t^X) = \sum_x t^x P(X=x)$$

$$E(X) = G'_X(1)$$

$$\text{var}(X) = G''_X(1) + G'_X(1) - (G'_X(1))^2$$

$$G_X(t) = P(X=0) + tP(X=1) + t^2P(X=2) + t^3P(X=3) + \dots$$

Moment Generating Functions

$$M_X(t) = E(e^{tX}) = \sum_x e^{tx} P(X=x) \quad \text{or} \quad \int_x e^{tx} f(x) dx$$

$$E(X) = M'_X(0)$$

$$\text{var}(X) = M''_X(0) - (M'_X(0))^2$$

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \dots$$

Cumulant Generating Functions

$$C_X(t) = \ln M_X(t)$$

$$E(X) = C'_X(0)$$

$$\text{var}(X) = C''_X(0)$$

$$\text{skew}(X) = C'''_X(0)$$

Linear Transformations

$$Y = aX + b \Rightarrow M_Y(t) = e^{bt} M_X(at)$$

Chapter 5 Solutions

Solution 5.1

The PGF is:

$$G(t) = \frac{1}{6}t + \frac{1}{6}t^2 + \cdots + \frac{1}{6}t^6$$

This is a geometric series with first term $a = \frac{1}{6}t$ and common ratio $r = t$. Summing this geometric series using the formula $S_n = \frac{a(1-r^n)}{1-r}$ gives:

$$G(t) = \frac{\frac{1}{6}t(1-t^6)}{1-t} = \frac{t}{6}(1-t^6)(1-t)^{-1}$$

Solution 5.2

The PGF is:

$$G_X(t) = E(t^X) = \sum_{x=1}^{\infty} t^x \frac{3}{4^x} = \frac{3}{4}t + \frac{3}{16}t^2 + \frac{3}{64}t^3 + \cdots$$

This is an infinite geometric series with first term $a = \frac{3}{4}t$ and common ratio $r = \frac{1}{4}t$.

Summing this geometric series to infinity using the formula $S_{\infty} = \frac{a}{1-r}$ gives:

$$G_X(t) = \frac{\frac{3}{4}t}{1 - \frac{1}{4}t} = \frac{3t}{4-t}$$

Remember this sum to infinity only exists when $-1 < r < 1$, ie when $-4 < t < 4$.

Solution 5.3

Expanding the brackets and simplifying gives:

$$G_Y(t) = 0.4 + 0.4t + 0.2t^2$$

Therefore by examining the coefficients we obtain:

$$P(Y = 0) = 0.4, P(Y = 1) = 0.4 \text{ and } P(Y = 2) = 0.2$$

Solution 5.4

Using the binomial expansion given on page 2 of the *Tables*:

$$\begin{aligned} G_N(t) &= p^k (1 - qt)^{-k} \\ &= p^k \left[1 + (-k)(-qt) + \frac{(-k)(-k-1)}{2!} (-qt)^2 + \frac{(-k)(-k-1)(-k-2)}{3!} (-qt)^3 + \dots \right] \\ &= p^k \left[1 + kqt + \frac{k(k+1)}{2!} (qt)^2 + \frac{k(k+1)(k+2)}{3!} (qt)^3 + \dots \right] \end{aligned}$$

$P(N = 3)$ is the coefficient of t^3 in this expansion, so:

$$P(N = 3) = \frac{k(k+1)(k+2)}{3!} p^k q^3$$

Solution 5.5

There are $b - a + 1$ values, each with probability $\frac{1}{b - a + 1}$.

So the PGF is given by:

$$\begin{aligned} G(t) &= \frac{1}{b-a+1} \left(t^a + t^{a+1} + t^{a+2} + \dots + t^b \right) \\ &= \frac{t^a}{b-a+1} \left(1 + t + t^2 + \dots + t^{b-a} \right) \end{aligned}$$

The terms in the bracket form a geometric series with first term $a = 1$ and common ratio $r = t$. Summing this geometric series using the formula $S_n = \frac{a(1-r^n)}{1-r}$ gives:

$$G(t) = \frac{t^a}{b-a+1} \left(\frac{1-t^{b-a+1}}{1-t} \right), \quad t \neq 1$$

In the notation of Core Reading, $a = 1$ and $b = k$. Notice that this is a generalisation of Question 5.1.

Solution 5.6

The probability distribution of Bernoulli random variable with parameter θ is:

$$P(X = 0) = 1 - \theta \quad \text{and} \quad P(X = 1) = \theta$$

Therefore, the PGF is given by:

$$G_X(t) = E\left(t^X\right) = \sum_x t^x P(X = x) = t^0(1 - \theta) + t^1\theta = (1 - \theta) + t\theta$$

Notice how the PGF of the binomial is the PGF of the Bernoulli random variable raised to the power of n . This connection is used in Chapter 6 to give an alternative method of obtaining the PGF of the binomial.

Solution 5.7

Now:

$$G_X(t) = \frac{3}{4-t} = 3(4-t)^{-1}$$

Differentiating this and substituting $t = 1$ into the resulting expressions gives:

$$G'_X(t) = 3(4-t)^{-2} \Rightarrow G'_X(1) = \frac{1}{3}$$

$$G''_X(t) = 6(4-t)^{-3} \Rightarrow G''_X(1) = \frac{2}{9}$$

Hence:

$$E(X) = G'_X(1) = \frac{1}{3}$$

$$\text{var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2 = \frac{2}{9} + \frac{1}{3} - \left(\frac{1}{3}\right)^2 = \frac{4}{9}$$

Solution 5.8

$$M_X(0) = E[e^0] = E[1] = 1$$

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 $t = 0$.

Solution 5.9

The MGF is:

$$M_X(t) = E(e^{tX}) = \sum_{x=1}^{\infty} e^{tx} \frac{3}{4^x} = \frac{3}{4}e^t + \frac{3}{16}e^{2t} + \frac{3}{64}e^{3t} + \dots$$

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 $S_{\infty} = \frac{a}{1-r}$ $-1 < r < 1$ gives:

$$M_X(t) = \frac{\frac{3}{4}e^t}{1 - \frac{1}{4}e^t} = \frac{3e^t}{4 - e^t}$$

where $-1 < \frac{1}{4}e^t < 1 \Rightarrow -4 < e^t < 4 \Rightarrow t < \ln 4$.

We are using the same distribution as in Question 5.2. Notice the similarity between the MGF of $\frac{3e^t}{4 - e^t}$ and the PGF of $\frac{3t}{4 - t}$. We will use this later to help us obtain the MGF if we have the PGF of a random variable.

Solution 5.10

The MGF is:

$$\begin{aligned}
 M_X(t) &= E(e^{tX}) = \int_x e^{tx} f(x) dx \\
 &= \int_{-1}^1 \frac{1}{2}(1-x)e^{tx} dx \\
 &= \int_{-1}^1 \frac{1}{2}e^{tx} dx - \int_{-1}^1 \frac{1}{2}xe^{tx} dx
 \end{aligned}$$

Now using integration by parts on the second integral we obtain:

$$\begin{aligned}
 M_X(t) &= \left[\frac{1}{2t} e^{tx} \right]_{-1}^1 - \left\{ \left[\frac{1}{2t} xe^{tx} \right]_{-1}^1 - \frac{1}{2t} \int_{-1}^1 e^{tx} dx \right\} \\
 &= \left[\frac{1}{2t} e^{tx} \right]_{-1}^1 - \left[\frac{1}{2t} xe^{tx} \right]_{-1}^1 + \frac{1}{2t} \left[\frac{1}{t} e^{tx} \right]_{-1}^1 \\
 &= \frac{1}{2t} e^t - \frac{1}{2t} e^{-t} - \left(\frac{1}{2t} e^t + \frac{1}{2t} e^{-t} \right) + \frac{1}{2t} \left(\frac{1}{t} e^t - \frac{1}{t} e^{-t} \right) \\
 &= -\frac{1}{t} e^{-t} + \frac{1}{2t^2} e^t - \frac{1}{2t^2} e^{-t}
 \end{aligned}$$

Solution 5.11

Now:

$$M_X(t) = \left(1 - \frac{t}{5}\right)^{-1}$$

Differentiating this and substituting $t = 0$ into the resulting expressions gives:

$$M'_X(t) = \frac{1}{5} \left(1 - \frac{t}{5}\right)^{-2} \Rightarrow E(X) = M'_X(0) = \frac{1}{5}$$

$$M''_X(t) = \frac{2}{25} \left(1 - \frac{t}{5}\right)^{-3} \Rightarrow E(X^2) = M''_X(0) = \frac{2}{25}$$

$$\Rightarrow \text{var}(X) = E(X^2) - E^2(X) = \frac{2}{25} - \left(\frac{1}{5}\right)^2 = \frac{1}{25}$$

Solution 5.12

Using the binomial expansion given on page 2 of the *Tables*:

$$\begin{aligned} M_X(t) &= \left(1 - \frac{t}{5}\right)^{-1} \\ &= \left[1 + (-1) \times \left(-\frac{t}{5}\right) + \frac{-1 \times -2}{2!} \left(-\frac{t}{5}\right)^2 + \frac{-1 \times -2 \times -3}{3!} \left(-\frac{t}{5}\right)^3 + \dots\right] \\ &= \left[1 + \frac{1}{5}t + \frac{1}{25}t^2 + \frac{1}{125}t^3 + \dots\right] \end{aligned}$$

Now recall that:

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \dots$$

Equating the coefficients gives:

$$E(X) = \frac{1}{5}$$

$$\frac{1}{2!}E(X^2) = \frac{1}{25} \Rightarrow E(X^2) = \frac{2}{25}$$

$$\frac{1}{3!}E(X^3) = \frac{1}{125} \Rightarrow E(X^3) = \frac{6}{125}$$

*Notice how the first two results match up with those found using differentiation in Question 5.11. The differentiation method for this particular distribution is easier than the series expansion and this is the method that will be used most of the time. However exam questions have been asked which specified that students should obtain the moments using a series expansion so make sure that you can use **both** methods!*

Solution 5.13

Examining the MGFs given in the *Tables* we want one which involves an exponential term. The normal distribution has the following MGF:

$$M(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$$

Equating coefficients, we see that $\mu = 5$ and $\sigma^2 = 6$. Hence by the uniqueness property since X has the same MGF as a $N(5,6)$, it means that X has a $N(5,6)$ distribution.

Solution 5.14

From the definition of the PGF:

$$G_X(e^t) = E[(e^t)^X] = E[e^{tX}] = M_X(t)$$

From the definition of the MGF:

$$M_X(\ln t) = E[e^{(\ln t)X}] = E[(e^{\ln t})^X] = E[t^X] = G_X(t)$$

Solution 5.15

Now $P(X = x) = \theta(1 - \theta)^{x-1}$, so the MGF is given by:

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \sum_{x=1}^{\infty} e^{tx} P(X = x) = \sum_{x=1}^{\infty} e^{tx} \theta(1 - \theta)^{x-1} \\ &= \theta \left(e^t + e^{2t}(1 - \theta) + e^{3t}(1 - \theta)^2 + \dots \right) \end{aligned}$$

The expression in the brackets is an infinite geometric series with $a = e^t$ and $r = e^t(1 - \theta)$. Summing it gives:

$$M_X(t) = \frac{\theta e^t}{1 - (1 - \theta)e^t} \quad \text{where } -1 < e^t(1 - \theta) < 1 \Rightarrow t < \ln\left(\frac{1}{1-\theta}\right)$$

Solution 5.16

$$M_X(t) = E[e^{tx}] = \int_a^b e^{tx} \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{1}{t} e^{tx} \right]_a^b = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

Solution 5.17

$$M_X(t) = E(e^{tX}) = \int_0^\infty \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-(\lambda-t)x} dx.$$

If $t \geq \lambda$, then the power in the exponential factor in the integral is positive and therefore the answer will be infinite. So the MGF does not exist in this case.

Solution 5.18

χ_v^2 is gamma with $\alpha = \frac{v}{2}$ and $\lambda = \frac{1}{2}$.

So it has the moment generating function:

$$M_X(t) = \frac{\left(\frac{1}{2}\right)^{\frac{v}{2}}}{\left(\frac{1}{2}-t\right)^{\frac{v}{2}}} = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{v}{2}} = \left(\frac{1}{1-2t}\right)^{\frac{v}{2}} = (1-2t)^{-\frac{v}{2}}$$

Solution 5.19

The moment generating function of the normal 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:

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[tx - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x^2 - 2\mu x - 2tx\sigma^2 + \mu^2)\right] dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x^2 - 2x(\mu + t\sigma^2) + \mu^2)\right] dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}((x - (\mu + t\sigma^2))^2 + \mu^2 - (\mu + t\sigma^2)^2)\right] dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x - (\mu + t\sigma^2))^2\right] \exp\left[-\frac{1}{2\sigma^2}(-2\mu t\sigma^2 - t^2\sigma^4)\right] dx \end{aligned}$$

Since the second term in the integral does not depend on x , we can take it outside the integral:

$$\begin{aligned} M_X(t) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(-2\mu t\sigma^2 - t^2\sigma^4)\right] \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x - (\mu + t\sigma^2))^2\right] dx \\ &= \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - (\mu + t\sigma^2)}{\sigma}\right)^2\right] dx \end{aligned}$$

The function now being integrated is equivalent to the PDF of a normal distribution with mean $\mu + t\sigma^2$ and standard deviation σ , so the integral must be 1, giving us:

$$M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right], \text{ as required}$$

Solution 5.20

The MGF of $Z = \frac{X - \mu}{\sigma}$ is:

$$\begin{aligned} M_Z(t) &= E[e^{tZ}] = E\left[e^{t\left(\frac{X-\mu}{\sigma}\right)}\right] \\ &= e^{-\frac{\mu t}{\sigma}} E\left[e^{\frac{t}{\sigma}X}\right] \\ &= e^{-\frac{\mu t}{\sigma}} M_X\left(\frac{t}{\sigma}\right) \end{aligned}$$

Using the formula for the MGF of a normal 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 $\frac{X - \mu}{\sigma}$ has a standard normal distribution.

Solution 5.21

$$C_X(t) = \ln M_X(t) = \ln(q + pe^t)^n = n \ln(q + pe^t)$$

Solution 5.22

For the normal distribution:

$$\begin{aligned} M_X(t) &= \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \\ \Rightarrow C_X(t) &= \ln M_X(t) = \mu t + \frac{1}{2}\sigma^2 t^2 \end{aligned}$$

Differentiating and setting $t = 0$ gives:

$$\begin{aligned} C'_X(t) &= \mu + \sigma^2 t \quad \Rightarrow \quad E(X) = C'_X(0) = \mu \\ C''_X(t) &= \sigma^2 \quad \Rightarrow \quad \text{var}(X) = C''_X(0) = \sigma^2 \\ C'''_X(t) &= 0 \quad \Rightarrow \quad \text{Skew}(X) = C'''_X(0) = 0 \end{aligned}$$

Since the skewness is zero, the coefficient of skewness is also 0.

Solution 5.23

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:

$$\begin{aligned} C'_X(t) &= \mu e^t \\ C''_X(t) &= \mu e^t \quad \Rightarrow \quad \kappa_2 = C''_X(0) = \mu e^0 = \mu \\ C'''_X(t) &= \mu e^t \quad \Rightarrow \quad \kappa_3 = C'''_X(0) = \mu \\ C''''_X(t) &= \mu e^t \quad \Rightarrow \quad \kappa_4 = C''''_X(0) = \mu \end{aligned}$$

So the 2nd, 3rd and 4th cumulants of the Poisson distribution are all equal to μ .

Solution 5.24

The probability generating function of the Poisson distribution is $G_X(t) = e^{\lambda(t-1)}$.

If $Y = a + bX$, then $G_Y(t) = t^a G_X(t^b)$.

In this question, $\lambda = 5$, $a = 3$ and $b = 2$, so:

$$G_Y(t) = t^3 G_X(t^2) = t^3 e^{5(t^2-1)}$$

Solution 5.25

The MGF of the gamma distribution is $M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-\alpha}$.

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 $\chi_{2\alpha}^2$ distribution.

If X is Gamma (20, 0.4), then this result tells us that $0.8X$ is χ_{40}^2 , which gives:

$$P(X > 75) = P(0.8X > 60) = P(\chi_{40}^2 > 60)$$

From the tables of the χ_{40}^2 distribution, this probability is just less than 0.025.

Solution 5.26

Now $C_Y(t) = \ln M_Y(t)$, so using the expression for $M_Y(t)$, we obtain:

$$\begin{aligned} C_Y(t) &= \ln M_Y(t) \\ &= \ln[e^{at} M_X(bt)] \\ &= at + \ln M_X(bt) \\ &= at + C_X(bt) \end{aligned}$$

Past Exam Question (Subject C1, April 1996, Q10)

- (i) Using the definition of an MGF:

$$\begin{aligned} M_X(t) &= E[e^{tX}] = \int_{\alpha}^{\infty} e^{tx} \lambda e^{-\lambda(x-\alpha)} dx \\ &= \int_{\alpha}^{\infty} \lambda e^{\lambda\alpha} e^{-(\lambda-t)x} dx \\ &= \left[-\frac{\lambda e^{\lambda\alpha}}{\lambda-t} e^{-(\lambda-t)x} \right]_{\alpha}^{\infty} \\ &= \frac{\lambda}{\lambda-t} e^{t\alpha} \quad \text{provided } t < \lambda \end{aligned}$$

- (ii) Re-writing the MGF to make it easier to differentiate:

$$\begin{aligned} M_X(t) &= \left(1 - \frac{t}{\lambda}\right)^{-1} e^{t\alpha} \\ M'_X(t) &= \frac{1}{\lambda} \left(1 - \frac{t}{\lambda}\right)^{-2} e^{t\alpha} + \alpha \left(1 - \frac{t}{\lambda}\right)^{-1} e^{t\alpha} \\ \Rightarrow E(X) &= M'_X(0) = \frac{1}{\lambda} + \alpha \end{aligned}$$

$$\begin{aligned}
 M''_X(t) &= \frac{2}{\lambda^2} \left(1 - \frac{t}{\lambda}\right)^{-3} e^{t\alpha} + \frac{2\alpha}{\lambda} \left(1 - \frac{t}{\lambda}\right)^{-2} e^{t\alpha} + \alpha^2 \left(1 - \frac{t}{\lambda}\right)^{-1} e^{t\alpha} \\
 \Rightarrow E(X^2) &= M''_X(0) = \frac{2}{\lambda^2} + \frac{2\alpha}{\lambda} + \alpha^2 \\
 \Rightarrow \text{var}(X) &= \frac{2}{\lambda^2} + \frac{2\alpha}{\lambda} + \alpha^2 - \left(\frac{1}{\lambda} + \alpha\right)^2 = \frac{1}{\lambda^2}
 \end{aligned}$$

Alternatively, we could notice that if $X \sim \text{Exp}(\lambda)$ then $Y = X + \alpha$ has a two-parameter exponential distribution. Therefore:

$$\begin{aligned}
 M_Y(t) &= E[e^{tY}] = E[e^{tX+t\alpha}] = e^{t\alpha} E[e^{tX}] = e^{t\alpha} M_X(t) = e^{t\alpha} \left(1 - \frac{t}{\lambda}\right)^{-1} \\
 E(Y) &= E(X + \alpha) = E(X) + \alpha = \frac{1}{\lambda} + \alpha \\
 \text{var}(Y) &= \text{var}(X + \alpha) = \text{var}(X) = \frac{1}{\lambda^2}
 \end{aligned}$$

Chapter 6

Joint distributions



Syllabus objectives

- (vi) 1. *Explain what is meant by jointly distributed random variables, marginal distributions and conditional distributions.*
2. *Define the probability function/density function of a marginal distribution and of a conditional distribution.*
3. *Specify the conditions under which random variables are independent.*
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.*
5. *Define the probability function/density function of the sum of two independent random variables as the convolution of two functions.*
6. *Derive the mean and variance of linear combinations of random variables.*
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.

A number of students find the notation in this chapter initially daunting, particular if the work of Chapter 3 was new. It is therefore helpful to notice the parallels with the single random variables covered in Chapter 3.

Firstly we will define a joint probability (density) function $P(X = x, Y = y)$ or $f(x, y)$. We will then see how we can obtain a single distribution *i.e.* $P(X = x)$ or $f(x)$ from the joint distribution. Lastly we will start to 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 Chapter 7, Section 1 with Section 1.3 of this chapter as it is common for exam questions to involve these two areas.

Once we have our joint distributions we will work out the mean and the equivalent of the variance (called the *covariance*). We will also look at the correlation between the two random variables. This work will be continued in Chapter 13, where we will attempt to estimate what the correlation is from a sample.

Finally, we will extend our work on PGFs and MGFs from Chapter 5 to combine distributions together. This will give us easier ways of obtaining results for the binomial, negative binomial and gamma distributions.

This is a long chapter with a lot of new results and will probably require two study sessions to cover it in detail. If you wish to do this then there is a natural break after Section 2.

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:

y	x		
	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:

$$f(x,y) \geq 0 \text{ for all values of } x \text{ and } y \text{ in the domain}$$

$$\sum_x \sum_y f(x,y) = 1$$

This parallels the results of Chapter 3, Section 1.4, where the probability function was $P(X = x)$ which had to satisfy $P(X = x) \geq 0 \quad \forall x$ and $\sum_x P(X = x) = 1$.

**Example 6.1**

The discrete random variables M and N have the joint probability function:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

Draw up a table showing the values of the joint probability function for M and N .

Solution

For example, $P(M = 1, N = 2) = \frac{1}{35 \times 2^0} = \frac{1}{35}$. Calculating this 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 6.1**

Use the table of probabilities given in 0 to calculate:

- (i) $P(M = 3, N = 1 \text{ or } 2)$
- (ii) $P(N = 3)$
- (iii) $P(M = 2 | N = 3)$.

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:

$$F(x,y) = P(X < x, Y < y)$$

and it is related to the joint density function by:

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$$

Note that the definition of $F(x,y)$ is $P(X \leq x, Y \leq y)$.

The conditions for a function to qualify as a joint probability density function of a pair of continuous random variables are:

$f(x,y) \geq 0$ for all values of x and y in the domain

$$\int \int_{x y} f(x,y) dx dy = 1$$

These results parallel those of Chapter 3, Section 2.2, where the probability density function was $f(x)$ which had to satisfy $f(x) \geq 0 \quad \forall x$ and $\int_x f(x) dx = 1$. Recall also

that probabilities were calculated using $P(a < X < b) = \int_{x=a}^b f(x) dx$.

If you are unfamiliar with double integrals and partial derivatives you might like to consult a pure mathematics textbook or see the FAC course.



Example 6.2

The continuous random variables U and V have the joint probability density function:

$$f_{U,V}(u,v) = \frac{2u+v}{3000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

Find $P(10 < U < 15, V > 0)$.

Solution

From the formula for the joint probability function:

$$P(10 < U < 15, V > 0) = \int_{u=10}^{15} \int_{v=0}^5 \frac{2u+v}{3,000} dudv$$

This can be integrated with respect to either u or v first. If we do v first, we get:

$$\begin{aligned} \int_{u=10}^{15} \int_{v=0}^5 \frac{2u+v}{3,000} dudv &= \int_{u=10}^{15} \left[\frac{2uv + \frac{1}{2}v^2}{3,000} \right]_{v=0}^5 du = \int_{u=10}^{15} \frac{10u + 12.5}{3,000} du \\ &= \left[\frac{5u^2 + 12.5u}{3,000} \right]_{10}^{15} \\ &= 0.229 \end{aligned}$$



Question 6.2

Let X and Y have joint density function given by:

$$f(x,y) = c(x+3y) \quad 0 < x < 2, 0 < y < 2$$

- (i) Calculate the value of c .
- (ii) Hence, calculate $P(X < 1, Y > 0.5)$.

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 Question 6.1(ii) when we calculated $P(N = 3)$, except now we will do it for every value of N to get the probability distribution.



Question 6.3

Let X and Y have the joint probability function given in the Core Reading:

		x		
		1	2	3
y		1	0.10	0.10
		2	0.15	0.10
	3	0.20	0.05	-
	4	0.15	0.05	-

- (i) Determine $P(X = 1)$, $P(X = 2)$ and $P(X = 3)$.
- (ii) Hence, give the marginal probability distribution of X .
- (iii) Obtain the marginal probability distribution of Y .

We can also do this given a function:



Example 6.3

Obtain the probability function for the marginal distribution of M from 0, where:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

Solution

Summing over the values of N gives:

$$f_M(m) = \sum_{n=1}^3 P(M = m, N = n) = \sum_{n=1}^3 \frac{m}{35 \times 2^{n-2}} = \frac{m}{35} \left(2 + 1 + \frac{1}{2} \right) = \frac{m}{10}$$



Question 6.4

Determine the probability function for the marginal distribution of N from 0, where:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

Continuous case

In the case of continuous variables the marginal density function 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 $f_X(x) = \int_y f(x,y) dy$.

The resulting $f_X(x)$ is a proper PDF – it integrates to 1. Similarly for $f_Y(y)$, by “integrating over x ” (for the given value of y).



Question 6.5

Determine the marginal probability density functions for U and V given in Example 6.2, where:

$$f_{U,V}(u,v) = \frac{2u+v}{3,000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

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.

For example:

$$f(x,y) = \begin{cases} 0.5 & x > 0, y > 0, x + y < 2 \\ 0 & \text{otherwise} \end{cases}$$

We obtain the marginal distribution of X as follows:

$$f(x) = \int_{y=0}^{2-x} 0.5 dy$$

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 $P_{X|Y=y}(x | y)$ for the conditional distribution of X given $Y = y$ for discrete random variables X and Y is:

$$P_{X|Y=y}(x, y) = P(X = x | Y = y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$$

for all values x in the range of X

This conditional distribution is only defined for those values of y for which $P_Y(y) > 0$.

This is what we were doing in Question 6.1(iii) when we calculated $P(M = 2 | N = 3)$, except now we calculate it for every value of M to get the probability distribution.



Question 6.6

A bivariate distribution has the following probability function:

		X			
		0	1	2	
Y		1	0.1	0.1	0
		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 $X|Y = 2$.



Example 6.4

Find the conditional probability function for M given $N = n$ from 0, where:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

Solution

$$P_{M|N=n}(m, n) = \frac{P_{M,N}(m,n)}{P_N(n)} = \left(\frac{m}{35 \times 2^{n-2}} \right) \div \left(\frac{1}{7 \times 2^{n-3}} \right) = \frac{m}{10}$$



Question 6.7

Find the conditional probability function for N given $M = m$ from 0, where:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

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 is only defined for those values of y for which $f_Y(y) > 0$.



Question 6.8

Let X and Y have joint density function given in Question 6.2:

$$f(x, y) = \frac{1}{16}(x + 3y) \quad 0 < x < 2, 0 < y < 2$$

Determine the conditional density function of X given $Y = y$.

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 $f_{X,Y}(x, y) = f_X(x)f_Y(y)$

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 6.9

Are the variables X and Y from Question 6.6 independent? Remember that:

		X			
		0	1	2	
		1	0.1	0.1	0
		2	0.1	0.1	0.2
		3	0.2	0.1	0.1



Question 6.10

Are the variables M and N from Q6.6 independent? Remember that:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

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:

$$P(x_1 < X < x_2, y_1 < Y < y_2) = P(x_1 < X < x_2)P(y_1 < Y < y_2)$$



Question 6.11

Are the variables U and V from Example 6.2 independent? Remember that:

$$f_{U,V}(u,v) = \frac{2u+v}{3000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

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 you think of independent 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).



Example 6.5

Consider the probability density 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 marginal density functions of X and Z , Y and Z , and the marginal distributions of X , Y and Z .

$$f_{X,Z}(x,z) = \int_0^1 (x+y)e^{-z} dy = \left[e^{-z}(xy + \frac{1}{2}y^2) \right]_0^1 = e^{-z}(x + \frac{1}{2})$$

$$f_{Y,Z}(y,z) = \int_0^1 (x+y)e^{-z} dx = \left[e^{-z}(\frac{1}{2}x^2 + yx) \right]_0^1 = e^{-z}(y + \frac{1}{2})$$

$$f_X(x) = \int_0^\infty e^{-z}(x + \frac{1}{2}) dz = \left[-e^{-z}(x + \frac{1}{2}) \right]_0^\infty = x + \frac{1}{2}$$

$$f_Z(z) = \int_0^1 e^{-z}(x + \frac{1}{2}) dx = \left[e^{-z}(\frac{1}{2}x^2 + \frac{1}{2}x) \right]_0^1 = e^{-z}$$

$$f_Y(y) = \int_0^\infty e^{-z}(y + \frac{1}{2}) dz = \left[-e^{-z}(y + \frac{1}{2}) \right]_0^\infty = y + \frac{1}{2}$$

By finding the product of the marginal distributions for X , Y and Z , ie $(x + \frac{1}{2})(y + \frac{1}{2})e^{-z}$, comparing it to the joint distribution, ie $(x+y)e^{-z}$, and noting that they are unequal, it can be seen that X , Y and Z are not independent.

However the product of the marginal distributions for X and Z , and for Y and Z do give the respective joint distributions, 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 \times 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

$$E[g(X,Y)] = \sum_x \sum_y g(x,y)p_{X,Y}(x,y) = \sum_x \sum_y g(x,y)P(X=x, Y=y)$$

where the summation is over all possible values of x and y .

This result parallels that of Chapter 3, Section 3.1, where the expected value of a function of a discrete random variable was defined to be $E[g(X)] = \sum_x g(x)P(X=x)$.


Example 6.6

Calculate the expected value of $\frac{N+1}{M}$, where M and N are from 0:

		M			
		1	2	3	4
		$\frac{2}{35}$	$\frac{4}{35}$	$\frac{6}{35}$	$\frac{8}{35}$
N	2	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{4}{35}$
	3	$\frac{1}{70}$	$\frac{1}{35}$	$\frac{3}{70}$	$\frac{2}{35}$

From the table of values, working across from the top left gives:

$$E\left[\frac{N+1}{M}\right] = 2 \times \frac{2}{35} + 1 \times \frac{4}{35} + \dots + \frac{4}{3} \times \frac{3}{70} + 1 \times \frac{2}{35} = \frac{36}{35}$$

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:

$$\begin{aligned} E\left[\frac{N+1}{M}\right] &= \sum_{m=1}^4 \sum_{n=1}^3 \frac{n+1}{m} P(M = m, N = n) \\ &= \sum_{m=1}^4 \sum_{n=1}^3 \frac{n+1}{m} \times \frac{m}{35 \times 2^{n-2}} \\ &= \frac{1}{35} \sum_{m=1}^4 \sum_{n=1}^3 \frac{n+1}{2^{n-2}} \\ &= \frac{1}{35} \times 4 \times \left(\frac{1+1}{2^{-1}} + \frac{2+1}{2^0} + \frac{3+1}{2^1} \right) = \frac{36}{35} \end{aligned}$$

**Question 6.12**

Calculate the expected value of $X^2 + 2XY$, where X and Y have the joint distribution given in Question 6.6:

		X			
		0	1	2	
		1	0.1	0.1	0
		2	0.1	0.1	0.2
		3	0.2	0.1	0.1

For continuous variables

$$E[g(X, Y)] = \int \int_{x y} g(x, y) f_{X,Y}(x, y) dx dy$$

where the integration is over all possible values of x and y .

This result parallels that of Chapter 3, Section 3.1, 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 6.13**

U and V have the joint distribution given in Example 6.2:

$$f_{U,V}(u, v) = \frac{2u + v}{3000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

(i) Calculate $E(U)$ and $E(V)$:

(a) using $f_{U,V}(u, v)$

(b) using $f_U(u)$ and $f_V(v)$ obtained in Question 6.5

(ii) Comment on your answers.

2.2 Expectation of a sum

It follows that:

$$E[ag(X) + bh(Y)] = aE[g(X)] + bE[h(Y)]$$

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 $E[g(X) + h(Y)] = E[g(X)] + E[h(Y)]$ – 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.



Question 6.14

Verify that $E[X^2 + 2Y] = E[X^2] + E[2Y]$, for the random variables X and Y given in Question 6.6:

		X			
		0	1	2	
Y		1	0.1	0.1	0
		2	0.1	0.1	0.2
		3	0.2	0.1	0.1

2.3 Expectation of a product

For independent random variables X and Y :

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

since the joint density function factorises into the two marginal density functions.

**Question 6.15**

Verify that $E\left[\frac{N+1}{M}\right] = E\left[\frac{1}{M}\right]E[N+1]$, where M and N are from 0:

$$P(M = m, N = n) = \frac{m}{35 \times 2^{n-2}}, \text{ where } m = 1, 2, 3, 4 \text{ and } n = 1, 2, 3$$

		M				
		1	2	3	4	
		1	$\frac{2}{35}$	$\frac{4}{35}$	$\frac{6}{35}$	$\frac{8}{35}$
		2	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{4}{35}$
		3	$\frac{1}{70}$	$\frac{1}{35}$	$\frac{3}{70}$	$\frac{2}{35}$

Use the result from 0.

If we take the functions to be $g(X) = X$ and $h(Y) = Y$, these last two results give us some simple relationships between any 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 $\text{cov}[X, Y]$ of two random variables X and Y is defined by:

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

which simplifies to:

$$\text{cov}[X, Y] = E[XY] - E[X]E[Y]$$

Notice the similarity between the covariance defined here and the variance defined in Chapter 3, Section 3.2, $\text{var}(X) = E[(X - E(X))^2] = E(X^2) - E^2(X)$.



Question 6.16

Show that the simplification $\text{cov}[X, Y] = E[XY] - E[X]E[Y]$ is correct.

Note that if we rearrange this 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 £x hours. Note also that $\text{cov}[X, X] = \text{var}[X]$.



Question 6.17

Calculate the covariance of the random variables X and Y given in Question 6.6:

		X			
		0	1	2	
Y		1	0.1	0.1	0
		2	0.1	0.1	0.2
		3	0.2	0.1	0.1

Useful results on handling covariances

(a) $\text{cov}[aX + b, cY + d] = ac\text{cov}[X, Y]$

Proof:

$$E[aX + b] = aE[X] + b \text{ and } E[cY + d] = cE[Y] + d$$

$$\text{so } aX + b - E[aX + b] = a(X - E[X]) \text{ and } cY + d - E[cY + d] = c(Y - E[Y])$$

$$\therefore \text{cov}[aX + b, cY + d] = E[a(X - E[X])c(Y - E[Y])] = ac\text{cov}[X, Y]$$

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 of Chapter 3, Section 3.3: $\text{var}[aX + b] = a^2 \text{var}[X]$.

(b) $\text{cov}[X, Y + Z] = \text{cov}[X, Y] + \text{cov}[X, Z]$

Proof:

$$E[X(Y + Z)] = E[XY] + E[XZ] \text{ and } E[Y + Z] = E[Y] + E[Z]$$

$$\begin{aligned} \therefore \text{cov}[X, Y + Z] &= E[XY] + E[XZ] - E[X](E[Y] + E[Z]) \\ &= E[XY] - E[X]E[Y] + E[XZ] - E[X]E[Z] \\ &= \text{cov}[X, Y] + \text{cov}[X, Z] \end{aligned}$$

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 6.18

Write down the formula for $\text{cov}[X + Y, W + Z]$.

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 from 0 will be 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 $E[X_1 \dots X_n] = E[X_1] \dots E[X_n]$.

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 6.19

Calculate the correlation coefficient of U and V in Example 6.2, where:

$$f_{U,V}(u, v) = \frac{2u + v}{3000}, \text{ where } 10 < u < 20 \text{ and } -5 < v < 5$$

and $E(U) = \frac{140}{9}$ and $E(V) = \frac{5}{18}$ from Question 6.13.

The correlation coefficient takes a value in the range $-1 \leq \rho \leq 1$. It reflects the degree of association between the two variables.

Use this range to do a reasonableness check when you get a 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).

In simple terms, you can consider that independent means that “probabilities factorise”, and uncorrelated means that “expectations factorise”!



Question 6.20

A bivariate distribution has the following probability function:

	$P = -1$	$P = 0$	$P = 1$
$Q = -1$	0.1	0.6	0.1
$Q = 1$	0.1	0	0.1

Show that P and Q are uncorrelated but not independent.

2.5 Variance of a sum

For any random variables X and Y :

$$\text{var}[X + Y] = \text{var}[X] + \text{var}[Y] + 2\text{cov}[X, Y]$$

This can be proved from the definitions of variance and covariance.

The proof is as follows:

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


Question 6.21

Provide an alternative proof of the above result starting from $\text{var}(X + Y) = \text{cov}(X + Y, X + Y)$.

For independent random variables, this can be simplified:

$$\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$$

since $\text{cov}[X, Y] = 0$.


Question 6.22

Show from first principles that the random variables in 0 satisfy $\text{var}[M + N] = \text{var}[M] + \text{var}[N]$.

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:

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

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 $P_Z(z) = \sum_x P(x, z - x)$.

You did this when you calculated the distribution of $m + n$ in Question 6.22.

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:

$$f_Z(z) = \int_x f_X(x) f_Y(z - x) dx$$



Example 6.7

If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent random variables, obtain the probability function of $Z = X + Y$.

Solution

Using the convolution formula for discrete random variables:

$$\begin{aligned}
 P(Z = z) &= \sum_{x=0}^z P(X = x)P(Y = z - x) \\
 &= \sum_{x=0}^z \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^{z-x} e^{-\mu}}{(z-x)!} \\
 &= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^z \frac{z!}{x!(z-x)!} \lambda^x \mu^{z-x} \\
 &= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^z \binom{z}{x} \lambda^x \mu^{z-x} \\
 &= \frac{e^{-(\lambda+\mu)}}{z!} (\lambda + \mu)^z
 \end{aligned}$$

Since this matches the probability function for a $\text{Poisson}(\lambda + \mu)$ distribution (and Z can take the values $Z = 0, 1, 2, \dots$), Z has a $\text{Poisson}(\lambda + \mu)$ distribution.



Question 6.23

If $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ are independent random variables, obtain the PDF of $Z = X + Y$.

We can also use MGFs to find the PDF of a sum of random variables. This will be dealt with later in this chapter. 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, \dots, X_n are any random variables (not necessarily independent), then:

$$E(c_1X_1 + c_2X_2 + \dots + c_nX_n) = c_1E(X_1) + c_2E(X_2) + \dots + c_nE(X_n)$$

$$\text{ie } E\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i E(X_i)$$

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 $Y = c_1X_1 + c_2X_2 + \dots + c_nX_n$, where the variables are not necessarily independent, and let us now consider the variance:

$$\begin{aligned} \text{var}(Y) &= \text{cov}(Y, Y) \\ &= \text{cov}(c_1X_1 + c_2X_2 + \dots + c_nX_n, c_1X_1 + c_2X_2 + \dots + c_nX_n) \\ &= \sum_i c_i^2 \text{cov}(X_i, X_i) + 2 \sum_{i < j} \sum_j c_i c_j \text{cov}(X_i, X_j) \end{aligned}$$

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, \dots, X_n are pairwise uncorrelated (and hence certainly if they are independent) random variables, then:

$$\text{var}(c_1X_1 + c_2X_2 + \dots + c_nX_n) = c_1^2 \text{var}(X_1) + c_2^2 \text{var}(X_2) + \dots + c_n^2 \text{var}(X_n)$$

$$\text{ie } \text{var}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \text{var}(X_i)$$

**Question 6.24**

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 $W = X - 2Y + 3Z$.

**Question 6.25**

If X_1, X_2, \dots, X_n are independent random variables with mean μ and variance σ^2 , what are the mean and variance of $S = X_1 + X_2 + \dots + X_n$ and $T = nX_1$?

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 $Y = c_1X_1 + c_2X_2 + \dots + c_nX_n$.

4.1 **Probability generating functions**

Suppose X_1 and X_2 are independent counting variables with PGFs $G_{X_1}(t)$ and $G_{X_2}(t)$ respectively, and let $S = c_1X_1 + c_2X_2$.

Then:

$$G_S(t) = E\left[t^{(c_1X_1+c_2X_2)}\right] = E\left[t^{c_1X_1}\right]E\left[t^{c_2X_2}\right] = G_{X_1}(t^{c_1})G_{X_2}(t^{c_2})$$

In the case of a simple sum $Y = X_1 + X_2$, we have:

$$G_Y(t) = G_{X_1}(t)G_{X_2}(t)$$

so the PGF of the sum of two independent variables is the product of the individual PGFs.

The result extends to the sum of more than two variables.

Let $Y = X_1 + X_2 + \dots + X_n$ where the X_i are independent and X_i has PGF $G_i(t)$, then:

$$G_Y(t) = G_1(t)G_2(t) \dots G_n(t)$$

and if X_i in the sum is replaced by cX_i then $G_i(t)$ in the product is replaced by $G_i(t^c)$.

If, in addition, the X_i 's are identically distributed, each with PGF $G(t)$, and $Y = X_1 + X_2 + \dots + X_n$, then $G_Y(t) = [G(t)]^n$.

4.2 Using PGFs to derive relationships among variables

Bernoulli/binomial

We will now derive the PGF of a $\text{Bin}(n, \theta)$ distribution using an alternative (and easier) method to that used in Chapter 5, Section 1.2. This method uses the fact that a $\text{Bin}(n, \theta)$ is the sum of n independent $\text{Bernoulli}(\theta)$ trials.

Let X_i , $i = 1, 2, \dots, n$, be independent Bernoulli (θ) variables.

Then each has PGF $G(t) = \theta t + (1 - \theta)$.

So $Y = X_1 + X_2 + \dots + X_n$ has PGF $[\theta t + (1 - \theta)]^n$ which is the PGF of a binomial (n, θ) variable.

So the binomial (n, θ) random variable is the sum of n independent Bernoulli (θ) random variables.

Physically, the number of successes in n trials is the sum of the numbers of successes (0 or 1) at each trial.

This method can also be used to obtain the MGF of a binomial distribution.

Further, the sum of two independent binomial variables, one (n, θ) and the other (m, θ) , is a binomial $(n + m, \theta)$ variable.



Question 6.26

Show that if $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, p)$ are independent random variables, then $X + Y$ also has a binomial distribution.

Geometric/negative binomial

We will now derive the PGF of a negative binomial with parameters (k, θ) distribution using an alternative (and easier) method to that used in Chapter 5, Section 1.2. This method uses the fact that a negative binomial (k, θ) is the sum of k independent geometric random variables with parameter θ .

Let X_i , $i = 1, 2, \dots, k$, be independent geometric (θ) variables.

Then each has PGF $G(t) = \frac{\theta t}{1 - (1 - \theta)t}$.

So $Y = X_1 + X_2 + \dots + X_k$ has PGF $\left(\frac{\theta t}{1 - (1 - \theta)t}\right)^k$, which is the PGF of a negative binomial (k, θ) variable.

So the negative binomial (k, θ) random variable is the sum of k independent geometric (θ) random variables.

Each geometric variable has mean $\frac{1}{\theta}$ and variance $\frac{1 - \theta}{\theta^2}$; hence the negative binomial has mean $\frac{k}{\theta}$ and variance $\frac{k(1 - \theta)}{\theta^2}$.

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.

This method can also be used to obtain the MGF of a negative binomial distribution.

Further, the sum of two independent negative binomial variables, one (k, θ) and the other (m, θ) , is a negative binomial $(k + m, \theta)$ variable.

Make sure that you can prove this result using either PGFs or MGFs.

Poisson

We will now find the distribution of the sum of two independent Poisson random variables using PGFs. This is an alternative method to convolutions.

Let X and Z be independent Poisson (λ) and Poisson (γ) variables.

Then X has PGF $G_X(t) = \exp\{\lambda(t-1)\}$, Z has PGF $G_Z(t) = \exp\{\gamma(t-1)\}$.

So the sum $X + Z$ has PGF $[\exp\{\lambda(t-1)\}][\exp\{\gamma(t-1)\}] = \exp\{(\lambda + \gamma)(t-1)\}$, which is the PGF of a Poisson ($\lambda + \gamma$) 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 = $\lambda + \gamma$.

ie “ $Poi(\lambda) + Poi(\gamma) \sim Poi(\lambda + \gamma)$ ”.

This is an important result to remember and is often quotable in the Subject CT3 exam.



Question 6.27

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. If you can assume that the calls are independent random variables having Poisson distributions, what is the probability that in half an hour the switchboard will receive:

- (i) at least 5 calls
- (ii) exactly 7 calls?

4.3 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:

$$\begin{aligned} M_S(t) &= E[e^{(c_1 X_1 + c_2 X_2)t}] \\ &= E[e^{c_1 X_1 t}] E[e^{c_2 X_2 t}] \\ &= M_{X_1}(c_1 t) M_{X_2}(c_2 t) \end{aligned}$$

In the case of a simple sum $Y = X_1 + X_2$, we have:

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t)$$

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 $Y = X_1 + X_2 + \dots + X_n$ where the X_i are independent and X_i has MGF $M_i(t)$, then:

$$M_Y(t) = M_1(t) M_2(t) \dots M_n(t)$$

and if X_i in the sum is replaced by cX_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 $Y = X_1 + X_2 + \dots + X_n$, then $M_Y(t) = [M(t)]^n$.

Both of the last two results are important to remember and are often quotable in the Subject CT3 exam.

4.4 Using MGFs to derive relationships among variables

Exponential/gamma

We will now derive the MGF of a $Ga(\alpha, \lambda)$ distribution using an alternative method to that used in Chapter 5, Section 2.2. This method uses the fact that a $Ga(\alpha, \lambda)$ is the sum of α independent $Exp(\lambda)$ random variables.

Let $X_i, i=1, 2, \dots, k$, be independent exponential (λ) variables.

Then each has MGF $M(t) = \lambda(\lambda - t)^{-1}$.

So $Y = X_1 + X_2 + \dots + X_k$ has MGF $[\lambda(\lambda - t)^{-1}]^k$, 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 exponential (λ) random variables.

Each exponential variable has mean $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$; hence the gamma (k, λ) has mean $\frac{k}{\lambda}$ and variance $\frac{k}{\lambda^2}$.

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.

Make sure that you can prove this result using MGFs.



Question 6.28

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, what is 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?

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 6.29

Prove, using MGFs, that if $X = X_1 + X_2$ where X_1 and X_2 are independent random variables such that $X_1 \sim \chi_m^2$ and $X_2 \sim \chi_n^2$ then $X \sim \chi_{m+n}^2$.

This result is essential for analysis of variance, which we will study in Chapter 14.

Normal

Let X_1 be a normal random variable with mean μ_{X_1} and standard deviation σ_{X_1} , and let X_2 be a normal random variable with mean μ_{X_2} and standard deviation σ_{X_2} . Let X_1 and X_2 be independent, and let $Y = X_1 + X_2$.

$$X_1 \text{ has MGF } M_{X_1}(t) = \exp\left(\mu_{X_1}t + \frac{1}{2}\sigma_{X_1}^2 t^2\right)$$

$$X_2 \text{ has MGF } M_{X_2}(t) = \exp\left(\mu_{X_2}t + \frac{1}{2}\sigma_{X_2}^2 t^2\right)$$

So the sum $Y = X_1 + X_2$ has MGF:

$$\exp\left(\mu_{X_1}t + \frac{1}{2}\sigma_{X_1}^2 t^2\right) \exp\left(\mu_{X_2}t + \frac{1}{2}\sigma_{X_2}^2 t^2\right) = \exp\left\{(\mu_{X_1} + \mu_{X_2})t + \frac{1}{2}(\sigma_{X_1}^2 + \sigma_{X_2}^2)t^2\right\}$$

which is the MGF of a normal variable with mean $\mu_{X_1} + \mu_{X_2}$ and variance $\sigma_{X_1}^2 + \sigma_{X_2}^2$.

So the sum of independent normal variables is a normal variable.

ie “ $N(\mu_X, \sigma_X^2) + N(\mu_Y, \sigma_Y^2) \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ ”

Similarly, it can be shown that:

$$N(\mu_X, \sigma_X^2) - N(\mu_Y, \sigma_Y^2) \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

These are important results to remember and are often quotable in the Subject CT3 exam.



Question 6.30

If X and Y are independent standard normal variables, what is the distribution of $2X - Y$?

5 Exam-type question



Exam-type question

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.

This page has been left blank so that you can keep the chapter summaries together for revision purposes.



Chapter 6 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:

$$F(x, y) = P(X < x, Y < y)$$

For continuous random variables $f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$.

Using the formulae overleaf, we can calculate the:

- Marginal distributions, eg $P(X = x)$ or $f_X(x)$
- Conditional distributions, eg $P(X = x | Y = y)$ or $f_{X|Y=y}(x | y)$
- Expectation of any function, $E[g(X, Y)]$
- Covariance, $\text{cov}(X, Y)$
- Correlation coefficient, $\rho(X, Y) = \text{corr}(X, Y)$

The random variables X and Y are uncorrelated if and only if:

$$\text{corr}(X, Y) = 0 \Leftrightarrow \text{cov}(X, Y) = 0 \Leftrightarrow E(XY) = E(X)E(Y)$$

The random variables X and Y are independent if, and only if:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Independent random variables are always uncorrelated. Uncorrelated random variables are not necessarily independent.

There are rules connecting sums and products of expectations and sums of variances.

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 on the formulae sheet. The convolution is written $f_Z = f_X * f_Y$.

Sums of independent random variables make other random variables. The full list is given on the formulae sheet.



Chapter 6 Formulae

Marginal probability (density) function

$$P(X = x) = \sum_y P(X = x, Y = y) \quad f_X(x) = \int_y f_{X,Y}(x, y) dy$$

Conditional probability (density) function

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \quad f_{X|Y=y}(x, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Expectation

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) P(X = x, Y = y) \quad \text{or} \quad \int_y \int_x g(x, y) f_{X,Y}(x, y) dx dy$$

Covariance

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

Correlation

$$\text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}$$

Sums and products of moments

$$\begin{aligned} E(X + Y) &= E(X) + E(Y) \\ E(XY) &= E(X)E(Y) + \text{Cov}(X, Y) \\ &= E(X)E(Y) \quad \text{if } X, Y \text{ independent} \end{aligned}$$

The above are also true for functions $g(X)$ and $h(Y)$ of the random variables.

$$\begin{aligned} \text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) \\ &= \text{var}(X) + \text{var}(Y) \quad \text{if } X, Y \text{ independent} \end{aligned}$$

Convolutions

$$f_Z = f_X * f_Y = \sum_x P(X=x)P(Y=z-x) \quad \text{or} \quad \int_x f_X(x)f_Y(z-x) dx$$

Linear combinations

For independent random variables X_1, \dots, X_n :

$$\begin{aligned} E(c_1X_1 + \dots + c_nX_n) &= c_1E(X_1) + \dots + c_nE(X_n) \\ \text{var}(c_1X_1 + \dots + c_nX_n) &= c_1^2 \text{ var}(X_1) + \dots + c_2^2 \text{ var}(X_n) \end{aligned}$$

PGFs and MGFs of sums of IID random variables

For independent and identically distributed random variables X_1, \dots, X_n :

$$\begin{aligned} Y = X_1 + \dots + X_n \Rightarrow G_Y(s) &= G_{X_1}(s) \dots G_{X_n}(s) \\ &= [G_X(s)]^n \quad X_i\text{'s identical} \\ \Rightarrow M_Y(t) &= M_{X_1}(t) \dots M_{X_n}(t) \\ &= [M_X(t)]^n \quad X_i\text{'s identical} \end{aligned}$$

Linear combinations of random variables

For independent distributions:

- “ $Bernoulli(p) + \dots + Bernoulli(p) \sim Bin(n, p)$ ”
- “ $Bin(n, \theta) + Bin(m, \theta) \sim Bin(n+m, \theta)$ ”
- “ $Geo(p) + \dots + Geo(p) \sim NBin(k, p)$ ”
- “ $NBin(k, \theta) + NBin(m, \theta) \sim NBin(k+m, \theta)$ ”
- “ $Exp(\lambda) + \dots + Exp(\lambda) \sim Ga(\alpha, \lambda)$ ”
- “ $Ga(\alpha, \lambda) + Ga(\delta, \lambda) \sim Ga(\alpha+\delta, \lambda)$ ”
- “ $\chi_m^2 + \chi_n^2 \sim \chi_{m+n}^2$ ”
- “ $Poi(\lambda) + Poi(\mu) \sim Poi(\lambda+\mu)$ ”
- “ $N(\mu_1, \sigma_1^2) \pm N(\mu_2, \sigma_2^2) \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$ ”

Please note that 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.

Chapter 6 Solutions

Solution 6.1

(i) Since the events are mutually exclusive:

$$P(M = 3, N = 1 \text{ or } 2) = \frac{6}{35} + \frac{3}{35} = \frac{9}{35}$$

(ii) We require $P(N = 3)$, since this does not depend on the value of M it the same as finding $P(N = 3, M = 1, 2, 3 \text{ or } 4)$, 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, $P(A | B) = \frac{P(A \cap B)}{P(B)}$, gives:

$$P(M = 2 | N = 3) = \frac{P(M = 2, N = 3)}{P(N = 3)} = \frac{\cancel{1/35}}{\cancel{1/7}} = \frac{1}{5}$$

Solution 6.2

(i) Using the result $\iint_{x,y} f(x,y) dy dx = 1$ gives:

$$\begin{aligned} \int_{x=0}^2 \int_{y=0}^2 c(x+3y) dx dy &= \int_{y=0}^2 c \left[\frac{1}{2}x^2 + 3xy \right]_{x=0}^2 dy \\ &= \int_{y=0}^2 c(2+6y) dy \\ &= c \left[2y + 3y^2 \right]_{y=0}^2 \\ &= 16c = 1 \\ \Rightarrow c &= \frac{1}{16} \end{aligned}$$

(ii) The probability is:

$$\begin{aligned}
 P(X < 1, Y > 0.5) &= \int_{x=0}^1 \int_{y=0.5}^2 \frac{1}{16}(x + 3y) dx dy \\
 &= \int_{y=0.5}^2 \frac{1}{16} \left[\frac{1}{2}x^2 + 3xy \right]_{x=0}^1 dy \\
 &= \int_{y=0.5}^2 \frac{1}{16} \left(\frac{1}{2} + 3y \right) dy \\
 &= \frac{1}{16} \left[\frac{1}{2}y + \frac{3}{2}y^2 \right]_{y=0.5}^2 \\
 &= \frac{51}{128} \simeq 0.398
 \end{aligned}$$

Solution 6.3

(i) $P(X = 1) = 0.1 + 0.15 + 0.2 + 0.15 = 0.6$

$$P(X = 2) = 0.1 + 0.1 + 0.05 + 0.05 = 0.3$$

$$P(X = 3) = 0.05 + 0.05 = 0.1$$

(ii) The probability distribution of X is:

x	1	2	3	
$P(X = x)$	0.6	0.3	0.1	

(iii) The probability distribution of Y is:

y	1	2	3	4	
$P(Y = y)$	0.25	0.3	0.25	0.2	

We can check that the sum of probabilities is 1 for each of the distributions.

Solution 6.4

Summing over all possible values of M gives:

$$\begin{aligned} f_N(n) &= \sum_{m=1}^4 P(M = m, N = n) \\ &= \sum_{m=1}^4 \frac{m}{35 \times 2^{n-2}} = \frac{1}{35 \times 2^{n-2}} (1 + 2 + 3 + 4) = \frac{1}{7 \times 2^{n-3}} \end{aligned}$$

Solution 6.5

To find the PDF of the marginal distribution of U , we integrate out V :

$$f_U(u) = \int_{v=-5}^5 \frac{2u+v}{3,000} dv = \frac{u}{150}$$

Therefore the marginal distribution of U is $f_U(u) = \frac{u}{150}$, $10 < u < 20$.

Similarly for V , we integrate out U :

$$f_V(v) = \int_{u=10}^{20} \frac{2u+v}{3,000} du = \frac{30+v}{300}$$

Therefore the marginal distribution of V is $f_V(v) = \frac{30+v}{300}$, $-5 < v < 5$.

If you want to check marginal distributions, you can integrate them over the appropriate range. The answer should be 1!

Solution 6.6

- (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 $P(X = 0 | Y = 2) = 0.1 \times 2.5 = 0.25$.

Solution 6.7

Dividing the joint probability function by the marginal probability function:

$$P_{N|M=m}(m, n) = \frac{P(N = n, M = m)}{P(M = m)} = \left(\frac{m}{35 \times 2^{n-2}} \right) \div \frac{m}{10} = \frac{1}{7 \times 2^{n-3}}$$

Solution 6.8

$$f_Y(y) = \int_{x=0}^2 \frac{1}{16} (x + 3y) dx = \frac{1}{16} \left[\frac{1}{2}x^2 + 3xy \right]_{x=0}^2 = \frac{1}{16} (2 + 6y)$$

$$\Rightarrow f_{X|Y=y}(x, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{\frac{1}{16}(x+3y)}{\frac{1}{16}(2+6y)} = \frac{x+3y}{2(1+3y)} \quad 0 < x < 2$$

Solution 6.9

No. For example, the conditional distribution of X given $Y = 2$ is not the same as the marginal distribution of X . (We worked these out in Question 6.6.)

Solution 6.10

Yes. This can be seen either because the joint probability function is the product of the marginal distributions, or because the conditional distributions are the same as the marginal distributions.

Solution 6.11

No. If they were it would be possible to factorise $\frac{2u+v}{3,000}$ into two functions of the form $g(u)h(v)$. As this is not possible, the random variables are not independent.

Solution 6.12

The expectation is given by $E(X^2 + 2XY) = \sum_{x=0}^2 \sum_{y=1}^3 (x^2 + 2xy)P(X = x, Y = y)$.

Adding up all the relevant terms, we get:

$$\begin{aligned} E(X^2 + 2XY) &= (0^2 + 2 \times 0 \times 1) \times 0.1 + (1^2 + 2 \times 1 \times 1) \times 0.1 \\ &\quad + \dots + (2^2 + 2 \times 2 \times 3) \times 0.1 = 5.5 \end{aligned}$$

Solution 6.13

(i)(a) Integrating with respect to v first and then with respect to u , we obtain:

$$\begin{aligned} E(U) &= \int_{u=10}^{20} \int_{v=-5}^5 u \frac{(2u+v)}{3,000} du dv = \int_{u=10}^{20} \int_{v=-5}^5 \frac{2u^2 + uv}{3,000} du dv = \int_{u=10}^{20} \left[\frac{2u^2 v + \frac{1}{2} u v^2}{3,000} \right]_{-5}^5 du \\ &= \int_{u=10}^{20} \frac{u^2}{150} du = \left[\frac{u^3}{450} \right]_{10}^{20} = \frac{140}{9} \end{aligned}$$

Similarly:

$$\begin{aligned} E(V) &= \int_{u=10}^{20} \int_{v=-5}^5 v \frac{(2u+v)}{3,000} du dv = \int_{u=10}^{20} \int_{v=-5}^5 \frac{2uv + v^2}{3,000} du dv = \int_{u=10}^{20} \left[\frac{uv^2 + \frac{1}{3} v^3}{3,000} \right]_{-5}^5 du \\ &= \int_{u=10}^{20} \frac{1}{36} du = \left[\frac{1}{36} u \right]_{10}^{20} = \frac{5}{18} \end{aligned}$$

(i)(b) From Question 6.5, we have $f_U(u) = \frac{u}{150}$ and $f_V(v) = \frac{30+v}{300}$. Hence:

$$\begin{aligned} E(U) &= \int_{u=10}^{20} u \frac{u}{150} du = \int_{u=10}^{20} \frac{u^2}{150} du = \left[\frac{u^3}{450} \right]_{10}^{20} = \frac{140}{9} \\ E(V) &= \int_{v=-5}^5 v \frac{30+v}{300} dv = \int_{v=-5}^5 \frac{30v + v^2}{300} dv = \left[\frac{15v^2 + \frac{1}{3} v^3}{300} \right]_{-5}^5 = \frac{5}{18} \end{aligned}$$

(ii) Both methods are equivalent.

Solution 6.14

Reading the values from the table, we have:

$$E[X^2 + 2Y] = (0^2 + 2 \times 1) \times 0.1 + (1^2 + 2 \times 1) \times 0.1 + \dots + (2^2 + 2 \times 3) \times 0.1 = 5.9$$

Looking at the terms on the right hand side:

$$E[X^2] = 0 \times 0.4 + 1 \times 0.3 + 4 \times 0.3 = 1.5$$

$$E[2Y] = 2 \times 0.2 + 4 \times 0.4 + 6 \times 0.4 = 4.4$$

Thus $E[X^2] + E[2Y] = 5.9$, and the result has been verified.

Solution 6.15

In 0, we calculated that $E\left(\frac{N+1}{M}\right) = \frac{36}{35}$.

We can calculate $E\left(\frac{1}{M}\right)$ and $E(N+1)$ using the marginal distributions:

$$E\left(\frac{1}{M}\right) = \sum_{m=1}^4 \frac{1}{m} f_M(m) = \sum_{m=1}^4 \frac{1}{m} \times \frac{m}{10} = \sum_{m=1}^4 \frac{1}{10} = \frac{2}{5}$$

$$E(N+1) = \sum_{n=1}^3 (n+1) f_N(n) = 2 \times \frac{4}{7} + 3 \times \frac{2}{7} + 4 \times \frac{1}{7} = \frac{18}{7}$$

This gives $E\left(\frac{1}{M}\right)E(N+1) = \frac{36}{35}$, which verifies the result.

Solution 6.16

If we expand the definition of the covariance we obtain:

$$\begin{aligned}\text{cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E(XY - XE[Y] - YE[X] + E[X]E[Y]) \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

Solution 6.17

We will use the formula $\text{cov}(X, Y) = E[XY] - E[X]E[Y]$.

From the table of values:

$$E[XY] = 0 \times 1 \times 0.1 + \dots + 2 \times 3 \times 0.1 = 2$$

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 $\text{cov}(X, Y) = 2 - 0.9 \times 2.2 = 0.02$.

Solution 6.18

$$\text{cov}(X+Y, W+Z) = \text{cov}(X, W) + \text{cov}(X, Z) + \text{cov}(Y, W) + \text{cov}(Y, Z)$$

Solution 6.19

First we need $E[UV]$:

$$E[UV] = \int_{u=10}^{20} \int_{v=-5}^5 uv \frac{2u+v}{3,000} dudv$$

Integrating first with respect to u :

$$E[UV] = \int_{v=-5}^5 \left[\frac{\frac{2}{3}u^3v + \frac{1}{2}u^2v^2}{3,000} \right]_{u=10}^{u=20} dv = \int_{v=-5}^5 \left(\frac{14v}{9} + \frac{v^2}{20} \right) dv = \left[\frac{14v^2}{18} + \frac{v^3}{60} \right]_{-5}^5 = \frac{25}{6}$$

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 :

$$\text{var}(U) = \int_{10}^{20} u^2 \times \frac{u}{150} du - \left(\frac{140}{9} \right)^2 = \left[\frac{u^4}{600} \right]_{10}^{20} - \left(\frac{140}{9} \right)^2 = \frac{650}{81}$$

Similarly:

$$\text{var}(V) = \int_{-5}^5 v^2 \times \frac{30+v}{300} dv - \left(\frac{5}{18} \right)^2 = \left[\frac{v^3}{30} + \frac{v^4}{1,200} \right]_{-5}^5 - \left(\frac{5}{18} \right)^2 = \frac{2675}{324}$$

So the correlation coefficient is $\text{corr}(U, V) = \frac{-\frac{25}{162}}{\sqrt{\frac{650}{81} \times \frac{2675}{324}}} = -\frac{1}{\sqrt{2782}} = -0.019$.

Solution 6.20

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.

Solution 6.21

$$\begin{aligned}\text{var}(X + Y) &= \text{cov}(X + Y, X + Y) \\ &= \text{cov}(X, X) + \text{cov}(X, Y) + \text{cov}(Y, X) + \text{cov}(Y, Y) \\ &= \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y)\end{aligned}$$

Solution 6.22

By adding up the probabilities from the table, the random variable $M + N$ has the distribution:

$m + n$	2	3	4	5	6	7
$P(M + N = m + n)$	$\frac{2}{35}$	$\frac{5}{35}$	$\frac{17}{70}$	$\frac{12}{35}$	$\frac{11}{70}$	$\frac{2}{35}$

So the expectation of $M + N$ is:

$$E[M + N] = 2 \times \frac{2}{35} + \dots + 7 \times \frac{2}{35} = \frac{32}{7}$$

So the variance of $M + N$ is:

$$\text{var}(M + N) = 2^2 \times \frac{2}{35} + \dots + 7^2 \times \frac{2}{35} - \left(\frac{32}{7}\right)^2 = \frac{75}{49}$$

Looking at the marginal distributions:

$$\text{var}(M) = 1^2 \times \frac{1}{10} + \dots + 4^2 \times \frac{4}{10} - 3^2 = 1$$

$$\text{var}(N) = 1^2 \times \frac{4}{7} + \dots + 3^2 \times \frac{1}{7} - \left(\frac{11}{7}\right)^2 = \frac{26}{49}$$

So M and N satisfy the given relationship.

Solution 6.23

Using the formula for continuous random variables (and assuming that $\lambda \neq \mu$):

$$\begin{aligned} f_Z(z) &= \int_0^z \lambda e^{-\lambda x} \mu e^{-\mu(z-x)} dx = \frac{\lambda \mu e^{-\mu z}}{\lambda - \mu} \int_0^z (\lambda - \mu) e^{-(\lambda - \mu)x} dx \\ &= \frac{\lambda \mu}{\lambda - \mu} e^{-\mu z} \left[1 - e^{-(\lambda - \mu)z} \right] = \frac{\lambda \mu}{\lambda - \mu} (e^{-\mu z} - e^{-\lambda z}) \end{aligned}$$

If $\lambda = \mu$, we get $\lambda^2 z e^{-\lambda z}$ ie a $\text{Gamma}(2, \lambda)$ distribution.

Solution 6.24

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 cannot simply use $\text{var}(W) = \text{var}(X - 2Y + 3Z) = \text{var}(X) + 4\text{var}(Y) + 9\text{var}(Z)$. So using covariances:

$$\begin{aligned}\text{var}(W) &= \text{var}(X - 2Y + 3Z) = \text{cov}(X - 2Y + 3Z, X - 2Y + 3Z) \\ &= \text{var}(X) + 4\text{var}(Y) + 9\text{var}(Z) - 4\text{cov}(X, Y) + 6\text{cov}(X, Z) - 12\text{cov}(Y, Z) \\ &= 1 + (4 \times 4) + (9 \times 3) - (4 \times -3) + (6 \times -2) - (12 \times 1) \\ &= 32\end{aligned}$$

Solution 6.25

The mean and variance of S (which is a sum of random variables) are:

$$E[S] = E[X_1 + X_2 + \dots + X_n] = E[X_1] + E[X_2] + \dots + E[X_n] = n\mu$$

$$\text{var}(S) = \text{var}(X_1 + X_2 + \dots + X_n) = \text{var}[X_1] + \text{var}[X_2] + \dots + \text{var}[X_n] = n\sigma^2$$

The mean and variance of T (which is a single random variable multiplied by a constant) are:

$$E[T] = E[nX_1] = nE[X_1] = n\mu$$

$$\text{var}(T) = \text{var}(nX_1) = n^2 \text{var}(X_1) = n^2\sigma^2$$

Note that the means are the same but that, for S , the variance is smaller because differences in the individual X values tend to “average out”.

Solution 6.26

The probability generating functions for X and Y are:

$$G_X(t) = (q + pt)^m \quad \text{and} \quad G_Y(t) = (q + pt)^n$$

Since X and Y are independent, we can write:

$$G_{X+Y}(t) = (q + pt)^m (q + pt)^n = (q + pt)^{m+n}$$

which we recognise as the PGF of a $\text{Bin}(m+n, p)$ distribution. Hence, by uniqueness of PGFs, $X + Y$ has a binomial distribution with parameters $m+n$ and p .

This result should be obvious. If you toss a coin 10 times in the morning and count the number of heads, the number would be distributed as $\text{Bin}(10, \frac{1}{2})$. If you toss the coin a further 20 times in the afternoon, the number of heads will be distributed as $\text{Bin}(20, \frac{1}{2})$. Adding the totals together is obviously the same as the $\text{Bin}(30, \frac{1}{2})$ distribution that you would expect for the whole day.

Solution 6.27

Summing the Poisson variables, the total number of telephone calls coming in is Poisson with mean $3.5 + 3.9 + 2.1 = 9.5$.

$$(i) \quad P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.04026 = 0.95974$$

$$(ii) \quad P(X = 7) = P(X \leq 7) - P(X \leq 6) = 0.26866 - 0.16495 = 0.10371$$

These figures are taken from the cumulative Poisson table on page 178 of the *Tables*.

Solution 6.28

- (i) The sum of two independent exponential distributions 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:

$$\begin{aligned} P(X > 8) &= \frac{1}{25} \left[\left[-5xe^{-\frac{1}{5}x} \right]_8^\infty + 5 \int_8^\infty e^{-\frac{1}{5}x} dx \right] \\ &= \frac{1}{25} \left[40e^{-\frac{8}{5}} - 25 \left[e^{-\frac{1}{5}x} \right]_8^\infty \right] \\ &= \frac{1}{25} \left[40e^{-\frac{8}{5}} + 25e^{-\frac{8}{5}} \right] \\ &= 0.525 \end{aligned}$$

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 $Y \sim \text{Poi}(0.2t)$. 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 $Y \sim \text{Poi}(0.2 \times 8)$, 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 distributions 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 easiest way is to use the gamma-chi relationship detailed in Section 2.2 of Chapter 4:

$$\begin{aligned} P(X > 15) &= P(2\lambda X > 30\lambda) \\ &= P(\chi_{2\alpha}^2 > 30\lambda) \end{aligned}$$

Substituting $\alpha = 3$ and $\lambda = \frac{1}{5}$, and using the χ^2 tables on page 165, we obtain:

$$P(X > 15) = P(\chi_6^2 > 6) = 1 - 0.5768 = 0.4232$$

Alternatively, we could use the Poisson process with $\lambda t = 0.2 \times 15$ and calculate the probability of 0, 1 or 2 tyres checked within 15 minutes.

Solution 6.29

Now $\chi_n^2 \equiv Ga\left(\frac{n}{2}, \frac{1}{2}\right)$ therefore $M_{X_1}(t) = (1-2t)^{-\frac{m}{2}}$ and $M_{X_2}(t) = (1-2t)^{-\frac{n}{2}}$.

Since X_1 and X_2 are independent:

$$M_X(t) = E(e^{tX}) = E(e^{tX_1+tX_2}) = E(e^{tX_1}e^{tX_2}) = E(e^{tX_1})E(e^{tX_2}) = M_{X_1}(t)M_{X_2}(t)$$

$$\text{So } M_X(t) = (1-2t)^{-\frac{m}{2}} \times (1-2t)^{-\frac{n}{2}} = (1-2t)^{-\frac{m+n}{2}}.$$

Since this is the MGF of χ_{m+n}^2 , by the uniqueness property, $X \sim \chi_{m+n}^2$.

Solution 6.30

The resulting distribution will be normal. So we just need to fill in the mean and variance to obtain:

$$2X - Y \sim N(2 \times 0 - 0, 2^2 \times 1 + (-1)^2 \times 1) = N(0, 5)$$

Exam-type question

Let X be the size of a home insurance claim and Y the size of a car insurance claim. Then:

$$X \sim N(800, 100^2) \quad Y \sim N(1200, 300^2)$$

We require:

$$\begin{aligned} & P((Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4) > 800) \\ &= P((Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4) - 800 > 0) \end{aligned}$$

So we need the distribution of $(Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4)$ which is:

$$\begin{aligned} (Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4) &\sim N(3 \times 1200 - 4 \times 800, 3 \times 300^2 + 4 \times 100^2) \\ &\sim N(400, 310\,000) \end{aligned}$$

Therefore:

$$\begin{aligned} P((Y_1 + Y_2 + Y_3) - (X_1 + X_2 + X_3 + X_4) > 800) &= P\left(Z > \frac{800 - 400}{\sqrt{310,000}}\right) \\ &= P(Z > 0.718) \\ &= 1 - P(Z < 0.718) = 0.23638 \end{aligned}$$

Chapter 7

Conditional expectation



Syllabus objectives

- (xiv) 1. Define the conditional expectation of one random variable given the value of another random variable, and calculate such a quantity.
2. Show how the mean and variance of a random variable can be obtained from expected values of conditional expected values, and apply this.
3. Derive the moment generating function of the sum of a random number of independent, identically distributed random variables (a compound distribution), and use the result to calculate the mean and variance of such a distribution.

0 **Introduction**

In this 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 Chapter 13 to define the regression line $E[Y | x] = \alpha + \beta x$. They will also feature in Subjects CT4, CT6 and CT8.

We also introduce the idea of a compound distribution (a random number of random variables) which will prove to be a key feature of Subject CT6 in finding the total monetary amount of claims received from a portfolio of insurance policies. Since this chapter is just the springboard into Subject CT6, we limit ourselves to finding the mean, variance and moment generating function of a compound distribution.

1 The conditional expectation $E[Y|X = x]$

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 $E[Y|X = x]$, or just $E[Y|x]$.

For a discrete distribution, this will be:

$$E[Y | X = x] = \sum_i y_i P[Y = y_i | X = x] = \sum_i y_i \frac{P[Y = y_i, X = x]}{P[X = x]}$$



Question 7.1

What is the equivalent expression for a continuous distribution?



Example 7.1

Two random variables X and Y have the following discrete joint distribution:

		Y	
		10	20
X	1	0.2	0.2
	2	0.2	0.3
			0.1

Calculate $E(Y|X = 1)$.

Solution

$$\begin{aligned} E(Y | X = 1) &= \sum y P(Y = y | X = 1) \\ &= 10P(Y = 10 | X = 1) + 20P(Y = 20 | X = 1) + 30P(Y = 30 | X = 1) \\ &= 10 \times \frac{0.2}{0.5} + 20 \times \frac{0.2}{0.5} + 30 \times \frac{0.1}{0.5} \\ &= 10 \times 0.4 + 20 \times 0.4 + 30 \times 0.2 \\ &= 18 \end{aligned}$$

**Question 7.2**

Calculate $E(X | Y = 10)$ for the joint distribution given in Example 7.1:

		Y		
		10	20	30
X	1	0.2	0.2	0.1
	2	0.2	0.3	0

**Question 7.3**

Let X and Y have joint density function given by:

$$f(x,y) = \frac{3}{5}x(x+y) \quad 0 < x < 1, \quad 0 < y < 2$$

Determine the conditional expectation $E[Y | X = x]$.

2 ***The random variable $E[Y|X]$***

The conditional expectation $E[Y | X = x] = g(x)$, 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 $E[Y | X]$.

We saw in Question 7.3 that $E[Y | X = x] = \frac{3x + 4}{3(x + 1)}$ which is a function of x .

Note: $E[Y|X]$ is also referred to as the regression of Y on X .

In Chapter 13 the regression line will be defined as $E[Y | x] = \alpha + \beta x$.

$E[Y | X]$, 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 $E[Y | X]$. The usefulness of considering this expected value, $E[E[Y | X]]$, comes from the following result, proved here in the case of continuous variables, but true in general.

Theorem: $E[E[Y | X]] = E[Y]$

Proof:

$$\begin{aligned} E[E[Y | X]] &= \int E[Y | x] f_X(x) dx \\ &= \int \left(\int y f(y | x) dy \right) f_X(x) dx \\ &= \int \int y f(x, y) dx dy = E[Y] \end{aligned}$$

Here $f(y | x)$ represents the density function of the conditional distribution of $Y | X = x$. This was written as $f_{Y|X}(x, y)$ in Chapter 6.

The last two steps follow by noting that $f(y | x) = \frac{f(x, y)}{f_X(x)}$ and $\int f(x, y) dx = f_Y(y)$ ie the marginal distribution of y .

**Question 7.4**

- (i) Calculate $E[Y]$ from first principles for Question 7.3, where:

$$f(x,y) = \frac{3}{5}x(x+y) \quad 0 < x < 1, \quad 0 < y < 2$$

- (ii) Given that $E[Y | X = x] = \frac{3x+4}{3(x+1)}$ for this distribution, find $E[E(Y | X)]$.

- (iii) Hence, confirm that $E[Y] = E[E(Y | X)]$ for this distribution.

**Question 7.5**

The random variable K has an $\text{Exp}(\lambda)$ distribution. For a given value of K , the random variable X has a $\text{Poisson}(K)$ distribution.

- (i) Obtain an expression for $E[X|K]$.
- (ii) Hence, calculate $E[X]$.

3 The random variable $\text{var}[Y | X]$ and the “ $E[V] + \text{var}[E]$ ” result

The variance of the conditional distribution of Y given $X = x$ is denoted $\text{var}[Y | x]$, where:

$$\text{var}[Y | x] = E\left[\{Y - E[Y | x]\}^2 | x\right] = E[Y^2 | x] - (E[Y | x])^2$$

$\text{var}[Y | x]$ is the observed value of a random variable $\text{var}[Y | X]$ where:

$$\text{var}[Y | X] = E[Y^2 | X] - (E[Y | X])^2 = E[Y^2 | X] - \{g(X)\}^2$$

Hence $E[\text{var}[Y | X]] = E[E[Y^2 | X]] - E[\{g(X)\}^2] = E[Y^2] - E[\{g(X)\}^2]$ and so:

$$E[Y^2] = E[\text{var}[Y | X]] + E[\{g(X)\}^2]$$

So the variance of Y , $\text{var}[Y] = E(Y^2) - [E(Y)]^2$, is given by:

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

i.e $\text{var}[Y] = E[\text{var}[Y | X]] + \text{var}[E[Y | X]]$.



Question 7.6

Evaluate $\text{var}[Y | X = 1]$ for the joint distribution given in Example 7.1:

		Y		
		10	20	30
X	1	0.2	0.2	0.1
	2	0.2	0.3	0



Question 7.7

Obtain an expression for $\text{var}[X | K]$, where X and K are as in Question 7.5.

4 Moment generating functions

4.1 Recap

The definition and properties of moment generating functions should be familiar and in particular the property that no two different distributions have the same moment generating function. Let X be a random variable whose moment generating function, denoted $M(t)$, exists. A property of moment generating functions, which should be familiar, is that for any positive integer n :

$$\left. \frac{d^n}{dt^n} M(t) \right|_{t=0} = E[X^n] \quad (4.1)$$

A less familiar property, but one of which considerable use shall be made, is that for $n = 2$ and 3 :

$$\left. \frac{d^n}{dt^n} \log M(t) \right|_{t=0} = E[(X - E[X])^n] \quad (4.2)$$

$\log M(t)$ is known as the cumulant generating function of X . The advantage of using (4.2) rather than (4.1) is that it is the moments of a random variable about its mean (ie the central moments), rather than about zero, that are of interest, and these, for $n = 2$ and 3 are given directly by (4.2).

So this is just saying that $C_X''(0) = \text{var}(X)$ and $C_X'''(0) = \text{skew}(X)$, where $C_X(t)$ is the cumulant generating function of X . We met this in Chapter 5.

5 Compound distributions

For convolutions of distribution functions suppose that $\{X_i\}_{i=1}^n$ are independent and identically distributed (iid) random variables with common distribution function $F(x)$. Then the distribution function of $\sum_{i=1}^n X_i$ is denoted $F^{n^*}(x)$, so that:

$$F^{n^*}(x) = P(X_1 + X_2 + \dots + X_n \leq x)$$

Let $S = X_1 + X_2 + \dots + X_N$ (and $S = 0$ if $N = 0$) where the X_i 's are independent, identically distributed (as a variable X) and are also independent of N . S is said to have a **compound distribution**.

Note that S is the sum of a *random* number of *random* quantities, which is the defining feature of a compound distribution.

Illustration: N is the number of claims which arise in a portfolio of business and X_i is the amount of the i th claim. S is the total claim amount.

Because compound distributions arise commonly in general insurance examples, the random variable N is often referred to as the “number of claims” and the distribution of the random variables X_1, X_2, \dots, X_N is referred to as the “individual claim size distribution”, even where the compound distribution arises in another context.

To define a compound distribution, you need to know two components:

- the distribution of N (which is a *discrete* distribution) and
- the distribution of the X_i 's (which may be *any* distribution).

When the particular distribution of N is known, a compound distribution is referred to by the name of this distribution *eg* a **compound Poisson distribution**.

5.1 Moments of compound distributions

The problem with finding the mean and variance of S is that we don't know what N is (as it is a random variable). This is resolved by conditioning on the value of N , ie saying N is some value n . We can then use the results $E[Y] = E[E(Y | X)]$ and $\text{var}[Y] = E[\text{var}(Y | X)] + \text{var}[E(Y | X)]$, replacing the Y with S and X with N , to obtain the mean and variance.

The mean and variance of S are easily found.

$$E(S|N = n) = E(X_1 + X_2 + \dots + X_N | N = n) = E(X_1 + X_2 + \dots + X_n) = nE(X)$$

Similarly:

$$\text{var}(S | N = n) = n \text{var}(X)$$

Therefore we have:

$$E(S) = E[E(S|N)] = E[NE(X)] = E(N)E(X)$$

or:

$$\mu_S = \mu_N \mu_X$$

and:

$$\begin{aligned} \text{var}(S) &= E[\text{var}(S | N)] + \text{var}[E(S | N)] \\ &= E[N \text{var}(X)] + \text{var}[NE(X)] = E(N) \text{var}(X) + \text{var}(N)[E(X)]^2 \end{aligned}$$

or:

$$\sigma_S^2 = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2$$



Question 7.8

If $S = \sum_{i=1}^N X_i$, where X_i are IID random variables from a $\text{Poisson}(5)$ distribution and $N \sim \text{Bin}(10, 0.2)$, calculate the mean and variance of S .

5.2 Generating functions of compound distributions

The MGF of S is given by:

$$M_S(t) = E(e^{tS}) = E[E(e^{tS}|N)]$$

$$\begin{aligned} E(e^{tS}|N=n) &= E[\exp\{t(X_1 + X_2 + \dots + X_N)\}|N=n] \\ &= E[\exp\{t(X_1 + X_2 + \dots + X_n)\}] \\ &= \prod E[\exp(tX_i)] \\ &= [M_X(t)]^n \end{aligned}$$

Therefore we have:

$$M_S(t) = E[\{M_X(t)\}^N] = E[\exp\{N \log M_X(t)\}] = M_N\{\log M_X(t)\}$$

ie to get the MGF of S we substitute $\log M_X(t)$ for t in the MGF of N .

Next we look at the properties of a compound Poisson distribution.

An important illustration is provided by the compound Poisson distribution, which is the case in which $N \sim \text{Poisson } (\lambda)$. In this case $\mu_N = \sigma_N^2 = \lambda$.

Properties:

$$E(S) = \lambda E(X)$$

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

$M_N(t) = \exp\{\lambda(e^t - 1)\}$ so $M_S(t) = \exp[\lambda\{M_X(t) - 1\}]$ from which the mean and variance can be obtained and the results above verified.



Question 7.9

Write down the MGF of a compound Poisson distribution with individual claim size distribution $\text{Gamma}(\alpha, \beta)$ and Poisson parameter λ .

**Question 7.10**

Let $S = X_1 + \dots + X_n$ (and $S = 0$ if $N = 0$) where the X_i 's are independent and identically distributed (as a variable X) and are also independent of N . If N and X are both discrete random variables, show that the PGF of S is given by:

$$G_S(t) = G_N[G_X(t)]$$

6 Exam-type question



Typical past exam question

The number of claims, N , which arise in a year from a group of policies, has a negative binomial distribution, where $P(N = n) = \binom{n+2}{n} 0.9^3 0.1^n$. The claim amounts (in £1,000s) are independent and identically distributed as a $\text{Gamma}(6, 2)$ and are also independent of N . Let Y be the total claim amount arising from these policies.

- (i) Obtain an expression for the moment generating function of Y . (You may assume the MGF of the negative binomial and gamma distributions).
- (ii) Calculate the standard deviation of the total claim amount.

This page has been left blank so that you can keep the chapter summaries together for revision purposes.



Chapter 7 Summary

$E(Y | X)$ is the mean of the conditional distribution of Y given X (which was defined in Chapter 6). The formulae for discrete and continuous distributions are given overleaf.

$\text{var}(Y | X)$ is the variance of the conditional distribution of Y given X , it is given by:

$$\text{var}(Y | X) = E(Y^2 | X) - E^2(Y | X)$$

The unconditional mean and variance can be found from the conditional mean and variance using the formulae given overleaf and on page 16 of the *Tables*.

A quantity that is the sum of a random number of random quantities has a compound distribution:

$$S = X_1 + \dots + X_N$$

We can find the mean, variance and MGF of a compound distribution using the formulae given overleaf.

We can find the skewness using the CGF.



Chapter 7 Formulae

Conditional expectation

$$E[Y | X = x] = \sum_i y_i P[Y = y_i | X = x] = \sum_i y_i \frac{P[Y = y_i, X = x]}{P[X = x]}$$

$$E[Y | X = x] = \int_y y f(y | x) dy = \int_y y \frac{f(x, y)}{f(x)} dy$$

Conditional variance

$$\text{var}[Y | X = x] = E[Y^2 | X = x] - E^2[Y | X = x]$$

Relationships between unconditional and conditional moments

$$E[Y] = E[E(Y | X)]$$

$$\text{var}[Y] = E[\text{Var}(Y | X)] + \text{var}[E(Y | X)]$$

Compound distributions

$$E(S) = E(N)E(X)$$

$$\text{var}(S) = E(N)\text{Var}(X) + \text{var}(N)E^2(X)$$

$$M_S(t) = M_N\{\log M_X(t)\}$$

Chapter 7 Solutions

Solution 7.1

$$E[Y | X = x] = \int_y y f(y | x) dy = \int_y y \frac{f(x, y)}{f(x)} dy$$

Solution 7.2

$$\begin{aligned} E(X | Y = 10) &= \sum x P(X = x | Y = 10) \\ &= 1P(X = 1 | Y = 10) + 2P(X = 2 | Y = 10) \\ &= 1 \times \frac{0.2}{0.4} + 2 \times \frac{0.2}{0.4} \\ &= 1 \times 0.5 + 2 \times 0.5 = 1.5 \end{aligned}$$

Alternatively, you can see this directly by noting that if you 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.

Solution 7.3

Using $E[Y | X = x] = \int_y y \frac{f(x, y)}{f(x)} dy$ and recalling from Chapter 6 that

$$f(x) = \int_y f(x, y) dy :$$

$$\begin{aligned} f(x) &= \int_{y=0}^2 \frac{3}{5} (x^2 + xy) dy \\ &= \frac{3}{5} \left[x^2 y + \frac{1}{2} x y^2 \right]_{y=0}^2 \\ &= \frac{3}{5} (2x^2 + 2x) = \frac{6}{5} x(x+1) \end{aligned}$$

Hence:

$$\begin{aligned}
 E[Y | X = x] &= \int_{y=0}^2 y \frac{\frac{3}{5}(x^2 + xy)}{\frac{6}{5}(x^2 + x)} dy \\
 &= \int_{y=0}^2 y \frac{x+y}{2(x+1)} dy \\
 &= \int_{y=0}^2 \frac{xy + y^2}{2(x+1)} dy \\
 &= \left[\frac{\frac{1}{2}xy^2 + \frac{1}{3}y^3}{2(x+1)} \right]_{y=0}^2 \\
 &= \frac{2x + \frac{8}{3}}{2(x+1)} \\
 &= \frac{x + \frac{4}{3}}{x+1} \\
 &= \frac{3x+4}{3(x+1)}
 \end{aligned}$$

Solution 7.4

(i) $E(Y) = \int_y yf(y) dy$, but $f(y) = \int_x f(x, y) dx$. So:

$$f(y) = \int_{x=0}^1 \frac{3}{5}(x^2 + xy) dx = \frac{3}{5} \left[\frac{1}{3}x^3 + \frac{1}{2}x^2y \right]_{x=0}^1 = \frac{1}{5} + \frac{3}{10}y$$

$$E(Y) = \int_{y=0}^2 \frac{1}{5}y + \frac{3}{10}y^2 dy = \left[\frac{1}{10}y^2 + \frac{1}{10}y^3 \right]_{y=0}^2 = \frac{6}{5} = 1.2$$

$$(ii) \quad E[E(Y | X)] = E\left[\frac{3x+4}{3(x+1)}\right] = \int_x^1 \frac{3x+4}{3(x+1)} f(x) dx$$

But $f(x) = \frac{6}{5}x(x+1)$ from Question 7.3, so:

$$\begin{aligned} E[E(Y | X)] &= \int_{x=0}^1 \frac{3x+4}{3(x+1)} \times \frac{6}{5}x(x+1) dx \\ &= \frac{2}{5} \int_{x=0}^1 3x^2 + 4x dx \\ &= \frac{2}{5} \left[x^3 + 2x^2 \right]_{x=0}^1 = \frac{6}{5} = 1.2 \end{aligned}$$

(iii) Comparing our answers in parts (i) and (ii), we can see that $E[Y] = E[E(Y | X)]$.

Solution 7.5

(i) If $K = k$, then X has a $Poisson(k)$ distribution, which has mean k . So $E[X | K = k] = k$, which can be written as $E[X | K] = K$.

$$(ii) \quad E[X] = E[E[X | K]] = E[K] = \frac{1}{\lambda}.$$

Solution 7.6

$$\text{var}[Y | X = 1] = E(Y^2 | X = 1) - E^2(Y | X = 1).$$

Now $E(Y | X = 1) = 18$ from Example 7.1.

$$\begin{aligned} E(Y^2 | X = 1) &= \sum y^2 P(Y = y | X = 1) \\ &= 10^2 P(Y = 10 | X = 1) + 20^2 P(Y = 20 | X = 1) + 30^2 P(Y = 30 | X = 1) \\ &= 100 \times \frac{0.2}{0.5} + 400 \times \frac{0.2}{0.5} + 900 \times \frac{0.1}{0.5} = 380 \end{aligned}$$

$$\text{So } \text{var}[Y | X = 1] = 380 - 18^2 = 56.$$

Solution 7.7

If $K = k$, X has a $\text{Poisson}(k)$ distribution, which has variance k .

So $\text{var}[X | K = k] = k$ which can be written as $\text{var}[X | K] = K$.

Solution 7.8

$$E[S] = 10 \times 0.2 \times 5 = 10$$

$$\text{var}(S) = (10 \times 0.2) \times 5 + (10 \times 0.2 \times 0.8) \times 5^2 = 50$$

Solution 7.9

$$M_S(t) = M_N(\ln M_X[t]) = \exp[\lambda(M_X[t] - 1)] = \exp\left(\lambda\left[\left(\frac{\beta}{\beta-t}\right)^\alpha - 1\right]\right)$$

Solution 7.10

$$G_S(t) = E(t^S) = E\left[E(t^S | N = n)\right]$$

$$E(t^S | N = n) = E(t^{X_1 + \dots + X_n})$$

$$= E(t^{X_1}) \dots E(t^{X_n})$$

$$= [G_X(t)]^n$$

$$\text{Therefore } G_S(t) = E\left[\left(G_X(t)\right)^N\right] = G_N[G_X(t)].$$

Past Exam Question

- (i) N has a type 2 negative binomial distribution with $k = 3$ and $p = 0.9$, hence:

$$M_N(t) = \left(\frac{0.9}{1 - 0.1e^t} \right)^3$$

If X is the random variable “amount of claim”, then:

$$M_X(t) = \left(1 - \frac{t}{2} \right)^{-6}$$

Therefore:

$$M_Y(t) = \left(\frac{0.9}{1 - 0.1 \left(1 - \frac{t}{2} \right)^{-6}} \right)^3$$

- (ii) Calculating the mean and variance for each of the random variables:

$$E(X) = \frac{6}{2} = 3 \quad \text{var}(X) = \frac{6}{2^2} = 1.5$$

$$E(N) = \frac{3 \times 0.1}{0.9} = \frac{1}{3} \quad \text{var}(N) = \frac{3 \times 0.1}{0.9^2} = \frac{10}{27}$$

Hence:

$$\text{var}(Y) = E(N) \text{var}(X) + E^2(X) \text{var}(N)$$

$$= \frac{1}{3} \times 1.5 + 3^2 \times \frac{10}{27} = 3\frac{5}{6}$$

$$\text{standard deviation}(Y) = \text{£}1,958$$

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.

Chapter 8

The Central Limit Theorem



Syllabus objectives

- (vii) 1. State the Central Limit Theorem for a sequence of independent, identically distributed random variables.
2. Apply the Central Limit Theorem to establish normal approximations to other distributions, and to calculate probabilities.
3. Explain and apply a continuity correction when using a normal approximation to a discrete distribution.

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 Chapter 12 (Hypothesis testing).

Basically, the Central Limit Theorem gives us an approximate distribution of the mean, \bar{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 $X \sim \text{Bin}(100, 0.3)$ would require us to work out 30 probabilities and then add them all up!

1 The Central Limit Theorem

1.1 Definition

If X_1, X_2, \dots, 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 $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ approaches the standard normal distribution, $N(0,1)$, as $n \rightarrow \infty$.

Remember that \bar{X} is the sample mean, calculated as $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

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 $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ and $\frac{\sum X_i - n\mu}{\sqrt{n}\sigma}$ are approximately distributed as $N(0,1)$ for large n .

Alternatively the unstandardised forms can be used. Thus \bar{X} is approximately $N(\mu, \sigma^2 / n)$ and $\sum X_i$ is approximately $N(n\mu, n\sigma^2)$.

In fact the mean and variance are exact (and this will be proved in Chapter 9). It is the shape of the curve that is approximate.

As a notation the symbol “ \doteq ” is used to mean “is approximately distributed”.

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.

**Example**

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 Core Reading, $\mu = 40$, $\sigma = 12$, $n = 50$.

The Central Limit Theorem states that $\bar{X} \sim N(40, 12^2/50)$.

We want $P(\bar{X} < 35)$:

$$\begin{aligned} P(\bar{X} < 35) &\doteq P\left(Z < \frac{35 - 40}{\sqrt{12^2/50}}\right) \\ &= P(Z < -2.946) = 1 - P(Z < 2.946) = 1 - 0.99839 = 0.00161 \end{aligned}$$

**Question 8.1**

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.

2 Normal approximations

Examples of common applications of the Central Limit Theorem follow.

2.1 Binomial distribution

Let X_i be iid Bernoulli random variables, that is, binomial $(1, \theta)$, so that

$$P(X_i = 1) = \theta$$

$$P(X_i = 0) = 1 - \theta$$

In other words X_i is the number of successes in a single Bernoulli trial.

Consider X_1, X_2, \dots, 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 $X = \sum X_i \sim \text{binomial}(n, \theta)$. Also note that $\frac{X}{n} = \bar{X}$. As a result of the Central Limit Theorem it can be said that, for large n :

$$\bar{X} \approx N(\mu, \sigma^2/n) \text{ or } \sum X_i \approx N(n\mu, n\sigma^2)$$

For the Bernoulli distribution:

$$\mu = E[X_i] = \theta \quad \text{and} \quad \sigma^2 = \text{var}[X_i] = \theta(1 - \theta)$$

Therefore $\sum X_i \approx N(n\theta, n\theta(1 - \theta))$ 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 8.2

What would be the distribution of \bar{X} ?

What is “large n ”? A commonly quoted rule of thumb is that the approximation can be used only when both $n\theta$ and $n(1-\theta)$ 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 $\theta = 0.5$ the Bernoulli distribution is symmetrical. In this case both $n\theta$ and $n(1-\theta)$ equal 5 when $n = 10$, and so the rule of thumb suggests that $n = 10$ is large enough.

As θ moves away from 0.5 towards either 0 or 1 the Bernoulli distribution becomes more severely skewed. For example, when $\theta = 0.2$ or 0.8 the rule of thumb gives $n = 25$ as large enough, but, when $\theta = 0.05$ or 0.95 the rule of thumb gives $n = 100$ as large enough.

Recall from Chapter 4, that the binomial distribution can also be approximated by the Poisson distribution. This approximation was valid when n was large and θ was small. This contrasts with the normal approximation, which requires n to be large and θ to be close to $\frac{1}{2}$ (although, as n gets larger the normal approximation works well even if θ is not close to $\frac{1}{2}$).

2.2 Poisson distribution

Let $X_i, i = 1, 2, \dots, n$ be iid Poisson (λ) random variables.

$$\therefore \mu = E[X_i] = \lambda \text{ and } \sigma^2 = \text{var}[X_i] = \lambda$$

The Central Limit Theorem implies that

$$\sum X_i \sim N(n\lambda, n\lambda) \text{ for large } n$$

But $\sum X_i \sim \text{Poisson}(n\lambda)$ and so, for large n , $\text{Poisson}(n\lambda) \sim N(n\lambda, n\lambda)$

or, equivalently, $\text{Poisson}(\lambda) \sim N(\lambda, \lambda)$ for large λ .

Again, we are approximating using a normal distribution, which has the same mean and variance as the Poisson distribution.

**Question 8.3**

Why is $\sum X_i \sim Poi(n\lambda)$?

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 $n \rightarrow \infty$ and $p \rightarrow 0$. 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 $X_i, i = 1, 2, \dots, n$ 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 , $Y = \sum X_i \approx N(n/\lambda, n/\lambda^2)$

$\therefore Y$, which is $Ga(n, \lambda)$, will have a normal approximation for large values of n .

Recall from Chapter 6, that if $X_i \sim Exp(\lambda)$ then $\sum X_i \sim Ga(n, \lambda)$.

Since $\chi_k^2 \equiv Ga(k/2, 1/2)$, 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 4 that if $X \sim Ga(\alpha, \lambda)$ 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.

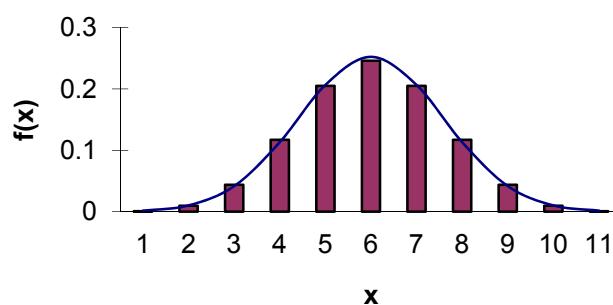
For an integer-valued discrete distribution, such as the binomial or Poisson, it is perfectly reasonable to consider individual probabilities such as $P(X = 4)$. However if X is continuous, such as the normal, $P(X = 4)$ is not meaningful and is taken to be zero. For a continuous variable it is sensible 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: this is why we say that it is not meaningful, eg:

$$P(X = 4) = P(4 \leq X \leq 4) = \int_4^4 f(x) dx = 0$$

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 $\text{Bin}(10, 0.5)$ 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,

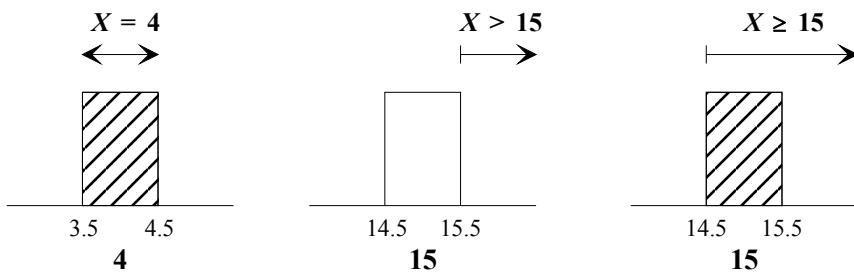
$X = 4$ is equivalent to " $3.5 < X < 4.5$ "

$X > 15$ 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 8.4

Draw the corresponding diagrams for:

- (i) $X < 8$ (ii) $X \leq 8$

Hence obtain the continuity-corrected inequalities.



Question 8.5

When using the continuity correction with a random variable X that can take any integer value, what expressions 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) $4 \leq 10X < 48$? (hard)

Example

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

$$X \sim \text{Poisson}(20) \therefore X \sim N(20, 20) \therefore \frac{X - 20}{\sqrt{20}} \sim N(0, 1)$$

$P(X \leq 15) \equiv P(X < 15.5)$: using continuity correction

$$= P\left(Z < \frac{15.5 - 20}{\sqrt{20}}\right) = P(Z < -1.006)$$

$= 1 - 0.84279$, 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:



Example

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.

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:

$$(i) \quad P(X = 36) = \frac{40^{36} e^{-40}}{36!} = 0.0539$$

- (ii) In order to calculate this, we sum the probabilities of getting 42, 43, 44 and 45:

$$\begin{aligned} P(42 \leq X \leq 45) &= \frac{40^{42} e^{-40}}{42!} + \frac{40^{43} e^{-40}}{43!} + \frac{40^{44} e^{-40}}{44!} + \frac{40^{45} e^{-40}}{45!} \\ &= 0.0585 + 0.0544 + 0.0495 + 0.0440 \\ &= 0.2064 \end{aligned}$$

The normal approximation to this Poisson distribution would be $N(40, 40)$. Calculating the probabilities again, and using continuity corrections:

$$\begin{aligned} (i) \quad P(X = 36) &= P(35.5 < X < 36.5) \\ &= P(-0.712 < Z < -0.553) \\ &= \Phi(0.712) - \Phi(0.553) \\ &= 0.7617 - 0.7099 = 0.0518 \end{aligned}$$

$$\begin{aligned}\text{(ii)} \quad P(42 \leq X \leq 45) &= P(41.5 < X < 45.5) \\ &= P(0.237 < Z < 0.870) \\ &= \Phi(0.870) - \Phi(0.237) \\ &= 0.8078 - 0.5937 = 0.2141\end{aligned}$$

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 8.6

Use a normal approximation to calculate an approximate value for the probability that an observation from a $\text{Gamma}(25, 50)$ random variable falls between 0.4 and 0.8.



Question 8.7

What is the approximate probability that the mean of a sample of 10 observations from a $\text{Beta}(10, 10)$ random variable falls between 0.48 and 0.52?



Question 8.8

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.

4 Exam-type question



Exam-type question

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.



Chapter 8 Formulae

The Central Limit Theorem

For X_1, \dots, X_n iid RV's with mean μ and variance σ^2 :

$$\sum X_i \stackrel{d}{\sim} N(n\mu, n\sigma^2) \Rightarrow \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \stackrel{d}{\sim} N(0,1) \quad \text{as } n \rightarrow \infty$$

$$\bar{X} \stackrel{d}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \stackrel{d}{\sim} N(0,1) \quad \text{as } n \rightarrow \infty$$

Normal approximations

$$\begin{aligned} \text{Bin}(n, p) &\stackrel{d}{\sim} N(np, npq) & np > 5, nq > 5 \\ \text{Poi}(\lambda) &\stackrel{d}{\sim} N(\lambda, \lambda) & \lambda \text{ large} \end{aligned} \Bigg\} \text{ with continuity correction}$$

$$Ga(\alpha, \lambda) \stackrel{d}{\sim} N\left(\frac{\alpha}{\lambda}, \frac{\alpha}{\lambda^2}\right) \quad \alpha \text{ large}$$

$$\chi_k^2 \stackrel{d}{\sim} N(k, 2k) \quad k \text{ large}$$

This page has been left blank so that you can keep the chapter summaries together for revision purposes.

Chapter 8 Solutions

Solution 8.1

Using the notation given in 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:

$$T \approx N(65 \times 6200, 65 \times 650^2) = N(403000, 5240^2)$$

So the probability is found as follows:

$$\begin{aligned} P(T > 400,000) &\doteq P\left(Z > \frac{400,000 - 403,000}{5,240}\right) \\ &= P(Z > -0.572) = P(Z < 0.572) = 0.71634 \end{aligned}$$

Solution 8.2

Since $\bar{X} = \frac{\sum X_i}{n}$, then:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n} E\left(\sum X_i\right) = \frac{1}{n} n\theta = \theta \\ \text{var}(\bar{X}) &= \text{var}\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2} \text{var}\left(\sum X_i\right) = \frac{1}{n^2} n\theta(1-\theta) = \frac{\theta(1-\theta)}{n} \end{aligned}$$

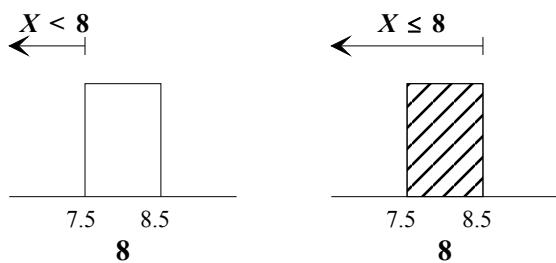
Therefore $\bar{X} \approx N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$.

Solution 8.3

Recall from Chapter 6 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 $\sum X_i \sim Poi(\lambda) + \dots + Poi(\lambda) \sim Poi(\lambda + \dots + \lambda) \sim Poi(n\lambda)$.

Solution 8.4

- (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 $X < 8.5$.

Solution 8.5

- (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. Using a continuity correction on these values, we get $5 < 10X < 45$, which gives an answer of $0.5 < X < 4.5$.

Solution 8.6

The mean and variance of a general gamma distribution are $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$, so here the mean and variance are 0.5 and 0.01 respectively. If X is the gamma RV, then we will use $X \sim N(0.5, 0.01)$:

$$\begin{aligned}
 P(0.4 < X < 0.8) &= P(-1 < Z < 3) \\
 &= \Phi(3) - \Phi(-1) \\
 &= \Phi(3) - [1 - \Phi(1)] \\
 &= 0.99865 - 0.15866 = 0.840
 \end{aligned}$$

No continuity correction is required, as we are dealing with a continuous distribution.

The exact answer is 0.8387.

Solution 8.7

The $Beta(10,10)$ distribution has mean $\frac{10}{10+10} = 0.5$ 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, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, so here $\bar{X} \sim N\left(0.5, \frac{0.01190}{10}\right)$:

$$\begin{aligned} P(0.48 < \bar{X} < 0.52) &= P(-0.5798 < Z < 0.5798) \\ &= \Phi(0.5798) - \Phi(-0.5798) \\ &= \Phi(0.5798) - (1 - \Phi(0.5798)) \\ &= 0.71897 - 0.28103 = 0.43794 \end{aligned}$$

No continuity correction is required, as we are dealing with a continuous distribution.

Solution 8.8

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:

$$\begin{aligned} P(X > 20) &\rightarrow P(X > 20.5) \\ &= P(Z > 1.540) \\ &= 1 - \Phi(1.54) \\ &= 1 - 0.93822 = 0.06178 \end{aligned}$$

So the approximate probability that more than 20 policies will lapse is 0.062.

The exact answer is 0.0663.

Exam-type question

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 $N(37.5, 37.5)$.

$$\begin{aligned} P(X \geq 50) &\rightarrow P(X > 49.5) && \text{continuity correction} \\ &= P\left(Z > \frac{49.5 - 37.5}{\sqrt{37.5}}\right) \\ &= P(Z > 1.960) \\ &= 1 - \Phi(1.960) \\ &= 0.025 \end{aligned}$$

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.

Chapter 9

Sampling and statistical inference



Syllabus objectives

- (viii) 1. Explain what is meant by a sample, a population and statistical inference.
2. Define a random sample from a distribution of a random variable.
3. Explain what is meant by a statistic and its sampling distribution.
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.
5. State and use the basic sampling distributions for the sample mean and the sample variance for random samples from a normal distribution.
6. State and use the distribution of the t-statistic for random samples from a normal distribution.
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 the mean and variance results form distributions as well. Naturally, we would then wish to know what the means and variances of the new distributions are and then, if possible, what the distribution actually is.

The distributions of the sample mean and sample variance are called sampling distributions and will be used extensively in Chapters 11 & 12 to construct confidence intervals and carry out hypothesis tests.

Part of this work will finally explain mathematically why the sample variance given in Chapter 1, Section 3.1 was defined to be $S^2 = \frac{1}{n-1} [\sum X^2 - n\bar{X}^2]$ rather than $S^2 = \frac{1}{n} [\sum X^2 - n\bar{X}^2]$.

We will also make use of the Central Limit Theorem from Chapter 8 to obtain the distribution of the sample mean.

Finally, this chapter will look at the *t*-distribution and the *F*-distribution in greater detail. You will require statistical tables to be able to progress through this chapter.

If you have still to obtain these, then recall that some of the tables required can be found at the end of Chapter 4 so that you are able to progress through the course. However, it is vital that you purchase a copy of the *Tables* as soon as possible so that you are thoroughly familiar with them before the examination. They are available directly from the Profession.

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, $\underline{X} = (X_1, X_2, \dots, X_n)$. An observed sample will be denoted by $\underline{x} = (x_1, x_2, \dots, x_n)$. The population distribution will be specified by a density (or probability function) denoted by $f(x; \theta)$, 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 9.1

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 claims might be for the whole year.

1.2 Definition of a statistic

A statistic is a function of \underline{X} only and does not involve any unknown parameters.

Thus $\bar{X} = \frac{\sum X_i}{n}$ and $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ are statistics whereas $\frac{1}{n} \sum (X_i - \mu)^2$ is not, unless of course μ is known.

Note here the difference between μ , which is the population mean (*i.e.* the mean for all possible observations, which is usually unknown) and \bar{X} , which is the sample mean (*i.e.* the mean of the sample values which we can calculate for any given sample).

We might also be interested in values 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 that we have n independent and identically distributed random variables, X_i , $i = 1, 2, \dots, n$, each with mean μ and variance σ^2 . Recall that the sample mean is $\bar{X} = \frac{\sum X_i}{n}$.

Consider first $\sum X_i$:

$$E[\sum X_i] = \sum E[X_i] = \sum \mu = n\mu \text{ since they are identically distributed}$$

$$\text{var}[\sum X_i] = \sum \text{var}[X_i] \text{ since they are independent}$$

$$= n\sigma^2 \text{ since they are identically distributed}$$

We are using the results from Chapter 6 that $E[X_1 + \dots + X_n] = E[X_1] + \dots + E[X_n]$, and if X_1, \dots, X_n are independent $\text{var}[X_1 + \dots + X_n] = \text{var}[X_1] + \dots + \text{var}[X_n]$.

As $\bar{X} = \frac{1}{n} \sum X_i$, we can now write down that $E[\bar{X}] = \mu$ and
 $\text{var}[\bar{X}] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$.

Note: $sd[\bar{X}] = \frac{\sigma}{\sqrt{n}}$ 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). This is a very important result and will be used extensively in Chapters 11 and 12.

A consequence of the variance 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 $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$.

Considering only the mean of S^2 , it can be proved as follows that $E[S^2] = \sigma^2$:

$$S^2 = \frac{1}{n-1} \left[\sum X_i^2 - n\bar{X}^2 \right]$$

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

$$\begin{aligned} E[S^2] &= \frac{1}{n-1} \left(\sum E[X_i^2] - nE[\bar{X}^2] \right) \\ &= \frac{1}{n-1} \left\{ \sum (\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) \right\} \\ &= \frac{1}{n-1} \left\{ n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2 \right\} \\ &= \frac{1}{n-1} \left\{ (n-1)\sigma^2 \right\} = \sigma^2 \end{aligned}$$

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.

The $n-1$ denominator is used so that the mean of S^2 equals the true value of σ^2 . This is the motivation behind the definition of the sample variance. Later in Chapter 10, 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 CT3 is for a normal population. This is covered in Section 3.2.

**Question 9.2**

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 $\lambda = 5$. What are the mean and variance of a sample mean based on 30 days figures?

**Question 9.3**

What are 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 $\lambda = 3,000$?

The results for the Pareto distribution are in the *Tables*. If you have yet to purchase your copy they are repeated below for your convenience:

$$f(x) = \alpha\lambda^\alpha(\lambda + x)^{-\alpha-1} \quad x > 0$$

$$E(X) = \frac{\lambda}{\alpha - 1}, \quad \text{var}(X) = \frac{\alpha\lambda^2}{(\alpha - 1)^2(\alpha - 2)}$$

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 :

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{d}{\sim} N(0, 1) \text{ or } \bar{X} \stackrel{d}{\sim} N(\mu, \sigma^2/n)$$

This result is often called the z result.

It transpires that the above result gives the exact sampling distribution of \bar{X} for random samples from a normal population.

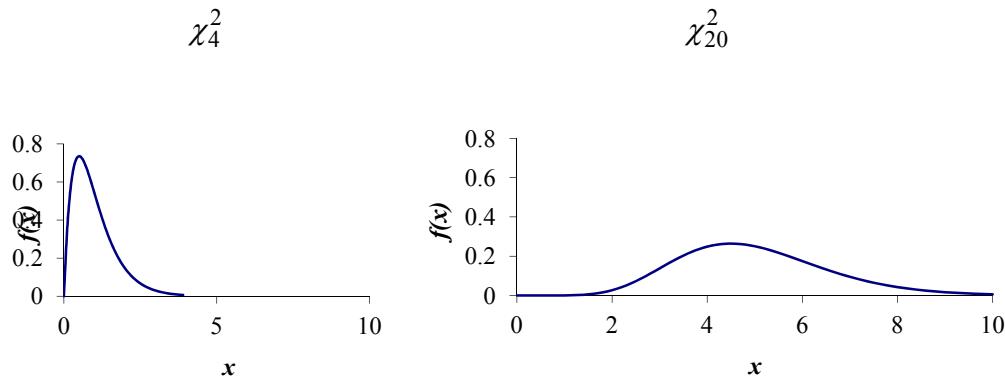
3.2 The sample variance

The sampling distribution of S^2 when sampling from a normal population, with mean μ and variance σ^2 , is:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

This is a more advanced result, the proof of which lies outside the Subject CT3 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 .



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_k^2 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, now being applied to the normal distribution.

$$\text{var}\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1) \Rightarrow \text{var}[S^2] = \frac{\sigma^4}{(n-1)^2} \cdot 2(n-1) = \frac{2\sigma^4}{n-1}.$$

These are important results to be able to derive, and so:



Question 9.4

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 . Derive from first principles $E(S^2)$ and $\text{var}(S^2)$.

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 9.5

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, and (ii) S will exceed 41.7.

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 $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$.

Other cases such as Poisson, binomial, gamma can be considered in a similar way, but only the normal has the independence property.



Question 9.6

What is the probability that, for the sample in Question 9.5, both conditions (i) and (ii) will hold?

4 The *t* result

The sampling distribution for \bar{X} , that is, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ 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 $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

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:

$t_k = \frac{N(0,1)}{\sqrt{\chi_k^2/k}}$ where the $N(0,1)$ and χ_k^2 random variables are independent.

Then the *t* result above follows from the sampling distributions of the last section, that is, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is the $N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$ is the χ_k^2 , together with their independence, to obtain $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$ when sampling from a normal population.

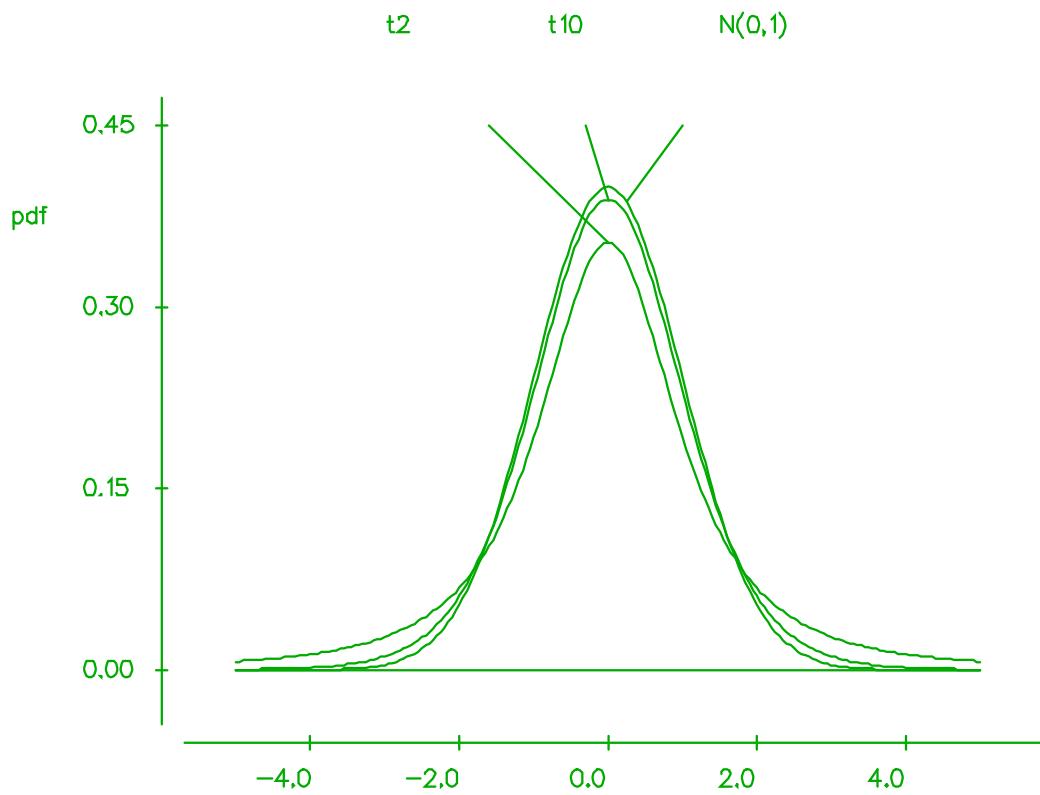
The *t* distribution is symmetrical about zero and its critical points are tabulated.

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 $k \rightarrow \infty$, $t_k \rightarrow N(0,1)$.

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 $k / (k - 2)$.

**Example**

State the distribution of $\frac{\bar{X} - 100}{S/\sqrt{5}}$ for a random sample of 5 values taken from a $N(100, \sigma^2)$ population. What is the probability that this quantity will exceed 1.533?

Solution

From previous results $\frac{\bar{X} - 100}{S/\sqrt{5}} \sim t_4$.

From the *Tables*, we see that the probability that this quantity will exceed 1.533 is 10%.

**Question 9.7**

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 $\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}$.
 - (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.

5 The *F* result for variance ratios

The *F* distribution is defined by $F = \frac{U/v_1}{V/v_2}$, 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 $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$.

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 $\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim F_{n_2-1, n_1-1}$.

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 make the larger value the numerator.

Alternatively, $F \sim F_{n_1-1, n_2-1} \Leftrightarrow \frac{1}{F} \sim F_{n_2-1, n_1-1}$.

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 11 in the work on confidence intervals and Chapter 12 in the work on hypothesis tests.

The percentage points for the *F* distribution can be found on pages 170-174 of the *Tables*.



Example

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 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 $\frac{1}{F_{m,n}}$ result:

$$\begin{aligned} P(F_{11,8} < 0.3392) &= P\left(\frac{1}{F_{11,8}} > \frac{1}{0.3392}\right) \\ &= P\left(F_{8,11} > \frac{1}{0.3392}\right) = P(F_{8,11} > 2.948) = 0.05 \end{aligned}$$

- (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 $\frac{1}{F_{m,n}}$ result again:

$$P(F_{14,6} < p) = P\left(F_{6,14} > \frac{1}{p}\right) = 0.01 \Rightarrow \frac{1}{p} = 4.456 \Rightarrow p = 0.2244$$



Question 9.8

- (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) $P(F_{18,9} > p) = 99\%$

We now apply the F result to problems involving sample variances:

**Example**

For random samples of size 10 and 25 from two normal populations with equal variances, use F tables to determine the values of α and β such that

$P\left(\frac{S_1^2}{S_2^2} > \alpha\right) = 0.05$ and $P\left(\frac{S_1^2}{S_2^2} < \beta\right) = 0.05$, where S_1^2 is the sample variance from the sample of size 10, and S_2^2 is the other sample variance.

Solution

Since the population variances are equal, $\frac{S_1^2}{S_2^2} \sim F_{9,24}$ and $\frac{S_2^2}{S_1^2} \sim F_{24,9}$.

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 $\frac{S_1^2}{S_2^2} < \beta$ is equivalent to $\frac{S_2^2}{S_1^2} > 1/\beta$ and $P(F_{24,9} > 2.900) = 0.05$, giving $\beta = \frac{1}{2.900} = 0.345$.

**Question 9.9**

What is 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?

6 Exam-type question



Exam-type question

A random sample of 10 observations is drawn from a normal distribution with mean μ and 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 $P(\bar{X} - \bar{Y} > 3)$.

7 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 6 to 9.
2. Attempt some of the questions in Part 2 of the **Question and Answer Bank**. 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. Students have said:

“The Flashcards were useful for highlighting what we needed to memorise and for revision on-the-go.”

“I found the Flashcards INCREDIBLY helpful. The method recommended really helped me bed down the facts. Thanks for providing a quality product.”

You can find lots more information on our website at www.ActEd.co.uk.

Buy online at www.ActEd.co.uk/estore

This page has been left blank so that you can keep the chapter summaries together for revision purposes.



Chapter 9 Summary

The sample mean and sample variance are given by:

$$\bar{X} = \frac{\sum X_i}{n} \quad S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2,$$

We can find their mean and variance – these are given overleaf.

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. These are given overleaf.

When sampling from a normal population, the sample mean and variance are independent.

The t and F distributions were defined, in Chapter 4, as:

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi^2_k/k}} \quad F_{m,n} \equiv \frac{\chi^2_m/m}{\chi^2_n/n}$$

Since $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ we can show that:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$$



Chapter 9 Formulae

Moments of statistics

$$E(\bar{X}) = \mu \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \quad E(S^2) = \sigma^2 \quad \text{any distribution}$$

$$\text{var}(S^2) = \frac{2\sigma^4}{n-1} \quad \text{normal distribution only}$$

t-distribution

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi_k^2/k}}$$

F-distribution

$$F_{m,n} \equiv \frac{\chi_m^2/m}{\chi_n^2/n}$$

Sampling distributions

$$\bar{X} \stackrel{\text{d}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{any distribution large } n \text{ (or normal any } n)$$

$$\Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \sigma^2 \text{ known}$$

$$\Rightarrow \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad \sigma^2 \text{ unknown}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \quad \text{normal distribution only}$$

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1} \quad \text{normal distributions only}$$

Chapter 9 Solutions

Solution 9.1

(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.

Solution 9.2

The Poisson distribution in the question has mean and variance of 5.

If the sample size is 30 then $E[\bar{X}] = 5$ and $\text{var}[\bar{X}] = \frac{5}{30} = 0.167$.

Solution 9.3

The Pareto distribution has a mean of $\frac{\lambda}{\alpha-1}$, and variance of $\frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)}$, so the distribution in the question has $\mu = 750$ and $\sigma^2 = 937,500$.

Thus $E[\bar{X}] = 750$ and $\text{var}[\bar{X}] = \frac{937,500}{110} = 8,522.7$.

Solution 9.4

The sampling distribution for S^2 is given by:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Now a χ_k^2 is a Gamma with $\alpha = \frac{k}{2}$ and $\lambda = \frac{1}{2}$. Therefore:

$$E[\chi_k^2] = \frac{k/2}{1/2} = k \quad \text{and} \quad \text{var}[\chi_k^2] = \frac{\alpha}{\lambda^2} = \frac{k/2}{(1/2)^2} = 2k$$

$$E\left(\frac{(n-1)S^2}{\sigma^2}\right) = E(\chi_{n-1}^2) = n-1$$

$$\Rightarrow \frac{(n-1)}{\sigma^2} E(S^2) = n-1$$

$$\Rightarrow E(S^2) = \sigma^2$$

$$\text{var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = \text{var}(\chi_{n-1}^2) = 2(n-1)$$

$$\Rightarrow \frac{(n-1)^2}{\sigma^4} \text{var}(S^2) = 2(n-1)$$

$$\Rightarrow \text{var}(S^2) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}$$

Solution 9.5

(i) Since $\bar{X} \sim N(100, 25^2 / 5) = N(100, 125)$:

$$\begin{aligned} P(80 < \bar{X} < 120) &= P\left(\frac{80-100}{\sqrt{125}} < Z < \frac{120-100}{\sqrt{125}}\right) \\ &= P(-1.789 < Z < 1.789) \\ &= \Phi(1.789) - \Phi(-1.789) \\ &= 0.96319 - (1 - 0.96319) = 0.926 \end{aligned}$$

(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:

$$P(S > 41.7) \approx 0.0253$$

Solution 9.6

Since \bar{X} and S^2 are independent, we can factorise the probability:

$$P(80 < \bar{X} < 120 \text{ } \& \text{ } S > 41.7) = P(80 < \bar{X} < 120) \times P(S > 41.7)$$

Referring back to the previous question, we have already found the probabilities. So:

$$P(80 < \bar{X} < 120 \text{ } \& \text{ } S > 41.7) = 0.023$$

Solution 9.7

(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 $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

(ii)(a) The variance of $\bar{X}_1 - \bar{X}_2$ is now $\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$ and so standardising gives:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0,1)$$

(ii)(b) As $\frac{(n_1-1)S_1^2}{\sigma^2} \sim \chi_{n_1-1}^2$ and $\frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi_{n_2-1}^2$ 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 (ie $\chi_m^2 + \chi_n^2 \sim \chi_{m+n}^2$), which we proved, in Chapter 6, Section 4.4.

(ii)(c) Using the definition of the t distribution:

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi_k^2/k}}$$

Part (ii)(a) had a $N(0,1)$ distribution and part (ii)(b) had a $\chi_{n_1+n_2-2}^2$ distribution.

So:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2}$$

$$\frac{\sqrt{(n_1-1)S_1^2 + (n_2-1)S_2^2}}{\sqrt{\sigma^2(n_1+n_2-2)}} \sim t_{n_1+n_2-2}$$

The σ^2 's cancel to give:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2}$$

We will see that $\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}$, 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.

Solution 9.8

(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 $\frac{1}{F_{m,n}}$ 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 the 10% tables gives:

$$P(F_{24,30} > p) = 0.10 \Rightarrow p = 1.638$$

(ii)(b) Since 99% of the distribution is greater than p it must be on the lower tail. So we need to use the $\frac{1}{F_{m,n}}$ result:

$$P(F_{18,9} > p) = P\left(F_{9,18} < \frac{1}{p}\right) = 0.99 \Rightarrow P\left(F_{9,18} > \frac{1}{p}\right) = 0.01$$

Hence reading off the 1% tables gives $\frac{1}{p} = 3.597 \Rightarrow p = 0.278$.

Solution 9.9

If X denotes the sample with 10 values and Y denotes the sample with 5 values, we know from the previous result that $\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{9,4}$.

Since the population variances are equal, this means that $S_X^2/S_Y^2 \sim F_{9,4}$.

So $P(S_X^2/S_Y^2 > 6) = P(F_{9,4} > 6)$.

From the *Tables* page 172 we see that the tabulated critical value of $F_{9,4}$ is 5.999. So the required probability is very nearly 5%.

Exam-type question

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:

$$\bar{X} \sim N\left(\mu, \frac{15^2}{10}\right) \quad \bar{Y} \sim N\left(\mu, \frac{12^2}{25}\right)$$

Using the work from Chapter 6 gives:

$$\bar{X} - \bar{Y} \sim N\left(0, \frac{15^2}{10} + \frac{12^2}{25}\right) = N(0, 28.26)$$

$$\begin{aligned} P(\bar{X} - \bar{Y} > 3) &= P(Z > 0.564) \\ &= 1 - P(Z < 0.564) \\ &= 1 - 0.71362 \\ &= 0.28638 \end{aligned}$$

Chapter 10

Point estimation



Syllabus objectives

- (ix) 1. *Describe the method of moments for constructing estimators of population parameters and apply it.*
2. *Describe the method of maximum likelihood for constructing estimators of population parameters and apply it.*
3. *Define the terms: efficiency, bias, consistency and mean square error.*
4. *Define the property of unbiasedness of an estimator and use it.*
5. *Define the mean square error of an estimator and use it to compare estimators.*
6. *Describe the asymptotic distribution of maximum likelihood estimators and use it.*

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:

Claims	0	1	2	3	4	5	6
Frequency	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 (from Chapter 1) to the population moments (from Chapters 3 and 4).

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 got. These are not the only methods of obtaining estimates (for example in Subject CT6 you will meet the method of percentiles) and 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 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 chapter very often contributes one large or two medium sized questions to the Subject CT3 paper and often these are for unfamiliar distributions to test that you can *apply* the principles of estimation rather than just regurgitating the notes.

1 The method of moments

The basic principle is to equate population moments (*i.e.* the means, variances, *etc* of the theoretical model) **to corresponding sample moments** (*i.e.* 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, *i.e.*:

$$E[X] = \frac{1}{n} \sum_{i=1}^n x_i$$



Example 10.1

A random sample from a $Exp(\lambda)$ 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 a $Exp(\lambda)$ from page 11 of the *Tables* is $E(X) = \frac{1}{\lambda}$.

The sample mean is $\bar{x} = \frac{14.84 + 0.19 + 11.75 + 1.18 + 2.44 + 0.53}{6} = 5.155$

Equating these gives us the method of moments estimate:

$$\frac{1}{\lambda} = 5.155 \Rightarrow \lambda = 0.1940$$

Because this is an estimate of λ rather than the true value, we distinguish this by putting a “hat” or similar over the parameter. So we have $\hat{\lambda} = 0.1942$.



Question 10.1

A random sample from a $Poisson(\mu)$ distribution is as follows:

4, 2, 7, 3, 1, 2, 5, 4, 0, 2

Calculate the method of moments estimate for μ .

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'd do is 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 (see below).

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 $\hat{\mu} = \bar{x}$ (the lower case denotes actual sample values being used). Whereas “estimator” refers to the *distribution* of all the results obtained from many samples, eg $\hat{\mu} = \bar{X}$.

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 the one-parameter case:

$$E[X] = \frac{1}{n} \sum_{i=1}^n x_i$$

The second-order equation is:

$$E[X^2] = \frac{1}{n} \sum_{i=1}^n x_i^2$$

or equivalently:

$$\text{var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$



Question 10.2

Show that these two second-order equations give the same answers for the parameter estimators.



Example 10.2

A random sample from a $\text{Bin}(n, p)$ distribution yields the following values:

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 $\text{Bin}(n, p)$ 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:

$$\hat{n}\hat{p} = 3.875 \tag{1}$$

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 these gives:

$$\hat{n}\hat{p}(1-\hat{p}) + (\hat{n}\hat{p})^2 = 17.875 \quad (2)$$

Substituting equation (1) into (2) gives:

$$3.875(1-\hat{p}) + 3.875^2 = 17.875 \Rightarrow \hat{p} = 0.2621$$

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 $\text{var}(X) = \hat{n}\hat{p}(1-\hat{p})$ and $\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{143}{8} - 3.875^2 = 2.859375$. Equating these gives:

$$\hat{n}\hat{p}(1-\hat{p}) = 2.859375 \quad (3)$$

Substituting equation (1) into (3) gives:

$$3.875(1-\hat{p}) = 2.859375 \Rightarrow \hat{p} = 0.2621$$

And hence $\hat{n} = 14.78$ as before.



Question 10.3

A random sample of size 10 from a Type 2 $NBin(k, p)$ distribution is as follows:

1, 1, 0, 1, 1, 1, 3, 2, 0, 5

Calculate method of moments estimates of k and p .

Note that s^2 with divisor $(n - 1)$ is often used in place of the second central sample moment. ie the definition of the sample variance given in Chapter 1 and quoted on page 22 of the *Tables*.

So our second-order equation is now:

$$\text{var}(X) = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right\}$$

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 in Section 3.



Example 10.3

A random sample from a $\text{Bin}(n, p)$ 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 .

Solution

We have sample mean and variance of:

$$\bar{x} = 3.875 \quad \text{and} \quad s^2 = \frac{1}{7} \left\{ 143 - 8 \times 3.875^2 \right\} = 3.26786$$

And population mean and variance of:

$$np \quad \text{and} \quad np(1-p)$$

Equating these gives:

$$\hat{np} = 3.875 \quad \text{and} \quad \hat{np}(1-\hat{p}) = 3.26786$$

Solving gives $\hat{p} = 0.1567$ and $\hat{n} = 24.73$ which are different to those calculated in Example 10.2.

**Question 10.4**

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 .

**Question 10.5**

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.

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:

$$E[X] = \frac{1}{n} \sum x_i \quad E[X^2] = \frac{1}{n} \sum x_i^2 \quad E[X^3] = \frac{1}{n} \sum x_i^3$$

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, \dots, 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 $f(x_1) \times f(x_2) \times f(x_3) \times \dots \times f(x_n)$. The

above statement is saying that the likelihood function is the product of the PDFs/PFs of the sample values.

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, \dots, 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:

$$\frac{n}{\lambda} - \sum x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$$

$$\therefore \text{MLE is } \hat{\lambda} = \frac{1}{\bar{x}}$$

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:

$$\frac{d^2}{d\lambda^2} \log L(\lambda) = -\frac{n}{\lambda^2} < 0 \Rightarrow \max$$

It is important that you do check, whether formally or through simple logic, *and* state this (together with your working/reasoning) in the exam to receive all the marks.

Note that, at the differentiation stage, any terms that do not contain the parameter (λ in this case) will disappear. So when you are writing down the log-likelihood, any terms that don't contain the parameter can be thought of as "a constant".



Question 10.6

A random sample of size n (ie x_1, x_2, \dots, x_n) is taken from a $Poisson(\mu)$ distribution.

- (i) Derive the maximum likelihood estimator of μ .
- (ii) A sample of 10 observations from a $Poisson(\mu)$ distribution had a total of 24. Calculate the maximum likelihood estimate $\hat{\mu}$.

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 10.7

The MLEs of the parameters of a lognormal distribution have been found to be $\hat{\mu} = 2$ and $\hat{\sigma}^2 = 0.25$. What can you say about the MLE of the mean of the lognormal distribution?

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 $\log L$ and simplify the resulting expression.
- Partially differentiate $\log L$ 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. If θ_1 and θ_2 are the parameters to be estimated, then the condition is that the Hessian matrix:

$$\begin{pmatrix} \frac{\partial^2 \ln L}{\partial \theta_1^2} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_2^2} \end{pmatrix}$$

ie the matrix of second derivatives, is negative definite. However, this goes beyond the scope of Subject CT3 and you would not be expected to check for a maximum in the two-parameter case. More detail on the second-order condition is included in the appendix at the end of this chapter if you are interested.



Example 10.4

Find 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:

$$L = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right] = \frac{1}{\sigma^n} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] \times \text{constant}$$

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:

$$\begin{aligned} \frac{\partial}{\partial \mu} \log L &= \frac{1}{2\sigma^2} \sum_{i=1}^n 2(x_i - \mu) = \frac{1}{\sigma^2} \left(\sum_{i=1}^n x_i - n\mu \right) \\ \frac{\partial}{\partial \sigma} \log L &= -\frac{n}{\sigma} - \frac{1}{2} \left(-\frac{2}{\sigma^3} \right) \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{\sigma} \left(\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - n \right) \end{aligned}$$

Setting these to zero gives:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

and:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \Rightarrow \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

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:

$$f(x) = \frac{1}{b-a} \quad a < x < b$$

We look at this in the next example – note how we specify when the likelihood is zero (*i.e.* it does not exist for the specified values of the parameter) and non-zero (*i.e.* where it does exist for specified values of the parameter).

The second important feature about this example is that the usual route of finding the maximum using differentiation breaks down.



Example 10.5

Find the maximum likelihood estimate of θ for $U(0, \theta)$ based on a sample x_1, x_2, \dots, x_n .

Solution

For a sample from the $U(0, \theta)$ distribution we must have $0 \leq x_1, \dots, x_n \leq \theta$. Hence $\max x_i \leq \theta$. Thus the likelihood for a sample of size n is:

$$L = \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq \max x_i \\ 0 & \text{otherwise} \end{cases}$$

Differentiation doesn't work because $\frac{d}{d\theta} \ln L(\theta) = -\frac{n}{\theta}$ which gives a turning point of $\theta \rightarrow \infty$! 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 $\theta \rightarrow \infty$.

So using common sense, we must find the θ that maximises $L(\theta) = \frac{1}{\theta^n}$. So 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, \dots, 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 between y and z and p above z in addition to the n known values, then we would use:

$$L(\theta) = \prod_{i=1}^n f(x_i, \theta) \times [P(y < X < z)]^m \times [P(X > z)]^p$$



Example 10.6 (censored data)

Claims (in £000s) on a particular policy have a distribution with PDF given by:

$$f(x) = 2cx e^{-cx^2} \quad x > 0$$

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 (and were passed onto a reinsurer). Find the maximum likelihood estimate of c .

Solution

We have 7 known claims and 3 claims greater than 6. So the likelihood is:

$$L(c) = \prod_{i=1}^7 f(x_i) \times [P(X > 6)]^3$$

$$\text{Since } P(X > 6) = \int_6^\infty 2cx e^{-cx^2} dx = \left[-e^{-cx^2} \right]_6^\infty = e^{-c \times 6^2} \text{ and } \sum_{i=1}^7 x_i^2 = 49.91$$

$$\begin{aligned} L(c) &= \prod_{i=1}^7 2cx_i e^{-cx_i^2} \times \left[e^{-c \times 6^2} \right]^3 \\ &= \text{constant} \times c^7 e^{-c \sum_{i=1}^7 x_i^2} \times e^{-108c} \\ &= \text{constant} \times c^7 e^{-157.91c} \end{aligned}$$

The log-likelihood is:

$$\ln L(c) = \text{constant} + 7 \ln c - 157.91c$$

Differentiating and setting equal to zero gives:

$$\frac{d}{dc} \ln L(c) = \frac{7}{c} - 157.91 = 0 \Rightarrow \hat{c} = 0.0443$$

Differentiating again to check we get a maximum:

$$\frac{d^2}{dc^2} \ln L(c) = -\frac{7}{c^2} < 0 \Rightarrow \text{max}$$


Question 10.8 (censored data)

Waiting times in a post office queue have an $Exp(\lambda)$ 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 find the maximum likelihood estimate of λ .


Example 10.7 (truncated data)

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 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 look at 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	≥ 3
$P(N = n)$	$\frac{3\theta}{1 - 5\theta}$	$\frac{\theta}{1 - 5\theta}$

The likelihood is:

$$L(\theta) = \text{constant} \times [P(N=1)]^{60} \times [P(N=2)]^{24} \times [P(N \geq 3)]^{16}$$

So:

$$\begin{aligned} L(\theta) &= \text{constant} \times \left(\frac{3\theta}{1-5\theta} \right)^{60} \times \left(\frac{\theta}{1-5\theta} \right)^{24} \times \left(\frac{1-9\theta}{1-5\theta} \right)^{16} \\ &= \text{constant} \times \frac{\theta^{84}(1-9\theta)^{16}}{(1-5\theta)^{100}} \end{aligned}$$

The constant arises from the fact that we don't know which 60 policies had 1 claim, etc and so there is some combinatorial factor to account for this.

The log-likelihood is:

$$\ln L(\theta) = \text{constant} + 84 \ln \theta + 16 \ln(1-9\theta) - 100 \ln(1-5\theta)$$

Differentiating and setting equal to zero gives:

$$\begin{aligned} \frac{d}{d\theta} \ln L(\theta) &= \frac{84}{\theta} - \frac{9 \times 16}{1-9\theta} + \frac{5 \times 100}{1-5\theta} = 0 \\ \Rightarrow 84(1-9\theta)(1-5\theta) - 144\theta(1-5\theta) + 500\theta(1-9\theta) &= 0 \\ \Rightarrow 84 - 820\theta &= 0 \\ \Rightarrow \hat{\theta} &= 0.102 \end{aligned}$$

Differentiating again to check we get a maximum:

$$\frac{d^2}{d\theta^2} \ln L(\theta) = -\frac{84}{\theta^2} - \frac{9 \times 9 \times 16}{(1-9\theta)^2} + \frac{5 \times 5 \times 100}{(1-5\theta)^2} < 0 \text{ when } \theta = 0.102 \Rightarrow \text{max}$$


Question 10.9 (truncated data)

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:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})} \quad x = 1, 2, 3, \dots$$

- (ii) Hence show that both the method of moments estimator 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.
- (iii) Solve this equation, by any means, for the given data and calculate the resulting estimate of λ to two decimal places.
- (iv) Hence, estimate the percentage of all policies with no claims during the year.

2.5 Independent samples

For independent samples from two populations which share a common parameter, the overall likelihood is the product of the two separate likelihoods.



Example 10.8

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. Find the maximum likelihood estimate of μ based on these data.

Solution

The likelihood for these 15 low-risk and 10 high-risk policies is:

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{15} P(X = x_i) \times \prod_{j=1}^{10} P(Y = y_j) = \prod_{i=1}^{15} \frac{\mu^{x_i}}{x_i!} e^{-\mu} \times \prod_{j=1}^{10} \frac{(2\mu)^{y_j}}{y_j!} e^{-2\mu} \\ &= \text{constant} \times \mu^{\sum_{i=1}^{15} x_i} e^{-15\mu} \times \mu^{\sum_{j=1}^{10} y_j} e^{-20\mu} \\ &= \text{constant} \times \mu^{48} e^{-15\mu} \times \mu^{59} e^{-20\mu} \\ &= \text{constant} \times \mu^{107} e^{-35\mu} \end{aligned}$$

The log-likelihood is:

$$\ln L(\mu) = \text{constant} + 107 \ln \mu - 35\mu$$

Differentiating and setting equal to zero gives:

$$\frac{d}{d\mu} \ln L(\mu) = \frac{107}{\mu} - 35 = 0 \Rightarrow \hat{\mu} = 3.057$$

Differentiating again to check we get a maximum:

$$\frac{d^2}{d\mu^2} \ln L(\mu) = -\frac{107}{\mu^2} < 0 \Rightarrow \text{max}$$

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 $\underline{X} = (X_1, X_2, \dots, X_n)$ 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.

Example 10.9

Show that the estimator for μ obtained in Question 10.6 is unbiased.

Solution

In Question 10.6, we had a $Poi(\mu)$ distribution and our estimator was $\hat{\mu} = \bar{X}$. To show that this is unbiased we need to show that $E(\hat{\mu}) = \mu$, ie $E(\bar{X}) = \mu$.

This proof is covered in Chapter 9, but here it is again:

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i)$$

Since $X_i \sim Poi(\mu)$ we have $E(X_i) = \mu$. Hence:

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$

So the estimator $\hat{\mu} = \bar{X}$ is unbiased as it gives the true value of μ on average.

If an estimator is biased, its bias is given by $E[g(\underline{X})] - \theta$.

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 10.10

The following are estimators for the variance of a distribution. Obtain the bias for each estimator:

$$(i) \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$(ii) \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

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 it 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:

$$\text{MSE}(g(\underline{X})) = E[(g(\underline{X}) - \theta)^2]$$

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:

$$\text{MSE} = \text{Variance} + \text{bias}^2$$

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 .)

$$\begin{aligned}\text{MSE}(g) &= E[(g - \theta)^2] \\ &= E[\{(g - E[g]) + (E[g] - \theta)\}^2] \\ &= E[(g - E[g])^2] + 2(E[g] - \theta)E[g - E[g]] + [E[g] - \theta]^2 \\ &= \text{var}[g] + 0 + \text{bias}^2[g] \text{ as required}\end{aligned}$$

Note: If the estimator $g(\underline{X})$ is unbiased, then $\text{MSE} = \text{variance}$.



Example 10.10

Obtain MSE of the estimator for μ obtained in Question 10.6.

Solution

In Question 10.6, we had a $Poi(\mu)$ distribution and our estimator was $\hat{\mu} = \bar{X}$. The MSE is given by:

$$MSE(\hat{\mu}) = \text{var}(\hat{\mu}) + \text{bias}^2(\hat{\mu})$$

In Example 10.9 we showed that the estimator was unbiased, ie $\text{bias}(\hat{\mu}) = 0$. So:

$$MSE(\hat{\mu}) = \text{var}(\hat{\mu}) + 0^2 = \text{var}(\hat{\mu}) = \text{var}(\bar{X})$$

This proof is covered in Chapter 9, but here it is again:

$$\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) \quad \text{since } X_i \text{ are independent}$$

As $X_i \sim Poi(\mu)$ we have $\text{var}(X_i) = \mu$. Hence:

$$\text{var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \mu = \frac{1}{n^2} \times n\mu = \frac{\mu}{n}$$

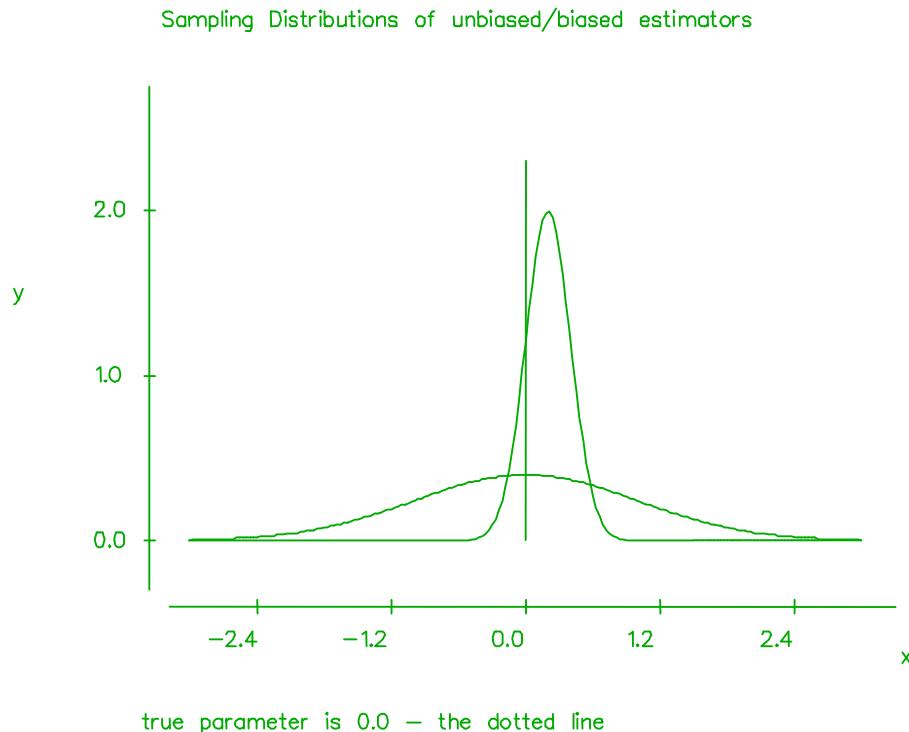
So the MSE is $\frac{\mu}{n}$.



Question 10.11

Determine the mean square error of $\hat{\mu} = \bar{X}$ which is used to estimate the mean of a $N(\mu, \sigma^2)$ distribution based on a random sample of n observations.

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.



The “dotted line” refers to the vertical line in the diagram.

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 $\text{MSE} \rightarrow 0$ as $n \rightarrow \infty$. This property is known as consistency.



Question 10.12

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:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- (i) Determine the mean square error of $\hat{\sigma}^2$.
- (ii) Is $\hat{\sigma}^2$ 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:

$$\hat{\theta} \stackrel{d}{\sim} N(\theta, \text{CRLB})$$

where $\text{CRLB} = \frac{1}{nE\left\{\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right\}}$.

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 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 Chapter 11.

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:

$$\text{CRLB} = \frac{1}{E\left\{\left[\frac{\partial}{\partial\theta}\log L(\theta, \underline{X})\right]^2\right\}} \quad \text{and} \quad \text{CRLB} = \frac{1}{-E\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta, \underline{X})\right]}$$

The second formula is normally easier to work (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*.



Example 10.11

Find the CRLB for estimators of μ , for a sample X_1, \dots, X_n from a $Poi(\mu)$ 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 $\hat{\mu} = \bar{X}$.

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:

$$CRLB = -1 \left/ E \left[\frac{d^2}{d\mu^2} \ln L(\mu) \right] \right. = \frac{\mu}{n}$$

In fact, in this case, the maximum likelihood estimator $\hat{\mu} = \bar{X}$ is unbiased and has variance μ/n (see previous example). So, the CRLB *can* be attained by the variance.



Question 10.13

- (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 $\frac{\sigma^2}{n}$.

Hint: Use the results from 0.

- (ii) Show that the variance of the maximum likelihood estimator $\hat{\mu} = \bar{X}$ attains the CRLB.

What follows now is a rather messy 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:

$$CRLB(\sigma) = -\frac{1}{E \left[\frac{d^2}{d\sigma^2} \ln L(\sigma) \right]}$$

whereas the formula for the CRLB of $v = \sigma^2$ is:

$$CRLB(v) = -\frac{1}{E \left[\frac{d^2}{dv^2} \ln L(v) \right]}$$

There will be no simple connection between the derivatives.



Example 10.12

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:

$$\frac{\partial^2}{\partial v^2} \ln L(v) = \frac{n}{2v^2} - \frac{1}{v^3} \sum_{i=1}^n (X_i - \mu)^2 \text{ ie } \frac{n}{2v^2} - \frac{1}{v^2} \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$$

We need to determine the expectation of this. We will use the fact that $X_i \sim N(\mu, \sigma^2)$, so $Z_i = \left(\frac{X_i - \mu}{\sigma}\right) \sim N(0,1)$ and hence:

$$E(Z_i^2) = \text{var}(Z_i) + E^2(Z_i) = 1 + 0^2 = 1$$

So we have:

$$\begin{aligned}
 E\left[\frac{\partial^2}{\partial v^2} \ln L(v)\right] &= \frac{n}{2v^2} - \frac{1}{v^2} \sum_{i=1}^n E\left[\left(\frac{X_i - \mu}{\sigma}\right)^2\right] \\
 &= \frac{n}{2v^2} - \frac{1}{v^2} \sum_{i=1}^n E[Z_i^2] \\
 &= \frac{n}{2v^2} - \frac{1}{v^2} \sum_{i=1}^n 1 \\
 &= \frac{n}{2v^2} - \frac{n}{v^2} = -\frac{n}{2v^2}
 \end{aligned}$$

Hence:

$$CRLB = -1 \left/ E\left[\frac{\partial^2}{\partial v^2} \log L\right]\right. = \frac{2v^2}{n} = \frac{2\sigma^4}{n}$$



Question 10.14

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.

The square root of the variance of an estimator, $\hat{\theta}$, is known as the standard error of $\hat{\theta}$ (which is denoted by $SE(\hat{\theta})$).

6 Final remarks on estimation

We now compare the method of moments and the method of maximum likelihood.

Essentially maximum likelihood is regarded as the better method.

However, there is perhaps an obvious restriction that flaws the method of moments. In the usual one-parameter case the method of moments estimator is always a function of the sample mean \bar{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 $2\bar{X}$ 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 $\bar{x} = 2.24$. Since the method of moments estimator is $\hat{\theta} = 2\bar{X}$, 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 Exam-type questions

The next question covers Sections 3 and 4.



Past Exam Question (Subject C1, April 1997, Q14)

Suppose that unbiased estimators X_1 and X_2 of a parameter θ have been determined by two 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) Find the relationship satisfied by α and β so that Y is also an unbiased estimator of θ .
- (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.

This question covers Sections 1 and 2. It is deliberately unusual, like many of the exam questions, to test your ability to *apply* the principles of estimation.



Exam-type question

A random sample x_1, x_2, \dots, x_n is taken from a population, which has the probability distribution 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} .

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} .
- (iii) Show that the probability density function for the distribution of X_{MIN} , is:

$$f_{X_{MIN}}(x) = \frac{n\alpha}{(1+x)^{n\alpha+1}} \quad x \geq 0$$

- (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 .
- (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.
- (vi) What further information would you need here in order to obtain a method of moments estimate of α ?

8 Appendix – second-order condition for maxima

In Section 2.2, we discussed maximum likelihood estimation in the two-parameter case. We noted that the second-order condition that we use to check for maxima involves the matrix of second derivatives, or Hessian matrix. In the two-parameter case, the Hessian matrix, H , is of the form:

$$H = \begin{pmatrix} \frac{\partial^2 \ln L}{\partial \theta_1^2} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_2^2} \end{pmatrix}$$

The estimates are maxima if the Hessian matrix is negative definite, *i.e.* if:

$$\underline{a}^T H \underline{a} < 0$$

for all non-zero vectors \underline{a} .

An equivalent condition is that all the eigenvalues of H are negative. λ is said to be an eigenvalue of an $n \times n$ matrix H if there is a non-zero vector \underline{x} , such that:

$$H\underline{x} = \lambda \underline{x}$$

The vector \underline{x} is known as an eigenvector. This equation is equivalent to:

$$(H - \lambda I) \underline{x} = 0$$

where I is the identity matrix. The equation $\det(H - \lambda I) = 0$ can be solved for λ to calculate the eigenvalues.

**Example**

Calculate the eigenvalues of the matrix $\begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$.

Solution

We have to solve:

$$\det\begin{pmatrix} 2-\lambda & 1 \\ 4 & 2-\lambda \end{pmatrix} = 0$$

Recall that the determinant of a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $ad - bc$. This gives:

$$(2-\lambda)^2 - 4 = \lambda^2 - 4\lambda = \lambda(\lambda - 4) = 0$$

Thus the eigenvalues are 0 and 4.

This page has been left blank so that you can keep the chapter summaries together for revision purposes.



Chapter 10 Summary

We have covered two methods here for estimating parameters.

The method of moments technique equates the population moments to the sample moments using the formulae detailed overleaf.

The method of maximum likelihood:

- find the likelihood $L(\theta) = \prod_{i=1}^n f(x_i; \theta)$
- $\log L$
- find θ that solves $\frac{\partial}{\partial \theta} \ln L(\theta) = 0$
- check for maximum $\frac{\partial^2}{\partial \theta^2} \ln L(\theta) < 0$.

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 $E[g(\underline{X})] = \theta$.

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 $\text{var}[g(\underline{X})] + \text{bias}^2[g(\underline{X})]$. An estimator is consistent if $MSE \rightarrow 0$ as $n \rightarrow \infty$, 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 found using the formulae overleaf. It can be used to obtain confidence intervals.

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.



Chapter 10 Formulae

Method of moments

$$1 \text{ parameter} \quad E(X) = \frac{1}{n} \sum_{i=1}^n X_i$$

$$2 \text{ parameters} \quad E(X) = \frac{1}{n} \sum_{i=1}^n X_i \quad E(X^2) = \frac{1}{n} \sum_{i=1}^n X_i^2 \quad \text{or} \quad \text{var}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\text{alternatively} \quad E(X) = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{var}(X) = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Method of maximum likelihood

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

$$\theta \text{ that solves } \frac{\partial}{\partial \theta} \ln L(\theta) = 0$$

Bias

$$\text{bias}[g(\underline{X})] = E[g(\underline{X})] - \theta$$

Mean square error

$$\text{MSE}[g(\underline{X})] = E[(g(\underline{X}) - \theta)^2] = \text{var}[g(\underline{X})] + \text{bias}^2[g(\underline{X})]$$

Cramér-Rao lower bound

$$\text{CRLB}(\theta) = -\frac{1}{E\left[\frac{\partial^2}{\partial \theta^2} \ln L(\theta, \underline{X})\right]}$$

Asymptotic distribution of MLE

$$\hat{\theta} \stackrel{d}{\sim} N(\theta, \text{CRLB})$$

Chapter 10 Solutions

Solution 10.1

The population mean for a $\text{Poisson}(\mu)$ from page 7 of the *Tables* is just μ .

The sample mean is $\bar{x} = \frac{30}{10} = 3$.

Equating population mean to sample mean gives:

$$\mu = 3$$

Since this is an estimate of the true value of μ we write $\hat{\mu} = 3$.

Solution 10.2

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:

$$\begin{aligned} \text{var}(X) &= \frac{1}{n} \sum x_i^2 - E^2(X) \\ \Rightarrow \text{var}(X) + E^2(X) &= \frac{1}{n} \sum x_i^2 \end{aligned}$$

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.

Solution 10.3

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 $E(X) = \frac{k(1-p)}{p}$. The sample mean is $\bar{x} = \frac{15}{10} = 1.5$. Equating these gives:

$$\frac{\hat{k}(1-\hat{p})}{\hat{p}} = 1.5 \quad (1)$$

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 $\frac{1}{n} \sum x_i^2 = \frac{43}{10} = 4.3$. Equating these gives:

$$\frac{\hat{k}(1-\hat{p})}{\hat{p}^2} + \left(\frac{\hat{k}(1-\hat{p})}{\hat{p}} \right)^2 = 4.3 \quad (2)$$

Substituting equation (1) into (2) gives:

$$\frac{1.5}{\hat{p}} + 1.5^2 = 4.3 \Rightarrow \hat{p} = 0.7317$$

Hence, equation (1) gives $\hat{k} = 4.091$.

Alternatively, if we use the second of the second-order equations, we would get $\text{var}(X) = \frac{\hat{k}(1-\hat{p})}{\hat{p}^2}$ and $\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{43}{10} - 1.5^2 = 2.05$. Equating these gives:

$$\frac{\hat{k}(1-\hat{p})}{\hat{p}^2} = 2.05 \quad (3)$$

Substituting equation (1) into (3) gives:

$$\frac{1.5}{\hat{p}} = 2.05 \Rightarrow \hat{p} = 0.7317$$

And hence $\hat{k} = 4.091$ as before.

Solution 10.4

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:

$$\hat{\lambda} = \frac{10}{25} = 0.4 \Rightarrow \hat{\alpha} = 10 \times 0.4 = 4$$

Solution 10.5

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:

$$\hat{\mu} = 126.6$$

$$\hat{\mu}^2 + \hat{\sigma}^2 = 16,041.8 \Rightarrow \hat{\sigma}^2 = 14.24$$

Alternatively, using $\bar{x} = 126.6$ and $s^2 = \frac{1}{4} \{80,209 - 5 \times 126.6^2\} = 17.8$ and equating these to the population moments of $E(X) = \mu$ and $\text{var}(X) = \sigma^2$ gives:

$$\hat{\mu} = \bar{x} = 126.6 \quad \text{and} \quad \hat{\sigma}^2 = s^2 = 17.8$$

Solution 10.6

- (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:

$$\hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

Differentiating again (to check that it is a maximum):

$$\frac{d^2}{d\mu^2} \ln L(\mu) = -\frac{\sum x_i}{\mu^2} < 0 \Rightarrow \max$$

So the estimate (the value obtained for a particular sample) is \bar{x} . But the estimator (the distribution of the estimates obtained) is \bar{X} . We simply use capital letters to show that we're talking about the distribution.

- (ii) We have $n = 10$ and $\sum x_i = 24$. Hence the estimate is simply:

$$\hat{\mu} = \bar{x} = \frac{24}{10} = 2.4$$

Solution 10.7

The formula for the mean θ (say) of a lognormal distribution is:

$$\theta = e^{\mu + \frac{1}{2}\sigma^2}$$

The invariance property tells us that the MLEs of θ, μ , 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:

$$\hat{\theta} = e^{2 + \frac{1}{2} \times 0.25} = 8.37$$

Solution 10.8

Using the likelihood formula given for censored data in Section 2.3:

$$L(\lambda) = \left[\prod_{i=1}^{10} f(x_i) \right] \times [P(X > 4)]^6 = \lambda^{10} e^{-\lambda \sum x_i} \times (e^{-4\lambda})^6$$

$$\text{since } f(x_i) = \lambda e^{-\lambda x_i} \text{ and } P(X > 4) = \int_4^\infty \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x} \right]_4^\infty = e^{-4\lambda}.$$

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 setting it equal to zero gives:

$$\frac{d}{d\lambda} \ln L(\lambda) = \frac{10}{\hat{\lambda}} - 42.3 = 0 \Rightarrow \hat{\lambda} = 0.2364$$

Checking that it's a maximum:

$$\frac{d^2}{d\lambda^2} \ln L(\lambda) = -\frac{10}{\lambda^2} < 0 \Rightarrow \text{max}$$

Solution 10.9

- (i) Since only policies with claims are included, we must use a truncated Poisson distribution:

$$P(X = x) = k \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 1, 2, 3, \dots$$

where k is the constant of proportionality to ensure that the sum of the probabilities is 1.

For the ordinary Poisson distribution:

$$\sum_x P(X = x) = 1 \Rightarrow \Rightarrow P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda}$$

So our probability function can be written as:

$$k \sum_{x=1}^{\infty} P(X = x) = 1 \Rightarrow k(1 - e^{-\lambda}) = 1 \Rightarrow k = \frac{1}{(1 - e^{-\lambda})}$$

- (ii) We will first use the method of moments technique, so we need the mean of the truncated Poisson distribution:

$$E[X] = \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})} = \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})} = \frac{1}{(1 - e^{-\lambda})} \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!}$$

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:

$$E[X] = \frac{1}{(1 - e^{-\lambda})} \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{(1 - e^{-\lambda})} \times \lambda = \frac{\lambda}{(1 - e^{-\lambda})}$$

So the method of moments equation is $\bar{x} = \frac{\lambda}{1 - e^{-\lambda}}$ or $\lambda = \bar{x}(1 - e^{-\lambda})$, as required.

Now using maximum likelihood, the likelihood function is:

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i! (1-e^{-\lambda})} = \text{constant} \times \frac{\lambda^{\sum x_i} e^{-n\lambda}}{(1-e^{-\lambda})^n}$$

where the constant incorporates the factorial factor.

Taking logs:

$$\log L = \text{constant} + (\sum x_i) \log \lambda - n\lambda - n \log(1 - e^{-\lambda})$$

Differentiating with respect to λ :

$$\begin{aligned} \frac{d}{d\lambda} \log L &= \frac{\sum x_i}{\lambda} - n - \frac{n e^{-\lambda}}{1 - e^{-\lambda}} = \frac{n \bar{x} (1 - e^{-\lambda}) - n \lambda (1 - e^{-\lambda}) - n \lambda e^{-\lambda}}{\lambda (1 - e^{-\lambda})} \\ &= \frac{n \bar{x} (1 - e^{-\lambda}) - n \lambda}{\lambda (1 - e^{-\lambda})} \end{aligned}$$

Equating to zero gives $\lambda = \bar{x}(1 - e^{-\lambda})$ as required.

(iii) From the data:

$$\bar{x} = \frac{174 \times 1 + 50 \times 2 + 10 \times 3 + 4 \times 4}{238} = \frac{320}{238}$$

$$\text{So } \frac{320}{238} (1 - e^{-\lambda}) = \lambda \text{ or } \frac{320}{238} = \frac{\lambda}{(1 - e^{-\lambda})}.$$

This equation must be solved by numerical methods. We can use trial and error with simple linear interpolation or a systematic method such as Newton-Raphson. The solution gives the estimate of λ as 0.62 to two decimal places.

(iv) Now $P(X = 0) = e^{-\lambda}$. By the invariance property, the maximum likelihood estimate of this probability is:

$$e^{-\hat{\lambda}} = e^{-0.62} = 0.538$$

So we estimate that 54% of policies have no claims.

Solution 10.10

- (i) The formula for the bias of S^2 is:

$$\text{bias}(S^2) = E(S^2) - \sigma^2$$

It was showed in Chapter 9 that $E(S^2) = \sigma^2$, but here is the proof again:

$$\begin{aligned} E(S^2) &= E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] = E\left[\frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)\right] \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right) \end{aligned}$$

Since:

$$E(X_i^2) = \text{var}(X_i) + E^2(X_i) = \sigma^2 + \mu^2$$

$$E(\bar{X}^2) = \text{var}(\bar{X}) + E^2(\bar{X}) = \frac{\sigma^2}{n} + \mu^2$$

We get:

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} \left(\sum_{i=1}^n (\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) \right) \\ &= \frac{1}{n-1} (n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) \\ &= \frac{1}{n-1} (n-1)\sigma^2 \\ &= \sigma^2 \end{aligned}$$

So the bias is:

$$\text{bias}(S^2) = E(S^2) - \sigma^2 = \sigma^2 - \sigma^2 = 0$$

(ii) Since $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ 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:

$$\text{bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{1}{n} \sigma^2$$

Solution 10.11

The MSE is given by:

$$\begin{aligned} \text{MSE}(\hat{\mu}) &= \text{var}(\hat{\mu}) + \text{bias}^2(\hat{\mu}) \\ &= \text{var}(\bar{X}) + \text{bias}^2(\bar{X}) \end{aligned}$$

where:

$$\text{bias}(\bar{X}) = E(\bar{X}) - \mu$$

When $X \sim N(\mu, \sigma^2)$ we have $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ so $E(\bar{X}) = \mu$ and $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$.

Hence:

$$\text{bias}(\bar{X}) = \mu - \mu = 0$$

Therefore:

$$\text{var}(\hat{\mu}) = \text{var}(\bar{X}) + 0^2$$

$$= \frac{\sigma^2}{n}$$

Alternatively, we could have proved that $E(\bar{X}) = \mu$ and $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$ in the same way that we did in Chapter 9 and in the examples in this chapter.

Solution 10.12

(i) Now:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{(n-1)}{n} S^2 \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \\ \Rightarrow \quad \frac{n\hat{\sigma}^2}{\sigma^2} &\sim \chi_{n-1}^2\end{aligned}$$

Hence the mean of $\hat{\sigma}^2$ is:

$$E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = E\left(\chi_{n-1}^2\right) = n-1 \Rightarrow E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$$

This gives a bias of:

$$\text{bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

We proved this result for any distribution in Question 10.10. However, it is only for the normal distribution that we can make use of the chi-square result.

The variance of $\hat{\sigma}^2$ is:

$$\text{var}\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = \text{var}\left(\chi_{n-1}^2\right) = 2(n-1) \Rightarrow \text{var}(\hat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4$$

Therefore the MSE of $\hat{\sigma}^2$ is given by:

$$\begin{aligned}\text{MSE}(\hat{\sigma}^2) &= \text{var}(\hat{\sigma}^2) + [\text{Bias}(\hat{\sigma}^2)]^2 \\ &= \frac{2(n-1)}{n^2}\sigma^4 + \left(-\frac{\sigma^2}{n}\right)^2 \\ &= \frac{2n-1}{n^2}\sigma^4\end{aligned}$$

(ii) Since the MSE, $\frac{2n-1}{n^2}\sigma^4$, tends to zero as $n \rightarrow \infty$, it is consistent.

Solution 10.13

(i) From 0 we had:

$$\frac{\partial}{\partial \mu} \ln L(\mu) = \frac{1}{2\sigma^2} \sum_{i=1}^n 2(X_i - \mu) = \frac{1}{\sigma^2} \left(\sum_{i=1}^n X_i - n\mu \right)$$

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, everything is a constant 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:

$$CRLB = -1 / E\left[\frac{\partial^2}{\partial \mu^2} \ln L(\mu) \right] = \frac{\sigma^2}{n}$$

(ii) From Chapter 9 we saw that if $X \sim N(\mu, \sigma^2)$ then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ so

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}. \text{ Hence the variance of the MLE attains the CRLB.}$$

Solution 10.14

- (i) Using the Core Reading example from Section 2.1, we have:

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^n X_i}$$

$$\Rightarrow \ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^n X_i$$

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 $\hat{\lambda} = \frac{1}{\bar{X}}$.

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, everything is a constant 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:

$$CRLB = -1 \Bigg/ E\left[\frac{d^2}{d\lambda^2} \ln L(\lambda)\right] = \frac{\lambda^2}{n}$$

It is not easy to show that $E(\hat{\lambda}) = E\left(\frac{1}{\bar{X}}\right)$ attains the CRLB as $E\left(\frac{1}{\bar{X}}\right) \neq \frac{1}{E(\bar{X})}$. So it has to be done from first principles.

- (ii) We are estimating the **mean** of an $Exp(\lambda)$ distribution, ie $\mu = \frac{1}{\lambda}$, therefore we need to work in terms of μ and differentiate with respect to μ .

The likelihood function for the sample is:

$$L(\mu) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \frac{1}{\mu^n} e^{-\sum X_i / \mu}$$

$$\Rightarrow \ln L(\mu) = -n \ln \mu - \frac{1}{\mu} \sum X_i$$

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 $E(X_i) = \mu = \frac{1}{\lambda}$, 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:

$$CRLB = -1 \left/ E\left[\frac{d^2}{d\mu^2} \log L\right]\right. = \frac{\mu^2}{n}$$

Comment

Although $\mu = \frac{1}{\lambda}$ we see that $CRLB(\mu) \neq \frac{1}{CRLB(\lambda)}$.

In fact we actually have $CRLB(\mu) = \frac{\mu^2}{n} = \frac{1}{n\lambda^2}$.

Past Exam Question (Subject C1, April 1997, Q14)

- (i) Since X_1 and X_2 are unbiased estimators of θ this means that:

$$E[X_1] = E[X_2] = \theta$$

$$\Rightarrow E(Y) = E(\alpha X_1 + \beta X_2) = \alpha E(X_1) + \beta E(X_2) = (\alpha + \beta)\theta$$

Hence, if Y is unbiased for θ , then $\alpha + \beta = 1$.

- (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:

$$\begin{aligned} \text{var}(Y) &= \text{var}(\alpha X_1 + \beta X_2) = \alpha^2 \text{var}(X_1) + \beta^2 \text{var}(X_2) \\ &= \sigma^2[\alpha^2 + (1-\alpha)^2\phi] \end{aligned}$$

To obtain the minimum, we set the derivative equal to zero:

$$\begin{aligned} \frac{d}{d\alpha} \text{var}(Y) &= \sigma^2[2\alpha - 2(1-\alpha)\phi] = 0 \\ \Rightarrow \alpha &= (1-\alpha)\phi \Rightarrow \alpha = \frac{\phi}{1+\phi} \end{aligned}$$

Checking it's a minimum:

$$\frac{d^2}{d\alpha^2} \text{var}(Y) = \sigma^2[2 + 2\phi] > 0 \Rightarrow \text{min}$$

So:

$$\begin{aligned} \text{var}(Y) &= \sigma^2 \left[\left(\frac{\phi}{1+\phi} \right)^2 + \left(1 - \frac{\phi}{1+\phi} \right)^2 \phi \right] \\ &= \sigma^2 \left[\frac{\phi^2}{(1+\phi)^2} + \frac{\phi}{(1+\phi)^2} \right] \\ &= \frac{\phi}{1+\phi} \sigma^2 \end{aligned}$$

Exam-type question

- (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 this is the distribution function of X_{MAX} .

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:

$$\begin{aligned} P(X_{MIN} \geq x) &= P(\text{all } X_i \geq x) = [1 - F(x)]^n \\ \Rightarrow F_{X_{MIN}}(x) &= 1 - [1 - F(x)]^n \end{aligned}$$

- (ii) The distribution function of X is given by:

$$F(x) = \int_0^x f(t) dt = \int_0^x \alpha(1+t)^{-\alpha-1} dt = \left[-(1+t)^{-\alpha} \right]_0^x = 1 - (1+x)^{-\alpha}$$

where $x \geq 0$.

$$\text{Hence } F_{X_{MAX}}(x) = [F(x)]^n = \left(1 - (1+x)^{-\alpha}\right)^n.$$

- (iii) Similarly $F_{X_{MIN}}(x) = 1 - [1 - F(x)]^n = 1 - \left[(1+x)^{-\alpha}\right]^n = 1 - (1+x)^{-n\alpha} \quad x \geq 0$

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:

$$f_{X_{MIN}}(x) = \frac{n\alpha}{(1+x)^{n\alpha+1}} \quad x \geq 0$$

(iv) The likelihood function for α , based on a *single value* of X_{MIN} , is:

$$\begin{aligned} L(\alpha) &= \frac{n\alpha}{(1+x)^{n\alpha+1}} \\ \Rightarrow \log L(\alpha) &= \log n + \log \alpha - (n\alpha + 1) \log(1+x) \\ \Rightarrow \frac{\partial}{\partial \alpha} \log L &= \frac{1}{\alpha} - n \log(1+x) = 0 \\ \Rightarrow \hat{\alpha} &= \frac{1}{n \log(1+x)} \end{aligned}$$

Substituting in $n = 25$ and $x = 23$, we get $\hat{\alpha} = 0.01259$.

(v) Applying the same approach to X_{MAX} , we have a likelihood function of:

$$\begin{aligned} L(\alpha) &= f_{X_{MAX}}(x) = n \left(1 - (1+x)^{-\alpha}\right)^{n-1} \alpha (1+x)^{-\alpha-1} \\ \Rightarrow \log L(\alpha) &= \log n + (n-1) \log \left[1 - (1+x)^{-\alpha}\right] + \log \alpha - (\alpha + 1) \log(1+x) \\ \Rightarrow \frac{\partial}{\partial \alpha} \log L &= \frac{1}{\alpha} + (n-1) \times \frac{(1+x)^{-\alpha} \log(1+x)}{1 - (1+x)^{-\alpha}} - \log(1+x) = 0 \end{aligned}$$

Substituting in $n = 25$ and $x = 770$ we get:

$$\frac{1}{\alpha} - \log 771 + 24 \times \frac{771^{-\alpha} \log 771}{1 - 771^{-\alpha}} = 0$$

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.

Chapter 11

Confidence intervals



Syllabus objectives

- (x) 1. Define in general terms a confidence interval for an unknown parameter of a distribution based on a random sample.
2. Derive a confidence interval for an unknown parameter using a given sampling distribution.
3. Calculate confidence intervals for the mean and the variance of a normal distribution.
4. Calculate confidence intervals for a binomial probability and a Poisson mean, including the use of the normal approximation in both cases.
5. Calculate confidence intervals for two-sample situations involving the normal distribution, and the binomial and Poisson distributions using the normal approximation.
6. Calculate confidence intervals for a difference between two means from paired data.

0 Introduction

In Chapter 10 we used the method of moments or the method of maximum likelihood to obtain estimates for the population parameter(s). For example, given the number of claims from a certain portfolio that we receive in a month:

Claims	0	1	2	3	4	5	6
Frequency	9	22	26	21	13	6	3

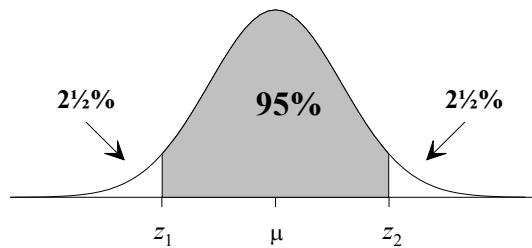
Assuming a Poisson distribution with parameter μ , our estimate of μ using the methods given in Chapter 10 would be $\mu = \bar{X} = 2.37$.

The natural question is “how confident are we that the value is correct?” Obviously, this will depend on a number of factors, including the sample size. This question will be addressed in this chapter, where we will construct “95% confidence intervals”, *ie* an interval based upon our estimate that has a 95% chance of containing the true parameter value. If this interval is wide we will not have that much confidence in our result, whereas if the interval is small we can be much more confident in our estimate.

We can also work backwards to see what sample size is required to get a confidence interval down to a certain width.

Confidence intervals will be constructed using the sampling distributions given in Chapter 9. For example, when sampling from a $N(\mu, \sigma^2)$ distribution where σ^2 is known:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



If we require a 95% confidence interval, then we can read off the 2½% and 97½% z -values from the normal tables of ± 1.96 and substitute these z -values into the equation, along with our values of \bar{X} , σ^2 and n . Rearranging this equation gives our confidence interval for μ of $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$.

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 $P(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})) = 1 - \alpha$.

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 $P(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})) = 0.95$ specifies $(\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X}))$ 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 .

With a known distribution you can write $P(g_1 < g(\underline{X}, \theta) < g_2) = 0.95$ where g_1, g_2 are such that:

$$\int_{g_1}^{g_2} f(t) dt = 0.95, \quad f(t) \text{ being the density of } g(\underline{X}, \theta)$$

g_1 and g_2 are usually constants.

If $g(\underline{X}, \theta)$ is monotonic increasing in θ , then:

$$g(\underline{X}, \theta) < g_2 \Leftrightarrow \theta < \theta_2 \quad \text{for some number } \theta_2$$

$$g_1 < g(\underline{X}, \theta) \Leftrightarrow \theta_1 < \theta \quad \text{for some number } \theta_1$$

and if $g(\underline{X}, \theta)$ is monotonic decreasing in θ , then:

$$g(\underline{X}, \theta) < g_2 \Leftrightarrow \theta < \theta_2$$

$$g_1 < g(\underline{X}, \theta) \Leftrightarrow \theta < \theta_1$$

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.

Example

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 $N(0,1)$.

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 you can write down the interval using the “ \pm ” notation, where the two values indicate the upper and lower limits. 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 11.1

Show that both this and the first interval given above are 95% confidence intervals. How long is each of these intervals?

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 11.2

The average IQ of a 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.

2.2 Confidence limits

The 95% confidence interval $\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$ 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.

This is an example of a two-sided symmetrical confidence interval. 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 $\pm 3\%$). 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 $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$.

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 11.3**

Guess how big a sample you would need to have a 99% confidence interval of width $\pm £1$. Now check your answer.

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:

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$$

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 the number such that the probability that a value from the t -distribution with $n - 1$ degrees of freedom is not less than this value is 0.025.

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.



Example 11.1

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:

$\frac{\bar{X} - \mu}{S/\sqrt{n}}$ has a t_{n-1} distribution, where S^2 is the sample variance.

From the *Tables* we find that $0.95 = P(-2.776 < t_4 < 2.776)$.

So $0.95 = P(-2.776 < \frac{\bar{X} - \mu}{S/\sqrt{n}} < 2.776)$.

Rearranging the inequality to isolate μ gives:

$$0.95 = P(\bar{X} - 2.776 S/\sqrt{n} < \mu < \bar{X} + 2.776 S/\sqrt{n})$$

Using the calculated values for the sample ($n = 5$, $\bar{X} = 126.6$, and $S^2 = 17.8$) gives:

$$0.95 = P(121.4 < \mu < 131.8)$$

So the confidence interval is (121.4, 131.8).



Question 11.4

An experiment was done 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. What assumptions did you have to make to find this confidence interval?

3.2 The variance

For the estimation of a normal variance σ^2 , there is again a pivotal quantity readily available:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

The resulting 95% confidence interval for the variance σ^2 is:

$$\left(\frac{(n-1)S^2}{\chi_{0.025,n-1}^2}, \frac{(n-1)S^2}{\chi_{0.975,n-1}^2} \right)$$

or for the standard deviation σ :

$$\left(\sqrt{\frac{(n-1)S^2}{\chi_{0.025,n-1}^2}}, \sqrt{\frac{(n-1)S^2}{\chi_{0.975,n-1}^2}} \right)$$

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.

**Question 11.5**

Calculate:

- (i) an equal-tailed 95% confidence interval and
- (ii) a 95% confidence interval of the form $\sigma < L$

for the standard deviation of the heights of the children in the population based on the information given in the last example.

4 Confidence intervals for binomial & Poisson

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:

$$P(h_1(\theta) < h(\underline{X}) < h_2(\theta)) \geq 0.95$$

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.

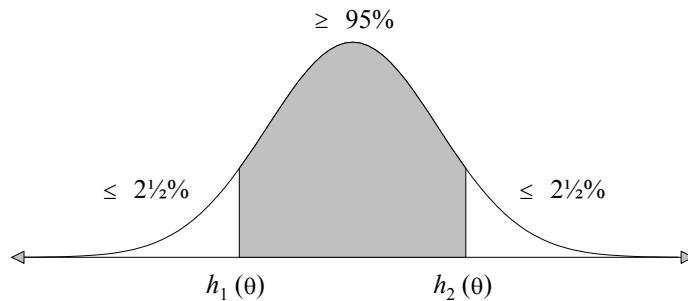
4.1 The binomial

If X is a single observation from $\text{binomial}(n, \theta)$, 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 $P(h_1(\theta) < X < h_2(\theta)) \geq 0.95$, where with equal tails $P(X \leq h_1(\theta)) \leq 0.025$ and $P(X \geq h_2(\theta)) \leq 0.025$.



You can have at most 2.5% in the lower (or upper) tail, so be very careful about finding the values of h_1 and h_2 .

Although explicit expressions for $h_1(\theta)$ and $h_2(\theta)$ cannot be written, clearly they do exist and are increasing (or at least non-decreasing) functions.

Example

For the binomial (20,0.3) case:

$$P(X \leq 1) = 0.0076 \text{ and } P(X \leq 2) = 0.0355 \quad \therefore h_1(\theta) = 1$$

Also:

$$P(X \geq 11) = 0.0171, P(X \geq 10) = 0.0480 \quad \therefore h_2(\theta) = 11$$

**Question 11.6**

For the binomial distribution with parameters $n = 20$, and $p = 0.4$, what are the values of h_1 and h_2 ?

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:

$$X \leq h_1(\theta) \Rightarrow \theta \geq \theta_1(X)$$

$$X \geq h_2(\theta) \Rightarrow \theta \leq \theta_2(X)$$

Now remember that these are the *tail* probabilities. So the inequalities involving θ_1 and θ_2 are defining the tails. So our confidence interval is the region *not* covered by these tail inequalities:

giving the 95% confidence interval $\theta_2(X) < \theta < \theta_1(X)$.

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 example 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 $X \sim \text{Bin}(n, \theta)$.

These can be expressed in terms of the distribution function $F(x; \theta)$:

$$1 - F(x-1; \theta_1) = 0.025 \quad \text{and} \quad F(x; \theta_2) = 0.025$$

Note: Equality can be attained as θ has a continuous range $(0, 1)$ and the “discrete” problem does not arise.



Example 11.2

We have obtained a value of 1 from the binomial distribution with parameters $n = 20$ and $p = \theta$. What is a 95% symmetrical confidence interval for θ ?

Solution

We need θ_1 such that $P(X \leq 1) = 0.025$ under binomial $(20, \theta_1)$, and θ_2 such that $P(X \geq 1) = 0.025$ under binomial $(20, \theta_2)$.

For the first equation, we have $(1 - \theta_1)^{20} + 20(1 - \theta_1)^{19} \theta_1 = 0.025$.

Solving this we obtain $\theta_1 = 0.249$. You will have to use a numerical method here. One approach would be to write the equation in the form $(1 - \theta_1)^{19}(1 + 19\theta_1) = 0.025$, then

iterate using $\theta_1 = 1 - \left(\frac{0.025}{1 + 19\theta_1} \right)^{\frac{1}{19}}$, 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 $0.00127 \leq \theta \leq 0.249$.



Question 11.7

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 a 95% confidence interval for the underlying proportion in favour of this idea.

The normal approximation

If n is “large”, confidence intervals for the binomial parameter θ are obtained using the normal approximation to the binomial distribution.

$\frac{X - n\theta}{\sqrt{n\theta(1 - \theta)}}$ can be used as a pivotal quantity.

But isolating the θ would be messy.

However $\frac{X - n\hat{\theta}}{\sqrt{n\hat{\theta}(1 - \hat{\theta})}}$, with $\hat{\theta}$ in place of θ , can be used in a simpler way and yields the standard 95% confidence interval used in practice, namely:

$$\frac{X \pm 1.96\sqrt{n\hat{\theta}(1 - \hat{\theta})}}{n}$$

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}, \text{ where } \hat{\theta} = \frac{X}{n}$$

**Example 11.3**

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 $\frac{X - nq}{\sqrt{nq(1-q)}} \stackrel{d}{\sim} N(0,1)$.

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:

$$P\left(\frac{X}{n} - 1.6449\sqrt{\frac{q(1-q)}{250}} < q < \frac{X}{n} + 1.6449\sqrt{\frac{q(1-q)}{250}}\right) = 0.90$$

From the question, we know that $X = 45$, and we can estimate q to be $\frac{45}{250}$.

Therefore a symmetrical 90% confidence interval for q is $0.140 < q < 0.220$.

**Question 11.8**

Repeat Example 11.2 using the normal approximation. Comment on your answer.

**Question 11.9**

An opinion poll of 1,000 voters found that 450 favoured Party P. Calculate a 99% confidence interval for the proportion of voters who favour Party P. Is it likely that more than 50% of the voters would vote for Party P in an election?

4.2 The Poisson

The Poisson situation can be tackled in a very similar way to the binomial for both large and small sample sizes.

If $X_i, i = 1, 2, \dots, n$ are independent Poisson(λ) variables, that is, a random sample of size n from Poisson (λ), then $\sum X_i \sim \text{Poisson}(n\lambda)$.

Using $\sum X_i$ as a single observation from Poisson($n\lambda$) is equivalent to the random sample of size n from Poisson(λ). This is similar to the single binomial situation.

Recall that a $\text{Bin}(n, p)$ distribution arises from the sum of n Bernoulli trials with probability p .

Given a single observation X from a Poisson(λ) distribution, then $P(h_1(\lambda) < X < h_2(\lambda)) \geq 0.95$, 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:

$$\sum_{r=x}^{\infty} p(r; \lambda_1) = 0.025 \text{ and } \sum_{r=0}^x p(r; \lambda_2) = 0.025$$

or:

$$1 - F(x-1; \lambda_1) = 0.025 \text{ and } F(x; \lambda_2) = 0.025$$

**Example 11.4**

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 $1 - e^{-\lambda_1} = 0.05 \Rightarrow e^{-\lambda_1} = 0.95$, 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 $\lambda = \log\left(\frac{1+\lambda}{0.05}\right)$, we obtain $\lambda_2 = 4.74$.

Therefore a symmetrical 90% confidence interval for λ is $0.0513 \leq \lambda \leq 4.74$. Not surprisingly this is very wide, since we only have 1 sample value.

The normal approximation

A normal approximation can be used either for a large sample from Poisson(λ), or for a single observation from Poisson(λ) where λ is large. Either from $\sum X_i \sim \text{Poisson}(n\lambda) \rightarrow N(n\lambda, n\lambda)$ or from the Central Limit Theorem as $\bar{X} \rightarrow N\left(\lambda, \frac{\lambda}{n}\right)$.

$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}}$ 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 $\frac{\bar{X} - \lambda}{\sqrt{\hat{\lambda}/n}}$ where $\hat{\lambda} = \bar{X}$.

This clearly gives $\bar{X} \pm 1.96 \sqrt{\frac{\bar{X}}{n}}$ as an approximate 95% confidence interval for λ .



Question 11.10

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 .

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:

$$\text{var}[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

when the samples are independent.

If the samples are not independent, then a covariance term will be included:

$$\text{var}[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\text{cov}[\bar{X}_1, \bar{X}_2]$$

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:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

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:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

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 or in Chapter 12, Section 4.2.

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 variance. S_1^2 and S_2^2 are the sample variances calculated in the usual way.



Question 11.11

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

Obtain 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.

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.

$$\text{It is } \frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}.$$

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 12, Section 4.2.

We'll also be using this result when we study analysis of variance in Chapter 14.



Question 11.12

For the fuel additive data in Question 11.11 obtain a 90% confidence interval for the ratio $\frac{\sigma_C^2}{\sigma_A^2}$ of the variances of the fuel consumption distributions with and without the additive and comment on the equality of variances assumption needed for the analysis in that question.

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 binomial (n_1, θ_1) and binomial (n_2, θ_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:

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}} \sim N(0, 1)$$

where $\hat{\theta}_1, \hat{\theta}_2$ are the MLEs $\frac{X_1}{n_1}, \frac{X_2}{n_2}$, respectively.



Question 11.13

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, calculate a symmetrical 95% confidence interval for the difference between male and female mortality rates at this age.

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 is an estimator of λ_i 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.



Question 11.14

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 Poisson distributions, calculate a 99% confidence interval for the difference between the two Poisson parameters.

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:

$$(X_{11}, X_{21}), (X_{12}, X_{22}), (X_{13}, X_{23}), \dots, (X_{1n}, X_{2n})$$

rather than two separate samples:

$$(X_{11}, X_{12}, X_{13}, \dots, X_{1n}) \text{ and } (X_{21}, X_{22}, X_{23}, \dots, X_{2n})$$

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 would 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 $\bar{D} \pm t_{0.025, n-1} \frac{S_D}{\sqrt{n}}$

**Question 11.15**

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. Do you think the diet is effective in reducing the patients' blood pressure? You are given that $\sum (B_i - A_i)^2 = 68$.

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.

**Question 11.16**

Should we have used a paired test in Question 11.11?

7 Past exam questions



Past Exam Question (Subject C1, September 1994, Q14)

- (i) 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):

A	102	98	93	86	92	94	89	97
B	86	88	92	95	98	97	94	92

- (a) Make an informative plot of these data.
 - (b) Comment on an 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.
- (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:

		<i>Property</i>							
		1	2	3	4	5	6	7	8
A	1	94	98	102	132	118	121	106	123
	2	92	96	111	129	111	122	101	118

- (a) Make an informative plot of these data.
- (b) Calculate a 90% confidence interval for the mean difference between valuations by A and B, and comment briefly on the result.

This next question applies confidence intervals to the results of Chapter 10, Section 5.



Past Exam Question (Subject C1, Specimen 1993, Q16)

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:

$$f(X; \lambda) = \lambda e^{-\lambda x} \quad x > 0$$

- (i) (a) Determine the maximum likelihood estimator $\hat{\lambda}$ of λ .
- (b) Obtain the large-sample approximate variance of $\hat{\lambda}$.
- (c) Hence calculate an approximate 95% confidence interval for λ .

- (ii) Using the fact that $2\lambda n \bar{X}$ has a χ^2_{2n} distribution, obtain an exact 95% confidence interval for λ , and comment briefly on how it compares with your interval in (i)(c).

8 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 10 and 11.
2. Attempt some of the questions in Part 3 of the **Question and Answer Bank**. 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. Students have said:

“I use the Revision Notes as a key part of my study. I find them incredibly useful for key reading questions and for studying topic by topic.”

“Grouping of past exam questions by topic - it takes ages to do this yourself.”

ASET – This contains past exam papers with detailed solutions and explanations, plus lots of comments about exam technique. Students have 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 on our website at www.ActEd.co.uk.

Buy online at www.ActEd.co.uk/estore



Chapter 11 Summary

Confidence intervals give us an interval in which we believe the true parameter value lies, together with an associated probability.

General confidence intervals for one or two samples can be found, using the pivotal method, using the formulae given overleaf. Some of these are given in the *Tables* on pages 22 and 23.

The confidence interval for two normal means (unknown variances) requires that the variances are the same and uses the pooled sample variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

We can test the equality of the variances by examining a dot plot, or more formally by using a confidence interval or a hypothesis test on σ_1^2/σ_2^2 .

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 11 Formulae

One-sample normal distribution

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \sigma^2 \text{ known} \quad \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1} \quad \sigma^2 \text{ unknown}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Two-sample normal distribution

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1) \quad \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

σ^2 known σ^2 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

$$\frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}} \stackrel{d}{\sim} N(0,1) \quad \text{or} \quad \frac{X - np}{\sqrt{np\hat{q}}} \stackrel{d}{\sim} N(0,1)$$

Two-sample binomial

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \stackrel{d}{\sim} N(0,1) \quad \text{where } \hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$$

One-sample Poisson

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/n}} \stackrel{d}{\sim} N(0,1) \quad \text{or} \quad \frac{\sum X - n\lambda}{\sqrt{n\hat{\lambda}}} \stackrel{d}{\sim} N(0,1)$$

Two-sample Poisson

$$\frac{(\hat{\lambda}_1 - \hat{\lambda}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\hat{\lambda}_1}{n_1} + \frac{\hat{\lambda}_2}{n_2}}} \stackrel{d}{\sim} N(0,1) \quad \text{where } \hat{\lambda}_1 = \bar{X}_1, \hat{\lambda}_2 = \bar{X}_2$$

Chapter 11 Solutions

Solution 11.1

For this one:

$$\begin{aligned}
 P(-1.8808 < Z < 2.0537) &= P(Z < 2.0537) - P(Z < -1.8808) \\
 &= P(Z < 2.0537) - (1 - P(Z < 1.8808)) \\
 &= 0.98000 - (1 - 0.97000) = 0.95
 \end{aligned}$$

This interval has length $3.9345 \frac{\sigma}{\sqrt{n}}$.

For the first one:

$$\begin{aligned}
 P(-1.96 < Z < 1.96) &= P(Z < 1.96) - P(Z < -1.96) \\
 &= 2 \times P(Z < 1.96) - 1 \\
 &= 2 \times 0.975 - 1 \\
 &= 0.95
 \end{aligned}$$

This interval has length $3.92 \frac{\sigma}{\sqrt{n}}$.

Solution 11.2

Since the sample comes from a normal distribution, we know that the quantity $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, 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 $126.5 < \mu < 137.5$.

Solution 11.3

I don't know what you guessed, but it might have been too small.

The answer can be calculated from the equation:

$$2.576 \times \frac{45}{\sqrt{n}} = 1 \Rightarrow n = 13,438$$

Note that the figure of 2.576 (rounded from 2.5758) can be found on page 162 in the *Tables*.

Solution 11.4

The sample mean and variance are 6.7 and 2.9 respectively.

So the confidence interval is given by $6.7 \pm t_9 \sqrt{\frac{2.9}{10}}$.

From the tables with $\alpha = 0.025$, $t_9 = 2.262$, so our confidence interval is $(5.48, 7.92)$.

We have assumed that the numbers of hours spent watching television have a normal distribution.

Solution 11.5

Since the sample comes from a normal distribution, we know that the quantity $\frac{4S^2}{\sigma^2}$, which equals $\frac{71.2}{\sigma^2}$ for this sample, has a χ_4^2 distribution.

- (i) From the *Tables* we find that:

$$0.95 = P(0.4844 < \chi_4^2 < 11.14)$$

So:

$$\begin{aligned} 0.95 &= P\left(0.4844 < \frac{4S^2}{\sigma^2} < 11.14\right) \\ &= P\left[\frac{71.2}{11.14} < \sigma^2 < \frac{71.2}{0.4844}\right] \\ &= P(6.39 < \sigma^2 < 147.0) \\ &= P(2.53 < \sigma < 12.1) \end{aligned}$$

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:

$$\begin{aligned} 0.95 &= P\left(0.7107 < \frac{4S^2}{\sigma^2}\right) \\ &= P\left[\sigma^2 < \frac{71.2}{0.7107}\right] \\ &= P(\sigma^2 < 100.2) \\ &= P(\sigma < 10.0) \end{aligned}$$

So, a one-sided 95% confidence interval for the standard deviation is $\sigma < 10.0$.

Solution 11.6

Under $\text{Bin}(20, 0.4)$, using the formula for the probability function, $P(X \leq 3) = 0.0160$ 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 $h_2 = 13$.

Note that h_1 and h_2 now have higher values than for the $\text{Bin}(20, 0.3)$ case.

Solution 11.7

The number in a sample of 8 who are in favour has a $\text{Bin}(8, p)$ 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 the iterative equation $p = \left(\frac{0.025}{8 - 7p}\right)^{\frac{1}{7}}$, 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 $(0.4735, 0.9968)$.

Solution 11.8

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 $n=20$, so 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 because 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 $np \geq 5$.

Solution 11.9

Assuming that the sample comes from a binomial distribution, we know that the quantity $\frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$ or $\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$. Here $n = 1,000$ and X 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 :

$$0.99 = P\left(\frac{X}{n} - 2.5758\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \frac{X}{n} + 2.5758\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Using $n = 1,000$, $X = 450$ and $\hat{p} = \frac{450}{1,000}$, we get the confidence interval to be $0.409 < p < 0.491$.

Since this 99% confidence interval doesn't contain the value $p = 0.5$, it is unlikely that Party P will gain more than 50% of the votes.

Solution 11.10

Since the sample comes from a Poisson distribution, we know that $\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \sim N(0,1)$ (approximately). Here $n = 5,000$.

From the *Tables* we find that $P(-1.6449 < Z < 1.6449) = 0.90$.

So:

$$P\left(-1.6449 < \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} < 1.6449\right) = 0.90$$

which we can rearrange to give:

$$P\left(\bar{X} - 1.6449\sqrt{\frac{\hat{\lambda}}{n}} < \lambda < \bar{X} + 1.6449\sqrt{\frac{\hat{\lambda}}{n}}\right) = 0.90$$

Using the observed values of $n = 5,000$, $\bar{X} = 0.16$, and using $\hat{\lambda} = \bar{X}$, the confidence interval is $0.151 < \lambda < 0.169$.

Solution 11.11

Assuming that the samples come from normal distributions with the same variance and that the samples are independent, we know that $\frac{(\bar{A} - \bar{C}) - (\mu_A - \mu_C)}{S \sqrt{\frac{1}{n_A} + \frac{1}{n_C}}} \sim t_{n_A+n_C-2}$, 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 is the pooled sample standard deviation.

Using the data given, $\bar{A} = 30.75$, $\bar{C} = 28.3$, $S^2 = 5.96$, $n_A = n_C = 8$. (The individual sample variances are $S_A^2 = \frac{48.06}{7}$ and $S_C^2 = \frac{35.38}{7}$.)

We also know that $t_{14} = 2.145$ for the required confidence interval.

Substituting these values in we get the symmetrical 95% confidence interval to be:

$$2.45 \pm 2.145 \sqrt{\frac{5.96}{4}} = (-0.168, 5.068)$$

Solution 11.12

For two independent random samples from $N(\mu_A, \sigma_A^2)$ and $N(\mu_C, \sigma_C^2)$, $\frac{S_A^2/\sigma_A^2}{S_C^2/\sigma_C^2} \sim F_{n_A-1, n_C-1}$, 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 $0.90 = P\left(\frac{1}{3.787} < F_{7,7} < 3.787\right)$, which gives us:

$$0.90 = P\left(\frac{1}{3.787} < \frac{S_A^2/\sigma_A^2}{S_C^2/\sigma_C^2} < 3.787\right)$$

Rearranging this to give $\frac{\sigma_C^2}{\sigma_A^2}$, we get $0.90 = P\left(0.1944 < \frac{\sigma_C^2}{\sigma_A^2} < 2.788\right)$.

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 Question 11.11 appears to be justified.

Solution 11.13

Since the samples come from binomial distributions, we know that, approximately:

$$\frac{\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) - (p_1 - p_2)}{\sqrt{\frac{X_1}{n_1} \left(1 - \frac{X_1}{n_1}\right) + \frac{X_2}{n_2} \left(1 - \frac{X_2}{n_2}\right)}} \sim N(0,1)$$

Calling $\frac{X_1}{n_1} = \hat{p}_1$, and $\frac{X_2}{n_2} = \hat{p}_2$, and using the *Tables*, we know that:

$$0.95 = P\left(-1.96 < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < 1.96\right)$$

From the data given in the question, $\hat{p}_1 = 0.25$ and $\hat{p}_2 = 0.133$. Substituting these in:

$$0.95 = P(0.016 < p_1 - p_2 < 0.218)$$

So a symmetrical 95% confidence interval for the difference in mortality rates is $(0.016, 0.218)$.

Solution 11.14

Since the samples come from independent Poisson distributions, we know that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\bar{X}_1}{n_1} + \frac{\bar{X}_2}{n_2}}} \sim N(0,1),$$

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:

$$0.99 = P\left(-2.5758 < \frac{(\bar{X}_1 - \bar{X}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\bar{X}_1}{n_1} + \frac{\bar{X}_2}{n_2}}} < 2.5758\right)$$

From the data in the question, $\bar{X}_1 = 0.3$, and $\bar{X}_2 = 0.1444$, so substituting these values in, we get $0.99 = P(0.0908 < \lambda_1 - \lambda_2 < 0.2204)$, which gives the confidence interval to be $(0.0908, 0.2204)$.

Solution 11.15

Since this is a paired sample from a normal distribution, we know that $\frac{\bar{D} - (\mu_A - \mu_B)}{S_D / \sqrt{n}} \sim t_{n-1}$, where $D = A - B$.

From the *Tables*, we know that $0.95 = (-2.262 < t_9 < 2.262)$, so:

$$0.95 = (-2.262 < \frac{\bar{D} - (\mu_A - \mu_B)}{S_D / \sqrt{n}} < 2.262)$$

From the data in the question, $n = 10$, $\bar{D} = -2.0$ and:

$$S_D = \sqrt{\frac{\sum (B_i - A_i)^2 - n(\bar{B} - \bar{A})^2}{n-1}} = \sqrt{\frac{68 - 10 \times 4}{9}} = 1.764$$

Substituting these values in we get $0.95 = P(-3.26 < \mu_A - \mu_B < -0.74)$, which gives the required confidence interval to be $(-3.26, -0.74)$.

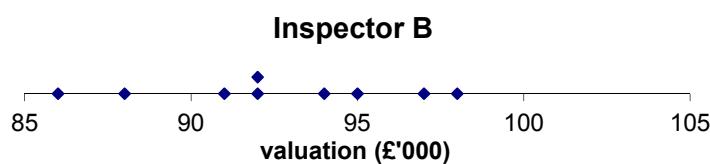
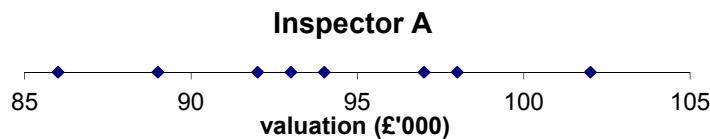
Since this interval does not include the value 0 (*ie* no difference in the average blood pressure before and after), the diet seems to be effective.

Solution 11.16

No. The vehicles were all the same make and age and separate cars were used in the control and additive tests, which were not matched up in any way.

Past Exam Question (Subject C1, September 1994, Q14)

- (i)(a) The dotplots are as follows:



- (i)(b) B appears to have a slightly smaller spread (but it is hard to tell on so few data points). It doesn't appear to be significantly different, so the assumption of equal variances could be taken.
- (i)(c) For inspector A, we have $n_A = 8$ $\sum x_A = 751$ $\sum x_A^2 = 70,683$ giving:

$$s_A^2 = \frac{1}{7} \left[70,683 - \frac{751^2}{8} \right] = 26.125$$

For inspector B, we have $n_B = 9$ $\sum x_B = 833$ $\sum x_B^2 = 77,223$ giving:

$$s_B^2 = \frac{1}{8} \left[77,223 - \frac{833^2}{9} \right] = 15.528$$

The common (or pooled) variance is given by:

$$s_P^2 = \frac{7 \times 26.125 + 8 \times 15.528}{7 + 8} = 20.473$$

The pivotal quantity is $\frac{15s_P^2}{\sigma_P^2} \sim \chi_{15}^2$. This gives a 95% confidence interval for

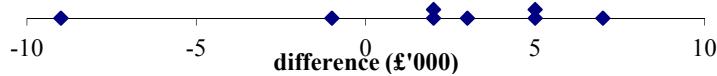
$$\sigma_P^2 \text{ of } \left(\frac{15 \times 20.473}{27.49}, \frac{15 \times 20.473}{6.262} \right) = (11.2, 49.0).$$

- (i)(d) The confidence interval is calculated using:

$$\begin{aligned}
 (\mu_A - \mu_B) &= (\bar{x}_A - \bar{x}_B) \pm t_{0.025,15} \sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\
 &= \left(\frac{751}{8} - \frac{833}{9} \right) \pm 2.131 \sqrt{20.473 \left(\frac{1}{8} + \frac{1}{9} \right)}
 \end{aligned}$$

This gives a confidence interval of $(-3.37, 6.00)$. Since this interval contains zero there is no evidence at the 5% level to suggest a significant difference in the valuations given by each of the two inspectors.

- (ii)(a) We are looking at paired data, so we need to examine the differences. The dotplot for the differences is as follows:



- (ii)(b) For the differences we have $n_D = 8$ $\sum x_D = 14$ $\sum x_D^2 = 198$ giving:

$$\bar{x}_D = 1.75 \quad s_D^2 = \frac{1}{7} \left[198 - \frac{14^2}{8} \right] = 24.786$$

The confidence interval is calculated using:

$$\begin{aligned}
 \mu_D &= \bar{x}_D \pm t_{0.05,7} \sqrt{\frac{s_D^2}{n_D}} \\
 &= \frac{14}{8} \pm 1.895 \sqrt{\frac{24.786}{8}}
 \end{aligned}$$

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.

Past Exam Question (Subject C1, Specimen 1993, Q16)

$$(i)(a) \quad L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

$$\Rightarrow \ln L(\lambda) = n \ln \lambda - \lambda \sum x_i \Rightarrow \frac{d}{d\lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum x_i$$

Setting the derivative equal to zero to obtain the MLE:

$$\Rightarrow \frac{n}{\hat{\lambda}} - \sum x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}}$$

Checking it's a maximum:

$$\frac{d^2}{d\lambda^2} \ln L(\lambda) = -\frac{n}{\lambda^2} < 0 \Rightarrow \text{max}$$

For these data, $\hat{\lambda} = \frac{1}{8.7} = 0.11494$.

$$(i)(b) \quad CRLB = -\frac{1}{E\left[\frac{d^2}{d\lambda^2} \ln L(\lambda)\right]} = \frac{1}{E\left[\frac{n}{\lambda^2}\right]} = \frac{1}{\frac{n}{\lambda^2}} = \frac{\lambda^2}{n}$$

For these data, our estimate of the CRLB is $\hat{\lambda}^2/n = 0.000661$.

- (i)(c) Since $\hat{\lambda} \sim N(\lambda, CRLB)$, the confidence interval will be given by $\hat{\lambda} \pm 1.96\sqrt{CRLB}$ which, using our CRLB estimate, gives $(0.06457, 0.1653)$.
- (ii) Since $2\lambda n \bar{X} \sim \chi^2_{2n}$, we have $2\lambda \times 20 \times 8.7 \sim \chi^2_{40}$. Reading off the χ^2_{40} values gives:

$$24.43 < 348\lambda < 59.34$$

So the confidence interval for λ is $(0.07020, 0.1705)$. 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.

Chapter 12

Hypothesis testing



Syllabus objectives

- (xi) 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.
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. Use a χ^2 test to test the hypothesis that a random sample is from a particular distribution, including cases where parameters are unknown.
4. Explain what is meant by a contingency (or two-way) table and use a χ^2 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:

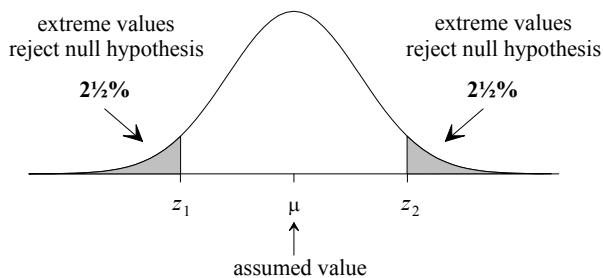
- Is drug A a more effective treatment for AIDS than drug B?
- Does training programme T lead to improved staff efficiency?
- Are the frequencies of large individual private motor insurance claims consistent with a lognormal distribution?

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 11. For example, when we were sampling from a $N(\mu, \sigma^2)$ distribution (σ^2 known) we used:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

By substituting in \bar{X}, σ^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 find the probability that our sample has this value of μ . If we find that our sample lies in the extreme ends of the distribution, we naturally conclude that since this is so rare it is likely that our sample does not have the assumed value of μ . In this case we would *reject* the “null” hypothesis. If, however our sample value 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 11, 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 10. For example, given the number of claims from a certain portfolio that we receive in a month:

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 10 would be $\mu = \bar{X} = 2.37$. We then obtained a confidence interval for this value in Chapter 11. But all of this work is great *only if* it is a Poisson distribution. Hence, we will now carry out a goodness-of-fit test to see if our sample does or does not conform to this distribution.

This chapter tends to form one of the longer questions of the Subject CT3 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 no 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:

$$H_0: \text{smoking makes no difference to life expectancy}$$

$$H_1: \text{smoking reduces life expectancy}$$

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:

$$H_0: \text{smoking makes no difference to life expectancy}$$

$$H_1: \text{smoking affects life expectancy}$$

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 0.

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 12.1

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 $\mu = \frac{1}{\lambda}$, a single value is observed from the distribution. If this value is less than 28, H_0 is accepted, otherwise H_0 is rejected. Find the probabilities of:

- (i) a Type I error
- (ii) a Type II error.

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 12.2

Obtain an expression in terms of μ for the power of the test in Question 12.1. Comment on how the power is affected by the value of μ .

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.

Formally, if C is a critical region of size α and there exists a constant k such that $\frac{L_0}{L_1} \leq k$ inside C and $\frac{L_0}{L_1} \geq k$ 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 12.3

Show that the test given in Question 12.1 is a Neyman-Pearson test.

Common tests are often such that the null hypothesis is simple, eg $H_0 : \theta = \theta_0$, 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 that are referred to as 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.

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:

$$\frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t_{n-1} \text{ under } H_0: \mu = \mu_0$$

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" approach, with a fixed predetermined value of α , a test will produce a decision as to whether or not 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:

$$P(X \leq 82) \text{ where } X \sim \text{binomial}(200, 0.5)$$

$$P[Z < (82.5 - 100)/\sqrt{50}] = P(Z < -2.475) = 0.0067$$

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 particular 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 12.4

Suppose the critical value of 28 in Question 12.1 is replaced by k , so that the test used is: if the value of X is less than k , H_0 is accepted, otherwise H_0 is rejected.

- (i) What is the value of k that gives a test of size 5%?
- (ii) What is the probability of a Type II error in this case?

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

Basic 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 $\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$ under H_0
- (b) σ unknown: test statistic is $\frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t_{n-1}$ 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.



Example 12.1

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:

$$H_0: \mu = 100 \quad \text{vs} \quad H_1: \mu > 100 \quad (\sigma \text{ known})$$

Under H_0 , $\frac{\bar{X} - 100}{\sigma/\sqrt{n}} \sim N(0,1)$.

The test statistic is $\frac{105 - 100}{20/\sqrt{50}} = 1.768$.

We need to form our conclusion and there are two ways of doing this.

Method 1:

Calculate the probability of getting a result as extreme as the test statistic (*i.e.* 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 12.5

Test using a 5% significance level whether the average IQ of university students is greater than 103, based on the sample in 0.



Question 12.6

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.

3.2 Testing the value of a population variance

Basic situation: random sample, size n , from $N(\mu, \sigma^2)$ – sample variance S^2 . If we are testing $H_0: \sigma^2 = \sigma_0^2$:

Test statistic is $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$ under H_0

For large samples, the test works well even if the population is not normally distributed.



Example 12.2

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:

$$H_0: \sigma = 3 \quad \text{vs} \quad H_1: \sigma \neq 3$$

Under H_0 :

$$\frac{4S^2}{3^2} \sim \chi^2_4$$

Using the data given in the question, the test statistic is $\frac{4 \times 17.8}{3^2} = 7.91$.

Our statistic of 7.91 is lies between 0.4844 and 11.14 (the lower and upper 2½% points of the χ^2_4 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 $P(\chi^2_4 > 7.91) \approx 0.0952$. 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 12.7

Making the same assumptions as previously, test at the 5% level whether the standard deviation of the annual rainfall at the weather station in Question 12.6 is equal to 4 cm.

3.3 Testing the value of a population proportion

Situation: n binomial trials with $P(\text{success}) = \theta$; we observe x successes.

Testing $H_0 : \theta = \theta_0$.

Test statistic is $X \sim \text{binomial}(n, \theta_0)$ under H_0 .

For large n , use the normal approximation to the binomial (with continuity correction), ie use:

$$\frac{\frac{X \pm \frac{1}{2}}{2} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \text{ approximately}$$

or:

$$\frac{\frac{X \pm \frac{1}{2} - np}{2} - p}{\sqrt{np(1-p)}} \sim N(0,1) \text{ approximately}$$

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.



Example 12.3

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 $\text{Bin}(250, q)$ distribution, test whether this result is consistent with a mortality rate of $q = 0.2$ for this age.

Solution

We are testing:

$$H_0 : q = 0.2 \quad \text{vs} \quad H_1 : q \neq 0.2$$

Under H_0 :

$$\frac{X/n - 0.2}{\sqrt{\frac{0.2 \times 0.8}{n}}} \sim N(0,1) \text{ approximately}$$

Using the observed values, $n = 250$ and $X = 45$, the test statistic with continuity correction is:

$$\frac{45.5/250 - 0.2}{\sqrt{\frac{0.2 \times 0.8}{250}}} = -0.712$$

Our statistic of -0.712 lies between ± 1.960 (the lower and upper $2\frac{1}{2}\%$ 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 $2 \times 0.238 = 0.48$, which is greater than 0.05 so we have insufficient evidence to reject H_0 at the 5% level.


Question 12.8

A new gene has been identified that makes carriers 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%.

3.4 Testing the value of the mean of a Poisson distribution

Situation: random sample, size n , from $\text{Poisson}(\lambda)$ distribution.

$$H_0 : \lambda = \lambda_0$$

Test statistic is sample sum $\sum X_i \sim \text{Poisson}(n\lambda_0)$ 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 $\sum X_i \sim \text{Poisson}(n\lambda) \rightarrow N(n\lambda, n\lambda)$.

Test statistic is \bar{X} , and:

$$\frac{\bar{X} - \lambda_0}{\sqrt{\lambda_0/n}} \sim N(0, 1) \text{ under } H_0$$

or we can use $\sum X_i$, and:

$$\frac{\sum X_i - n\lambda_0}{\sqrt{n\lambda_0}} \sim N(0, 1) \text{ under } H_0$$

Using the second version it is easier to incorporate a continuity correction.

The first version has continuity correction $\frac{0.5}{n}$, whereas the second version has continuity correction 0.5.



Example 12.4

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 Poisson (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 :

$$\frac{\bar{X} - 0.175}{\sqrt{0.175/n}} \sim N(0,1)$$

Using the observed values, $n = 5,000$ and $\bar{X} = 0.16$, 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, so we have sufficient evidence to reject H_0 at the 1% level.



Question 12.9

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.

4 Basic tests – two independent samples

4.1 Testing the value of the difference between two population means

Basic situation: independent random samples, sizes n_1 and n_2 from $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ respectively.

$$H_0 : \mu_1 - \mu_2 = \delta$$

(a) σ_1^2, σ_2^2 known

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

(b) σ_1^2, σ_2^2 unknown – much the more usual situation

Large samples: use S_i^2 to estimate σ_i^2 . You can then calculate the same test statistic as in (a).

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 $\sigma_1^2 = \sigma_2^2 (= \sigma^2$ say), this common variance is estimated by S_p^2 , and the test statistic is distributed as t with $n_1 + n_2 - 2$ degrees of freedom under H_0 . So

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Remember that $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$.



Example 12.5

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 $\sum_{i=1}^{10} C_i^2 = 59,420$ and $\sum_{i=1}^{10} T_i^2 = 56,390$.

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 :

$$\frac{(\bar{C} - \bar{T}) - (0)}{S \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

Using the observed values of $m = 10$, $n = 10$, $\bar{T} = 75.0$, $\bar{C} = 77.0$, and $S^2 = 3.873^2$, the test statistic is:

$$\frac{(77.0 - 75.0)}{3.873 \sqrt{\frac{1}{10} + \frac{1}{10}}} = 1.15$$

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.

Alternatively, using probability values, we have $P(t_{18} > 1.15) \approx 0.134$. Since this is greater than 0.05, we have insufficient evidence to reject H_0 at the 5% level.



Question 12.10

Verify the calculation of the pooled variance s^2 in this example.



Question 12.11

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:

$$(i) \quad H_0 : \mu_A - \mu_C = 0 \quad vs \quad H_1 : \mu_A - \mu_C > 0$$

$$(ii) \quad H_0 : \mu_A - \mu_C = 6 \quad vs \quad H_1 : \mu_A - \mu_C \neq 6$$

You may have noticed that we've been using some of the same examples in this chapter as in Chapter 11. 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.

4.2 Testing the value of the ratio of two population variances

Basic 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 .

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

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 11 that $\frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}$, so it follows that if we are testing the hypothesis $\sigma_1^2 = \sigma_2^2$, we can use the test statistic $\frac{S_1^2}{S_2^2}$ and compare it with the critical

points on the appropriate F table.



Question 12.12

Test whether the variances in the two populations in 0 can be considered to be equal.

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.

Situation:

n_1 (large) trials with $P(\text{success}) = \theta_1$; observe x_1 successes.

n_2 (large) trials with $P(\text{success}) = \theta_2$; observe x_2 successes.

Testing $H_0 : \theta_1 = \theta_2$.

Test statistic is:

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2)}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n_1} + \frac{\hat{\theta}(1-\hat{\theta})}{n_2}}} \sim N(0,1)$$

under H_0 where $\hat{\theta}_1, \hat{\theta}_2$ are the MLEs of θ_1 and θ_2 respectively, (the sample proportions $\frac{X_1}{n_1}, \frac{X_2}{n_2}$), and $\hat{\theta}$ is the MLE of the common θ under the null hypothesis, which is the overall sample proportion, namely $\frac{X_1 + X_2}{n_1 + n_2}$.

In some textbooks you may see an alternative test statistic, namely:

$$\frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}} \sim N(0,1) .$$

The denominator in the Core Reading expression is found by pooling the sample proportions, whereas in the alternative version, the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ are used separately.

Since the test is approximate, both approximations are valid.



Example 12.6

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:

$$H_0 : q_M = q_F \quad \text{vs} \quad H_1 : q_M \neq q_F$$

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 :

$$\frac{\left(\frac{X_M}{m} - \frac{X_F}{f} \right) - 0}{\sqrt{\frac{\hat{q}(1-\hat{q})}{m} + \frac{\hat{q}(1-\hat{q})}{f}}} \sim N(0,1)$$

Using the observed values of $m = 100$, $f = 150$, $X_M = 25$, $X_F = 20$, and $\hat{q} = \frac{45}{250}$, the value of the test statistic is:

$$\frac{(0.25 - 0.133)}{\sqrt{(0.18 \times 0.82)/100 + (0.18 \times 0.82)/150}} = 2.36$$

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 $P(Z > 2.36) = 0.0091$. Since this test is two-sided, the probability of obtaining a more extreme value than the one actually obtained is $2 \times 0.0091 = 0.018$. Since this is less than 0.05 we have sufficient evidence to reject H_0 at the 5% level.


Question 12.13

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.

4.4 Testing the value of the difference between two Poisson means

Situation: independent random samples, sizes n_1 and n_2 , from $\text{Poisson}(\lambda_1)$ and $\text{Poisson}(\lambda_2)$ 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:

$$H_0 : \lambda_1 = \lambda_2 .$$

Test statistic is:

$$\frac{(\hat{\lambda}_1 - \hat{\lambda}_2)}{\sqrt{\frac{\hat{\lambda}}{n_1} + \frac{\hat{\lambda}}{n_2}}} \sim N(0,1)$$

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:

$$\frac{(\hat{\lambda}_1 - \hat{\lambda}_2)}{\sqrt{\frac{\hat{\lambda}_1}{n_1} + \frac{\hat{\lambda}_2}{n_2}}} \sim N(0,1) .$$

Similarly to the last section, the Core Reading version has a pooled value for the parameter, whereas the alternative version doesn't.

**Question 12.14**

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.

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.

Assumption: differences constitute a random sample from a normal distribution.

$$H_0 : \mu_D (= \mu_1 - \mu_2) = \delta$$

Test statistic is $\frac{\bar{D} - \delta}{S_D / \sqrt{n}} \sim t_{n-1}$ under H_0 .

You use $N(0,1)$ for t , and do not require the “normal” assumption, if n is large.



Example 12.7

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 $\sum_{i=1}^{10} (B_i - A_i)^2 = 68.0$.

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:

$$\bar{D} = \bar{A} - \bar{B} = 75.0 - 77.0 = -2$$

$$\begin{aligned} S_D^2 &= \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n D_i^2 - n\bar{D}^2 \right] \\ &= \frac{1}{9} \left[68.0 - 10(-2.0)^2 \right] = 3.111 = 1.764^2 \end{aligned}$$

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 $P(t_9 < -3.59) \approx 0.0037$, 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 the two-sample t-test in 0 was unable to reach this conclusion because the reduction was masked by other factors.



Question 12.15

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.

6 Tests and confidence intervals

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. And in fact this is generally 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. So in this, and similar cases, there is only an approximate match (but still a good match) 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 12.16

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). Do you think that people are getting sufficient vitamin C in each country? (The recommended daily allowance is 60mg.)

7 χ^2 tests

These tests are relevant to the situation in which the data are 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 i^{th} category/cell, and the summation is taken over all categories/cells involved. This statistic has, approximately, a 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.

7.1 Goodness of fit

This is investigating whether it is reasonable to regard a random sample as coming from a particular 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 has to 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 χ^2 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.)

Example

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:

$$H_0: \text{Number thrown has distribution (†)}$$

$$H_1: \text{Number thrown does not have distribution (†)}$$

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

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 (*i.e.* when the observed and expected values are 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 12.17

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



Question 12.18

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.



Question 12.19

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.

7.2 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 or not 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 $f_{i\cdot} / f$ so if the criteria are independent, the number expected in cell (i, j) is $(f_{i\cdot} / f) \times f_{\cdot j}$.)

We will explain the “dot notation” used here when we study analysis of variance in Chapter 14.

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.

Example

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:

$$\begin{aligned}\chi^2 &= \frac{0.25^2}{22.75} + \frac{5.25^2}{22.75} + \frac{5.5^2}{45.50} + \frac{0.25^2}{77.25} + \frac{5.25^2}{77.25} + \frac{5.5^2}{154.50} \\ &= 0.003 + 1.212 + 0.665 + 0.001 + 0.357 + 0.196 \\ &= 2.43\end{aligned}$$

on 2df.

where df stands for “degrees of freedom”.

This is an unremarkable value for χ^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 $\chi^2 = 7.3$.

$p\text{-value} = P(\chi_2^2 > 7.3)$, 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 ($p\text{-value}$ 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 when based on sample sizes which are considerably bigger (in particular sizes 300, 300, and 600).



Question 12.20

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.

**Question 12.21**

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

8 Exam-type questions

The next question applies the goodness-of-fit test to the results of Chapter 10.



Past Exam Question (Subject C1, September 1994, Q15)

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.

<i>No. of houses burgled</i>	<i>x</i>	0	1	2	3	4	5	6
<i>No. of streets</i>	<i>f</i>	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 sampled 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) Fit the binomial model using your estimate of p , and, without doing a formal test, comment on the fit.
- (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.

This question is typical of the joint Chapter 11 and 12 questions:



Exam-type question

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:

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) Illustrate 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.
- (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.
- (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.
- (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.
- (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.



Chapter 12 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 on the formulae summary sheet.

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”.



Chapter 12 Formulae

One-sample normal distribution

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \sigma^2 \text{ known} \quad \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \quad \sigma^2 \text{ unknown}$$

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Two-sample normal distribution

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1) \quad \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

σ^2 known

σ^2 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

$$\frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \stackrel{d}{\sim} N(0,1) \quad \text{or} \quad \frac{X - np_0}{\sqrt{np_0 q_0}} \stackrel{d}{\sim} N(0,1) \quad \text{with continuity correction}$$

Two-sample binomial

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}} \stackrel{d}{\sim} N(0,1) \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad \text{is the overall sample proportion}$$

One-sample Poisson

$$\frac{\bar{X} - \lambda_0}{\sqrt{\lambda_0/n}} \stackrel{d}{\sim} N(0,1) \quad \text{or} \quad \frac{\sum X - n\lambda_0}{\sqrt{n\lambda_0}} \stackrel{d}{\sim} N(0,1) \quad \text{with continuity correction}$$

Two-sample Poisson

$$\frac{(\hat{\lambda}_1 - \hat{\lambda}_2) - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\hat{\lambda}}{n_1} + \frac{\hat{\lambda}}{n_2}}} \stackrel{d}{\sim} N(0,1) \quad \hat{\lambda} = \frac{n_1 \hat{\lambda}_1 + n_2 \hat{\lambda}_2}{n_1 + n_2} \quad \text{is the overall sample mean}$$

Chapter 12 Solutions

Solution 12.1

- (i) The probability of a Type I error is given by:

$$\begin{aligned} P(\text{reject } H_0 \text{ when } H_0 \text{ true}) &= P(X > 28 \text{ when } X \sim \text{Exp}(1/20)) \\ &= 1 - F_X(28) = e^{-28/20} = 0.2466 \end{aligned}$$

The CDF of the exponential distribution is given on Page 12 of the Tables.

- (ii) The probability of a Type II error is given by:

$$\begin{aligned} P(\text{do not reject } H_0 \text{ when } H_0 \text{ false}) &= P(X < 28 \text{ when } X \sim \text{Exp}(1/30)) \\ &= F_X(28) = 1 - e^{-28/30} = 0.6068 \end{aligned}$$

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.

Solution 12.2

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:

$$P(X > 28 | X \sim \text{Exp}(1/\mu)) = 1 - F_X(28) = e^{-28/\mu}$$

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.

Solution 12.3

Given a single value from an exponential distribution, the Neyman-Pearson criterion is “reject H_0 if $\frac{L_0}{L_1} < \text{critical value.}$ ” Using the null and alternative hypotheses given in Question 12.1, the test becomes:

$$\frac{\frac{1}{20}e^{-\frac{x}{20}}}{\frac{1}{30}e^{-\frac{x}{30}}} < \text{constant}$$

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.

Solution 12.4

(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:

$$\int_0^k \frac{1}{30} e^{-\frac{x}{30}} dx = \left[-e^{-\frac{x}{30}} \right]_0^k = 1 - e^{-1.997} = 0.864$$

Solution 12.5

We are testing:

$$H_0 : \mu = 103 \quad \text{vs} \quad H_1 : \mu > 103$$

Under H_0 :

$$\frac{\bar{X} - 103}{\sigma/\sqrt{n}} \sim N(0,1)$$

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 $P(Z > 0.707) \approx 0.24$. This is greater than 0.05, so we have insufficient evidence to reject H_0 at the 5% level.

Solution 12.6

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 :

$$\frac{\bar{X} - 22}{S/\sqrt{n}} \sim t_{n-1}$$

We have:

$$s^2 = \frac{1}{9}(6895.73 - 10 \times 25.87^2) = 22.57$$

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), so 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 $P(t_9 > 2.576) \approx 0.0166$. This is less than 0.05, so we have sufficient evidence to reject H_0 at the 5% level.

Solution 12.7

We are testing:

$$H_0 : \sigma = 4 \quad \text{vs} \quad H_1 : \sigma \neq 4$$

The test is two-sided. Assuming independence and normality, then under H_0 :

$$\frac{9S^2}{4^2} \sim \chi^2_9$$

The observed value of the test statistic is:

$$\frac{9 \times 22.57}{16} = 12.69$$

This is between the upper and lower $2\frac{1}{2}\%$ points of χ^2_9 (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 $P(\chi^2_9 > 12.69) \approx 0.1775$. 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 so we have insufficient evidence to reject H_0 at the 5% level.

Solution 12.8

We are testing:

$$H_0 : p = 0.1 \quad vs \quad H_1 : p < 0.1$$

Under H_0 :

$$\frac{X/n - 0.1}{\sqrt{\frac{0.1 \times 0.9}{n}}} \sim N(0,1) \text{ approximately}$$

The observed value of the test statistic, with continuity correction, is:

$$\frac{8.5/250 - 0.1}{\sqrt{\frac{0.1 \times 0.9}{250}}} = -3.479$$

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 $P(Z < -3.479) \approx 0.00025$. 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.

Solution 12.9

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 :

$$\frac{\bar{X} - 0.18}{\sqrt{0.18/n}} \sim N(0,1) \text{ approximately}$$

The observed value of the test statistic, with continuity correction, is:

$$\frac{0.231 - 0.18}{\sqrt{0.18/500}} = 2.688$$

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 level. Therefore it is reasonable to conclude that the true claim frequency is more than 0.18.

Alternatively, using probability values, we have $P(Z > 2.688) \approx 0.0036$, 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.

Solution 12.10

$$\begin{aligned} S^2 &= \frac{1}{m+n-2} \left[(m-1)S_C^2 + (n-1)S_T^2 \right] \\ &\quad \frac{1}{m+n-2} \left[\sum_{i=1}^m (C_i - \bar{C})^2 + \sum_{i=1}^n (T_i - \bar{T})^2 \right] \\ &= \frac{1}{m+n-2} \left[\sum_{i=1}^m C_i^2 - m\bar{C}^2 + \sum_{i=1}^n T_i^2 - n\bar{T}^2 \right] \\ &= \frac{1}{10+10-2} \left[59,420 - 10 \times 77.0^2 + 56,390 - 10 \times 75.0^2 \right] = 15.00 = 3.873^2 \end{aligned}$$

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.

Solution 12.11

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:

$$\sum C_i = 226.4, \sum C_i^2 = 6442.5, \sum A_i = 246, \sum A_i^2 = 7612.56$$

Our estimate of the pooled sample variance is:

$$\begin{aligned} S^2 &= \frac{1}{m+n-2} \left[\sum C_i^2 - n\bar{C}^2 + \sum A_i^2 - m\bar{A}^2 \right] \\ &= \frac{1}{14} (6442.5 - 8 \times 28.3^2 + 7612.56 - 8 \times 30.75^2) = 5.96 \end{aligned}$$

(i) We are testing:

$$H_0: \mu_A - \mu_C = 0 \quad vs \quad H_1: \mu_A - \mu_C > 0$$

Assuming that the underlying distributions are normal, then under H_0 :

$$\frac{(\bar{A} - \bar{C}) - 0}{S \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

The observed value of the test statistic is:

$$\frac{30.75 - 28.3}{\sqrt{5.96} \sqrt{\frac{1}{8} + \frac{1}{8}}} = 2.007$$

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 $P(t_{14} > 2.007) \approx 0.0340$. This is less than 0.05, so we have sufficient evidence to reject H_0 at the 5% level.

(ii) We are now testing:

$$H_0: \mu_A - \mu_C = 6 \quad \text{vs} \quad H_1: \mu_A - \mu_C \neq 6$$

Making the same assumptions as before, under H_0 :

$$\frac{(\bar{A} - \bar{C}) - 6}{S \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

The observed value of the test statistic is now:

$$\frac{(30.75 - 28.3) - 6}{\sqrt{5.96} \sqrt{\frac{1}{8} + \frac{1}{8}}} = -2.908$$

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 $P(t_{14} < -2.908) \approx 0.00598$. 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.

Solution 12.12

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:

$$S_T^2 = \frac{1}{9} (56,390 - 10 \times 75^2) = 15.56$$

$$S_C^2 = \frac{1}{9} (59,420 - 10 \times 77^2) = 14.44$$

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 $\frac{1}{4.026} = 0.2484$ (the upper and lower $2\frac{1}{2}\%$ 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 in 0 which presupposes equal variances.

Note that had we used $\frac{S_C^2}{S_T^2} = \frac{14.44}{15.56} = 0.9280$, we would have reached the same conclusion.

Solution 12.13

We are testing:

$$H_0 : \theta_1 = \theta_2 \quad vs \quad H_1 : \theta_2 > \theta_1 \quad (ie \theta_1 - \theta_2 < 0)$$

where θ_1 and θ_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 :

$$\frac{(N_1/200 - N_2/100) - 0}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{200} + \frac{\hat{\theta}(1-\hat{\theta})}{100}}} \sim N(0,1) \text{ approximately}$$

The observed value of the test statistic is:

$$\frac{(115/200 - 62/100) - 0}{\sqrt{\frac{0.59(1-0.59)}{200} + \frac{0.59(1-0.59)}{100}}} = -0.747$$

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 $P(Z < -0.747) \approx 0.228$. Since this is greater than 0.05, we have insufficient evidence to reject H_0 at the 5% level.

Solution 12.14

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 :

$$\frac{(\bar{Y} - \bar{O}) - 0}{\sqrt{\frac{\hat{\lambda}}{m} + \frac{\hat{\lambda}}{n}}} \sim N(0,1) \text{ approximately}$$

The observed value of the test statistic is:

$$\frac{0.300 - 0.144}{\sqrt{\frac{0.16}{500} + \frac{0.16}{4,500}}} = 8.25$$

We are carrying out a two-sided test and our statistic is much greater than ± 2.5758 (the upper and lower $\frac{1}{2}\%$ points 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 $P(Z > 8.25) << 0.0005\%$. Doubling this (as this test is two-sided) gives a p-value that is still less than 0.0005%. So we have sufficient evidence to reject H_0 , even at the 0.0005% 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.

Solution 12.15

We are testing:

$$H_0: \mu_A = \mu_B \quad vs \quad H_1: \mu_A < \mu_B \quad (ie \mu_A - \mu_B < 0)$$

where A is “After” and B is “Before”.

Taking the differences $D = A - B$ (so that a positive value of D represents an improvement in performance), we have:

$$-4 \quad -14 \quad -5 \quad -3 \quad 3$$

Applying a one-sample t -test to the D values (and assuming that the underlying distributions are normal):

$$\frac{\bar{D} - (\mu_A - \mu_B)}{S_D / \sqrt{n}} \sim t_{n-1}$$

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 lower 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 $P(t_4 < -1.684) \approx 0.0874$, which is greater than 0.05. So we have insufficient evidence to reject H_0 at the 5% level.

Solution 12.16

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.

Solution 12.17

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^2_4 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 $P(\chi^2_4 > 8.29) \approx 0.0819$, which is greater than 0.05. So we have insufficient evidence to reject H_0 at the 5% level.

Solution 12.18

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 :

$$\sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_f$$

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 *i.e.* 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

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^2_1 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.

Solution 12.19

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 $\hat{\theta} = \frac{\bar{x}}{6}$, where \bar{x} is the mean number of impure bottles per crate. From the data, $\bar{x} = \frac{301}{200} = 1.505$, so, given that there are six bottles in each crate, $\hat{\theta} = \frac{\bar{x}}{6} = 0.25083333$.

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, \dots, x_{200} . Each x_i comes from a $Bin(6, \theta)$ distribution, and so the likelihood function for θ is:

$$\begin{aligned} L(\theta) &= \binom{6}{x_1} \theta^{x_1} (1-\theta)^{6-x_1} \cdots \binom{6}{x_{200}} \theta^{x_{200}} (1-\theta)^{6-x_{200}} \\ &= \text{constant} \times \theta^{\sum x_i} (1-\theta)^{1200-\sum x_i} \end{aligned}$$

Taking logs, differentiating with respect to θ and setting the result equal to zero:

$$\log L = \sum x_i \log \theta + (1200 - \sum x_i) \log(1-\theta)$$

$$\frac{\partial}{\partial \theta} \log L = \frac{\sum x_i}{\theta} - \frac{1200 - \sum x_i}{1-\theta} = 0$$

Solving this we get $\hat{\theta} = \frac{\sum x_i}{1200} = \frac{301}{1200} = 0.25083$.

We can now calculate the expected frequencies. We calculate the probabilities from a $\text{Bin}(6, 0.25083)$ 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:

$$\begin{aligned}\chi^2 &= \frac{(38 - 35.36)^2}{35.36} + \frac{(70 - 71.03)^2}{71.03} + \frac{(58 - 59.46)^2}{59.46} + \frac{(25 - 26.54)^2}{26.54} + \frac{(9 - 7.61)^2}{7.61} \\ &= 0.59\end{aligned}$$

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!

Solution 12.20

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

So we can now calculate the value of the chi square statistic:

$$\chi^2 = \frac{(3 - 31.5)^2}{31.5} + \dots + \frac{(29 - 48.8)^2}{48.8} = 85.39$$

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 distribution. In fact we could have stopped after working out the first term in the χ^2 value which is already 25.78! 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.

Solution 12.21

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 or columns, if you calculate the entries in one of the rows or columns first, you can work out the entries in the other row or column 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.

$$\chi^2 = \frac{(10 - 3.5)^2}{3.5} + \dots + \frac{(5 - 6.1)^2}{6.1} = 19.2$$

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^2_4 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 done 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.

$$\chi^2 = \frac{(22 - 18.30)^2}{18.30} + \dots + \frac{(30 - 28.32)^2}{28.32} = 1.89$$

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.

Past Exam Question (Subject C1, September 1994, Q15)

(i)(a) Each house independently must have the same probability of being burgled.

$$(i)(b) \quad L(p) = [P(X=0)]^{39} [P(X=1)]^{38} [P(X=2)]^{18} [P(X=3)]^4 P(X=5)$$

Using a $\text{Bin}(6, p)$ distribution to calculate the probabilities:

$$\begin{aligned} L(p) &= c[(1-p)^6]^{39} [p(1-p)^5]^{38} [p^2(1-p)^4]^{18} [p^3(1-p)^3]^4 p^5(1-p) \\ &= cp^{91}(1-p)^{509} \\ \Rightarrow \ln L(p) &= \ln c + 91\ln p + 509\ln(1-p) \\ \Rightarrow \frac{\partial}{\partial p} \ln L(p) &= \frac{91}{p} - \frac{509}{1-p} \end{aligned}$$

Setting the differential equal to zero to obtain the maximum:

$$\Rightarrow \frac{91}{\hat{p}} - \frac{509}{1-\hat{p}} = 0 \Rightarrow \hat{p} = \frac{91}{600}$$

Checking it's a maximum:

$$\frac{\partial^2}{\partial p^2} \ln L(p) = -\frac{91}{p^2} - \frac{509}{(1-p)^2} < 0 \Rightarrow \max$$

Alternatively, since the binomial distribution is additive, we could've looked at a single $\text{Bin}(600, p)$ distribution instead.

(i)(c) Using the estimate $\hat{p} = \frac{91}{600}$ we get frequencies of 37.3, 40.0, 17.9, 4.3, 0.6, 0.0, 0.0, using $P(X=x) = \binom{6}{x} \hat{p}^x (1-\hat{p})^{6-x}$. These are very similar to the observed frequencies – implying that it is a good fit.

(ii) 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

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

Calculating our statistic:

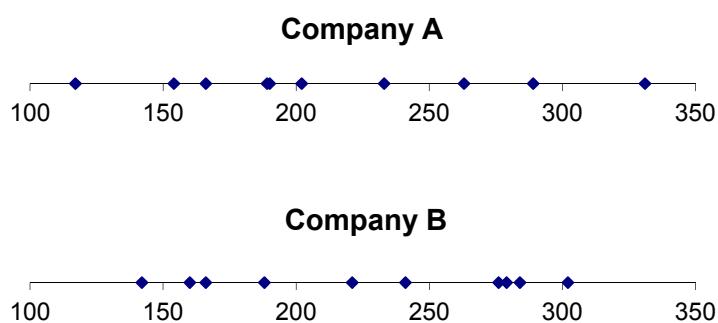
$$\chi^2 = \frac{(39 - 30.40)^2}{30.40} + \dots + \frac{(5 - 7.58)^2}{7.58} = 4.13$$

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.

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.

Exam-type question

- (i) We need a line plot showing the sample values for the two companies:



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.

(ii) We require the variances to be equal, so we are testing:

$$H_0: \sigma_A^2 = \sigma_B^2 \quad vs \quad H_1: \sigma_A^2 \neq \sigma_B^2$$

$$s_A^2 = \frac{1}{9} \left(494,126 - \frac{2,134^2}{10} \right) = 4,303.4 \quad s_B^2 = \frac{1}{9} \left(541,463 - \frac{2,259^2}{10} \right) = 3,461.7$$

Using $\frac{s_A^2/s_B^2}{\sigma_A^2/\sigma_B^2} \sim F_{n_A-1, n_B-1}$ we obtain a test statistic of:

$$\frac{4,303.4/3,461.7}{1} = 1.243$$

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 $\sigma_A^2 = \sigma_B^2$.

(iii) To test whether the premiums charged by Company B are higher than those charged by Company A, we carry out a two-sample t test. We are testing:

$$H_0: \mu_B = \mu_A \quad vs \quad H_1: \mu_B > \mu_A$$

Under this null hypothesis, we use:

$$\frac{\mu_B - \mu_A}{\sqrt{s_P^2 \left(\frac{1}{n_B} + \frac{1}{n_A} \right)}} \sim t_{n_A+n_B-2}$$

Substituting in the values, we get a test statistic of:

$$\frac{225.9 - 213.4}{\sqrt{\frac{9 \times 4303.4 + 9 \times 3461.7}{18} \left(\frac{1}{10} + \frac{1}{10} \right)}} = 0.4486$$

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.

- (iv) Using the pivotal value, from Chapter 11 of:

$$\frac{(\hat{p}_A - \hat{p}_B) - (p_A - p_B)}{\sqrt{\frac{\hat{p}_A \hat{q}_A}{n_A} + \frac{\hat{p}_B \hat{q}_B}{n_B}}} \stackrel{d}{\sim} N(0,1)$$

We have:

$$\hat{p}_A = 0.5, \quad \hat{q}_A = 0.5, \quad \hat{p}_B = 0.6, \quad \hat{q}_B = 0.4, \quad n_A = n_B = 10$$

We obtain a 95% confidence interval of:

$$-0.1 \pm 1.96 \sqrt{\frac{0.25}{10} + \frac{0.24}{10}} = (-0.53, 0.33)$$

Since this confidence interval contains zero, we cannot conclude that the proportions of premiums in excess of £200 are different for the companies.

- (v) We now carry out a single sample *t*-test on the data for Company A. We are testing:

$$H_0: \mu_A = 170 \quad \text{vs} \quad H_1: \mu_A > 170$$

Our test statistic is:

$$\frac{213.4 - 170}{\sqrt{4303.4/10}} = 2.092$$

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.

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.

Chapter 13

Correlation and regression

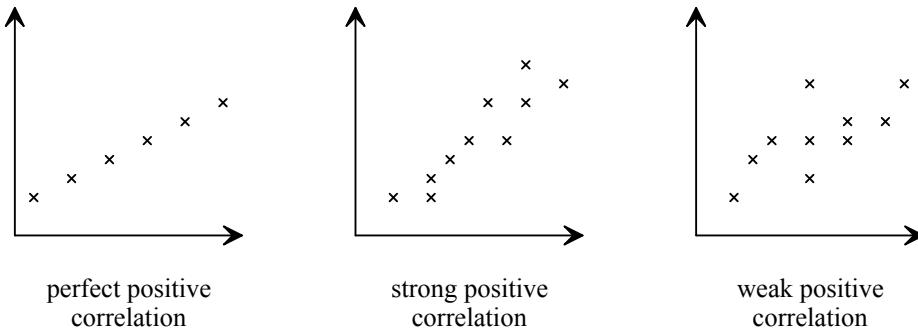


Syllabus objectives

- (xii) 1. Draw scatterplots for bivariate data and comment on them.
2. Define and calculate the correlation coefficient for bivariate data, explain its interpretation and perform statistical inference as appropriate.
3. Explain what is meant by response and explanatory variables.
4. State the usual simple regression model (with a single explanatory variable).
5. Derive and calculate the least squares estimates of the slope and intercept parameters in a simple linear regression model.
6. Perform statistical inference on the slope parameter in a simple linear regression.
7. Calculate R^2 (coefficient of determination) and describe its use to measure the goodness of fit of a linear regression model.
8. Use a fitted linear relationship to predict a mean response or an individual response with confidence limits.
9. Use residuals to check the suitability and validity of a linear regression model.
10. State the usual multiple linear regression model (with several explanatory variables).

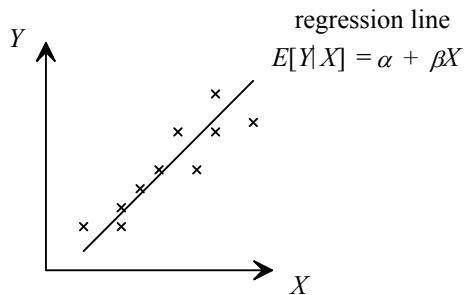
0 Introduction

This chapter examines the linear correlation between two variables, that is, it looks at how strong a linear relationship there is between them. The diagrams below show the various degrees of positive correlation:



You might recall that we met correlation in Chapter 6 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, \bar{X} , and the population mean, μ , in Chapters 9 to 12.

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 11. Due to the mathematically rigorous nature of this work, there are a number of results that are quoted without proof. Whilst students are required to memorise and apply these results, their proof lies outside the syllabus. Interested students should refer to the further reading material for more details.

This is a long chapter and will probably require two study sessions to cover it in detail. It would be wise to take your time, as the material often forms one of the longer questions on the Subject CT3 exam.

1 Bivariate data

1.1 Introduction

In this chapter relationships between variables are considered. This is one of the most important areas of statistical theory and practice – the methods of this chapter are perhaps more widely applied than any other statistical method.

There can be many variables involved, but initially in this chapter only the bivariate case (X, Y) will be considered. In addition, only linear relationships will be considered – the models assume that the expected value of Y , for any given value x of X , is a linear function of that value x ie:

$$E[Y | x] = \alpha + \beta x$$

Recall from Chapter 7 that $E[Y|x]$ is a conditional mean, which represents the average value of Y corresponding to a given value of x .

Correlation analysis

In a bivariate correlation analysis the problem of interest is an assessment of the strength of the linear relationship between Y and X .

Regression analysis

In a bivariate regression analysis the problem of interest is the nature of the relationship itself between Y , the response (or dependent) variable, and X , the explanatory (or independent, or regressor) variable. The analysis consists of choosing and fitting an appropriate 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. In this chapter only linear relationships will be considered.

Scatterplots

In any analysis, it is assumed that measurements (or counts) have been made, and are available, on the variables, giving us bivariate data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. The starting point with such data is always the same – draw a scatterplot and get a feel for the relationship (if any) between the variables as revealed/suggested by the data.

If a linear function for the conditional means is plausible, the variation of Y for fixed x values should be looked at to get a feel for how strong this linear relationship is.

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, *i.e.* into the form $Y = a + bX$.



Example

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:

$$Y' = \log a + x \log b$$

Example

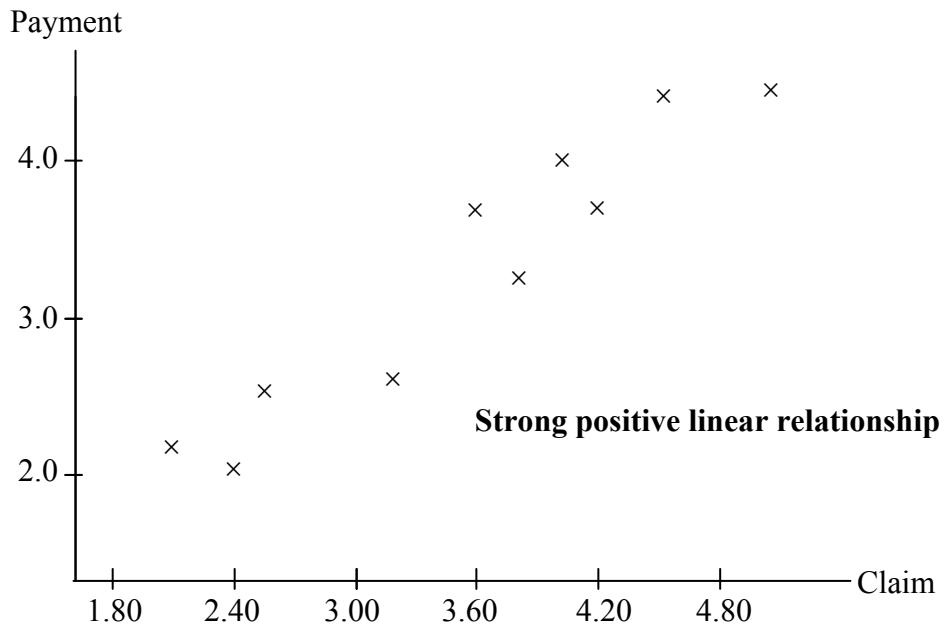
A sample of ten claims and corresponding payments on settlement for household policies is taken from the business of an insurance company.

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 for these data is as follows:



You can see from the graph that there appears to be a linear (*i.e.* straight line) relationship between the x and y values in this case.

In regression questions, $\sum_{i=1}^n x_i^2$ is often abbreviated to $\sum x^2$, etc to simplify the notation.

If a linear relationship (even a weak one) is indicated by the data, the methods of this chapter can be used. If so, the following calculations should be carried out – their values are needed in everything that follows.

Calculate the sums $\sum x_i$, $\sum y_i$ and the sums of squares $\sum x_i^2$, $\sum y_i^2$; calculate the sum of products $\sum x_i y_i$ then calculate the following three sums:

$$s_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum x_i^2 - n\bar{x}^2$$

$$s_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = \sum y_i^2 - n\bar{y}^2$$

$$s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = \sum x_i y_i - n\bar{x}\bar{y}$$


Question 13.1

Show that the first of these relationships is true, ie that:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum x_i^2 - n\bar{x}^2$$

These formulae are given on page 24 of the *Tables* in the $S_{xx} = \sum x_i^2 - n\bar{x}^2$ format.

For the example given, the following calculations can be made:

Number of pairs of observations $n = 10$.

$$\sum x = 35.4, \sum x^2 = 133.76, \sum y = 32.87, \sum y^2 = 115.2025, \sum xy = 123.81$$

$$S_{xx} = 133.76 - \frac{35.4^2}{10} = 8.444$$

$$S_{yy} = 115.2025 - \frac{32.87^2}{10} = 7.1588$$

$$S_{xy} = 123.81 - \frac{(35.4 \times 32.87)}{10} = 7.4502$$

1.2 Examples

The data given above, relating to insurance claims and payments, will continue to be used in the Core Reading. In addition, the following two sets of data will be used throughout this chapter in further examples and questions:

Foetal weights

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 foetus at fortnightly intervals during the pregnancy.

Gestation period (weeks)	30	32	34	36	38	40
Estimated foetal weight (kg)	1.6	1.7	2.5	2.8	3.2	3.5

AIDS cases

The numbers of new AIDS cases recorded in the US in successive years during the early part of the AIDS epidemic are shown in the table below.

Year	81	82	83	84	85	86	87	88
Number of cases (000s)	0.34	1.20	3.15	6.37	12.04	19.40	29.11	36.13

Source: www.avert.org

$$\sum x = 676 \quad \sum x^2 = 57,164 \quad \sum y = 107.74 \quad \sum y^2 = 2,726.1456 \quad \sum xy = 9,326.28$$



Question 13.2

Calculate S_{xx} , S_{yy} and S_{xy} for the foetal weights example.

2 Correlation analysis

2.1 Data summary

The association between the x and y values is summarised by the value of the sample correlation coefficient, which can be introduced as follows.

The sample correlation coefficient r is given by:

$$r = \frac{S_{xy}}{(S_{xx} S_{yy})^{1/2}}$$

Note:

1. r is such that $-1 \leq r \leq 1$.
2. r is a measure of linear association, and does not of itself indicate “cause and effect”. Two variables can be strongly associated, but the correlation may be “spurious”. The fact that the variables “move together” may be attributable to the effects on each of one or more other variables eg the variables may exhibit strong correlation simply because both are increasing with time. 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.

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.

For the claims settlement data:

$$r = \frac{7.4502}{(8.444 \times 7.1588)^{1/2}} = 0.958$$



Question 13.3

Calculate the linear correlation coefficient for the foetal weights example and comment.

Recall from Question 13.2 that $S_{xx} = 70$, $S_{yy} = 3.015$ and $S_{xy} = 14.3$.

Recall from Chapter 6 that the population correlation coefficient was defined to be:

$$\rho = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}}$$

Notice the similarity between the two formulae. The 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*.

2.2 The normal model and 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). 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 .

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, but you won’t be expected to know this in the exam.

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:

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\Big/2(1-\rho^2)\right)$$

where ρ is the correlation parameter, which takes values in the range $-1 < \rho < 1$.

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 $\rho \rightarrow \pm 1$, the bivariate distribution degenerates into a single line $\frac{Y-\mu_Y}{\sigma_Y} = \pm \frac{X-\mu_X}{\sigma_X}$ 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:

$$E(Y|X=x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)$$

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 skew (to the left ie negatively) – and has high spread/variability.

Remember skewed to the left means negatively skewed.

Two results are available (both of which are given on page 25 of the *Tables*).

Result 1

Under $H_0: \rho = 0$, $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ 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 .

Result 2 (Fisher's transformation of r)

This is a more general result – it is not restricted to the case $\rho = 0$.

If $W = \frac{1}{2} \ln \frac{1+r}{1-r}$, then W has (approximately) a normal distribution with mean $\frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ and standard deviation $\frac{1}{\sqrt{n-3}}$.

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 [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.

Example

Considering the data on claims and settlements, the test $H_0: \rho = 0.9$ vs $H_1: \rho > 0.9$ for the population of all claims/payments of this type is carried out as follows.

For the given data:

$$n = 10, r = 0.958, \text{ observed value of } W = 1.921$$

Under H_0 , W has a normal distribution with mean 1.472 and standard deviation 0.378. So $P(W > 1.921) = P\left(Z > \frac{(1.921 - 1.472)}{0.378}\right) = P(Z > 1.19) \approx 0.12$. 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.



Question 13.4

Carry out a test of $H_0: \rho = 0$ vs $H_1: \rho > 0$ for the data in the foetal weights example, using Result 1 and then Result 2.

Recall from Question 13.3 that $r = 0.984$.



Question 13.5

Test whether the data from the foetal weights example supports the hypothesis that the correlation parameter is greater than 0.9.

Recall from Question 13.3 that $r = 0.984$.

The proofs of these results are *not* required for the Subject CT3 exam, however the derivation of the t result and a reference for the Fisher’s transformation are included in the Appendix for interested students.

3 Regression analysis – the simple linear model

3.1 Introduction

The principal variable of interest is referred to as the response variable. The values of the response variable depend on, or are, in part, explained by, the values of the other variable, which is referred to as the explanatory variable.

In the notation we're using here the response variable is the Y value and the explanatory variable is the x value.

Ideally, the values used for the explanatory variable 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).

The 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. A prediction of the value of an individual response may also be needed. 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.

Given a set of n pairs of data (x_i, y_i) , $i = 1, 2, \dots, n$, 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 the 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:

$$Y_i = \alpha + \beta x_i + e_i \quad i = 1, 2, \dots, n$$

where the e_i are uncorrelated error variables with mean 0 and common variance σ^2 .

So $E[e_i] = 0$, $V[e_i] = \sigma^2$, $i = 1, 2, \dots, n$.

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.

3.2 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 $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

These are the equations you would use to calculate the “best” values of α and β .

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:

$$q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2$$

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 this 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 $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$.

The first of the two normal equations gives $\hat{\alpha}$ as:

$$\hat{\alpha} = \frac{\sum_{i=1}^n y_i - \hat{\beta} \sum_{i=1}^n x_i}{n}$$

This is just $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$.



Question 13.6

Produce a full derivation for the results $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$.

Note that a fitted line will pass through the point (\bar{x}, \bar{y}) .

$\hat{\beta}$ is the observed value of a statistic \hat{B} whose sampling distribution has the following properties:

$$E[\hat{B}] = \beta, \quad \text{var}[\hat{B}] = \frac{\sigma^2}{S_{xx}}$$

The derivation of these results is *not* required, but is included in the Appendix 6.2.

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 written as $\hat{\sigma}^2 = \frac{1}{n-2} (S_{yy} - S_{xy}^2 / S_{xx})$ and is given on page 24 of the *Tables*. We will see later that this is an unbiased estimator of σ^2 .



Question 13.7

Calculate the least squares estimators of the slope and intercept parameters for the foetal weights example.

Recall from Question 13.2 that $S_{xx} = 70$, $S_{yy} = 3.015$ and $S_{xy} = 14.3$.



Question 13.8

Calculate $\hat{\sigma}^2$ for the foetal weights example.

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 13.9

What would you expect the baby’s weight to be at 42 weeks (assuming it hasn’t been born by then)?

3.3 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 $S_{yy} = \sum (y_i - \bar{y})^2$ 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), \bar{y} (the overall average of the y values) and \hat{y}_i (the “predicted” value of y).

Squaring and summing both sides of:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

gives:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

the cross-product term vanishing.

This proof is *not* required but is included in Appendix 6.3 for interested students.

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 $\frac{SS_{RES}}{n - 2}$.

So:

$$SS_{TOT} = SS_{RES} + SS_{REG}$$

You may well see different abbreviations from these. For example, SS_{RES} is often written as SS_{ERR} ("error") or SS_U ("unexplained").

For computational purposes $SS_{TOT} = S_{yy}$ and:

$$SS_{REG} = \sum \left[(\hat{\alpha} + \hat{\beta} x_i) - (\hat{\alpha} + \hat{\beta} \bar{x}) \right]^2 = \hat{\beta}^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

The last step uses the fact that $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$.

$$\text{So } SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

Question 13.10

Split the total variation in the foetal weights model between the residual sum of squares and the regression sum of squares.

Recall from Question 13.2 that $S_{xx} = 70$, $S_{yy} = 3.015$ and $S_{xy} = 14.3$.

It can then be shown that:

$$E[SS_{TOT}] = (n - 1)\sigma^2 + \beta^2 S_{xx} \quad E[SS_{REG}] = \sigma^2 + \beta^2 S_{xx}$$

from which it follows that $E[SS_{RES}] = (n - 2)\sigma^2$.

The proofs of these results are *not* required but are included in Appendices 6.3, 6.4 and 6.5 for interested students.

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.

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.

In this case (the simple linear regression model), note that the value of the coefficient of determination is the square of the correlation coefficient for the data – since $r = \frac{S_{xy}}{(S_{xx}S_{yy})^{1/2}}$.

Considering again the data on claims and payments above, the linear regression of payments on claims and the coefficient of determination can be calculated as follows.

From the data given:

$$n = 10, \sum x = 35.4, \sum y = 32.87$$

$$S_{xx} = 8.444, S_{yy} = 7.1588, S_{xy} = 7.4502$$

$$\hat{\beta} = \frac{7.4502}{8.444} = 0.88231$$

$$\hat{\alpha} = 3.287 - (0.88231 \times 3.54) = 0.164$$

So the fitted regression line is $\hat{y} = 0.164 + 0.8823x$. Also:

$$SS_{TOT} = 7.1588$$

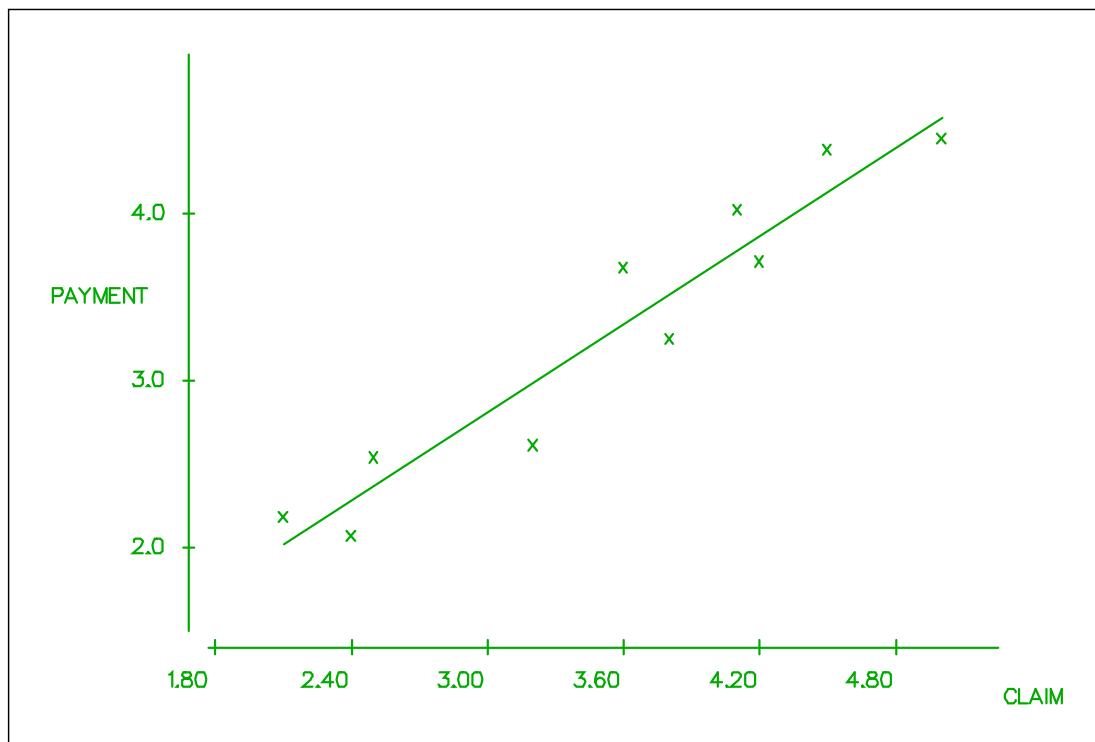
$$SS_{REG} = \frac{7.4502^2}{8.444} = 6.5734$$

$$\therefore SS_{RES} = 0.5854$$

$$\text{so } \hat{\sigma}^2 = \frac{0.5854}{8} = 0.0732.$$

$$R^2 = \frac{6.5734}{7.1588} = 0.918 \text{ (91.8%)}$$

The graph showing the fitted regression line and a plot of the data is as follows:



Question 13.11

Calculate R^2 for the foetal weights model.

3.4 The full normal model and inference

To go further, and make inferences concerning the responses based on the fitted model, the model must be specified further. In particular information is needed on the distributions of the Y_i 's.

In the full model, we now assume that the errors, e_i , are independent and identically distributed $N(0, \sigma^2)$. 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 $V[Y_i] = \sigma^2$.

\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) $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ has a χ^2 distribution with $v = n - 2$.

The derivation of these results is *not* required, however a reference for them is included in the Appendix for interested students.

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 , β , and σ^2 (since maximum likelihood estimation requires us to know the distribution whereas least squares estimation does not).

**Question 13.12** (hard)

Show that the maximum likelihood estimators of α and β are the same as before, but the MLE of σ^2 has a different denominator from before.

3.5 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 3.2:

$$M = (\hat{\beta} - \beta) / (\sigma^2 / S_{xx})^{1/2} \text{ is a standard normal variable}$$

Repeating result (2) from Section 3.4:

$$N = (n - 2)\hat{\sigma}^2 / \sigma^2 \text{ is a } \chi^2 \text{ variable with } v = n - 2 \text{ degrees of freedom}$$

Now, since $\hat{\beta}$ and $\hat{\sigma}^2$ are independent, it follows that $M / \{N / (n - 2)\}^{1/2}$ has a t distribution with $v = n - 2$, ie:

$$(\hat{\beta} - \beta) / \text{se}(\hat{\beta}) \text{ has a t distribution with } v = n - 2 \quad (*)$$

where the symbol $\text{se}(\hat{\beta})$ denotes the estimated standard error of $\hat{\beta}$, namely $(\hat{\sigma}^2 / S_{xx})^{1/2}$.

Result (*) 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 $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, 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*. A more detailed derivation is given in Appendix 6.7 for interested students.

Example

Considering again the claims/settlements example, it is possible to:

- (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 $H_1 : \beta \neq 1$.

Solution

(a) $se(\hat{\beta}) = (0.0732 / 8.444)^{1/2} = 0.0931$

95% confidence interval for β is $\hat{\beta} \pm \{t_{0.025, 8} \times se(\hat{\beta})\}$

i.e. $0.8823 \pm (2.306 \times 0.0931)$ i.e. 0.8823 ± 0.2147

So 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.

**Question 13.13**

Carry out a test of $H_0 : \beta = 0$ vs $H_1 : \beta > 0$ for the data in the foetal weights example, assuming a linear model is appropriate.

Recall from Question 13.7 that $\hat{\beta} = 0.2043$ and from Question 13.8 that $\hat{\sigma}^2 = 0.0234$.

**Question 13.14**

Carry out a test based on the linear regression model to determine whether the underlying rate of increase in AIDS cases exceeds 3,000 pa. The data for this example was given in Section 1.2.

3.6 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 $\mu_0 = E[Y|x_0] = \alpha + \beta x_0$), 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:

$$\text{var}(\hat{\mu}_0) = \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right\} \sigma^2$$

This result is given on page 25 of the *Tables*. Its derivation is *not* required, but is included in Appendix 6.8 for interested students.

The distribution actually used is a *t* distribution – the argument is similar to that described in Section 3.5

$$(\hat{\mu}_0 - \mu_0) / \text{se}[\hat{\mu}_0] \text{ has a } t \text{ distribution with } v = n - 2 \quad (**)$$

where $\text{se}[\hat{\mu}_0]$ denotes the estimated standard error of the estimate, namely:

$$\text{se}[\hat{\mu}_0] = \left[\left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right\} \hat{\sigma}^2 \right]^{\frac{1}{2}}$$

Result () can then 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 $E[Y|x_0]$ 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 in to the expression for the variance of the estimator of a mean response.

In other words, the variance of the individual response estimator is:

$$\text{var}(\hat{y}_0) = \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} \sigma^2$$

The result is:

$$(\hat{y} - y_0) / \text{se}[\hat{y}_0] \text{ has a } t \text{ distribution with } v = n - 2 \quad (\ast \ast \ast)$$

where $\text{se}[\hat{y}_0]$ denotes the estimated standard error of estimate, namely:

$$\text{se}[\hat{y}_0] = \left[\left\{ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} \hat{\sigma}^2 \right]^{1/2}$$

Result ($\ast \ast \ast$) 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 .



Question 13.15

- (i) Estimate the mean weight of a foetus at 35 weeks. What is the variance of this mean predicted response?
- (ii) Estimate the actual weight of an individual foetus at 35 weeks. What is the variance of this individual predicted response?

Recall that $\hat{\alpha} = -4.60$, $\hat{\beta} = 0.2043$, $\hat{\sigma}^2 = 0.0234$ and $S_{xx} = 70$.

Example

Consider again the claims/settlements example. It is now possible to find:

- (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.

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)

**Question 13.16**

Calculate a 90% confidence interval for the actual weight of the foetus at 33 weeks.

**Question 13.17**

Calculate a 95% confidence interval for the number of new AIDS cases recorded during 1994, assuming that the linear regression model is applicable for the period from 1981 to 1994. Comment on your answer.

Recall that $\hat{\beta} = 5.2917$ and $\hat{\sigma}^2 = 16.5140$ from Question 13.14.

3.7 Checking the model

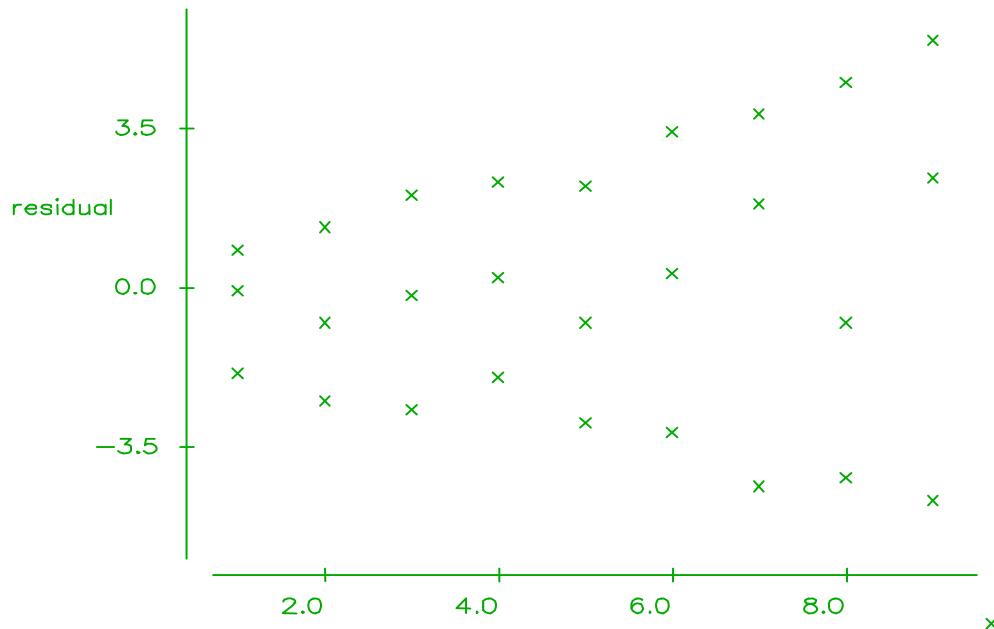
The residual from the fit at x_i is the estimated error, the difference between the response y_i and the fitted value ie:

$$\text{residual at } x_i \text{ is } \hat{e}_i = y_i - \hat{y}_i$$

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.

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:



This plot is for an imaginary set of data that we have not seen before.

In the above plot, 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.)



Question 13.18

Plot the residuals for the AIDS cases example using the simple linear regression model, and comment on your graph.

Recall that $\hat{\alpha} = -433.678$ and $\hat{\beta} = 5.2917$.

3.8 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 – $E[Y_i|x_i] = \alpha \exp(\beta x_i)$.

In such a case the response data can be transformed using $w_i = \log y_i$ and the linear model:

$$W_i = \eta + \beta x_i + e_i \text{ (where } \eta = \log \alpha \text{)}$$

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 study in Subject CT6.

4 The multiple linear regression model

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:

$$E[Y|X_1, X_2, \dots, X_k] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

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, \dots, x_k .

$\beta_1, \beta_2, \dots, \beta_k$ are known as the multiple regression coefficients. They are numerical constants which can be determined from observed data.

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 $x_{i1}, x_{i2}, \dots, x_{ik}$ by:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + e_i \quad i = 1, \dots, n$$

and so the least squares estimates of $\alpha, \beta_1, \beta_2, \dots, \beta_k$ are the values $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ for which:

$$q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})]^2$$

is minimised.

As for the simple linear model to find the estimates the above is differentiated partially with respect to α and $\beta_1, \beta_2, \dots, \beta_k$ in turn and the results are equated to zero.

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.

**Question 13.19**

A senior actuary wants to analyse the salaries of the 50 actuarial students employed by his company, using a linear model based on number of exam passes and years of experience. How would you express this model and the available data in terms of the notation given here?

5 Exam-type question



Exam-type question

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) Calculate the fitted linear regression equation of y on x .
- (ii) Assuming the full normal model, calculate an estimate of the error variance σ^2 and obtain a 90% confidence interval for σ^2 .
- (iii) By considering the slope parameter, formally test whether the data is positively correlated.
- (iv) Calculate a 95% confidence interval for the mean finals paper score corresponding to an individual entrance score of 53.
- (v) Test whether this data could come from a population with correlation coefficient equal to 0.75.
- (vi) Calculate the proportion of variation explained by the model. Hence, comment on the fit of the model.

6 Appendix – proofs

There are a number of results that are simply quoted in the Core Reading of this chapter (and so the proofs are *not* required) as the focus is on *applying* these results rather than deriving them.

This appendix details the various proofs for those students who may still wish to understand the mathematics behind the results. These proofs do not form part of the Core Reading for Subject CT3 and hence are *not* examinable.

The results included are as follows:

Section	Proof
1	$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$
2	$E[\hat{B}] = \beta$ and $\text{var}[\hat{B}] = \frac{\sigma^2}{S_{xx}}$
3	$SS_{TOT} = SS_{RES} + SS_{REG}$
4	$E[SS_{TOT}] = (n-1)\sigma^2 + \beta^2 S_{xx}$
5	$E[SS_{REG}] = \sigma^2 + \beta^2 S_{xx}$
6	$E[SS_{RES}] = (n-2)\sigma^2$
7	$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$
8	$\text{var}(\hat{\mu}_0) = \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} \sigma^2$

The results not included here are $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, \hat{B} and $\hat{\sigma}^2$ are independent and Fisher's transformation. The first two proofs can be found in:

Wilks, S. S., *Mathematical Statistics*. New York: John Wiley & Sons, Inc., 1962.

Fisher's transformation can be found in:

Kendall, M.G., and Stuart, A., *The Advanced Theory of Statistics*, Vol. 1, 4th ed. New York: Macmillan Publishing Co., Inc., 1977.

6.1 Proof of $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$

This result only holds true when $\rho = 0$, which is equivalent to saying that $\beta = 0$. It actually requires the “full normal model” from Section 3.4, in which case the distribution of $\hat{\beta}$ is:

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

This result is derived in Section 6.7. Now when $\beta = 0$ we have:

$$\frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

Using the definitions of $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\sigma}^2 = \frac{1}{n-2} \left[S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right]$ gives:

$$\frac{\frac{S_{xy}}{S_{xx}}}{\sqrt{\frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{(n-2)S_{xx}}}} = \frac{\frac{S_{xy}}{S_{xx}} \sqrt{n-2} \sqrt{S_{xx}}}{\sqrt{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}} = \frac{\frac{S_{xy}}{\sqrt{S_{xx}}} \sqrt{n-2}}{\sqrt{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}} \sim t_{n-2}$$

Dividing the numerator and denominator by $\sqrt{S_{yy}}$ and using $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ gives:

$$\frac{\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \sqrt{n-2}}{\sqrt{1 - \frac{S_{xy}^2}{S_{xx}S_{yy}}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

6.2 Proof of $E[\hat{\beta}] = \beta$ and $\text{var}[\hat{\beta}] = \frac{\sigma^2}{S_{xx}}$

Firstly, we need to rewrite the definition of $\hat{\beta}$ to make it easier to work with:

$$\begin{aligned}\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} \\ &= \frac{\sum (x_i - \bar{x})y_i - \bar{y} \sum (x_i - \bar{x})}{S_{xx}} \\ &= \frac{\sum (x_i - \bar{x})y_i - \bar{y}(\sum x_i - n\bar{x})}{S_{xx}} \\ &= \frac{\sum (x_i - \bar{x})y_i}{S_{xx}}\end{aligned}$$

Now recall that $\hat{\beta}$ is the random variable that has $\hat{\beta}$ as its realisation, so $\hat{\beta} = \frac{\sum (x_i - \bar{x})Y_i}{S_{xx}}$. We also need to recall that $E(Y_i) = \alpha + \beta x_i$. Hence:

$$\begin{aligned}E[\hat{\beta}] &= E\left[\frac{\sum (x_i - \bar{x})Y_i}{S_{xx}}\right] = \frac{\sum (x_i - \bar{x})E[Y_i]}{S_{xx}} \\ &= \frac{\sum (x_i - \bar{x})(\alpha + \beta x_i)}{S_{xx}} \\ &= \frac{\alpha \sum (x_i - \bar{x}) + \beta \sum x_i(x_i - \bar{x})}{S_{xx}} \\ &= \frac{\alpha(\sum x_i - n\bar{x}) + \beta(\sum x_i^2 - \bar{x}\sum x_i)}{S_{xx}}\end{aligned}$$

But since $\sum x_i = n\bar{x}$ the first term on the numerator is zero and we also have $\sum x_i^2 - \bar{x}\sum x_i = \sum x_i^2 - n\bar{x}^2 = S_{xx}$. Therefore:

$$E[\hat{\beta}] = \frac{0 + \hat{\beta}S_{xx}}{S_{xx}} = \beta$$

Using the fact that the Y_i 's are uncorrelated so that $\text{var}(\sum Y_i) = \sum \text{var}(Y_i)$ and recalling that $\text{var}(Y_i) = \sigma^2$ we get:

$$\begin{aligned}\text{var}[\hat{B}] &= \text{var}\left[\frac{\sum (x_i - \bar{x})Y_i}{S_{xx}}\right] \\ &= \frac{\sum (x_i - \bar{x})^2 \text{var}[Y_i]}{S_{xx}^2} \\ &= \frac{\sum (x_i - \bar{x})^2 \sigma^2}{S_{xx}^2} \\ &= \frac{\sigma^2}{S_{xx}^2} \sum (x_i - \bar{x})^2 \\ &= \frac{\sigma^2}{S_{xx}^2} \\ &= \frac{\sigma^2}{S_{xx}}\end{aligned}$$

6.3 Proof of $SS_{TOT} = SS_{RES} + SS_{REG}$

The *total* variation of the responses – the variance of the y values (*i.e.* how far the y values are from the mean) is given by:

$$SS_{TOT} = \sum (y_i - \bar{y})^2 = S_{yy}$$

We are trying to split this total variation up into the sum of two parts:

$$SS_{TOT} = SS_{REG} + SS_{RES}$$

SS_{REG} is the variation due to the *regression* model (*i.e.* the variance of the predicted y values, \hat{y}_i) given by:

$$SS_{REG} = \sum (\hat{y}_i - \bar{y})^2$$

and SS_{RES} is the variation that is left over (*i.e.* the *residual* variance) which tells us how far out the predicted y values, \hat{y}_i , are from the real y values. This is given by:

$$SS_{RES} = \sum (y_i - \hat{y}_i)^2$$

We show two methods of proving this result; the first way is harder but is the one hinted at in the Core Reading:

$$\begin{aligned} SS_{TOT} &= \sum (y_i - \bar{y})^2 \\ &= \sum [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2 \\ &= \sum (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum (y_i - \hat{y}_i)^2 \\ &= SS_{REG} + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + SS_{RES} \end{aligned}$$

All we have to do is show that the “cross product” term vanishes, *i.e.* we want to show that $2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$.

Now using our linear regression model:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \quad \text{and} \quad \bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$

Substituting these into the cross product term gives:

$$\begin{aligned} 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2\sum\left[y_i - (\hat{\alpha} + \hat{\beta}x_i)\right]\left[\hat{\alpha} + \hat{\beta}x_i - (\hat{\alpha} + \hat{\beta}\bar{x})\right] \\ &= 2\hat{\beta}\sum(y_i - \hat{\alpha} - \hat{\beta}x_i)(x_i - \bar{x}) \\ &= 2\hat{\beta}\left[\sum x_i y_i - \bar{x}\sum y_i - \hat{\alpha}\sum x_i + n\hat{\alpha}\bar{x} - \hat{\beta}\sum x_i^2 + \hat{\beta}\bar{x}\sum x_i\right] \end{aligned}$$

Now using $\sum x_i = n\bar{x}$ and $\sum y_i = n\bar{y}$ gives:

$$\begin{aligned} 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2\hat{\beta}\left[\sum x_i y_i - n\bar{x}\bar{y} - n\hat{\alpha}\bar{x} + n\hat{\alpha}\bar{x} - \hat{\beta}\sum x_i^2 + \hat{\beta}n\bar{x}^2\right] \\ &= 2\hat{\beta}\left(\sum x_i y_i - n\bar{x}\bar{y}\right) - 2\hat{\beta}^2\left(\sum x_i^2 - n\bar{x}^2\right) \end{aligned}$$

But using the definitions of S_{xy} and S_{xx} from page 24 of the *Tables* gives:

$$2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 2\hat{\beta}S_{xy} - 2\hat{\beta}^2S_{xx}$$

But $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$, so this gives:

$$\begin{aligned} 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2\frac{S_{xy}}{S_{xx}}S_{xy} - 2\frac{S_{xy}^2}{S_{xx}^2}S_{xx} \\ &= 2\frac{S_{xy}^2}{S_{xx}} - 2\frac{S_{xy}^2}{S_{xx}} \\ &= 0 \end{aligned}$$

The alternative, and frankly simpler proof, is as follows:

$$S_{TOT} = \sum (y_i - \bar{y})^2 = S_{yy}$$

We need to show that $SS_{REG} + SS_{RES} = S_{yy}$.

Since (\bar{x}, \bar{y}) lies on the regression line this means that $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$. Recall also that $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. This gives:

$$\begin{aligned} SS_{REG} &= \sum (\hat{y}_i - \bar{y})^2 = \sum [(\hat{\alpha} + \hat{\beta}x_i) - (\hat{\alpha} + \hat{\beta}\bar{x})]^2 \\ &= \sum [\hat{\beta}(x_i - \bar{x})]^2 \\ &= \hat{\beta}^2 \sum (x_i - \bar{x})^2 \\ &= \left(\frac{S_{xy}}{S_{xx}} \right)^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

So all we have to do is to show that $SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$. Using the results $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

and $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ gives:

$$\begin{aligned} SS_{RES} &= \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \\ &= \sum (y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i)^2 \\ &= \sum [(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})]^2 \\ &= \sum [(y_i - \bar{y})^2 - 2\hat{\beta}(x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}^2(x_i - \bar{x})^2] \\ &= \sum (y_i - \bar{y})^2 - 2\hat{\beta}\sum (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}^2 \sum (x_i - \bar{x})^2 \\ &= S_{yy} - 2\hat{\beta}S_{xy} + \hat{\beta}^2 S_{xx} \\ &= S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

6.4 Proof of $E[SS_{TOT}] = (n-1)\sigma^2 + \beta^2 S_{xx}$

Now $SS_{TOT} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$, so we require:

$$E[SS_{TOT}] = \sum E(y_i^2) - nE(\bar{y}^2)$$

Recall that:

$$E[y_i] = E[\alpha + \beta x_i + e_i] = \alpha + \beta x_i$$

$$\text{var}[y_i] = \text{var}[\alpha + \beta x_i + e_i] = \text{var}[e_i] = \sigma^2$$

$$\Rightarrow E(y_i^2) = \text{var}(y_i) + E^2(y_i) = \sigma^2 + (\alpha + \beta x_i)^2$$

Now:

$$E(\bar{y}) = E\left(\frac{1}{n} \sum y_i\right) = \frac{1}{n} \sum E(y_i) = \frac{1}{n} \sum (\alpha + \beta x_i) = \frac{1}{n} (n\alpha + \beta \sum x_i) = \alpha + \beta \bar{x}$$

$$\text{var}(\bar{y}) = \text{var}\left(\frac{1}{n} \sum y_i\right) = \frac{1}{n^2} \sum \text{var}(y_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$\Rightarrow E(\bar{y}^2) = \text{var}(\bar{y}) + E^2(\bar{y}) = \frac{\sigma^2}{n} + (\alpha + \beta \bar{x})^2$$

Substituting these into our equation for $E[SS_{TOT}]$ gives:

$$\begin{aligned} E[SS_{TOT}] &= \sum \left[\sigma^2 + (\alpha + \beta x_i)^2 \right] - n \left[\frac{\sigma^2}{n} + (\alpha + \beta \bar{x})^2 \right] \\ &= n\sigma^2 + \sum (\alpha + \beta x_i)^2 - \sigma^2 - n(\alpha + \beta \bar{x})^2 \\ &= (n-1)\sigma^2 + \sum (\alpha^2 + 2\alpha\beta x_i + \beta^2 x_i^2) - n(\alpha^2 + 2\alpha\beta \bar{x} + \beta^2 \bar{x}^2) \\ &= (n-1)\sigma^2 + n\alpha^2 + 2\alpha\beta \sum x_i + \beta^2 \sum x_i^2 - n\alpha^2 - 2\alpha\beta n\bar{x} - n\beta^2 \bar{x}^2 \end{aligned}$$

Since $\sum x_i = n\bar{x}$ we get:

$$E[SS_{TOT}] = (n-1)\sigma^2 + \beta^2 \left(\sum x_i^2 - n\bar{x}^2 \right)$$

$$= (n-1)\sigma^2 + \beta^2 S_{xx}$$

6.5 Proof of $E[SS_{REG}] = \sigma^2 + \beta^2 S_{xx}$

Rewriting SS_{REG} by recalling that $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ and that (\bar{x}, \bar{y}) lies on the regression line, so $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$ gives:

$$\begin{aligned} SS_{REG} &= \sum (\hat{y}_i - \bar{y})^2 = \sum [(\hat{\alpha} + \hat{\beta}x_i) - (\hat{\alpha} + \hat{\beta}\bar{x})]^2 \\ &= \sum [\hat{\beta}(x_i - \bar{x})]^2 = \hat{\beta}^2 \sum (x_i - \bar{x})^2 \\ &= \hat{\beta}^2 S_{xx} = \left(\frac{S_{xy}}{S_{xx}} \right)^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

So $E[SS_{REG}] = \frac{E[S_{xy}^2]}{S_{xx}}$, where $E[S_{xy}^2] = \text{var}[S_{xy}] + E^2[S_{xy}]$. To find these we need to rewrite the definition of S_{xy} as follows:

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x})y_i - \bar{y} \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x})y_i - \bar{y}(\sum x_i - n\bar{x}) = \sum (x_i - \bar{x})y_i \end{aligned}$$

Recalling that $E[y_i] = \alpha + \beta x_i$, $\text{var}[y_i] = \sigma^2$ and $\sum x_i = n\bar{x}$, we get:

$$\begin{aligned} E[S_{xy}] &= E\left[\sum (x_i - \bar{x})y_i\right] = \sum (x_i - \bar{x})E[y_i] = \sum (x_i - \bar{x})(\alpha + \beta x_i) \\ &= \alpha \sum (x_i - \bar{x}) + \beta \sum x_i(x_i - \bar{x}) \\ &= \alpha(\sum x_i - n\bar{x}) + \beta(\sum x_i^2 - \bar{x} \sum x_i) \\ &= \beta(\sum x_i^2 - n\bar{x}^2) = \beta S_{xx} \end{aligned}$$

$$\text{var}[S_{xy}] = \text{var}\left[\sum (x_i - \bar{x})y_i\right] = \sum (x_i - \bar{x})^2 \text{var}[y_i] = \sigma^2 \sum (x_i - \bar{x})^2 = \sigma^2 S_{xx}$$

$$\begin{aligned} \Rightarrow E[SS_{REG}] &= \frac{E[S_{xy}^2]}{S_{xx}} = \frac{\text{var}[S_{xy}] + E^2[S_{xy}]}{S_{xx}} = \frac{\sigma^2 S_{xx} + (\beta S_{xx})^2}{S_{xx}} \\ &= \sigma^2 + \beta^2 S_{xx} \end{aligned}$$

6.6 Proof of $E[SS_{RES}] = (n - 2)\sigma^2$

Now since $SS_{TOT} = SS_{RES} + SS_{REG}$, we have $SS_{RES} = SS_{TOT} - SS_{REG}$. Hence:

$$\begin{aligned} E[SS_{RES}] &= E[SS_{TOT}] - E[SS_{REG}] \\ &= \left[(n-1)\sigma^2 + \beta^2 S_{xx} \right] - \left[\sigma^2 + \beta^2 S_{xx} \right] \\ &= (n-2)\sigma^2 \end{aligned}$$

6.7 Proof of $\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$

Now $\hat{\beta}$ can be rewritten:

$$\begin{aligned}\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} = \frac{\sum (x_i - \bar{x})y_i - \bar{y} \sum (x_i - \bar{x})}{S_{xx}} \\ &= \frac{\sum (x_i - \bar{x})y_i - \bar{y}(\sum x_i - n\bar{x})}{S_{xx}} = \frac{\sum (x_i - \bar{x})y_i}{S_{xx}}\end{aligned}$$

Now recall that $\hat{\beta}$ is the random variable that has $\hat{\beta}$ as its realisation, so $\hat{\beta} = \frac{\sum (x_i - \bar{x})Y_i}{S_{xx}}$. In the full normal model, we have $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, since $\hat{\beta}$ is a linear combination of Y_i 's it also has a normal distribution and since $E[\hat{\beta}] = \beta$ and $\text{var}[\hat{\beta}] = \frac{\sigma^2}{S_{xx}}$ we have:

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

Hence by standardising, we get:

$$\frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1)$$

From Chapter 9, $t_k \equiv \frac{N(0,1)}{\sqrt{\chi^2_k / k}}$, where the $N(0,1)$ and χ^2 are independent. Now using the results that $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$ and that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent gives:

$$\frac{\frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / S_{xx}}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} / (n-2)}} = \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$$

$$6.8 \quad \text{Proof of } \text{var}(\hat{\mu}_0) = \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} \sigma^2$$

Recall that $\mu_0 = E[Y | x_0] = \alpha + \beta x_0$ is just shorthand for the expected value of y when $x = x_0$. Hence $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta}x_0$, where $\text{var}[\hat{B}] = \frac{\sigma^2}{S_{xx}}$ from section 6.2.

Let \hat{A} is the random variable that has $\hat{\alpha}$ as its realisation and recall that \hat{B} is the random variable that has $\hat{\beta}$ as its realisation. For this proof we will require the following two results:

$$\text{var}(\hat{A}) = \frac{\sigma^2}{nS_{xx}}(S_{xx} + n\bar{x}^2) \quad (1)$$

$$\text{cov}(\hat{A}, \hat{B}) = -\frac{\bar{x}\sigma^2}{S_{xx}} \quad (2)$$

So:

$$\begin{aligned} \text{var}(\hat{\mu}_0) &= \text{var}(\hat{A} + \hat{B}x_0) \\ &= \text{var}(\hat{A}) + \text{var}(\hat{B}x_0) + 2\text{cov}(\hat{A}, \hat{B}x_0) \\ &= \text{var}(\hat{A}) + x_0^2 \text{var}(\hat{B}) + 2x_0 \text{cov}(\hat{A}, \hat{B}) \\ &= \frac{\sigma^2}{nS_{xx}}(S_{xx} + n\bar{x}^2) + x_0^2 \frac{\sigma^2}{S_{xx}} - 2x_0 \frac{\bar{x}\sigma^2}{S_{xx}} \quad \text{using (1) and (2)} \\ &= \left\{ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{x_0^2}{S_{xx}} - 2 \frac{x_0\bar{x}}{S_{xx}} \right\} \sigma^2 \\ &= \left\{ \frac{1}{n} + \frac{x_0^2 - 2x_0\bar{x} + \bar{x}^2}{S_{xx}} \right\} \sigma^2 \\ &= \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} \sigma^2 \end{aligned}$$

We shall now derive the two results required for this proof.

6.9 Proof of $\text{var}(\hat{A}) = \frac{\sigma^2}{nS_{xx}}(S_{xx} + n\bar{x}^2)$

First we shall rewrite $\hat{\alpha}$:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{1}{n} \sum y_i - \frac{S_{xy}}{S_{xx}} \bar{x} = \frac{S_{xx} \sum y_i - nS_{xy}\bar{x}}{nS_{xx}}$$

Recall we can rewrite S_{xy} as follows:

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x})y_i - \bar{y} \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x})y_i - \bar{y}(\sum x_i - n\bar{x}) = \sum (x_i - \bar{x})y_i \end{aligned}$$

Therefore:

$$\hat{\alpha} = \frac{S_{xx} \sum y_i - n\bar{x} \sum (x_i - \bar{x})y_i}{nS_{xx}}$$

Now since \hat{A} is the random variable that has $\hat{\alpha}$ as its realisation we have $\hat{A} = \frac{S_{xx} \sum Y_i - n\bar{x} \sum (x_i - \bar{x})Y_i}{nS_{xx}}$. Hence using the fact that the Y_i 's are uncorrelated so

that $\text{var}(\sum Y_i) = \sum \text{var}(Y_i)$ and recalling that $\text{var}(Y_i) = \sigma^2$, we get:

$$\begin{aligned} \text{var}[\hat{A}] &= \text{var} \left[\frac{S_{xx} \sum Y_i - n\bar{x} \sum (x_i - \bar{x})Y_i}{nS_{xx}} \right] \\ &= \frac{S_{xx}^2 \sum \text{var}[Y_i] + n^2 \bar{x}^2 \sum (x_i - \bar{x})^2 \text{var}[Y_i]}{n^2 S_{xx}^2} \\ &= \frac{nS_{xx}^2 \sigma^2 + n^2 \bar{x}^2 \sigma^2 \sum (x_i - \bar{x})^2}{n^2 S_{xx}^2} \\ &= \frac{nS_{xx}^2 \sigma^2 + n^2 \bar{x}^2 \sigma^2 S_{xx}}{n^2 S_{xx}^2} \\ &= \frac{\sigma^2}{nS_{xx}} (S_{xx} + n\bar{x}^2) \end{aligned}$$

6.10 Proof of $\text{cov}(\hat{A}, \hat{B}) = -\frac{\bar{x}\sigma^2}{S_{xx}}$

Now $\hat{A} = \frac{S_{xx} \sum Y_i - n\bar{x} \sum (x_i - \bar{x})Y_i}{nS_{xx}}$ from above and $\hat{B} = \frac{\sum (x_i - \bar{x})Y_i}{S_{xx}}$ from

Section 6.7. Since the Y_i 's are uncorrelated we have $\text{cov}(Y_i, Y_j) = 0$ $i \neq j$. So:

$$\begin{aligned}\text{cov}(\hat{A}, \hat{B}) &= \text{cov}\left(\frac{S_{xx} \sum Y_i - n\bar{x} \sum (x_i - \bar{x})Y_i}{nS_{xx}}, \frac{\sum (x_i - \bar{x})Y_i}{S_{xx}}\right) \\ &= \frac{S_{xx}}{nS_{xx}^2} \sum (x_i - \bar{x}) \text{cov}(Y_i, Y_i) - \frac{n\bar{x}}{nS_{xx}^2} \sum (x_i - \bar{x})^2 \text{cov}(Y_i, Y_i)\end{aligned}$$

Now $\text{cov}(Y_i, Y_i) = \text{var}(Y_i) = \sigma^2$, so we get:

$$\begin{aligned}\text{cov}(\hat{A}, \hat{B}) &= \frac{S_{xx}}{nS_{xx}^2} \sigma^2 \sum (x_i - \bar{x}) - \frac{n\bar{x}}{nS_{xx}^2} \sigma^2 \sum (x_i - \bar{x})^2 \\ &= \frac{\sigma^2}{nS_{xx}^2} \left[S_{xx} (\sum x_i - n\bar{x}) - n\bar{x} S_{xx} \right]\end{aligned}$$

Recall that $\sum x_i = n\bar{x}$, which gives:

$$\text{cov}(\hat{A}, \hat{B}) = \frac{\sigma^2}{nS_{xx}^2} [0 - n\bar{x}S_{xx}] = -\frac{\sigma^2 \bar{x}}{S_{xx}}$$



Chapter 13 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 sample correlation coefficient, r , measures the strength of the linear relationship between x and y . The formula is given overleaf and on page 25 of the *Tables*.

We can carry out hypothesis tests on the population correlation coefficient, ρ , using the t result or the Fisher-Z test. Both of these results are given overleaf and on page 25 of the *Tables*.

The linear regression model is given by:

$$Y_i = \alpha + \beta x_i + e_i \quad \text{where } e_i \sim N(0, \sigma^2)$$

The parameters α , β and σ^2 can be estimated using the formulae overleaf and on page 24 of the *Tables*.

Confidence intervals can be obtained for β and the predicted individual (or mean) response y using the formulae given overleaf and on pages 24 and 25 of the *Tables*.

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 coefficient of determination, R^2 , gives the percentage of this variance which is explained by the model. The formula is given overleaf but not in the *Tables*.

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 can be examined using diagrams.



Chapter 13 Formulae

Correlation

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$H_0 : \rho = 0 \quad \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

$$\text{otherwise} \quad \frac{1}{2} \ln \frac{1+r}{1-r} \sim N \left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right)$$

Regression

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)$$

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$$

$$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \hat{\sigma}^2}} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{y}_0 - y_0}{\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \hat{\sigma}^2}} \sim t_{n-2}$$

Fit of model

$$SS_{TOT} = SS_{RES} + SS_{REG}$$

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

$$\hat{e}_i = y_i - \hat{y}_i$$

Chapter 13 Solutions

Solution 13.1

Proving the first of these relationships:

$$\begin{aligned} S_{xx} &= \sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2 \\ &= \sum x_i^2 - \frac{2(\sum x_i)^2}{n} + n\frac{(\sum x_i)^2}{n^2} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \end{aligned}$$

Now since $\sum x_i = n\bar{x}$, we get:

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum x_i^2 - \frac{(n\bar{x})^2}{n} = \sum x_i^2 - n\bar{x}^2$$

Solution 13.2

$$S_{xx} = \sum x^2 - \frac{1}{n}(\sum x)^2 = 7,420 - \frac{1}{6} \times 210^2 = 70$$

$$S_{yy} = \sum y^2 - \frac{1}{n}(\sum y)^2 = 42.03 - \frac{1}{6} \times 15.3^2 = 3.015$$

$$S_{xy} = \sum xy - \frac{1}{n}(\sum x)(\sum y) = 549.8 - \frac{1}{6} \times 210 \times 15.3 = 14.3$$

Solution 13.3

Using the results from Question 13.2:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{14.3}{\sqrt{70 \times 3.015}} = 0.984$$

There is a strong linear association between gestation period and foetal weight.

Solution 13.4

Using Result 1:

We are testing $H_0 : \rho = 0$ vs $H_1 : \rho > 0$.

If H_0 is true, then the test statistic $\frac{r\sqrt{4}}{\sqrt{1-r^2}}$ has a t_4 distribution.

The observed value of this statistic is $\frac{0.984 \times 2}{\sqrt{1 - 0.984^2}} = 11.05$. This is much greater than 4.604, the upper 0.5% point of the t_4 distribution.

So, we reject H_0 at the 0.5% 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 time.

Using Result 2:

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 $\frac{1}{2} \log \frac{1+0.984}{1-0.984} = 2.41$, which corresponds to a value of $\frac{2.41}{\sqrt{\frac{1}{3}}} = 4.17$ 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 linear correlation between the baby's weight and time.

Solution 13.5

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 $z_\rho = \frac{1}{2} \log \frac{1+0.9}{1-0.9} = 1.47$.

The observed value of this statistic is $\frac{1}{2} \log \frac{1+0.984}{1-0.984} = 2.41$, which corresponds to a value of $\frac{2.41-1.47}{\sqrt{\frac{1}{3}}} = 1.63$ 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 *i.e.* the data does not provide enough evidence to conclude that the correlation parameter between the baby's weight and time exceeds 0.9.

Solution 13.6

We need to minimise $q = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2$. To do this, we partially differentiate q

with respect to α and β and set the derivatives equal to zero. Dropping the subscripts for convenience:

$$\begin{aligned}\frac{\partial q}{\partial \alpha} &= \sum 2[y - (\alpha + \beta x)] \times -1 \\ &= 0 \text{ when } \sum y - n\hat{\alpha} - \hat{\beta} \sum x = 0\end{aligned}\tag{1}$$

$$\begin{aligned}\frac{\partial q}{\partial \beta} &= \sum 2[y - (\alpha + \beta x)] \times (-x) \\ &= 0 \text{ when } \sum xy - \hat{\alpha} \sum x - \hat{\beta} \sum x^2 = 0\end{aligned}\tag{2}$$

Solving these simultaneously:

$$(2) \times n - (1) \times \sum x :$$

$$n \sum xy - \sum x \sum y + \hat{\beta} (\sum x)^2 - \hat{\beta} n \sum x^2 = 0$$

$$\Rightarrow \hat{\beta} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

By dividing this expression by n in the numerator and denominator, and noting that $\sum y = n \bar{y}$:

$$\hat{\beta} = \frac{\sum xy - n \bar{x} \bar{y}}{\sum x^2 - n \bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

Returning to equation (1), and dividing by n :

$$\bar{y} - \hat{\alpha} - \hat{\beta} \bar{x} = 0 \Rightarrow \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

We could also solve this using substitution. Since this proof crops up quite regularly in the exams it is worth spending time learning it.

Solution 13.7

Using the values previously calculated for S_{xy} and S_{xx} :

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{14.3}{70} = 0.2043$$

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:

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = 2.55 - 0.2043 \times 35 = -4.60$$

Solution 13.8

Using the second form of the formula:

$$\hat{\sigma}^2 = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \frac{1}{4} \left(3.015 - \frac{14.3^2}{70} \right) = 0.0234$$

Solution 13.9

Using the estimated values of α and β :

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = -4.60 + 0.2043 \times 42 = 3.98\text{kg}$$

Solution 13.10

For the foetal weights, $S_{xx} = 70$, $S_{yy} = 3.015$ and $S_{xy} = 14.3$. So:

$$SS_{TOT} = 3.015$$

$$SS_{REG} = \frac{14.3^2}{70} = 2.921$$

$$SS_{RES} = 3.015 - 2.921 = 0.094$$

So we see that, in this case, most of the variation is explained by the model.

Solution 13.11

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{2.921}{3.015} = 0.969 \text{ or } 96.9\%$$

Alternatively, using $r = 0.984$ from Question 13.3:

$$R^2 = r^2 = 0.984^2 = 96.8\%$$

The slight difference is due to using a rounded value of r .

Solution 13.12

Each Y_i has a $N(\alpha + \beta x_i, \sigma^2)$ distribution, so the joint likelihood function is:

$$L = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{y_i - \alpha - \beta x_i}{\sigma} \right)^2 \right] = \frac{1}{\sigma^n (\sqrt{2\pi})^n} \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2} \right]$$

Taking logs we get:

$$\log L = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 + \text{constant}$$

Differentiating with respect to β and then with respect to α :

$$\frac{\partial \log L}{\partial \beta} = -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(y_i - \alpha - \beta x_i) \times (-x_i)$$

$$\frac{\partial \log L}{\partial \alpha} = -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(y_i - \alpha - \beta x_i) \times (-1)$$

By setting $\frac{\partial \log L}{\partial \alpha}$ equal to 0 we get $\sum_{i=1}^n y_i - n\hat{\alpha} - \hat{\beta} \sum_{i=1}^n x_i = 0$.

By setting $\frac{\partial \log L}{\partial \beta}$ equal to 0 we get $\sum_{i=1}^n y_i x_i - \hat{\alpha} \sum_{i=1}^n x_i - \hat{\beta} \sum_{i=1}^n x_i^2 = 0$.

These are the same normal equations that we got before, so the MLEs are as before, ie:

$$\hat{\beta} = \frac{n \left(\sum_{i=1}^n x_i y_i \right) - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2} = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\alpha} = \frac{\sum_{i=1}^n y_i - \hat{\beta} \sum_{i=1}^n x_i}{n} = \bar{y} - \hat{\beta} \bar{x}$$

Checking that they give the maximum, we differentiate again:

$$\frac{\partial^2 \log L}{\partial \beta^2} = -\frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 < 0 \Rightarrow \max \quad \frac{\partial \log L}{\partial \alpha} = -\frac{1}{\sigma^2} \sum_{i=1}^n 1 < 0 \Rightarrow \max$$

Now differentiating the log likelihood with respect to σ :

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma} &= -\frac{n}{\sigma} - \frac{1}{2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \times -2\sigma^{-3} \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \end{aligned}$$

By setting $\frac{\partial \log L}{\partial \sigma}$ equal to 0 and substituting $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, we obtain:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n [(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})]^2 \\ &= \frac{1}{n} \left[\sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{\beta} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= \frac{1}{n} \left[S_{yy} - 2\hat{\beta}S_{xy} + \hat{\beta}^2 S_{xx} \right] \\ &= \frac{1}{n} \left[S_{yy} - 2 \frac{S_{xy}}{S_{xx}} S_{xy} + \left(\frac{S_{xy}}{S_{xx}} \right)^2 S_{xx} \right] \\ &= \frac{1}{n} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) \end{aligned}$$

which has a different denominator from before (and therefore is an *biased*) estimator.

Solution 13.13

If H_0 is true, then the test statistic $\frac{\hat{\beta} - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$ has a t_4 distribution.

The observed value of this statistic is $\frac{0.2043 - 0}{\sqrt{0.0234/70}} = 11.2$, which is much greater than 4.604, the upper 0.5% point of the t_4 distribution.

So, we reject H_0 at the 0.5% level and conclude that there is very strong evidence that $\beta > 0$ ie that the baby's weight is increasing over time.

Solution 13.14

An annual rate of increase of 3,000 corresponds to $\beta = 3$.

So, we need to test $H_0 : \beta = 3$ vs $H_1 : \beta > 3$.

If H_0 is true, then the test statistic $\frac{\hat{\beta} - 3}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$ has a t_6 distribution.

We have:

$$S_{xx} = 42, S_{yy} = 1,275.157, S_{xy} = 222.25$$

So:

$$\hat{\beta} = \frac{222.25}{42} = 5.2917 \text{ and } \hat{\sigma}^2 = \frac{1}{6} \left(1,275.157 - \frac{222.25^2}{42} \right) = 16.5140$$

So, the observed value of this statistic is $\frac{5.2917 - 3}{\sqrt{16.5140/42}} = 3.65$, which is certainly greater than 1.943 the upper 5% point of the t_6 distribution and it's even greater than 3.143, the upper 1% point.

So, we reject H_0 at even the 1% level and conclude that there is very strong evidence that $\beta > 3$ ie that the rate of increase exceeds 3,000 cases per year.

Solution 13.15

- (i) The parameter values were estimated to be:

$$\hat{\beta} = 0.2043 \text{ and } \hat{\alpha} = -4.60$$

So the least squares regression line is:

$$\hat{y} = -4.60 + 0.2043x$$

When $x_0 = 35$, $\hat{\mu}_0 = -4.60 + 0.2043 \times 35 = 2.55$ ie the mean weight of a foetus at 35 weeks is expected to be 2.55kg.

The variance of this estimate is calculated as:

$$\begin{aligned}\text{var}(\hat{\mu}_0) &= \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} \sigma^2 \\ &= \left\{ \frac{1}{6} + \frac{(35 - 35)^2}{70} \right\} \times 0.0234 = 0.0039\end{aligned}$$

- (ii) The *individual* predicted response is also $\hat{y}_0 = 2.55$ kg.

The variance of this estimate is calculated as:

$$\begin{aligned}\text{var}(\hat{y}_0) &= \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} \sigma^2 \\ &= \left\{ 1 + \frac{1}{6} + \frac{(35 - 35)^2}{70} \right\} \times 0.0234 = 0.0273\end{aligned}$$

Solution 13.16

Here we require a confidence interval for the *individual* predicted response:

$$\hat{y}_0 \pm t_{0.05,4} \sqrt{\left(1 + \frac{1}{6} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \hat{\sigma}^2}$$

Using the values found previously:

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0 = -4.60 + 0.2043 \times 33 = 2.1419$$

So the confidence interval is given by:

$$2.1419 \pm 2.132 \sqrt{\left(1 + \frac{1}{6} + \frac{(33 - 35)^2}{70}\right) \times 0.0234} = 2.14 \pm 0.36$$

So the required 90% confidence interval is (1.78, 2.50).

Solution 13.17

Here we require a confidence interval for the *individual* predicted response:

$$\hat{y}_0 \pm t_{0.025,6} \sqrt{\left(1 + \frac{1}{8} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \hat{\sigma}^2}$$

$\hat{\beta} = 5.2917$, $S_{xx} = 42$ and $\hat{\sigma}^2 = 16.5140$ (calculated earlier), so:

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = 13.4675 - 5.2917 \times 84.5 = -433.678$$

$$\Rightarrow \hat{y}_0 = -433.678 + 5.2917 \times 94 = 63.74$$

The required confidence interval is found as follows:

$$63.74 \pm 2.447 \sqrt{\left(1 + \frac{1}{8} + \frac{(94 - 84.5)^2}{42}\right) \times 16.5140} = 63.4 \pm 17.99$$

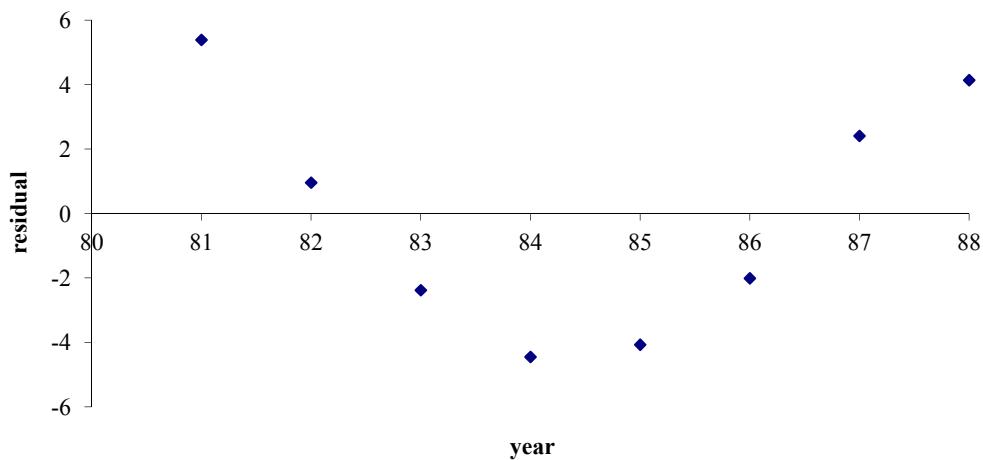
So the required 95% confidence interval is (45.7, 81.7).

This is a prediction outside the range of values of the data. It is unlikely to be reliable since increased awareness of AIDS is likely to have affected the pattern of incidence of new cases.

Solution 13.18

The residuals for the simple linear regression model, $\hat{y} = -433.678 + 5.2917x$, show a definite pattern and some of them are quite large. This shows that the linear regression model is *not* a good fit.

x_i	81	82	83	84	85	86	87	88
y_i	0.34	1.20	3.15	6.37	12.04	19.40	29.11	36.13
\hat{y}_i	-5.05	0.24	5.53	10.82	16.11	21.41	26.70	31.99
$\hat{e}_i = y_i - \hat{y}_i$	5.39	0.96	-2.38	-4.45	-4.07	-2.01	2.41	4.14



The model is not a good fit as the growth in AIDS cases is exponential and not linear.

Solution 13.19

The basic model would be:

$$E[Y | x_1, x_2] = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Here x_1 represents a certain number of exam passes, x_2 represents a certain number of years experience and Y would represent the corresponding salary.

α , β_1 and β_2 are constants.

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.

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.

Exam-type question

- (i) Calculating the sum of squares:

$$S_{xx} = 60,016 - \frac{836^2}{12} = 1,774.67$$

$$S_{xy} = 1,112$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1,122}{1,774.67} = 0.63223$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 72.25 - 0.63223 \times 69.667 = 28.205$$

Hence, the fitted regression equation of y on x is $y = 28.205 + 0.63223x$.

(ii) We have $S_{yy} = 63,603 - \frac{867^2}{12} = 962.25$, so:

$$\hat{\sigma}^2 = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \frac{1}{10} \left(962.25 - \frac{1,122^2}{1,774.6} \right) = 25.289$$

Now $\frac{10\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{10}$, which gives a confidence interval for σ^2 of $\left(\frac{10 \times 25.289}{18.31}, \frac{10 \times 25.289}{3.94} \right) = (13.8, 64.2)$.

(iii) We are testing $H_0 : \beta = 0$ vs $H_1 : \beta > 0$.

Now $\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{10}$. Our observed value here is:

$$\frac{0.63223 - 0}{\sqrt{25.289 / 1774.67}} = 5.296$$

This is a highly significant result, compared which exceeds the 0.5% critical value 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 is positively correlated).

(iv) The variance of the distribution of the mean finals score corresponding to an entrance score of 53 is:

$$\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \hat{\sigma}^2 = \left[\frac{1}{12} + \frac{(53 - 69.667)^2}{1,774.67} \right] \times 25.289 = 6.0657$$

The predicted value is $28.205 + 0.63223 \times 53 = 61.713$. We have a t_{10} distribution, so the 95% confidence intervals is:

$$61.713 \pm 2.228 \times \sqrt{6.0657} = (56.2, 67.2)$$

- (v) We are testing $H_0 : \rho = 0.75$ vs $H_1 : \rho \neq 0.75$

If H_0 is true, then the test statistic Z_r has a $N\left(z_\rho, \frac{1}{9}\right)$ distribution, where

$$z_\rho = \frac{1}{2} \log \frac{1+0.75}{1-0.75} = 0.97296.$$

The observed value of this statistic is $\frac{1}{2} \log \frac{1+0.85860}{1-0.85860} = 1.2880$, which corresponds to a value of $\frac{1.2880 - 0.97296}{\sqrt{\frac{1}{9}}} = 0.945$ on the $N(0,1)$ distribution.

This is clearly less than 1.96, the upper 2.5% point of the standard normal distribution.

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.

- (vi) The proportion of variability explained by the model is given by R^2 :

$$R^2 = r^2 = \left(\frac{1,122}{\sqrt{1,774.67 \times 962.25}} \right)^2 = 0.85860^2 = 73.7\%$$

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.

Chapter 14

Analysis of variance



Syllabus objectives

- (xiii) 1. *Describe the circumstances in which a one-way analysis can be used.*
2. *State the usual model for a one-way analysis of variance and explain what is meant by the term treatment effects.*
3. *Perform a simple one-way analysis of variance.*

0 Introduction

In this chapter the work of Chapter 12, Section 4 will be generalised. We will investigate the problem of deciding whether the observed differences between more than two sample means are purely random or whether there are actually real differences between the sample means.

We will look at experiments designed to enable the effects of a number of different “treatments” on the mean value of some response variable to be compared. In many situations the effects to be investigated run the risk of being masked by variability present in extraneous factors not of direct concern – in such cases a sensible experimental design is essential to obtain really useful data.

A good design will incorporate randomization (eg the various treatments may be allocated to the experimental units at random) and replication (using each treatment more than once). This enables the investigation to concentrate on the effects of interest and the analysis of the results provides estimates of the treatment effects and of the extraneous, uncontrolled, or residual variation present.

The technique of analysis of variance consists of separating the total variability in a set of experimental results into components associated with the different sources of that variability. These components are then compared and this enables us to test the null hypothesis that no differences exist between the (population) treatment means.

Despite its name, analysis of variance (frequently abbreviated to ANOVA) is actually testing whether means are the same. To do this we examine the variance in a similar way to our regression analysis from Chapter 13.

Recall that the total variability was separated as $SS_{TOT} = SS_{REG} + SS_{RES}$, where SS_{REG} was the variability explained by the model and SS_{RES} was the residual variability. In ANOVA, the total variability is separated as $SS_T = SS_R + SS_B$, where SS_R corresponds to the variability explained by the model and SS_B corresponds to the variability between the different treatments.

In Chapter 13 we used a simple ratio, $\frac{SS_{REG}}{SS_{TOT}}$, to see what proportion of the variability was due to the model. In ANOVA, we wish to see if the variability between the different treatments is significant, so we can say that the means are different. For this,

we will use the two population variance test from Chapter 12, ie $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$.

This chapter starts by dealing with the theory behind the test and then shows how the test is used in practice. We would advise that you read this chapter through twice. On the first reading just concentrate on being able to carry out the ANOVA test numerically. Then, with this understanding, re-read the chapter looking at the algebra behind it more closely.

1 One-way analysis of variances

1.1 The model

A one-way analysis of variance is used to compare k treatments when the experiment provides n_i responses for treatment i , $i = 1, 2, \dots, k$. The data available are then $n = \sum_i n_i$ responses y_{ij} , where y_{ij} is the j th observation using treatment i .

To help get to grips with this new notation, consider a company providing pet insurance. The claim amounts over the last month for four types of household pet are given in the table below:

Crocodile	Tarantula	Piranha	Lion
85	65	88	124
76	82	97	80
90	77	72	90
54	91	83	
85	54		
	63		
	46		
	66		

There are four “treatments” (*ie* four different samples that we wish to compare), so $k = 4$. The “treatment” term is a hang-over from its original use in studying fertilisers!

The first “treatment” (crocodile) has 5 results (called responses), so $n_1 = 5$. Similarly, the second treatment (tarantula) has 8 results, so $n_2 = 8$. Finally $n_3 = 4$ and $n_4 = 3$.

The total number of responses is simply the sum of the treatment totals and is given by $n = \sum_i n_i = 5 + 8 + 4 + 3 = 20$.

We use y_{ij} to stand for the j th result in the i th treatment. For example, y_{12} stands for the 1st treatment 2nd result which is 76. Similarly, y_{31} is the 3rd treatment 1st result, which is 88.



Question 14.1

The health of employees working in three different departments of a company is being compared. The resting pulse rates are given below:

Sales	72	81	96	75	69	88	74	102
Accounts	65	73	78	66				
Admin	61	79	67	74	65			

- (i) Find k , n and the n_i 's.
- (ii) Find y_{32} , y_{21} and y_{13} .

Now the pet insurance company wants to see if there is a significant difference between the mean claim amounts for the four pets. This is just the same as the two-sample t -test except we now have four samples.

We'll call the treatment means μ_1, μ_2, μ_3 and μ_4 . Now when we carried out our two-sample t -test we assumed that each of the samples came from a normal distribution with the *same* variance. We shall make exactly the same assumption here. We shall call the common variance σ^2 .

Therefore, the crocodile results, y_{1j} , come from $Y_{1j} \sim N(\mu_1, \sigma^2)$. Similarly, the tarantula results, y_{2j} , come from $Y_{2j} \sim N(\mu_2, \sigma^2)$ and so on. In general, the result y_{ij} comes from $Y_{ij} \sim N(\mu_i, \sigma^2)$.

Suppose for simplicity that the means for each pet are just the sample means. In which case:

$$\text{crocodile } \mu_1 = \frac{85 + 76 + 90 + 54 + 85}{5} = 78$$

$$\text{tarantula } \mu_2 = 68 \quad \text{piranha } \mu_3 = 85 \quad \text{lion } \mu_4 = 98$$

Now the aim of ANOVA is to compare the means – is there any significant difference between them? Or is it just random variation? *ie* is the high lion claim amount due to the fact that lions are more ferocious household pets or is it just pure chance?

However, instead of comparing the treatment means of 78, 68, 85 and 98 we are going to work with the “treatment effect”. Basically, the treatment effect (denoted by τ_i) is simply how different the treatment mean is from the overall mean.

$$\text{overall mean } \mu = \frac{85 + 76 + 90 + \dots + 124 + 80 + 90}{20} = 78.4$$

So the treatment effects for each of our pets are as follows:

$$\text{crocodile} \quad \tau_1 = 78 - 78.4 = -0.4$$

(ie the crocodile’s mean is 0.4 below the overall mean).

$$\text{tarantula} \quad \tau_2 = 68 - 78.4 = -10.4$$

$$\text{piranha} \quad \tau_3 = 85 - 78.4 = 6.6$$

$$\text{lion} \quad \tau_4 = 98 - 78.4 = 19.6$$

So we are now comparing $-0.4, -10.4, 6.6$ and 19.6 to decide if there is any significant difference between the treatment means. If there was no difference between the treatment means then they would all be the same as the overall mean and so the treatment effects would all be zero.



Question 14.2

Using the pulse rates from Question 14.1:

Sales	72	81	96	75	69	88	74	102
Accounts	65	73	78	66				
Admin	61	79	67	74	65			

- (i) Calculate the overall mean μ , assuming it is the same as the sample mean.
- (ii) Calculate the treatment means, μ_i , assuming they are the same as the sample treatment means.
- (iii) Hence, calculate the treatment effects, τ_i .



Question 14.3

What is the general formula for calculating the treatment effect, τ_i ?

Rearranging the formula from Question 14.3, we can see that we have essentially split up the treatment means into two parts – the overall mean and the treatment effect, ie $\mu_i = \mu + \tau_i$. So for the crocodile, its treatment mean of 78 is the same as the overall mean of 78.4 plus the treatment effect of -0.4 .

Recall that we said the crocodile's results, y_{1j} , came from $Y_{1j} \sim N(\mu_1, \sigma^2)$. Splitting up the treatment mean, we can now say that the crocodile's results come from $Y_{1j} \sim N(\mu + \tau_1, \sigma^2)$ and so on for the other pets.

In general, the result y_{ij} comes from $Y_{ij} \sim N(\mu_i, \sigma^2)$ which can now be written as $Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$. Splitting this up we get $Y_{ij} \sim \mu + \tau_i + N(0, \sigma^2)$ or $Y_{ij} = \mu + \tau_i + e_{ij}$, where $e_{ij} \sim N(0, \sigma^2)$. All we are saying is that any result is the treatment mean plus some random variation. Hence we have:

The mathematical model is:

$$Y_{ij} = \mu + \tau_i + e_{ij}, \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, n_i$$

where the errors e_{ij} are independent $N(0, \sigma^2)$ random variables. Under this model the error variance does not depend on the treatment concerned, the Y_{ij} 's are independent, and Y_{ij} is distributed $N(\mu + \tau_i, \sigma^2)$.

$\mu = \frac{1}{n} \sum_i \sum_j E(Y_{ij})$ is the “overall” population mean.

τ_i is the deviation of the i th treatment mean from μ , ie the i th treatment effect, and $\sum_i n_i \tau_i = 0$.

You may be wondering where the $\sum_i n_i \tau_i = 0$ result came from. We expect each of the results τ_i to be the relevant treatment difference from the overall mean. So when we total up all of these differences we'll get zero, all treatments combined will equal the overall mean (ie there is no overall treatment effect).

Consider our pet insurance example again. We had 5 crocodile results with a treatment effect of -0.4 . This gives a total of $5 \times -0.4 = -2$. Similarly for the 8 tarantula results with a treatment effect of -10.4 , we have $8 \times -10.4 = -83.2$. For the piranha results we get a total of $4 \times 6.6 = 26.4$ and for the lion we get $3 \times 19.6 = 58.8$. The total of these is $-2 - 83.2 + 26.4 + 58.8 = 0$.



Question 14.4

Show that $\sum_i n_i \tau_i = 0$ holds for the results calculated in Question 14.2.

The proof of $\sum_i n_i \tau_i = 0$ is *not* expected, but we have included it in the appendix for interested students.

Assumptions

There are three assumptions underlying analysis of variance, namely:

- (1) The populations must be normal.
- (2) The populations have a common variance.
- (3) The observations are independent.

We can “test” the first two of these assumptions, just like we did for the two-sample t -test, by drawing a dot plot for each of the samples. Then see if they appear to come from a normal distribution, and if they appear to have the same variance (*i.e* spread).

1.2 Estimation of the parameters

Before we estimate the μ and τ_i 's in our model we need to familiarise ourselves with the “dot notation” shorthand. Where a subscript is replaced by a dot, this just means that we sum over all the values of that subscript. So, for example, $Y_{i\cdot}$ means $\sum_{j=1}^{n_i} Y_{ij}$. If the symbol includes a bar, the dot represents averaging over all values of the replaced subscript. So, for example, $\bar{Y}_{i\cdot}$ means $\frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$.



Question 14.5

Write down the meaning of each of the following:

- | | | | |
|------|-------------------|-------------------------|----------------------------|
| (i) | (a) $Y_{i\cdot}$ | (b) $\bar{Y}_{i\cdot}$ | (c) $Y_{\cdot\cdot}$ |
| (ii) | (a) $Y_{\cdot j}$ | (b) $\bar{Y}_{\cdot j}$ | (c) $\bar{Y}_{\cdot\cdot}$ |

We shall now estimate the μ and τ_i 's in our model using the method of least squares, just like we did when we estimated α and β in our regression model.

The parameters μ and τ_i , $i = 1, 2, \dots, k$ can be estimated using least squares by finding values for μ , τ_i , $i = 1, 2, \dots, k$ such that:

$$q = \sum_i \sum_j e_{ij}^2 = \sum_i \sum_j (Y_{ij} - \mu - \tau_i)^2$$

is minimised.

Differentiating this partially with respect to μ and τ_i , $i = 1, 2, \dots, k$, equating to zero and solving gives the normal equations:

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} \quad \text{where } \bar{Y}_{\cdot\cdot} = \frac{1}{n} \sum_i \sum_j Y_{ij} \text{ the overall mean of the observed responses,}$$

and:

$$\hat{\tau}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \quad \text{where } \bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_j Y_{ij} \text{ the mean of the } i\text{th treatment responses.}$$

These results are just what we should expect; we use the overall sample mean to estimate the overall mean, μ , and we use the difference between overall sample mean and the sample treatment mean to estimate the treatment effect, τ_i . This is what we used in Question 14.2 for the pulse rates.

It should also be noted that both of these estimates are unbiased. Make sure you can prove this.



Question 14.6

A student who regularly goes tenpin bowling thinks that he scores higher with one particular bowling ball than he does with his other two. For the next 15 games that he bowls, he chooses a ball at random and notes down the scores as follows:

Ball	Score	sum	sum of squares
1	173, 166, 179, 183, 199	900	162,616
2	195, 206, 188, 195, 210, 221	1,215	246,771
3	169, 155, 172, 188	684	117,514

From the data obtained, calculate unbiased estimates of:

- (i) the overall mean, μ
- (ii) the “treatment effect” for each ball, τ_i .



Question 14.7

Using least squares, prove that $\hat{\mu} = \bar{Y}_{..}$ and $\hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}$.

In our model we have $k+1$ parameters for the means: the overall mean, μ , and the k treatment effects, τ_i .

Since $\sum_i n_i \tau_i = 0$ the number of *independent* parameters specifying the treatment means is k , not $k+1$. The weighted sum of the estimated effects is zero, ie $\sum_i n_i \hat{\tau}_i = 0$.

We are now going to estimate the common variance σ^2 .

The i th treatment responses provide:

$$S_i^2 = \frac{1}{n_i - 1} \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 \text{ as an unbiased estimator of } \sigma^2$$

and:

$$\frac{1}{\sigma^2} \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 \text{ is } \chi^2 \text{ with } (n_i - 1) \text{ degrees of freedom.}$$

There are n_i data values for treatment i and these have a sample mean of $\bar{Y}_{i\cdot}$. So this is just the usual definition of the sample variance and the corresponding χ^2 distribution

$$ie S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2 \text{ and } \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_i (X_i - \bar{X})^2 \sim \chi^2_{n-1}.$$

Combining the information from within each treatment gives:

$$E \left[\sum_i (n_i - 1) S_i^2 \right] = \sum_i (n_i - 1) \sigma^2 = (n - k) \sigma^2$$

and so:

$$\hat{\sigma}^2 = \frac{1}{n - k} \sum_i (n_i - 1) S_i^2 = \frac{1}{n - k} \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$$

provides a pooled unbiased estimator of σ^2 and $\frac{1}{\sigma^2} \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$ is χ^2 with $(n - k)$ degrees of freedom.

This is the estimator we will *always* use for the common underlying variance of each of the treatments. We will see shortly that $\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$ is denoted by SS_R and so our estimate for the variance is $\hat{\sigma}^2 = \frac{SS_R}{n - k}$. This is the formula given on page 26 of the *Tables*.


Question 14.8

Use $\hat{\sigma}^2 = \frac{1}{n-k} \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$ to find an unbiased estimate of the underlying common variance, σ^2 , for the bowling data from Question 14.6, which is reproduced below for convenience.

Ball	Score	sum	sum of squares
1	173, 166, 179, 183, 199	900	162,616
2	195, 206, 188, 195, 210, 221	1,215	246,771
3	169, 155, 172, 188	684	117,514

We now have estimates for all the unknowns in our model and are ready to look at how we carry out our one-way analysis of variance. First we need to state our hypotheses:

The null hypothesis is that the treatment means are equal, ie the treatment effects are zero, so:

$$H_0: \tau_i = 0, \quad i = 1, 2, \dots, k$$

(H_1 is the general alternative : $\tau_i \neq 0$ for at least one i).

1.3 Partitioning the variability

The total variability can be partitioned into two components, one measuring the inherent variability within the treatments and the other measuring the variability between the treatment means $\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \dots, \bar{y}_{k\cdot}$. The result is:

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

say $SS_T = SS_R + SS_B$

This relationship is very similar to the one we saw in the regression chapter. Its proof is *not* expected, but is included in the appendix for interested students.

Recall from earlier that $Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$. So in our model we *expect* there to be some variation within each of the treatments as they are normally distributed with variance σ^2 . The whole point of partitioning the variability is to see how much of the overall variance is made of this expected “within treatment” variance, SS_R , and how much is made up of variance *between* the means, SS_B . The larger the *between*-means variance is, the less likely it is that we can assume that they all have the *same* mean.

The definitions for SS_T , SS_R and SS_B are a bit long-winded to use for calculating, so we shall rewrite them in the same way that we rewrote the sample variance in Chapter 1:

$$S^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_i x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{n-1} \left[\sum_i x_i^2 - n\bar{x}^2 \right]$$

So we have:

$$\begin{aligned} SS_T &= \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = \sum_i \sum_j y_{ij}^2 - \frac{(\sum \sum y_{ij})^2}{n} = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{n} \\ SS_B &= \sum_i n_i (\bar{y}_{i\bullet} - \bar{y}_{..})^2 = \sum_i \frac{y_{i\bullet}^2}{n_i} - \frac{(\sum \sum y_{ij})^2}{n} = \sum_i \frac{y_{i\bullet}^2}{n_i} - \frac{y_{..}^2}{n} \end{aligned}$$

$$SS_R = SS_T - SS_B$$

These formulae are given on page 26 of the *Tables* and it is vital that you can calculate these for any data set.

The Core Reading has an example of how to calculate these for a data set given below. First we need to consider how we can test whether the between means variance is significant or not (and so whether there is a significant difference between the treatment means or not). To do this we need to know the distributions of SS_T , SS_R and SS_B . Since these are effectively “sample variances” they have χ^2 distributions.

SS_R is the within-treatments or residual sum of squares – it is just the sum of squares of the residuals from the fit (the estimated errors $\hat{e}_{ij} = Y_{ij} - \bar{Y}_i.$) and is based on $(n-1)-(k-1) = n-k$ degrees of freedom, the degrees of freedom remaining after estimating the parameters for the means. As noted above $\hat{\sigma}^2 = SS_R/(n-k)$ is an unbiased estimator of σ^2 and SS_R/σ^2 is χ^2_{n-k} .

SS_B is the between-treatments sum of squares.

When H_0 is true, $SS_T/(n-1)$ is the overall sample variance and so SS_T/σ^2 is χ^2_{n-1} .

$SS_T/(n-1)$ is just the sample variance for the entire data set, ignoring the fact that the data contains subsets with different treatments.

Since SS_R and SS_B are in fact independent and SS_R/σ^2 is χ^2_{n-k} it follows that SS_B/σ^2 is χ^2_{k-1} . $SS_B/(k-1)$ is another unbiased estimator of σ^2 .

We are making use here of a special property of the χ^2 distribution which states that if X_1 and X_2 are independent, $X_3 = X_1 + X_2$, $X_1 \sim \chi^2_{n_1}$ and $X_3 \sim \chi^2_{n_3}$, then $X_2 \sim \chi^2_{n_3-n_1}$. You can prove this using MGFs:



Question 14.9

Prove, using MGFs, that if $X = X_1 + X_2$ where $X_1 \sim \chi^2_m$ and $X_2 \sim \chi^2_n$ then $X \sim \chi^2_{m+n}$.

We are now ready to formally test whether the between variances result is significant or not. To do that we use the fact that the F distribution is defined as $F_{m,n} \equiv \frac{\chi^2_m/m}{\chi^2_n/n}$.

Finally:

$$\frac{SS_B/(k-1)}{SS_R/(n-k)} = \frac{\text{between treatments mean square}}{\text{residual mean square}}$$

is $F_{k-1,n-k}$ and H_0 is rejected for “large” values of this ratio.

This result is given on page 26 of the *Tables*. Since we are only interested in whether the between-means variance, SS_B , is significantly larger than the expected variance, SS_R , we carry out a one-sided test.

The results are usually set out in an ANOVA table:

Source of variation	Degrees of Freedom	Sums of Squares	Mean Squares
Between treatments	$k - 1$	SS_B	$SS_B / (k - 1)$
Residual	$n - k$	SS_R	$SS_R / (n - k)$
Total	$n - 1$	SS_T	

This completes the theory behind the one-way analysis of variance. It is now essential for exam questions that you are able to calculate the values of SS_T , SS_B , SS_R and then carry out the F test.

1.4 Example

A random sample of insurance policies on the contents of private houses was examined for each of three insurance companies and the sum insured under each policy noted. The results (in units of £100) were as follows:

Company 1: 36, 28, 32, 43, 30, 21, 33, 37, 26, 34

Company 2: 26, 21, 31, 29, 27, 35, 23, 33

Company 3: 39, 28, 45, 37, 21, 49, 34, 38, 44

$$y_{1\cdot} = 320, y_{2\cdot} = 225, y_{3\cdot} = 335, y_{\cdot\cdot} = 880, n_1 = 10, n_2 = 8, n_3 = 9, \sum \sum y_{ij}^2 = 30152$$

Recall that SS_T can be worked out from the formula $SS_T = \sum_i \sum_j y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{n}$.

$$\therefore SS_T = 30152 - \frac{880^2}{27} = 1470.52$$

Recall that SS_B can be worked out from the formula $SS_B = \sum_i \frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{n}$.

$$SS_B = \left(\frac{320^2}{10} + \frac{225^2}{8} + \frac{335^2}{9} \right) - \frac{880^2}{27} = 356.09$$

Recall that SS_R can be found by subtraction using $SS_R = SS_T - SS_B$.

$$\therefore SS_R = 1470.52 - 356.09 = 1114.43$$

Source of variation	df	SS	MSS
Between companies	2	356.09	178.04
Residual	24	1114.43	46.43
Total	26	1470.52	

Under $H_0: \tau_1 = \tau_2 = \tau_3 = 0$, $F = \frac{178.04}{46.43} = 3.83$ on (2,24) degrees of freedom.

Prob-value of $F = 3.83$ is less than 0.05, so H_0 is rejected at the 5% level.

Recall that this is a one-sided test on the upper end on the F distribution. Hence, the 5% critical value for $F_{2,24}$ given on page 172 of the *Tables* is 3.403.



Question 14.10

Using the pulse rates from Question 14.1:

Sales	72	81	96	75	69	88	74	102
Accounts	65	73	78	66				
Admin	61	79	67	74	65			

(i) Find $y_{1\cdot}$, $y_{2\cdot}$, $y_{3\cdot}$, $y_{\cdot\cdot}$ and $\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2$.

(ii) Hence calculate SS_T , SS_B and SS_R .

**Question 14.11**

An investigation was carried out into the number of hours lost due to stress-related illnesses at various companies. The results for a random sample of individuals from five companies are shown below.

		company				
		A	B	C	D	E
	30	42	65	67	70	
	25	57	46	58	63	
	12	47	55	81	80	
	23	30	27			
	16					
<i>totals</i>		106	176	193	206	213

$$\sum \sum y_{ij}^2 = 50,354$$

Use analysis of variance to test for differences between companies. You should present your results in the form of an ANOVA table.

**Question 14.12**

For the tenpin bowling scores in Question 14.6:

Ball	Score	sum	sum of squares
1	173, 166, 179, 183, 199	900	162,616
2	195, 206, 188, 195, 210, 221	1,215	246,771
3	169, 155, 172, 188	684	117,514

- (i) Construct an ANOVA table to test the whether there is a difference between the performances of the three bowling balls.
- (ii) What assumption does your conclusion depend on?

1.5 Checking the model

Recall the assumptions we listed in Section 1.1, that observations from each treatment are normally distributed with equal variance. We stated that these could be examined by drawing a dot plot of the sample values. In this section we will draw dot plots using the residuals instead. The residual is defined in the same way as in Chapter 13, *ie* as the difference between the observed result and the expected result.

Since $Y_{ij} \sim N(\mu + \tau_i, \sigma^2) = N(\mu_i, \sigma^2)$, we expect Y_{ij} to be $\mu + \tau_i = \mu_i$. Therefore, the residuals are defined as $e_{ij} = Y_{ij} - \mu_i$.

Examining the residuals from the fitted model can reveal inadequacies in the model.

The residuals, r_{ij} , are the estimated errors, *ie*:

$$r_{ij} = \hat{e}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i = y_{ij} - \bar{y}_i.$$

ie the residual of a response is the difference between it and the appropriate observed treatment mean. Plotting the residuals for each treatment may reveal a pattern (lack of randomness – some uncontrolled factor at work), or reveal non-normality (if serious a transformation of the data may be required eg a log transformation if there is evidence of a pronounced positive skew), or reveal that the error variance is not independent of the treatment (again a transformation may be helpful in making the variance homogeneous among treatments – in particular, if the variance increases with the mean then a log transformation is recommended).

However it is worth noting that the F-test procedure used is quite robust against departures from normality and homogeneity of variance.

This means that we don't need to worry *too* much about whether these assumptions are met.

Let us consider again the insurance policies.

For our example above:

$$\bar{y}_{1\cdot} = 32.0, \bar{y}_{2\cdot} = 28.1, \bar{y}_{3\cdot} = 37.2$$

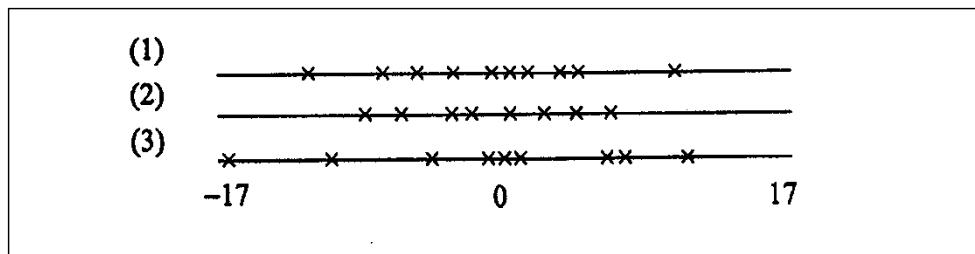
So the residuals are:

$$r_{1j}: 4.0, -4.0, 0.0, 11.0, -2.0, -11.0, 1.0, 5.0, -6.0, 2.0$$

$$r_{2j}: -2.1, -7.1, 2.9, 0.9, -1.1, 6.9, -5.1, 4.9$$

$$r_{3j}: 1.8, -9.2, 7.8, -0.2, -16.2, 11.8, -3.2, 0.8, 6.8$$

A simple plot of the residuals $r_{ij} = y_{ij} - \bar{y}_{i\cdot}$ gives:



The normality of population 2 might be questioned and the homogeneity of error variance (note the sample variances are in the order $S_2^2 < S_1^2 < S_3^2$, the same order as the sample means $\bar{y}_{2\cdot} < \bar{y}_{1\cdot} < \bar{y}_{3\cdot}$.)



Question 14.13

Draw a plot of the residuals for the company stress levels example in Question 14.11.

		company				
		A	B	C	D	E
	30	42	65	67	70	
	25	57	46	58	63	
	12	47	55	81	80	
	23	30	27			
	16					
totals		106	176	193	206	213

$$\sum \sum y_{ij}^2 = 50,354$$

2 Examining the treatment means

In this section we will be finding confidence intervals for a single treatment mean or between a pair of treatment means using the results of Chapter 11. However, we will be using our estimate for the overall common variance, $\hat{\sigma}^2 = \frac{SS_R}{n-k}$, rather than s^2 or s_P^2 .

2.1 Confidence intervals for a single treatment mean

In the situation where interest is focused on a particular treatment, say treatment i , σ^2 can be estimated using the residual mean square $\hat{\sigma}^2$ and a confidence interval for $\mu + \tau_i$ (ie for treatment mean μ_i) is given by:

$$\bar{y}_{i\bullet} \pm t\hat{\sigma}/\sqrt{n_i}$$

where t is based on $(n - k)$ degrees of freedom.

Recall from Chapter 11, Section 3.1 that a confidence interval for the mean from a normal distribution was based on:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

We are going to use this to obtain a confidence interval for treatment mean i , μ_i , using our estimate $\bar{y}_{i\bullet}$. We will use $\hat{\sigma}^2$ instead of s^2 , so the t -distribution will have $n - k$ degrees of freedom:

$$\frac{\bar{y}_{i\bullet} - \mu_i}{\hat{\sigma}/\sqrt{n_i}} \sim t_{n-k}$$

$$\text{where } \hat{\sigma}^2 = \frac{SS_R}{n - k}.$$

This rearranges to the confidence interval given in the Core Reading above.

2.2 Confidence intervals for a pair of treatment means

In the situation where interest is focused on a particular pair of treatments, say treatments 1 and 2 for convenience, then:

$$\text{var}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

and a confidence interval for $\mu_1 - \mu_2 = (\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$ is given by:

$$(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) \pm t\hat{\sigma} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{1/2}$$

where t is again based on $(n - k)$ degrees of freedom.

Recall from Chapter 11, Section 5.1 that a confidence interval for the difference between two means from normal distributions was based on:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_P \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1+n_2-2}$$

This result can be adapted to obtain a confidence interval for the difference between treatment means 1 and 2, $\mu_1 - \mu_2$, using our estimate $\bar{y}_{1\bullet} - \bar{y}_{2\bullet}$. We will use $\hat{\sigma}^2$, which is a generalisation of the pooled variance to the k -sample case. We have already seen that:

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

We now extend this to define:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \dots + (n_k - 1)S_k^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_k - 1)} \\ &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \dots + (n_k - 1)S_k^2}{n - k} \end{aligned}$$

Note that the numerator of this expression is the residual sum of squares, SS_R , and that:

$$\frac{(n - k)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-k}^2$$

Using the definition of the t distribution (from Chapter 9), it follows that:

$$\frac{(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) - (\mu_1 - \mu_2)}{\hat{\sigma} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n-k}$$

and from this result we get the confidence interval given in Core Reading above.



Question 14.14

Using the pulse rates from Question 14.1

Sales	72	81	96	75	69	88	74	102
Accounts	65	73	78	66				
Admin	61	79	67	74	65			

In Question 14.2, we found that $\bar{y}_{1\bullet} = 82.125$, $\bar{y}_{2\bullet} = 70.5$, $\bar{y}_{3\bullet} = 69.2$. In Question 14.10, we found that $SS_T = 1,986.12$, $SS_B = 649.44$, $SS_R = 1,336.68$.

- (i) Show using ANOVA that there is no significant difference between the treatment means.
- (ii)
 - (a) Show that the 95% confidence interval for the difference between the mean pulse rates for Sales and Admin is $(0.976, 24.9)$
 - (b) What does this confidence interval mean?
 - (c) Explain the apparent contradiction between the confidence interval and the ANOVA result.

2.3 Analysing treatment means using a least significant difference approach

Constructing such intervals for all possible pairs of treatments is not recommended – the interpretation of them becomes difficult as the overall level of confidence of the whole set of intervals has to be considered. However, if $H_0: \tau_i = 0, i = 1, 2, \dots, k$ has been rejected, a good idea as to whether the treatments fall into several reasonably homogeneous groups can be obtained as follows.

So if there *is* a significant difference between the treatment means we can compare treatment means to see which are the same and which are different.

Step 1

List the observed treatments in order, eg with $k = 4$ we might have:

$$\bar{y}_{2\bullet} < \bar{y}_{3\bullet} < \bar{y}_{1\bullet} < \bar{y}_{4\bullet}$$

Step 2

We will now examine each of the pairs in order to see whether the means are the same or not. We do this by using a two-sample test. For example on the first pair:

$$H_0: \mu_2 = \mu_3$$

$$H_1: \mu_2 \neq \mu_3$$

Note that Core Reading carries out a *two*-sided test and this is what is expected in the exam. It would be natural to choose a one-sided test (to see if $\mu_3 > \mu_2$) as we have arranged the means in increasing order, however, this is not the approach used.

Our statistic is:

$$\frac{(\bar{Y}_{3\bullet} - \bar{Y}_{2\bullet}) - (\mu_3 - \mu_2)}{\hat{\sigma} \sqrt{\left(\frac{1}{n_3} + \frac{1}{n_2}\right)}} \sim t_{n-k}$$

Now, under H_0 , $\mu_3 - \mu_2 = 0$. Since $\bar{y}_{3.} > \bar{y}_{2.}$ there will be a significant difference between the means if the statistic is greater than the upper 2½% critical value of the appropriate t distribution:

$$\frac{(\bar{y}_{3.} - \bar{y}_{2.}) - (\mu_3 - \mu_2)}{\hat{\sigma} \sqrt{\left(\frac{1}{n_3} + \frac{1}{n_2}\right)}} > t_{0.025, n-k}$$

Rearranging:

$$(\bar{y}_{3.} - \bar{y}_{2.}) > t_{0.025, n-k} \times \hat{\sigma} \sqrt{\left(\frac{1}{n_3} + \frac{1}{n_2}\right)}$$

The value on the right hand side is the “least significant difference”, ie the value that the difference between the sample means needs to exceed to say that there is a significant difference.

If the n_i ’s are the same for each of the treatments, then the “least significant difference” for each treatment will be the same. This makes it a quick way to test each of the pairs of means. A standard two-sample test is also fine.

For a given level of significance, say 5%, calculate the least difference between $\bar{y}_{3.}$ and $\bar{y}_{2.}$ which would be significant, namely:

$$t \hat{\sigma} \left(\frac{1}{n_2} + \frac{1}{n_3} \right)^{1/2} \text{ where } t = t_{0.025, n-k}$$

ie the value of a t_{n-k} variable which is exceeded with probability 0.025. If the difference $\bar{y}_{3.} - \bar{y}_{2.}$ is less than this least significant difference then it can be indicated that the treatment means fall into the same group, for example by underlining the pair. This process can be repeated for $\bar{y}_{3.}$ and $\bar{y}_1.$ and then for $\bar{y}_1.$ and $\bar{y}_4.$ As an example the results may give:

$$\underline{\bar{y}_{2.}} < \underline{\bar{y}_{3.}} < \bar{y}_1. < \bar{y}_4.$$

This indicates that treatment 4 is on its own.

Since $\bar{y}_{2\bullet}$, $\bar{y}_{3\bullet}$ fall into the same group and $\bar{y}_{3\bullet}$, $\bar{y}_{1\bullet}$ fall into the same group, it is worth checking to see if $\bar{y}_{2\bullet}$ and $\bar{y}_{1\bullet}$ fall into the same group. If they were this would mean all three of these means fall into the same group and we would show this as:

$$\underline{\bar{y}_{2\bullet} < \bar{y}_{3\bullet} < \bar{y}_{1\bullet} < \bar{y}_4}$$

Core reading now analyses its example from earlier:

Company 1: 36, 28, 32, 43, 30, 21, 33, 37, 26, 34 $n_1 = 10$ $\bar{y}_{1\bullet} = 32$

Company 2: 26, 21, 31, 29, 27, 35, 23, 33 $n_2 = 8$ $\bar{y}_{2\bullet} = 28.125$

Company 3: 39, 28, 45, 37, 21, 49, 34, 38, 44 $n_3 = 9$ $\bar{y}_{3\bullet} = 37.2$

$$SS_R = 1,114.43 \Rightarrow \hat{\sigma}^2 = \frac{SS_R}{n-k} = \frac{1,114.43}{24} = 46.43$$

Step 1

In the example above, $\bar{y}_{2\bullet} < \bar{y}_{1\bullet} < \bar{y}_{3\bullet}$.

Step 2

For a 5% level, the least significant difference between $\bar{y}_{1\bullet}$ and $\bar{y}_{2\bullet}$ is:

$$2.064\sqrt{46.43}\left(\frac{1}{10} + \frac{1}{8}\right)^{\frac{1}{2}} = 6.67$$

Using the fact that $t_{0.025, 24} = 2.064$.

$\bar{y}_{1\bullet} - \bar{y}_{2\bullet} = 3.9$ (1 dp) so $\underline{\bar{y}_{2\bullet} < \bar{y}_{1\bullet}}$

For $\bar{y}_{3\bullet}$ and $\bar{y}_{1\bullet}$ the least significant difference is 6.46 and $\bar{y}_{3\bullet} - \bar{y}_{1\bullet} = 5.2$ (1 dp)
so:

$$\underline{\bar{y}_{2\bullet} < \bar{y}_{1\bullet} < \bar{y}_{3\bullet}}$$



Question 14.15

- (i) Show that the least significant difference between $\bar{y}_{3.}$ and $\bar{y}_{1.}$ is 6.46.
- (ii) By calculating the least significant difference between $\bar{y}_{2.}$ and $\bar{y}_{3.}$, confirm that we should *not* underline all three treatments.

Note that this result does not contradict the earlier rejection of H_0 . If in advance it had been decided to concentrate on companies 2 and 3 (while at the same time using the information from company 1) a 95% confidence interval for $\tau_3 - \tau_2$ would be obtained as:

$$(37.2 - 28.1) \pm 2.064\sqrt{46.43} \left(\frac{1}{9} + \frac{1}{8} \right)^{1/2}$$

ie 9.1 ± 6.8 ie (2.3, 15.9)

and this interval does not contain zero.

So there *is* a significant difference between the means of companies 2 and 3. However, just because the biggest and smallest company means are different from each other doesn't necessarily imply that *all* of the treatment means are different from the *overall mean*.



Question 14.16

The bowling ball scores in Question 14.6 were:

Ball	Score	sum	sum of squares
1	173, 166, 179, 183, 199	900	162,616
2	195, 206, 188, 195, 210, 221	1,215	246,771
3	169, 155, 172, 188	684	117,514

Where $SS_R = 1,899.5$.

Analyse the means of the bowling ball scores in Question 14.6 using a least significant difference approach.

3 Further comments

ANOVA for a completely randomized comparative experiment on only $k = 2$ treatments is equivalent to the two-sample t test of Chapter 12. To verify this, one must be aware that $t_d^2 \equiv F_{1,d}$.

It is no coincidence that the sums of squares can be split up in the same way for ANOVA and linear regression. In fact linear regression is just a special type of ANOVA and the results of a regression analysis can be presented in an ANOVA table.

A linear regression analysis of:

$$Y_i = a + bx_i + e_i, \quad i = 1, 2, \dots, n$$

can be represented in the ANOVA framework as follows:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$SS_T = SS_R + SS_{REG}$$

Slightly different abbreviations have been used here for the sum of squares.

Source of variation	Degrees of Freedom	Sums of Squares	Mean Squares
Regression	1	SS_{REG}	SS_{REG}
Residual	$n - 2$	SS_R	$SS_R / (n - 2)$
Total	$n - 1$	SS_T	

Under $H_0 : \beta = 0$, $\frac{SS_{REG}}{SS_R / (n - 2)}$ is $F_{1,n-2}$.

A one-way ANOVA analysis of k treatments is equivalent to a regression in which the response is regressed on $(k - 1)$ “dummy” explanatory variables, each of which takes values 0 or 1 only.

4 Exam-type question



Exam-type question

Five students have compared their scores in some practice papers that they sat before their exam. Their marks were as follows:

Student 1	72, 75, 62, 71, 60, 59
Student 2	78, 82, 64, 72
Student 3	90, 78, 67, 71, 83
Student 4	80, 77, 76, 81, 64
Student 5	95, 88, 62

Consider the model:

$$Y_{ij} = \mu + \tau_i + e_{ij} \quad i = 1, 2, 3, 4, 5 \quad j = 1, 2, \dots, n_i \quad e_{ij} \sim N(0, \sigma^2)$$

where y_{ij} is the score of the i th student on the j th practice paper and n_i is the number of papers taken by student i . The e_{ij} are independent and identically distributed and $\sum_i n_i \tau_i = 0$.

- (i) Calculate the least squares estimates of μ and τ_i , $i = 1, 2, 3, 4, 5$.
- (ii) Perform an analysis of variance on these results stating clearly the null hypothesis and your conclusion.
- (iii) Calculate a 95% confidence interval for the underlying common standard deviation for all students assuming that the null hypothesis holds true.

5 Appendix – proofs

This section contains the proofs of $\sum_i n_i \tau_i = 0$ and $SS_T = SS_R + SS_B$. Both of these

results are simply quoted in Core Reading of this chapter (and so the proofs are *not* required) as the focus is on *applying* these results rather than deriving them. They are included here for those students who want to understand the mathematics behind the results.

5.1 Proof of $\sum_i n_i \tau_i = 0$

The model is $Y_{ij} = \mu + \tau_i + e_{ij}$.

Taking expected values and using the fact that $e_{ij} \sim N(0, \sigma^2)$:

$$\begin{aligned} E[Y_{ij}] &= E[\mu + \tau_i + e_{ij}] \\ &= \mu + \tau_i + E[e_{ij}] \\ &= \mu + \tau_i \end{aligned}$$

Summing over all values of i and j :

$$\begin{aligned} \sum_i \sum_j E[Y_{ij}] &= \sum_i \sum_j (\mu + \tau_i) \\ &= \mu \sum_i \sum_j 1 + \sum_i (\tau_i \sum_j 1) \\ &= n\mu + \sum_i n_i \tau_i \end{aligned}$$

But since $\mu = \frac{1}{n} \sum_i \sum_j E[Y_{ij}]$, the LHS is $n\mu$ and we get $\sum_i n_i \tau_i = 0$.

5.2 Proof of $SS_T = SS_R + SS_B$

We are trying to show that:

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 + \sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

say $SS_T = SS_R + SS_B$

Rewriting the SS_T term:

$$\begin{aligned} SS_T &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} [(Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..})]^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 \end{aligned}$$

Since all the terms in the third summation are independent of j , we can take them out of the summation:

$$\begin{aligned} SS_T &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) + \sum_{i=1}^k (\bar{Y}_{i.} - \bar{Y}_{..})^2 \sum_{j=1}^{n_i} 1 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) + \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ &= SS_R + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) + SS_B \end{aligned}$$

All we have to do is show that the “cross product” term vanishes, ie we want to show

$$2 \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) = 0.$$

Expanding out the brackets gives:

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet}) (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} \bar{Y}_{i\bullet} - Y_{ij} \bar{Y}_{\bullet\bullet} - \bar{Y}_{i\bullet}^2 + \bar{Y}_{i\bullet} \bar{Y}_{\bullet\bullet}) \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} \bar{Y}_{i\bullet} - \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} \bar{Y}_{\bullet\bullet} - \sum_{i=1}^k \sum_{j=1}^{n_i} \bar{Y}_{i\bullet}^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} \bar{Y}_{i\bullet} \bar{Y}_{\bullet\bullet} \end{aligned}$$

Now $\bar{Y}_{i\bullet}$ does not depend on j and so it can be taken out of the j summation. Similarly, $\bar{Y}_{\bullet\bullet}$ does not depend on either i or j and so it can be taken out of both summations. So:

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet}) (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) &= \sum_{i=1}^k \bar{Y}_{i\bullet} \sum_{j=1}^{n_i} Y_{ij} - \bar{Y}_{\bullet\bullet} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} - \sum_{i=1}^k \bar{Y}_{i\bullet}^2 \sum_{j=1}^{n_i} 1 + \bar{Y}_{\bullet\bullet} \sum_{i=1}^k \bar{Y}_{i\bullet} \sum_{j=1}^{n_i} 1 \\ &= \sum_{i=1}^k \bar{Y}_{i\bullet} \sum_{j=1}^{n_i} Y_{ij} - \bar{Y}_{\bullet\bullet} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} - \sum_{i=1}^k n_i \bar{Y}_{i\bullet}^2 + \bar{Y}_{\bullet\bullet} \sum_{i=1}^k n_i \bar{Y}_{i\bullet} \end{aligned}$$

Recall $\bar{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ and $\bar{Y}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$, hence $\sum_{j=1}^{n_i} Y_{ij} = n_i \bar{Y}_{i\bullet}$ and $\sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} = n \bar{Y}_{\bullet\bullet}$.

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet}) (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) &= \sum_{i=1}^k n_i \bar{Y}_{i\bullet}^2 - n \bar{Y}_{\bullet\bullet}^2 - \sum_{i=1}^k n_i \bar{Y}_{i\bullet}^2 + \bar{Y}_{\bullet\bullet} \sum_{i=1}^k n_i \bar{Y}_{i\bullet} \\ &= -n \bar{Y}_{\bullet\bullet}^2 + \bar{Y}_{\bullet\bullet} \sum_{i=1}^k n_i \bar{Y}_{i\bullet} \end{aligned}$$

We can also write $\bar{Y}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^k Y_{i\bullet}$, hence $\sum_{i=1}^k Y_{i\bullet} = n \bar{Y}_{\bullet\bullet}$ or $\sum_{i=1}^k n_i \bar{Y}_{i\bullet} = n \bar{Y}_{\bullet\bullet}$.

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet}) (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) = -n \bar{Y}_{\bullet\bullet}^2 + \bar{Y}_{\bullet\bullet} \sum_{i=1}^k n_i \bar{Y}_{i\bullet} = -n \bar{Y}_{\bullet\bullet}^2 + n \bar{Y}_{\bullet\bullet}^2 = 0$$

6 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 12 to 14.
2. Attempt some of the questions in Part 4 of the **Question and Answer Bank**. 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 A / AMP and Marking – There are three separate mock exam papers that you can attempt and get marked. A recent student survey 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.”

“Fantastic marking – a lot of time was spent and for that I am very grateful. Loads of hint/tips on answering questions and feedback detailed methodically and separately from my script – a great way to mark.”

You can find lots more information on our website at www.ActEd.co.uk.

Buy online at www.ActEd.co.uk/estore

And finally ...

All the very best in your exam. May your hard work be richly rewarded!

This page has been left blank so that you can keep the chapter summaries together for revision purposes.



Chapter 14 Summary

Analysis of variance (ANOVA) is testing for a difference between treatment means. The model is:

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

where e_{ij} 's are independent and identically distributed $N(0, \sigma^2)$.

The parameters μ (total mean), τ_i (treatment effect) and σ^2 can be estimated using the formulae overleaf. The variance estimate is given on page 26 of the *Tables*.

The total variance (SS_T) is split into variance within each treatment (SS_R) and variance between treatment means (SS_B):

$$SS_T = SS_R + SS_B$$

These can be calculated using the formulae given overleaf and on page 26 of the *Tables*.

The significance of the variance between treatment means (SS_B) is established using the F test given overleaf and on page 26 of the *Tables*.

Residuals can be plotted to check the adequacy of the model – *ie* the normality and equality of variances assumptions.

Confidence intervals for means should use the results of Chapter 11, but with the variance estimate $\hat{\sigma}^2$ instead.

If it is found under ANOVA that the treatment means are *not* the same, we can analyse them further using the least significant difference approach. This is essentially testing each of the pairs of treatments to see if they have the same mean or not. The results of this can be shown on a diagram like the one below:

$$\underline{\bar{y}_2.} < \underline{\bar{y}_3.} < \underline{\bar{y}_1.} < \underline{\bar{y}_4.}$$



Chapter 14 Formulae

Parameter Estimates

$$\hat{\mu} = \bar{y}_{..} = \frac{1}{n} \sum_i \sum_j y_{ij}$$

$$\hat{\tau}_i = \bar{y}_{i\cdot} - \bar{y}_{..} = \frac{1}{n_i} \sum_j y_{ij} - \frac{1}{n} \sum_i \sum_j y_{ij}$$

$$\hat{\sigma}^2 = \frac{1}{n-k} SS_R$$

Sum of Squares

$$SS_T = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{n}$$

$$SS_B = \sum_i \frac{y_{i\cdot}^2}{n_i} - \frac{y_{..}^2}{n}$$

$$SS_R = SS_T - SS_B$$

Statistical Test

$$\frac{SS_B}{k-1} / \frac{SS_R}{n-k} \sim F_{k-1, n-k}$$

Chapter 14 Solutions

Solution 14.1

(i) There are 3 “treatments”: sales, accounts and admin. So $k = 3$.

The first treatment, sales, has 8 results. So $n_1 = 8$.

The second treatment, accounts, has 4 results. So $n_2 = 4$.

The third treatment, admin, has 5 results. So $n_3 = 5$.

Hence, $n = 8 + 4 + 5 = 17$.

(ii) y_{32} is the 3rd treatment (admin) 2nd result, which is 79.

y_{21} is the 2nd treatment (accounts) 1st result, which is 65.

y_{13} is the 1st treatment (sales) 3rd result, which is 96.

Solution 14.2

$$(i) \text{ overall mean } \mu = \frac{72 + 81 + 96 + \dots + 67 + 74 + 65}{17} = 75.588$$

$$(ii) \text{ sales } \mu_1 = \frac{675}{8} = 82.125$$

$$\text{accounts } \mu_2 = \frac{282}{4} = 70.5$$

$$\text{admin } \mu_3 = \frac{346}{5} = 69.2$$

$$(iii) \text{ sales } \tau_1 = 82.125 - 75.588 = 6.537$$

$$\text{accounts } \tau_2 = 70.5 - 75.588 = -5.088$$

$$\text{admin } \tau_3 = 69.2 - 75.588 = -6.388$$

Solution 14.3

We are taking the treatment mean, μ_i , and subtracting the overall mean, μ . Hence:

$$\tau_i = \mu_i - \mu$$

Solution 14.4

Totalling up the treatment effects gives:

$$(8 \times 6.537) + (4 \times -5.088) + (5 \times -6.388) = 0.00$$

The slight error is due to our rounding of our answers in Solution 14.2.

Solution 14.5

$$(i) \quad (a) \quad Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$$

$$(b) \quad Y_{\bullet j} = \sum_{i=1}^k Y_{ij}$$

$$(c) \quad Y_{\bullet\bullet} = \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$$

$$(ii) \quad (a) \quad \bar{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

$$(b) \quad \bar{Y}_{\bullet j} = \frac{1}{k} \sum_{i=1}^k Y_{ij}$$

$$(c) \quad \bar{Y}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$$

Solution 14.6

(i) Since we are given the treatment sums, the mean is just:

$$\hat{\mu} = \bar{y}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} = \frac{900 + 1,215 + 684}{15} = \frac{2,799}{15} = 186.6$$

(ii) Our estimates for the sample means and treatment effects are:

$$\hat{\mu}_1 = \bar{y}_{1\bullet} = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} = \frac{900}{5} = 180 \Rightarrow \hat{\tau}_1 = 180 - 186.6 = -6.6$$

$$\hat{\mu}_2 = \bar{y}_{2\bullet} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} = \frac{1,215}{6} = 202.5 \Rightarrow \hat{\tau}_2 = 202.5 - 186.6 = 15.9$$

$$\hat{\mu}_3 = \bar{y}_{3\bullet} = \frac{1}{n_3} \sum_{j=1}^{n_3} y_{3j} = \frac{684}{4} = 171 \Rightarrow \hat{\tau}_3 = 171 - 186.6 = -15.6$$

Solution 14.7

If we differentiate the expression for q with respect to μ , we get:

$$\frac{\partial q}{\partial \mu} = - \sum_i \sum_j 2(Y_{ij} - \mu - \tau_i)$$

Setting this to zero gives us:

$$\sum_i \sum_j (Y_{ij} - \mu - \tau_i) = 0 \text{ or } \sum_i \sum_j Y_{ij} - n\mu - \sum_i n_i \tau_i = 0$$

We have already shown that $\sum_i n_i \tau_i = 0$, so:

$$\hat{\mu} = \frac{1}{n} \sum_i \sum_j Y_{ij} = \bar{Y}_{\bullet\bullet}$$

Differentiating with respect to τ_i :

$$\frac{\partial q}{\partial \tau_i} = -\sum_j 2(Y_{ij} - \mu - \tau_i)$$

Note that terms involving a “different” τ don’t contribute anything to the derivative.

Setting this equal to zero gives:

$$\sum_j (Y_{ij} - \mu - \tau_i) = 0 \Rightarrow \sum_j Y_{ij} - n_i \mu - n_i \tau_i = 0$$

Dividing by n_i :

$$\bar{Y}_{i\bullet} - \mu - \tau = 0 \Rightarrow \hat{\tau} = \bar{Y}_{i\bullet} - \mu = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$$

These results are exactly what we would expect. μ is the average for all treatments and we are estimating this using $\bar{Y}_{\bullet\bullet}$ ie the average over all the data. τ_i represents the difference between the mean for treatment i and the overall mean, and we are estimating this based on the corresponding difference in the sample means.

Solution 14.8

We want to calculate $\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2$. So we require the treatment means:

$$\bar{y}_{1\bullet} = \frac{900}{5} = 180 , \quad \bar{y}_{2\bullet} = \frac{1,215}{6} = 202.5 , \quad \bar{y}_{3\bullet} = \frac{684}{4} = 171$$

Using these to calculate the square deviations:

$$\sum_j (y_{1j} - \bar{y}_{1\bullet})^2 = (173 - 180)^2 + \dots + (199 - 180)^2 = 616$$

$$\sum_j (y_{2j} - \bar{y}_{2\bullet})^2 = (195 - 202.5)^2 + \dots + (221 - 202.5)^2 = 733.5$$

$$\sum_j (y_{3j} - \bar{y}_{3\bullet})^2 = (169 - 171)^2 + \dots + (188 - 171)^2 = 550$$

Therefore:

$$\hat{\sigma}^2 = \frac{1}{15-3} \times (616 + 733.5 + 550) = 158.29$$

Solution 14.9

Now $\chi_n^2 \equiv Ga\left(\frac{n}{2}, \frac{1}{2}\right)$ therefore $M_{X_1}(t) = (1-2t)^{-\frac{m}{2}}$ and $M_{X_2}(t) = (1-2t)^{-\frac{n}{2}}$.

We need $M_X(t)$:

$$M_X(t) = E(e^{tX}) = E(e^{tX_1+tX_2}) = E(e^{tX_1} e^{tX_2}) = E(e^{tX_1})E(e^{tX_2}) = M_{X_1}(t)M_{X_2}(t)$$

$$\text{So } M_X(t) = (1-2t)^{-\frac{m}{2}} \times (1-2t)^{-\frac{n}{2}} = (1-2t)^{-\frac{m+n}{2}}.$$

Since this is the MGF of χ_{m+n}^2 , by the uniqueness property $X \sim \chi_{m+n}^2$.

Solution 14.10

- (i) Recall that $y_{i\bullet}$ is simply the sum of the results for the j th treatment. Hence:

$$y_{1\bullet} = \sum_{j=1}^8 y_{1j} = 72 + 81 + 96 + 75 + 69 + 88 + 74 + 102 = 657$$

$$y_{2\bullet} = \sum_{j=1}^4 y_{2j} = 65 + 73 + 78 + 66 = 282$$

$$y_{3\bullet} = \sum_{j=1}^5 y_{3j} = 61 + 79 + 67 + 74 + 65 = 346$$

Recall that $y_{\bullet\bullet}$ is the total sum of all the results.

$$y_{\bullet\bullet} = 657 + 282 + 346 = 1,285$$

Finally:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 = 72^2 + 81^2 + 96^2 + \dots + 67^2 + 74^2 + 65^2 = 99,117$$

- (ii) Recalling from Question 14.1 that $n_1 = 8$, $n_2 = 4$, $n_3 = 5$ and $n = 17$:

$$SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{n} = 99,117 - \frac{1,285^2}{17} = 1,986.12$$

$$SS_B = \sum_{i=1}^3 \frac{y_{i\bullet}^2}{n_i} - \frac{y_{\bullet\bullet}^2}{n} = \left(\frac{657^2}{8} + \frac{282^2}{4} + \frac{346^2}{5} \right) - \frac{1,285^2}{17} = 649.44$$

$$SS_R = SS_T - SS_B = 1,986.12 - 649.44 = 1,336.68$$

Solution 14.11

For this data:

Number of treatments: $k = 5$

Sample sizes:

$$n_1 = 5 \quad n_2 = 4 \quad n_3 = 4 \quad n_4 = 3 \quad n_5 = 3 \quad n = 19$$

We are given the sample totals:

$$y_{1\bullet} = 106 \quad y_{2\bullet} = 176 \quad y_{3\bullet} = 193 \quad y_{4\bullet} = 206 \quad y_{5\bullet} = 213 \quad y_{\bullet\bullet} = 894$$

$$\sum \sum y_{ij}^2 = 50,354$$

$$SS_T = 50,354 - \frac{894^2}{19} = 8,288.9$$

$$SS_B = \left(\frac{106^2}{5} + \frac{176^2}{4} + \frac{193^2}{4} + \frac{206^2}{3} + \frac{213^2}{3} \right) - \frac{894^2}{19} = 6,506.7$$

$$SS_R = 8,288.9 - 6,506.7 = 1,782.2$$

The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between treatments	4	6,506.7	1,626.7
Residual	14	1,782.2	127.3
Total	18	8,288.9	

The variance ratio is $F = \frac{1,626.7}{127.3} = 12.78$.

Under H_0 , this has a $F_{4,14}$ distribution.

The 1% critical point is 5.035, so we have very strong evidence to reject H_0 , and conclude that the mean stress levels in the companies are different.

Solution 14.12

- (i) Calculating the quantities required:

$$\sum_i \sum_j Y_{ij}^2 = 162,616 + 246,771 + 117,514 = 526,901$$

$$SS_T = 52,6901 - \frac{2,799^2}{15} = 4,607.6$$

$$SS_B = \left(\frac{900^2}{5} + \frac{1215^2}{6} + \frac{684^2}{4} \right) - \frac{2,799^2}{15} = 2,708.1$$

$$SS_R = 4,607.6 - 2,708.1 = 1,899.5$$

The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between treatments	2	2,708.1	1,354.05
Residual	12	1,899.5	158.29
Total	14	4,607.6	

The variance ratio is $F = \frac{1,354.05}{158.29} = 8.554$.

Under H_0 , this has a $F_{2,12}$ distribution.

The 1% critical point is 6.927, so we have strong evidence to reject H_0 and conclude that the mean performance of the balls is different.

- (ii) The conclusion is dependent on the assumption that the scores conform to a normal distribution with equal variances. Otherwise the χ^2 and F results must be considered to be only approximations.

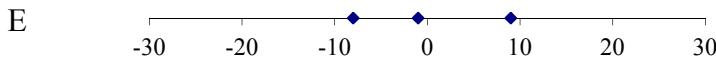
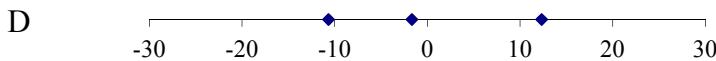
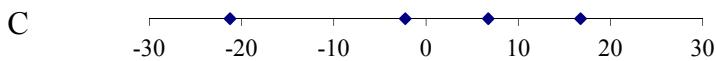
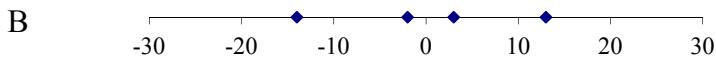
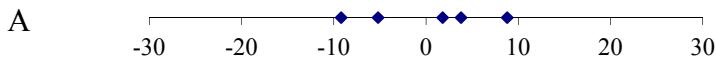
Solution 14.13

We have $\bar{y}_1 = 21.2$, $\bar{y}_2 = 44$, $\bar{y}_3 = 48.25$, $\bar{y}_4 = 68.67$, $\bar{y}_5 = 71$.

So the residuals are:

- A: 8.8 3.8 -9.2 1.8 -5.2
- B: -2 13 3 -14
- C: 16.75 -2.25 6.75 -21.25
- D: -1.67 -10.67 12.33
- E: -1 -8 9

The plots are:



It is difficult to judge normality due to the small amount of data.

Solution 14.14

- (i) The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between treatments	2	649.44	324.72
Residual	14	1,336.68	95.477
Total	16	1,986.12	

The variance ratio is $F = \frac{324.72}{95.477} = 3.40$.

Under H_0 , this has a $F_{2,14}$ distribution.

The 5% critical point is 3.739, so we have no evidence to reject H_0 , and we conclude that the mean pulse rates between the three departments are the same.

- (ii) (a) The confidence is given by:

$$(\bar{y}_{1\bullet} - \bar{y}_{3\bullet}) \pm t\hat{\sigma} \left(\frac{1}{n_1} + \frac{1}{n_3} \right)^{\frac{1}{2}}$$

Our variance estimate is given by:

$$\hat{\sigma}^2 = \frac{SS_R}{n-k} = \frac{1,336.68}{14} = 95.477$$

The 2½% value for t_{14} is 2.145. So our confidence interval is:

$$(82.125 - 69.2) \pm 2.145 \times \sqrt{95.477} \times \sqrt{\left(\frac{1}{8} + \frac{1}{5}\right)} = (0.976, 24.9)$$

(b) Since the confidence interval is greater than zero, the mean sales pulse rate is greater than the mean admin pulse rate.

(c) The ANOVA compares *all* of the treatment means with the overall treatment mean of 75.6. This is different to just comparing the biggest and smallest treatment means.

Solution 14.15

- (i) The least significant difference between $\bar{y}_{3\bullet}$ and $\bar{y}_{1\bullet}$ is given by:

$$t_{0.025,24} \times \hat{\sigma} \sqrt{\left(\frac{1}{n_3} + \frac{1}{n_1}\right)} = 2.064 \times \sqrt{46.43} \times \sqrt{\left(\frac{1}{9} + \frac{1}{10}\right)} = 6.46$$

- (ii) The least significant difference between $\bar{y}_{2\bullet}$ and $\bar{y}_{3\bullet}$ is given by:

$$t_{0.025,24} \times \hat{\sigma} \sqrt{\left(\frac{1}{n_2} + \frac{1}{n_3}\right)} = 2.064 \times \sqrt{46.43} \times \sqrt{\left(\frac{1}{8} + \frac{1}{9}\right)} = 6.83$$

Now we have $\bar{y}_{3\bullet} - \bar{y}_{2\bullet} = 37.2 - 28.125 = 9.10$. Since this is greater than the least significant difference, this means that there *is* a significant difference between them. Hence, we should *not* underline all three treatment means together.

Solution 14.16

Since, $\bar{y}_{1\bullet} = 180$, $\bar{y}_{2\bullet} = 202.5$, $\bar{y}_{3\bullet} = 171$ we can write $\bar{y}_{3\bullet} < \bar{y}_{1\bullet} < \bar{y}_{2\bullet}$.

Now $\hat{\sigma}^2 = \frac{SS_R}{n-k} = \frac{1,899.5}{15-3} = 158.29$ (which we calculated in 0).

The least significant difference between $\bar{y}_{3\bullet}$ and $\bar{y}_{1\bullet}$ is:

$$t_{0.025,12} \hat{\sigma} \sqrt{\left(\frac{1}{n_3} + \frac{1}{n_1}\right)} = 2.179 \times \sqrt{158.29} \times \sqrt{\frac{1}{4} + \frac{1}{5}} = 18.39$$

$\bar{y}_{1\bullet} - \bar{y}_{3\bullet} = 9$, so there is *no* significant difference between balls 1 and 3. Hence:

$$\underline{\bar{y}_{3\bullet}} < \bar{y}_{1\bullet} < \bar{y}_{2\bullet}$$

For $\bar{y}_{1\bullet}$ and $\bar{y}_{2\bullet}$, we have the least significant difference as:

$$2.179 \times \sqrt{158.29} \times \sqrt{\frac{1}{6} + \frac{1}{5}} = 16.60$$

$\bar{y}_{2\bullet} - \bar{y}_{1\bullet} = 22.5$, so there *is* a significant difference between balls 1 and 2 (and hence balls 2 and 3). So we don't underline $\bar{y}_{1\bullet}$ and $\bar{y}_{2\bullet}$. The diagram remains as before.

Exam-type question

$$(i) \quad \hat{\mu} = \bar{Y}_{..} = \frac{1}{n} \sum_i \sum_j Y_{ij} \Rightarrow \hat{\mu} = \frac{1}{23} \times 1,707 = 74.217$$

$$\hat{\tau}_i = \bar{Y}_{i\cdot} - \bar{Y}_{..} = \frac{1}{n_i} \sum_j Y_{ij} - \frac{1}{n} \sum_i \sum_j Y_{ij}$$

$$\Rightarrow \hat{\tau}_1 = \frac{399}{6} - 74.217 = -7.717$$

$$\Rightarrow \hat{\tau}_2 = \frac{296}{4} - 74.217 = -0.217$$

$$\Rightarrow \hat{\tau}_3 = \frac{389}{5} - 74.217 = 3.583$$

$$\Rightarrow \hat{\tau}_4 = \frac{378}{5} - 74.217 = 1.383$$

$$\Rightarrow \hat{\tau}_5 = \frac{245}{3} - 74.217 = 7.449$$

- (ii) If we assume the observations are from normal populations with the same variance, we can apply an ANOVA test of the hypotheses:

H_0 : Each student has the same mean score.

H_1 : There are differences between the mean marks obtained by the different students.

For this data:

Number of treatments: $k = 5$

Sample sizes:

$$n_1 = 6 \quad n_2 = 4 \quad n_3 = 5 \quad n_4 = 5 \quad n_5 = 3 \quad n = 23$$

Sample totals:

$$y_{1\cdot} = 399 \quad y_{2\cdot} = 296 \quad y_{3\cdot} = 389 \quad y_{4\cdot} = 378 \quad y_{5\cdot} = 245 \quad y_{..} = 1,707$$

$$\sum \sum y_{ij}^2 = 128,841$$

$$SS_T = 128,841 - \frac{1,707^2}{23} = 2,151.9$$

$$SS_B = \left(\frac{399^2}{6} + \frac{296^2}{4} + \frac{389^2}{5} + \frac{378^2}{5} + \frac{245^2}{3} \right) - \frac{1,707^2}{23} = 597.75$$

$$SS_R = SS_T - SS_B = 1,554.2$$

The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between treatments	4	597.75	149.44
Residual	18	1,554.2	86.34
Total	22	2,151.9	

The variance ratio is $F = \frac{149.44}{86.34} = 1.731$. Under H_0 , this has a $F_{4,18}$ distribution.

The 5% critical point is 2.928, so we have insufficient evidence to reject H_0 at the 5% level. Therefore we conclude that there is no difference between the students mean scores at the 5% level.

- (iii) The unbiased estimate of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-k} SS_R = 86.34$$

Now:

$$\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k} \Rightarrow \frac{1,554.2}{\sigma^2} \sim \chi^2_{18}$$

So, using the critical values of 8.231 and 31.53, we obtain a confidence interval for σ^2 of $\left(\frac{1,554.2}{31.53}, \frac{1,554.2}{8.231} \right) = (49.3, 189)$.

Hence, the confidence interval for σ is (7.02, 13.7).

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.

Part 1 – Questions

Introduction

The Question and Answer Bank is divided into five parts. The first four parts of the Question and Answer Bank are split into two sections:

- Section 1 – Development questions. The aim of these questions is to build on your understanding, test key Core Reading and bring your knowledge and skills to the level required to tackle exam-style questions.
- Section 2 – Exam-style questions. These questions are of the level of difficulty you are likely to face in the examination. It is very important that you focus on these questions as preparation for the exam.

The last part contains a set of exam-style questions covering the whole course.

For each part the questions may require knowledge from earlier parts of the course.

We strongly recommend that you use these questions to practise the techniques necessary to pass the exam. Do not use them as a set of material to *learn* but attempt the questions for yourself under strict exam-style conditions, before looking at the solutions provided.

This distinction represents the difference between active studying and passive studying. Given that the examiners will be aiming to set questions to make you think (and in doing so they will be devising questions you have not seen before) it is much better if you practise the skills that they will be testing.

It may also be useful to you if you group a number of the questions together to attempt under exam conditions. Ideally three hours would be set aside, but anything from one hour (*ie* 35 marks) upwards will help your time management.

Note that the split between development questions and exam-style questions is somewhat subjective. For example, there have been past CT3 exam questions that test knowledge of the Core Reading, and so are similar to what we've included here as development questions. The exam-style questions involve more application and a wider range of ideas and are typically the more challenging questions in the exam.

1 ***Development questions***

Question 1.1

Calculate three commonly used statistics that provide a measure of spread, and are expressed in terms of the original unit of measure, for the following set of ten observations:

5.1, 2.6, 7.3, 4.4, 4.6, 2.9, 3.4, 3.2, 4.4, 5.0 [4]

Question 1.2

For a particular set of data x_1, x_2, \dots, x_{100} , $\Sigma x_i = 550$, $\Sigma x_i^2 = 3,250$ and $\Sigma x_i^3 = 19,750$. The data set is:

- A symmetrical
- B positively skewed
- C negatively skewed
- D cannot be determined. [3]

Question 1.3

A life insurance company examines the ages last birthday of the last 100 policyholders to take out endowment assurance with them. The results are shown below:

Age (years)	0 – 14	15 – 19	20 – 24	25 – 34	35 – 54	55 – 79
Frequency	9	28	21	16	14	12

- (i) Draw a histogram for these data and use it to comment on the shape of the distribution. [4]
- (ii) Estimate the mean age for these policyholders. [2]
- (iii) Estimate the median age for these policyholders. [2]
- (iv) How do the result of parts (ii) and (iii) confirm your observations in part (i)? [1]
[Total 9]

Question 1.4

A small shop has a “pick-n-mix” counter where customers may choose wine gums, jelly beans or cola bottles. The probability that a customer purchases cola bottles is 0.45, jelly beans and wine gums 0.19, cola bottles and jelly beans 0.15, cola bottles and wine gums 0.25, cola bottles **or** jelly beans 0.6, cola bottles **or** wine gums 0.84, and at least one of them 0.9.

Calculate the probability that a customer purchases:

- (i) jelly beans [2]
 - (ii) wine gums [2]
 - (iii) all three [3]
 - (iv) none of them. [1]
- [Total 8]

Question 1.5

A discrete random variable X has probability function given by:

x	0	1	2
$P(X = x)$	0.3	0.5	0.2

Calculate:

- (i) $E(X)$ [1]
 - (ii) $\text{var}(X)$ [2]
 - (iii) the coefficient of skewness. [2]
- [Total 5]

Question 1.6

Calculate $P(X < 8)$ if:

- (i) X is the number of claims reported in a year by 20 policyholders. Each policyholder makes claims at the rate of 0.2 per year independently of the other policyholders. [2]
 - (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. [2]
 - (iii) X is the number of deaths amongst a group of 500 policyholders. Each policyholder has a 0.01 probability of dying independently of any other policyholder. [2]
 - (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. [2]
- [Total 8]

Question 1.7

Calculate $P(X < 8)$ if:

- (i) $X \sim U(5,10)$ [1]
 - (ii) $X \sim N(10,5)$ [2]
 - (iii) $X \sim Exp(0.5)$ [1]
 - (iv) $X \sim \chi^2_5$ [1]
 - (v) $X \sim Gamma(8,2)$ [2]
 - (vi) $X \sim \log N(2,5)$ [2]
- [Total 9]

Question 1.8

Derive an iterative formula for the probability function of the Poisson distribution. [3]

Question 1.9

A random variable X has a $Poi(3.6)$ distribution.

- (i) Calculate the mode of the probability distribution. [2]
 - (ii) Calculate the standard deviation of the distribution. [1]
 - (iii) State, with reasons, whether the distribution is positively or negatively skewed. [1]
- [Total 4]

Question 1.10

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 the variance of U and V . Comment on your answers.

[6]

Question 1.11

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. [5]

Question 1.12

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. [3]

Question 1.13

X and Y are independent exponential random variables with parameters λ and μ , respectively. By first considering the distribution function of $W = \min(X, Y)$, find the PDF of W . Hence identify the distribution of W . [4]

Question 1.14

If the random variable X has a Poisson distribution with mean λ , derive an expression for the expected value of $1/(X + 1)$. [3]

Question 1.15

Show that the variance of a discrete random variable X is given by:

$$\text{var}(X) = G''(1) + G'(1) - [G'(1)]^2$$

where $G(t)$ denotes the probability generating function of X . [4]

Question 1.16

Use cumulant generating functions to find expressions for the mean, variance and skewness of the Poisson distribution with parameter μ . [3]

Question 1.17

- (i) If the moment generating function of X is $M_X(t)$, then derive an expression for the moment generating function of $2X + 3$. [2]
- (ii) Hence, if X is normally distributed with mean μ and variance σ^2 , derive the distribution of $2X + 3$. [2]
- [Total 4]

Question 1.18

X is normally distributed with mean μ and variance σ^2 . Use generating functions to find the fourth central moment of X . [3]

2 Exam-style questions

Question 1.19

In a claims department of a motor insurance company, a sample of 12 claims had a mean of £845 and a standard deviation of £208.

It was then discovered that two mistakes were made; a claim of £526 was classified wrongly and is removed from the set of 12 claims, and a new claim of £1,034 is added to make the sample back up to 12.

Calculate the sample mean and standard deviation of the modified sample of 12 claims.

[4]

Question 1.20

An actuarial recruitment company places adverts in three publications with probabilities of 0.2, 0.3 and 0.5 respectively.

The probability that the recruitment company gets an enquiry from an advert in the first publication is 0.001. The probabilities for the other two publications are 0.002 and 0.004 respectively.

Given that the company has just received an enquiry, calculate the probability that it came from an advert in the first publication. [3]

Question 1.21

A continuous random variable Y has PDF:

$$f(y) = \begin{cases} y(y-1)(y-2) + 0.4 & 0 \leq y \leq 2 \\ c & 2 < y \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant.

Determine:

- (i) the value of c [3]
 - (ii) $E[Y]$ [2]
 - (iii) the standard deviation of Y . [2]
- [Total 7]

Question 1.22

The random variable U has PDF:

$$f(u) = 1, \quad 0 < u < 1$$

Determine the PDF of U^2 . [3]

Question 1.23

Let X be a random variable with an exponential distribution, so that the probability density function is given by:

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

- (i) Determine the probability density function of the random variable Y , where $Y = X^2$. [3]
 - (ii) Show that Y has a Weibull distribution, stating clearly the parameters of the distribution. [1]
- [Total 4]

Question 1.24

The random variable X has a normal distribution with mean μ and variance σ^2 , such that the probability density function of X is:

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

- (i) Determine the probability density function of the random variable Y , where $Y = 2X$. [3]
 - (ii) Hence, show that Y is normally distributed with mean 2μ and variance $4\sigma^2$. [1]
 - (iii) Assuming the moment generating function of a $N(\mu, \sigma^2)$ random variable, prove the result obtained in part (ii). [3]
- [Total 7]

Question 1.25

An insurance company issues policies covering the loss of legs on pet spiders. Policies are issued to the owners and cover any number of spiders. The number of claims per policy in a year has a Poisson distribution with mean 0.6.

- (i) Calculate the probability that a particular policyholder makes at least 3 claims this year. [2]
 - (ii) The policyholder in part (i) is known to have already made at least 1 claim this year. Calculate a revised probability that this policyholder makes at least 3 claims in total over this year. [2]
 - (iii) Calculate the probability that a portfolio of 5 independent policies gives rise to at least 3 claims this year. [2]
 - (iv) Obtain the probability that there is a wait of at least 4 months between claims on a single randomly selected policy. State any assumptions that you make. [3]
- [Total 9]

Question 1.26

Simulate two observations from the distribution with probability density function:

$$f(x) = \frac{50}{(5+x)^3}, \quad x > 0$$

using the random numbers 0.863 and 0.447 selected from the uniform distribution on the interval (0,1). [4]

Question 1.27

The probability generating function for a discrete random variable X is given by:

$$G_X(t) = 0.1(1+t)(4+t^2)$$

Calculate:

- (i) $E(X)$ [1]
 - (ii) the standard deviation X [2]
 - (iii) $P(X \geq 2)$. [2]
- [Total 5]

Question 1.28

The moment generating function, $M_Y(t)$, of a random variable, Y , is given by:

$$M_Y(t) = (1 - 4t)^{-2} \quad t < 0.25$$

Calculate:

- (i) $E(Y)$ [1]
 - (ii) the standard deviation Y [2]
 - (iii) $E(Y^6)$. [2]
- [Total 5]

Question 1.29

The random variable U has a geometric distribution with probability function:

$$P(U = u) = pq^{u-1} \quad u = 1, 2, 3, \dots \quad \text{where } p + q = 1$$

- (i) Derive the probability generating function of U . [2]
 - (ii) Derive the moment generating function of U . [2]
 - (iii) Explain the connection between the PGF and MGF. [1]
- [Total 5]

Question 1.30

A random variable X has probability density function:

$$f(x) = ke^{-2x} \quad x > R$$

where R is a positive constant and k is a constant.

- (i) (a) Derive a formula for the moment generating function of X .
 (b) State the values of t for which your formula in (i)(a) is valid. [4]
 - (ii) Hence determine the value of the constant k in terms of R . [1]
- [Total 5]

Question 1.31

- (i) Derive, from first principles, the moment generating function of a $\text{Gamma}(\alpha, \lambda)$ 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]

Part 1 – Solutions

Solution 1.1

The three most commonly used measures are the standard deviation, the range and the interquartile range. [1]

(a) ***Standard deviation***

For the data values given: $\sum x_i = 42.9$ and $\sum x_i^2 = 201.15$

So the sample variance is:

$$s^2 = \frac{1}{9} \left\{ 201.15 - 10 \times 4.29^2 \right\} = 1.901$$

So the sample standard deviation (which is measured in the original units) is:

$$s = \sqrt{1.901} = 1.379 \quad [1]$$

(b) ***Range***

Arranging the values in ascending order gives:

2.6, 2.9, 3.2, 3.4, 4.4, 4.4, 4.6, 5.0, 5.1, 7.3

So the range of the values is $s = 7.3 - 2.6 = 4.7$. [1]

(c) ***IQR***

The lower quartile is the $\frac{10+2}{4} = 3$ rd value: 3.2

The upper quartile is the $\frac{3 \times 10 + 2}{4} = 8$ th value: 5.0

So the interquartile range is $5.0 - 3.2 = 1.8$. [1]

Using the alternative quartile definitions gives a lower quartile of $\frac{10+1}{4} = 2\frac{3}{4}$ th value which is 3.125 and an upper quartile of $\frac{3 \times 10 + 3}{4} = 8\frac{1}{4}$ th value which is 5.025. So the interquartile range under these definitions is $5.025 - 3.125 = 1.9$.

Solution 1.2

The sign of the third central sample moment will indicate the direction of the skewness:

$$\begin{aligned}\sum (x_i - \bar{x})^3 &= \sum x_i^3 - 3\bar{x}\sum x_i^2 + 2n\bar{x}^3 \\ &= 19,750 - 3 \times 5.5 \times 3,250 + 2 \times 100 \times 5.5^3 = -600\end{aligned}$$

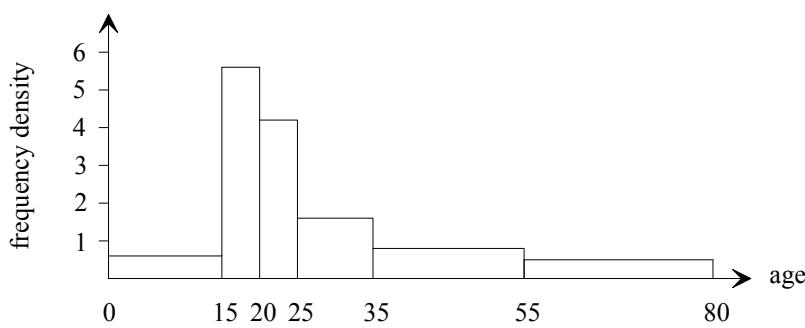
So the data set is negatively skewed.

[3]

Solution 1.3(i) **Histogram**

To draw a histogram we calculate the heights using $\frac{\text{frequency}}{\text{width}}$. The width of the 0–14 group is 15, as someone aged 14 could be aged just under 15 years.

The heights are: $\frac{9}{15} = 0.6$, $\frac{28}{5} = 5.6$, $\frac{21}{5} = 4.2$, $\frac{16}{10} = 1.6$, $\frac{14}{20} = 0.7$, $\frac{12}{25} = 0.48$ [1]



[2]

The distribution appears to be positively skewed.

[1]

(ii) **Mean**

Using $\bar{x} = \frac{\sum fx}{\sum f}$ where x is the midpoint of the groups gives:

$$\begin{aligned}\bar{x} &= \frac{9 \times 7.5 + 28 \times 17.5 + 21 \times 22.5 + 16 \times 30 + 14 \times 45 + 12 \times 67.5}{9 + 28 + 21 + 16 + 14 + 12} \\ &= \frac{2,950}{100} \\ &= 29.5 \text{ years}\end{aligned}$$

[2]

(iii) **Median**

The median splits the distribution into two equal halves, so that half of the observations (*ie* 50) are on one side of the median and half are on the other. We can see that 37 policyholders are aged under 20; and 42 are aged 25 and over. The median lies in the age group 20 to 25. Using interpolation gives:

$$\text{median} = 20 + \frac{13}{21} \times 5 = 23.1 \text{ years} \quad [2]$$

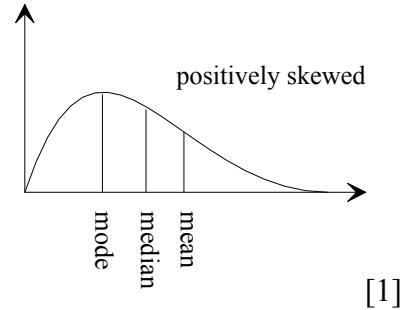
Alternatively, using the $\frac{1}{2}(n+1)$ definition for the median, we would be 13.5 observations into the 20-24 group. This would give a median of:

$$\text{median} = 20 + \frac{13.5}{21} \times 5 = 23.2 \text{ years}$$

(iv) **Comment**

The mean is greater than the median indicating the distribution is positively skewed.

The mean is “pulled” to the right by the few extreme positive values, whereas the median is not affected by these but just stays in the middle of the distribution.



[1]

Solution 1.4

Let C be the event “cola bottles are purchased”, W the event “wine gums are purchased” and J the event “jelly beans are purchased”.

We are given:

$$\begin{aligned} P(C) &= 0.45 & P(J \cap W) &= 0.19 & P(C \cap J) &= 0.15 & P(C \cap W) &= 0.25 \\ P(C \cup J) &= 0.6 & P(C \cup W) &= 0.84 & P(C \cup J \cup W) &= 0.9 \end{aligned}$$

(i) **Probability purchase jelly beans**

We require $P(J)$. Using the addition rule for two sets:

$$\begin{aligned} P(C \cup J) &= P(C) + P(J) - P(C \cap J) \\ 0.6 &= 0.45 + P(J) - 0.15 \quad \Rightarrow \quad P(J) = 0.3 \end{aligned}$$

(ii) ***Probability purchase wine gums***

We require $P(W)$. Using the addition rule for two sets again:

$$P(C \cup W) = P(C) + P(W) - P(C \cap W)$$

$$0.84 = 0.45 + P(W) - 0.25$$

So:

$$P(W) = 0.64$$

(iii) ***Probability purchase all three***

We require $P(C \cap J \cap W)$. Using the additional rule for three sets:

$$P(C \cup J \cup W) = P(C) + P(J) + P(W) - P(C \cap J) - P(C \cap W) - P(J \cap W)$$

$$+ P(C \cap J \cap W)$$

$$0.9 = 0.45 + 0.3 + 0.64 - 0.15 - 0.25 - 0.19 + P(C \cap J \cap W)$$

So:

$$P(C \cap J \cap W) = 0.1$$

(iv) ***Probability purchase none of them***

We require $P(C' \cup J' \cup W')$. The easiest way to get this is to realise that the probability of buying none is the complement of choosing at least one:

$$P(C' \cup J' \cup W') = 1 - P(C \cup J \cup W) = 1 - 0.9 = 0.1$$

Solution 1.5(i) ***E(X)***

$$E(X) = \sum_x x P(X=x) = (0 \times 0.3) + (1 \times 0.5) + (2 \times 0.2) = 0.9 \quad [1]$$

(ii) ***Variance***

$$E(X^2) = \sum_x x^2 P(X=x) = (0^2 \times 0.3) + (1^2 \times 0.5) + (2^2 \times 0.2) = 1.3 \quad [1]$$

$$\Rightarrow \text{var}(X) = E(X^2) - E^2(X) = 1.3 - 0.9^2 = 0.49 \quad [1]$$

(iii) ***Coefficient of skewness***

The skewness is given by:

$$\begin{aligned} \text{skew}(X) &= E[(X - \mu)^3] = \sum_x (x - \mu)^3 P(X=x) \\ &= (0 - 0.9)^3 \times 0.3 + (1 - 0.9)^3 \times 0.5 + (2 - 0.9)^3 \times 0.2 = 0.048 \end{aligned} \quad [1]$$

Alternatively, we could use $\text{skew}(X) = E(X^3) - 3\mu E(X^2) + 2\mu^3$ where $E(X^3) = 2.1$.

The coefficient of skewness is given by:

$$\frac{\text{skew}(X)}{[\text{var}(X)]^{1.5}} = \frac{0.048}{0.49^{1.5}} = 0.140 \quad [1]$$

Solution 1.6(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. [2]

Alternatively, we could use $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$ to calculate the values of $P(X = 0), P(X = 1), \dots, P(X = 7)$ (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 $p = 0.3$.

$$P(X = x) = \binom{x-1}{3} 0.3^4 0.7^{x-4} \quad x = 4, 5, 6, \dots$$

So $P(X < 8) = P(X = 4) + \dots + P(X = 7)$.

$$P(X = 4) = \binom{3}{3} 0.3^4 = 0.0081$$

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, $P(X < 8) = 0.0081 + 0.02268 + 0.03969 + 0.05557 = 0.12604$. [2]

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)).

$$\text{Bin}(500, 0.01) \doteq \text{Poi}(5)$$

Using the cumulative Poisson tables gives $P(X < 8) = P(X \leq 7) = 0.86663$. [2]

Alternatively, we could calculate this accurately:

$$P(X = 0) = \binom{500}{0} \times 0.99^{500} = 0.00657$$

Now using the iterative formula $P(X = x) = \frac{n-x+1}{x} \times \frac{p}{q} \times P(X = x-1)$:

$$P(X = 1) = \frac{500}{1} \times \frac{0.01}{0.99} \times 0.00657 = 0.03318$$

$$P(X = 2) = \frac{499}{2} \times \frac{0.01}{0.99} \times 0.03318 = 0.08363$$

$$P(X = 3) = \frac{498}{3} \times \frac{0.01}{0.99} \times 0.08363 = 0.14023$$

$$P(X = 4) = \frac{497}{4} \times \frac{0.01}{0.99} \times 0.14023 = 0.17600$$

$$P(X = 5) = \frac{496}{5} \times \frac{0.01}{0.99} \times 0.17600 = 0.17635$$

$$P(X = 6) = \frac{495}{6} \times \frac{0.01}{0.99} \times 0.17635 = 0.14696$$

$$P(X = 7) = \frac{494}{7} \times \frac{0.01}{0.99} \times 0.14696 = 0.10476$$

Hence, $P(X = 8) = P(X = 0) + \dots + P(X = 7) = 0.86768$.

(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 $p = 0.01$.

$$P(X = x) = 0.01 \times 0.99^x \quad x = 0, 1, 2, \dots$$

Now:

$$\begin{aligned} P(X < 8) &= P(X \leq 7) \\ &= P(X = 0) + \dots + P(X = 7) \\ &= 0.01 + 0.01 \times 0.99 + 0.01 \times 0.99^2 + \dots + 0.01 \times 0.99^7 \end{aligned} \quad [1]$$

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 $S_n = \frac{a(1 - r^n)}{1 - r}$. This gives:

$$P(X < 8) = \frac{0.01 \times (1 - 0.99^8)}{1 - 0.99} = 0.07726 \quad [1]$$

Solution 1.7(i) **Uniform**

$$P(X < 8) = \int_5^8 0.2 \, dx = [0.2x]_5^8 = 0.6 \quad [1]$$

Alternatively, we could simply use the DF given on page 13.
 $P(X < 8) = F(8) = \frac{8-5}{10-5} = 0.6$.

(ii) **Normal**

$$\begin{aligned} P(X < 8) &= P\left(Z < \frac{8-10}{\sqrt{5}}\right) = P(Z < -0.894) \\ &= 1 - P(Z < 0.894) \\ &= 1 - 0.81434 \\ &= 0.18566 \end{aligned} \quad [2]$$

(iii) ***Exponential***

$$P(X < 8) = \int_0^8 0.5e^{-0.5x} dx = \left[-e^{-0.5x} \right]_0^8 = 1 - e^{-4} = 0.98168 \quad [1]$$

Alternatively, we could simply use the DF given on page 11.
 $P(X < 8) = F(8) = 1 - e^{-0.5 \times 8} = 0.98168.$

(iv) ***Chi-squared***

Using the χ^2 tables on page 165 of the *Tables* gives $P(X < 8) = 0.8438$. [1]

(v) ***Gamma***

The only practical way in an exam to calculate probabilities of a $X \sim \text{Gamma}(\alpha, \lambda)$ distribution is to use the exact relationship $2\lambda X \sim \chi^2_{2\alpha}$ and then read off the probability on the χ^2 tables.

$$P(X < 8) = P(2\lambda X < 16\lambda) = P(4X < 32) = P(\chi^2_{16} < 32) = 0.99 \quad [2]$$

(vi) ***Lognormal***

Using the fact that if $X \sim \log N(\mu, \sigma^2)$ then $\ln X \sim N(\mu, \sigma^2)$:

$$\begin{aligned} P(X < 8) &= P(\ln X < \ln 8) = P\left(Z < \frac{\ln 8 - 2}{\sqrt{5}}\right) \\ &= P(Z < 0.036) \\ &= 0.51436 \end{aligned} \quad [2]$$

Solution 1.8

For the $Poisson(\lambda)$ distribution:

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \Rightarrow P(X = x - 1) = \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda}$$

$$\Rightarrow \frac{P(X = x)}{P(X = x - 1)} = \frac{\lambda^x e^{-\lambda}}{x!} \Big/ \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} = \frac{\lambda}{x}$$

Therefore the iterative formula is:

$$P(X = 0) = e^{-\lambda} \quad [1]$$

$$P(X = x) = \frac{\lambda}{x} P(X = x - 1) \quad x = 1, 2, \dots \quad [2]$$

Solution 1.9(i) **Mode**

We can find the mode simply by calculating probabilities and seeing which value has the highest probability.

$$P(X = 0) = e^{-3.6} = 0.02732$$

Using the iterative formula derived in Solution 1.8 gives:

$$P(X = 1) = \frac{3.6}{1} \times 0.02732 = 0.09837$$

$$P(X = 2) = \frac{3.6}{2} \times 0.09837 = 0.17706$$

$$P(X = 3) = \frac{3.6}{3} \times 0.17706 = 0.21247$$

$$P(X = 4) = \frac{3.6}{4} \times 0.21247 = 0.19122$$

$$P(X = 5) = \frac{3.6}{5} \times 0.19122 = 0.13768 \quad [1]$$

etc

We can clearly see that 3 is the mode.

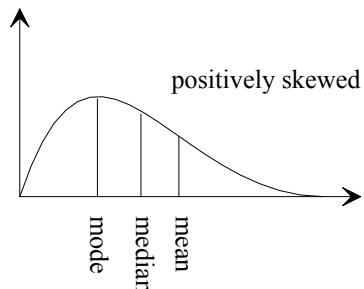
[1]

(ii) **Standard deviation**

The variance of the $\text{Poisson}(\lambda)$ distribution is λ . So the standard deviation of the $\text{Poisson}(3.6)$ distribution is $\sqrt{3.6} = 1.8974$. [1]

(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.



[1]

Solution 1.10

The probability density function of U is constant, ie $f_U(u) = \frac{1}{2}$, $-1 < u < 1$.

The probability function of V is constant ie $f_V(v) = \frac{1}{5}$, $v = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$.

By symmetry the mean value of both variables is zero.

[1]

Alternatively:

$$E(U) = \int u f(u) du = \int_{-1}^1 \frac{1}{2} u du = \left[\frac{1}{4} u^2 \right]_{-1}^1 = \frac{1}{4} - \frac{1}{4} = 0$$

$$\begin{aligned} E(V) &= \sum v P(V=v) \\ &= (-1 \times \frac{1}{5}) + (-\frac{1}{2} \times \frac{1}{5}) + (0 \times \frac{1}{5}) + (\frac{1}{2} \times \frac{1}{5}) + (1 \times \frac{1}{5}) = 0 \end{aligned}$$

So the variance of U is:

$$\begin{aligned} E(U^2) &= \int_{-1}^1 \frac{1}{2} u^2 du = \left[\frac{1}{6} u^3 \right]_{-1}^1 = \frac{1}{3} \\ \Rightarrow \quad \text{var}(U) &= \frac{1}{3} - 0^2 = \frac{1}{3} \end{aligned} \quad [2]$$

Alternatively, you could use the formula $\frac{1}{12}(b-a)^2$ from page 13 of the Tables.

So the variance of V is:

$$\begin{aligned} E(V^2) &= \sum v^2 P(V=v) = \frac{1}{5} \left[(-1)^2 + \left(\frac{1}{2}\right)^2 + 0^2 + \left(\frac{1}{2}\right)^2 + 1^2 \right] = \frac{1}{2} \\ \Rightarrow \text{var}(V) &= \frac{1}{2} - 0^2 = \frac{1}{2} \end{aligned} \quad [1]$$

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 $+1$. [2]

Solution 1.11

Let X denote the random variable and recalling that the mean and variance of a lognormal distribution are *not* μ and σ^2 .

Using the formulae for the mean and variance of a lognormal distribution:

$$E[X] = e^{\mu + \frac{1}{2}\sigma^2} = 10 \quad (1)$$

$$\text{var}(X) = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1) = 4 \quad (2)$$

Squaring equation (1) and substituting into equation (2):

$$\begin{aligned} \text{var}(X) &= 10^2 \left(e^{\sigma^2} - 1 \right) = 4 \\ \Rightarrow e^{\sigma^2} - 1 &= 0.04 \\ \Rightarrow \sigma^2 &= \log 1.04 = 0.03922 \end{aligned} \quad [2]$$

Substituting this into equation (1) gives:

$$\mu = \log 10 - \frac{1}{2}\sigma^2 = 2.2830 \quad [1]$$

So the required probability is:

$$\begin{aligned}
 P(7.5 < X < 12.5) &= P(X < 12.5) - P(X < 7.5) \\
 &= P(\ln X < \ln 12.5) - P(\ln X < \ln 7.5) \\
 &= P\left(Z < \frac{\log 7.5 - 2.2830}{\sqrt{0.03922}}\right) - P\left(Z < \frac{\log 12.5 - 2.2830}{\sqrt{0.03922}}\right) \\
 &= \Phi(1.226) - \Phi(-1.354) \\
 &= 0.88990 - 0.08787 \\
 &= 0.802
 \end{aligned} \tag{2}$$

Solution 1.12

The conditional probability is:

$$P(N = 1 | N \geq 1) = \frac{P(N = 1)}{P(N \geq 1)} = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} = \frac{\lambda}{e^\lambda - 1} \tag{1}$$

$$\text{Trial and improvement gives } \frac{1.62}{e^{1.62} - 1} = 0.3997. \text{ So } \lambda \approx 1.62. \tag{2}$$

Solution 1.13

The distribution function of W is:

$$\begin{aligned}
 P(W < w) &= P[\min(X, Y) < w] = 1 - P[\min(X, Y) \geq w] \\
 &= 1 - P[X \geq w \text{ & } Y \geq w]
 \end{aligned} \tag{1}$$

Since X and Y are independent, we can factorise this probability:

$$P(W < w) = 1 - P(X \geq w)P(Y \geq w) \tag{1/2}$$

Since $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$, this gives:

$$P(W < w) = 1 - e^{-\lambda w}e^{-\mu w} = 1 - e^{-(\lambda + \mu)w} \tag{1}$$

Differentiating to find the PDF of W :

$$f_W(w) = (\lambda + \mu)e^{-(\lambda + \mu)w} \quad [1]$$

So, since W can take any positive value, it has an $\text{Exp}(\lambda + \mu)$ distribution. [½]

Solution 1.14

From the definition of the expected value of a function:

$$\begin{aligned} E\left[\frac{1}{X+1}\right] &= \sum_{x=0}^{\infty} \frac{1}{x+1} P(X=x) = \sum_{x=0}^{\infty} \frac{1}{x+1} \frac{e^{-\lambda} \lambda^x}{x!} \\ &= \frac{e^{-\lambda}}{\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x+1}}{(x+1)!} = \frac{e^{-\lambda}}{\lambda} (e^\lambda - 1) = \frac{1-e^{-\lambda}}{\lambda} \end{aligned} \quad [3]$$

Solution 1.15

The definition of the PGF is:

$$G(t) = E[t^X]$$

Differentiating with respect to t :

$$G'(t) = E[Xt^{X-1}] \quad [1]$$

$$G''(t) = E[X(X-1)t^{X-2}] \quad [1]$$

Putting $t = 1$ into each of the above expressions:

$$G'(1) = E[X]$$

$$G''(1) = E[X(X-1)] = E[X^2] - E[X]$$

So the variance is:

$$\begin{aligned} \text{var}(X) &= E[X^2] - (E[X])^2 \\ &= E[X^2] - E[X] + E[X] - (E[X])^2 \\ &= G''(1) + G'(1) - [G'(1)]^2 \end{aligned} \quad [2]$$

Alternatively, you can start from the Taylor series for $[1 + (t - 1)]^X$.

Solution 1.16

From the Tables, the moment generating function of the Poisson distribution is $M_X(t) = e^{\mu(e^t - 1)}$. So the cumulant generating function is found as follows:

$$C_X(t) = \ln M_X(t) = \mu(e^t - 1) \quad [1]$$

Differentiating:

$$C'_X(t) = \mu e^t$$

Repeated differentiation will give the same expression, so we can substitute $t = 0$ into the expressions to find that the mean, variance and skewness are all equal to μ . [2]

Solution 1.17

(i) **MGF**

The MGF of X is:

$$M_X(t) = E(e^{tX})$$

So the MGF of $2X + 3$ is:

$$M_{2X+3}(t) = E[e^{t(2X+3)}] = E[e^{(2t)X+3t}] = e^{3t} E[e^{(2t)X}] = e^{3t} M_X(2t) \quad [2]$$

(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]

Solution 1.18

The MGF of $X - \mu$, which has a $N(0, \sigma^2)$ distribution, is:

$$M_{X-\mu}(t) = e^{\frac{1}{2}\sigma^2 t^2} = 1 + \frac{1}{2}\sigma^2 t^2 + \frac{1}{2}\left(\frac{1}{2}\sigma^2 t^2\right)^2 + \dots \quad [1]$$

So $E[(X - \mu)^4]$ is the coefficient of $\frac{t^4}{4!}$ in this series, ie:

$$E[(X - \mu)^4] = 3\sigma^4 \quad [2]$$

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 $E[(X - \mu)^4]$:

$$E[(X - \mu)^4] = E(X^4) - 4\mu E(X^3) + 6\mu^2 E(X^2) - 3\mu^4$$

Note: 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 $\text{skew}(X) = C'''_X(0)$. However, thereafter the CGF does **not** give central moments.

Solution 1.19

We are given that $n = 12$, $\bar{x} = 845$ and $s = 208$.

$$\bar{x} = \frac{\sum x}{n} \Rightarrow \sum x = 12 \times 845 = 10,140$$

$$s^2 = \frac{1}{n-1} \left(\sum x^2 - n\bar{x}^2 \right) \Rightarrow \sum x^2 = 11 \times 208^2 + 12 \times 845^2 = 9,044,204 \quad [1]$$

Now subtracting the £526 claim and adding the £1,034 claim gives:

$$\begin{aligned} \sum x &= 10,140 - 526 + 1034 = 10,648 \\ \sum x^2 &= 9,044,204 - 526^2 + 1,034^2 = 9,836,684 \end{aligned} \quad [1]$$

So the new results are:

$$\bar{x} = \frac{10,648}{12} = £887.33 \quad [1]$$

$$s^2 = \frac{1}{11} (9,836,684 - 12 \times 887.33^2) = 35,305.33$$

$$\Rightarrow s = \sqrt{35,305.33} = £187.90 \quad [1]$$

Solution 1.20

Let A be the event “an advert is placed in publication A”.

Let B be the event “an advert is placed in publication B”.

Let C be the event “an advert is placed in publication C”.

Let E be the event “an enquiry is received”.

We have:

$$\begin{aligned} P(A) &= 0.2 & P(B) &= 0.3 & P(C) &= 0.5 \\ P(E | A) &= 0.001 & P(E | B) &= 0.002 & P(E | C) &= 0.004 \end{aligned} \quad [1]$$

We want $P(A | E)$:

$$P(A | E) = \frac{P(E | A)P(A)}{P(E)} = \frac{P(E | A)P(A)}{P(E | A)P(A) + P(E | B)P(B) + P(E | C)P(C)} \quad [1]$$

Putting in the values from the question, we get:

$$P(A | E) = \frac{0.001 \times 0.2}{0.001 \times 0.2 + 0.002 \times 0.3 + 0.004 \times 0.5} = 0.0714 \quad [1]$$

Solution 1.21(i) **Determine c**

To find the value of c we need to integrate over the whole range of y and set the value of the integral equal to 1. Since the definition is different for different ranges of y , we need to carry out separate integrations:

$$\begin{aligned} & \int_0^2 (y(y-1)(y-2) + 0.4) dy + \int_2^4 c dy = 1 & [1] \\ \Rightarrow & \int_0^2 (y^3 - 3y^2 + 2y + 0.4) dy + \int_2^4 c dy = 1 \\ \Rightarrow & \left[\frac{y^4}{4} - y^3 + y^2 + 0.4y \right]_0^2 + [cy]_2^4 = 1 & [1] \\ \Rightarrow & (4 - 8 + 4 + 0.8) + (4c - 2c) = 1 \\ \Rightarrow & c = 0.1 & [1] \end{aligned}$$

(ii) **Expectation**

To find $E[Y]$ we need to multiply the density function by y and then integrate:

$$\begin{aligned} E[Y] &= \int_0^2 (y^4 - 3y^3 + 2y^2 + 0.4y) dy + \int_2^4 0.1y dy \\ &= \left[\frac{y^5}{5} - \frac{3y^4}{4} + \frac{2y^3}{3} + 0.2y^2 \right]_0^2 + \left[\frac{0.1y^2}{2} \right]_2^4 & [1] \\ &= \left(\frac{32}{5} - 12 + \frac{16}{3} + 0.8 \right) + (0.8 - 0.2) = 1.1333 & [1] \end{aligned}$$

(iii) **Standard deviation**

First we need $E[Y^2]$:

$$\begin{aligned} E[Y^2] &= \int_0^2 (y^5 - 3y^4 + 2y^3 + 0.4y^2) dy + \int_2^4 0.1y^2 dy \\ &= \left[\frac{y^6}{6} - \frac{3y^5}{5} + \frac{2y^4}{4} + \frac{0.4}{3}y^3 \right]_0^2 + \left[\frac{0.1y^3}{3} \right]_2^4 \\ &= \left(\frac{64}{6} - \frac{96}{5} + 8 + \frac{3.2}{3} \right) + \left(\frac{6.4}{3} - \frac{0.8}{3} \right) = 2.4 \end{aligned} \quad [\frac{1}{2}]$$

So the standard deviation is given by:

$$\sqrt{\text{var}(X)} = \sqrt{E(X^2) - E^2(X)} = \sqrt{2.4 - 1.1333^2} = \sqrt{1.1156} = 1.056 \quad [1]$$

Solution 1.22

Let $V = U^2$. Since U only takes positive values, the distribution function of V is:

$$F_V(v) = P(V \leq v) = P(U^2 \leq v) = P(U \leq \sqrt{v}) \quad [1]$$

We can calculate this using integration:

$$P(U < \sqrt{v}) = \int_0^{\sqrt{v}} 1 du = [u]_0^{\sqrt{v}} = \sqrt{v} \quad [\frac{1}{2}]$$

Alternatively, we could use the DF of a $U(0,1)$ from page 13 of the Tables:
 $F_U(u) = \frac{u-0}{1-0} = u$. Hence, $P(U \leq \sqrt{v}) = F_U(\sqrt{v}) = \sqrt{v}$.

So the PDF of V is:

$$f_V(v) = F'_V(v) = \frac{d}{dv} v^{\frac{1}{2}} = \frac{1}{2}v^{-\frac{1}{2}} \quad [1]$$

Since U can take values in the range $0 < U < 1$, V can take values in the range $0 < V < 1$. [\frac{1}{2}]

Alternatively, using the function of a random variable formula:

$$f_V(v) = \left| \frac{d}{dv} g^{-1}(v) \right| \times f_U(g^{-1}(v)) \quad \text{for } v = g(u)$$

This is written as $f(y) = f_x[w(y)] \left| \frac{dw(y)}{dy} \right|$ in Chapter 3, Section 4.2.

$$v = g(u) = u^2 \Rightarrow u = g^{-1}(v) = v^{1/2} \Rightarrow \frac{du}{dv} = \frac{1}{2}v^{-1/2}$$

$$\Rightarrow f_V(v) = \frac{1}{2}v^{-1/2} \times f_U(v^{1/2}) = \frac{1}{2}v^{-1/2} \times 1 \\ = \frac{1}{2}v^{-1/2}$$

If we write the PDF in the form $f_V(v) = \frac{1}{2}v^{1/2-1}(1-v)^0$, $0 < v < 1$, we can see that this is actually the PDF of a Beta($\frac{1}{2}, 1$) distribution.

Solution 1.23

(i) **PDF of Y**

Working from first principles by considering the distribution function of Y :

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq y^{1/2}) \quad \text{since } x > 0 \quad [1]$$

Using integration:

$$P(X \leq y^{1/2}) = \int_0^{y^{1/2}} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x} \right]_0^{y^{1/2}} = 1 - e^{-\lambda y^{1/2}} \quad [1]$$

Alternatively, we could use the DF of a $\text{Exp}(\lambda)$ from page 11 of the Tables:
 $F(x) = 1 - e^{-\lambda x}$. Hence $P(X \leq y^{1/2}) = F(y^{1/2}) = 1 - e^{-\lambda y^{1/2}}$ and:

$$f_Y(y) = F'_Y(y) = \frac{1}{2}\lambda y^{-1/2} e^{-\lambda y^{1/2}} \quad [1]$$

Alternatively, using the function of a random variable formula:

$$f_Y(y) = \left| \frac{d}{dy} u^{-1}(y) \right| \times f_X(u^{-1}(y)) \quad \text{for } y = u(x)$$

This is written as $f(y) = f_x[w(y)] \left| \frac{dw(y)}{dy} \right|$ in Chapter 3, Section 4.2.

$$y = u(x) = x^2 \Rightarrow x = u^{-1}(y) = y^{1/2} \Rightarrow \frac{dx}{dy} = \frac{1}{2}y^{-1/2}$$

$$\begin{aligned} \Rightarrow f_Y(y) &= \frac{1}{2}y^{-1/2} \times f_X\left(y^{1/2}\right) = \frac{1}{2}y^{-1/2} \times \lambda e^{-\lambda y^{1/2}} \\ &= \frac{1}{2}\lambda y^{-1/2} e^{-\lambda y^{1/2}} \end{aligned}$$

(ii) **Show distribution is a Weibull**

Comparing this to the Weibull pdf $f(x) = c\gamma x^{\gamma-1}e^{-cx^\gamma}$, we see that Y has a Weibull distribution with parameters $c = \lambda$ and $\gamma = \frac{1}{2}$. [1]

Solution 1.24(i) **PDF of Y**

For $y = u(x)$ then $f_Y(y) = f_X[w(y)] \left| \frac{dw(y)}{dy} \right|$, where $w(y) = u^{-1}(y)$. We have:

$$y = u(x) = 2x \Rightarrow x = u^{-1}(y) = \frac{1}{2}y \Rightarrow \frac{dx}{dy} = \frac{1}{2} \quad [1]$$

$$\Rightarrow f_Y(y) = \frac{1}{2} \times f_X\left(\frac{1}{2}y\right) = \frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\frac{1}{2}y-\mu)^2}{2\sigma^2}} \quad [2]$$

Alternatively, from first principles, we could consider the DF of Y :

$$F_Y(y) = P(Y \leq y) = P(2X \leq y) = P(X \leq \frac{1}{2}y) = \int_0^{\frac{1}{2}y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

We cannot easily work out this integral and we need to obtain $f_Y(y)$:

$$f_Y(y) = F'_Y(y) = \frac{\partial}{\partial y} \int_0^{\frac{1}{2}y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\frac{1}{2}y-\mu)^2}{2\sigma^2}}$$

We differentiated the integral using Leibnitz' formula given on page 3 of the Tables.

(ii) **Show Y is normally distributed**

$$f_Y(y) = \frac{1}{\sqrt{8\pi\sigma^2}} e^{-\frac{[\frac{1}{2}(y-2\mu)]^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi(4\sigma^2)}} e^{-\frac{(y-2\mu)^2}{2(4\sigma^2)}}$$

This is the pdf of a $N(2\mu, 4\sigma^2)$, so Y has a $N(2\mu, 4\sigma^2)$ distribution. [1]

(iii) **Prove part (ii) using MGFs**

Now $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, so:

$$M_Y(t) = E(e^{tY}) = E(e^{2tX}) = M_X(2t) = e^{2\mu t + 2\sigma^2 t^2} = e^{(2\mu)t + \frac{1}{2}(4\sigma^2)t^2} \quad [2]$$

This is the MGF of a $N(2\mu, 4\sigma^2)$. Therefore by the uniqueness property of MGFs, Y has a $N(2\mu, 4\sigma^2)$ distribution. [1]

Solution 1.25(i) **Probability of at least 3 claims**

We have $X \sim Poi(0.6)$, so using the Poisson tables on page 175 we get:

$$P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 1 - 0.97688 = 0.0231 \quad [2]$$

Alternatively, from first principles:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) \\ &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - e^{-0.6} - 0.6e^{-0.6} - 0.18e^{-0.6} \\ &= 1 - 0.5488 - 0.3293 - 0.0988 = 0.0231 \end{aligned}$$

(ii) **Probability of at least 3 claims given at least 1 claim**

We require:

$$P(X \geq 3 | X \geq 1) = \frac{P(X \geq 3 \text{ and } X \geq 1)}{P(X \geq 1)} = \frac{P(X \geq 3)}{P(X \geq 1)}$$

$$\text{where } P(X \geq 1) = 1 - P(X = 0) = 1 - 0.54881 = 0.45119. \quad [1]$$

$$\Rightarrow P(X \geq 3 | X \geq 1) = \frac{0.0231}{0.45119} = 0.0512 \quad [1]$$

(iii) **Probability portfolio has at least 3 claims**

Recall the fact that the sum of two independent Poisson variables is a Poisson variable, ie if $X_1 \sim Poi(\lambda_1)$ and $X_2 \sim Poi(\lambda_2)$ then $X_1 + X_2 \sim Poi(\lambda_1 + \lambda_2)$. So here the number of claims on 5 policies is distributed as $Poi(3)$. [1]

Using the Poisson tables on page 176 gives:

$$P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 1 - 0.42319 = 0.577 \quad [1]$$

Alternatively, from first principles:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - e^{-3} - 3e^{-3} - 4.5e^{-3} = 1 - 0.0498 - 0.1494 - 0.2240 = 0.577 \end{aligned}$$

(iv) **Probability of waiting at least 4 months**

If we assume that claims are equally likely to occur at any time during the year, then we are dealing with a Poisson process. [1]

The waiting time between events occurring in a $Poi(0.6)$ process has an $Exp(0.6)$ distribution. We must ensure that we are working in years, so we need to convert 4 months into $\frac{1}{3}$ of a year. We require:

$$P(X > \frac{1}{3}) = \int_{\frac{1}{3}}^{\infty} 0.6e^{-0.6x} dx = \left[-e^{-0.6x} \right]_{\frac{1}{3}}^{\infty} = e^{-0.2} = 0.819 \quad [2]$$

Solution 1.26

To simulate a random variable we require the distribution function, $F(x)$:

$$F(x) = P(X \leq x) = \int_0^x 50(5+t)^{-3} dt = \left[-25(5+t)^{-2} \right]_0^x = 1 - \frac{25}{(5+x)^2} \quad [1]$$

We can now use the inverse transform method:

$$u = 1 - \frac{25}{(5+x)^2} \Rightarrow x = \sqrt{\frac{25}{1-u}} - 5 \quad [1]$$

Note that we only use the positive root as we are told that $x > 0$.

Substituting in our values of u , we obtain:

$$x_1 = x = \sqrt{\frac{25}{1-0.863}} - 5 = 8.51 \quad [1]$$

$$x_2 = x = \sqrt{\frac{25}{1-0.447}} - 5 = 1.72 \quad [1]$$

Solution 1.27

Expanding the expression for the PGF gives:

$$G_X(t) = 0.4 + 0.4t + 0.1t^2 + 0.1t^3$$

(i) ***Expectation***

$$G'_X(t) = 0.4 + 0.2t + 0.3t^2$$

$$\Rightarrow E(X) = G'_X(1) = 0.4 + 0.2 + 0.3 = 0.9 \quad [1]$$

(ii) ***Standard deviation***

Using the formula for the variance derived in Solution 1.15.

$$G''_X(t) = 0.2 + 0.6t \Rightarrow G''_X(1) = 0.8$$

$$\Rightarrow \text{var}(X) = 0.8 + 0.9 - 0.9^2 = 0.89 \Rightarrow \text{sd}(X) = \sqrt{0.89} = 0.943 \quad [2]$$

(iii) ***Probability***

Recalling that $P(X = k)$ is the coefficient of t^k in the expansion of $G_X(t)$.

$$P(X \geq 2) = 0.1 + 0.1 = 0.2 \quad [2]$$

Solution 1.28(i) **Expectation**

$$M'_Y(t) = 8(1-4t)^{-3} \Rightarrow E(Y) = M'_X(0) = 8 \quad [1]$$

(ii) **Standard deviation**

$$M''_Y(t) = 96(1-4t)^{-4} \Rightarrow E(Y^2) = 96$$

$$\Rightarrow \text{var}(Y) = 96 - 8^2 = 32 \Rightarrow \text{standard deviation} = \sqrt{32} = 5.6569 \quad [2]$$

(iii) **Sixth moment**

Recalling that $E(Y^6)$ is the coefficient of $\frac{t^k}{k!}$ in the expansion of $M_Y(t)$. From the binomial expansion of the MGF, $(1-4t)^{-2}$, (using the formula given on page 1 of the *Tables*) the term is:

$$\frac{-2 \times -3 \times -4 \times -5 \times -6 \times -7}{6!} (-4t)^6 \quad [1]$$

$$\text{Hence, } E(Y^6) = -2 \times -3 \times -4 \times -5 \times -6 \times -7 \times (-4)^6 = 20,643,840. \quad [1]$$

Alternatively, we can use $E(Y^6) = M_Y^{(6)}(0)$ but this requires us to differentiate the MGF six times.

Solution 1.29(i) **PGF**The PGF of U is:

$$\begin{aligned} G_U(t) &= E[t^U] = \sum_{u=1}^{\infty} t^u P(U=u) = \sum_{u=1}^{\infty} t^u pq^{u-1} \\ &= pt + pqt^2 + pq^2t^3 + \dots \end{aligned} \quad [1]$$

This is an infinite geometric series with $a = pt$ and $r = qt$ so using the formula

$$S_{\infty} = \frac{a}{1-r}$$
 gives:

$$G_U(t) = \frac{pt}{1-qt} \quad [1]$$

(ii) **MGF**The MGF of U is:

$$\begin{aligned} M_U(t) &= E[e^{tU}] = \sum_{u=1}^{\infty} e^{tu} P(U=u) = \sum_{u=1}^{\infty} e^{tu} pq^{u-1} \\ &= pe^t + pqe^{2t} + pq^2e^{3t} + \dots \end{aligned} \quad [1]$$

This is an infinite geometric series with $a = pe^t$ and $r = qe^t$ so using the formula

$$S_{\infty} = \frac{a}{1-r}$$
 gives:

$$M_U(t) = \frac{pe^t}{1-qe^t} \quad [1]$$

(iii) **Connection between PGF and MGF**The MGF is just the PGF with the t 's replaced by e^t 's. This is because $G_X(t) = E(t^X)$, so:

$$G_X(e^t) = E[(e^t)^X] = E[e^{tX}] = M_X(t) \quad [1]$$

Solution 1.30(i)(a) **MGF**The MGF of X is:

$$M_X(t) = E[e^{tX}] = \int_R^\infty e^{tx} k e^{-2x} dx \quad [1]$$

$$= k \int_R^\infty e^{-(2-t)x} dx = k \left[\frac{e^{-(2-t)x}}{-(2-t)} \right]_R^\infty \quad [1]$$

$$= \frac{k e^{-(2-t)R}}{(2-t)} \quad [1]$$

(i)(b) **Values of t for which valid**The integral converges as $x \rightarrow \infty$ only if $2 - t$ is positive. So 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]

Solution 1.31(i) **MGF**

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x} dx \end{aligned} \quad [1]$$

The integral looks like the PDF of a $\text{Gamma}(\alpha, \lambda - t)$, so putting in the appropriate constants:

$$\begin{aligned} M_X(t) &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \int_0^\infty \frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} dx \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \quad \text{provided } t < \lambda \\ &= \left(\frac{\lambda}{\lambda-t} \right)^\alpha = \left(\frac{\lambda-t}{\lambda} \right)^{-\alpha} = \left(1 - \frac{t}{\lambda} \right)^{-\alpha} \end{aligned} \quad [2]$$

Since the integral of a Gamma PDF over the whole range is 1.

Alternatively, we can use the substitution method given in Chapter 5.

(ii) **Mean and variance**

Using the results $E(X) = M'_X(0)$ and $E(X^2) = M''_X(0)$:

$$M'_X(t) = \frac{\alpha}{\lambda} \left(1 - \frac{t}{\lambda} \right)^{-\alpha-1} \Rightarrow E(X) = M'_X(0) = \frac{\alpha}{\lambda} \quad [1]$$

$$M''_X(t) = \frac{\alpha(\alpha+1)}{\lambda^2} \left(1 - \frac{t}{\lambda} \right)^{-\alpha-2} \Rightarrow E(X^2) = M''_X(0) = \frac{\alpha(\alpha+1)}{\lambda^2}$$

$$\text{var}(X) = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2} \quad [1]$$

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.

Part 2 – Questions

Note that the split between development questions and exam-style questions is somewhat subjective. For example, there have been past CT3 exam questions that test knowledge of the Core Reading, and so are similar to what we've included here as development questions. The exam-style questions involve more application and a wider range of ideas and are typically the more challenging questions in the exam.

1 *Development questions*

Question 2.1

If X and Y denote any random variables, f denotes any function and c denotes any constant, state whether the following are true:

- A always
- B if X and Y are independent
- C never
- D none of the above.

- (i) $E[X + Y] = E[X] + E[Y]$ [1]
 - (ii) $E[XY] = E[X] E[Y]$ [1]
 - (iii) $E[cX] = cE[X]$ [1]
 - (iv) $E[f(X)] = f(E[X])$ [1]
 - (v) $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$ [1]
 - (vi) $\text{var}(XY) = \text{var}(X) \text{var}(Y)$ [1]
 - (vii) $\text{var}(cX) = c \text{var}(X)$ [1]
 - (viii) $\text{var}(-X) = -\text{var}(X)$ [1]
- [Total 8]

Question 2.2

Two discrete random variables, X and Y , have the following joint probability function:

		X		
		1	2	3
		1	0.2	0.2
Y	2	0	0.2	0
	3	0.2	0	0.2

Determine:

- (i) $E(X)$ [1]
 - (ii) the probability distribution of $Y | X = 1$ [1]
 - (iii) if X and Y are correlated or not [2]
 - (iv) if X and Y are independent or not. [1]
- [Total 5]

Question 2.3

Let X and Y have joint density function:

$$f_{X,Y}(x,y) = \frac{4}{5}(3x^2 + xy) \quad 0 < x < 1, 0 < y < 1$$

Determine:

- (i) the marginal density function X [2]
 - (ii) the conditional density function of Y given $X = x$ [1]
 - (iii) the covariance of X and Y . [5]
- [Total 8]

Question 2.4

The random variables X and Y have joint density function given by:

$$kx^{-\alpha}e^{-y/\beta} \quad 1 < x < \infty, 1 < y < \infty$$

where $\alpha > 1$, $\beta > 0$, and k is a constant.

Calculate the value of k .

[3]

Question 2.5

Two actuaries are interested in combining Poisson distributions.

John states that the difference between two independent Poisson distributions has a Poisson distribution, ie if $X_1 \sim Poi(\lambda)$ and $X_2 \sim Poi(\mu)$ then $X_1 - X_2 \sim Poi(\lambda - \mu)$.

Jennie states that the sum of two independent Poisson distributions has a Poisson distribution, ie if $X_1 \sim Poi(\lambda)$ and $X_2 \sim Poi(\mu)$ then $X_1 + X_2 \sim Poi(\lambda + \mu)$.

Use moment generating functions to examine John and Jennie's claims.

[5]

Question 2.6

The random variable U has a geometric distribution with probability function:

$$P(U = u) = pq^{u-1} \quad u = 1, 2, 3, \dots \quad \text{where } p + q = 1$$

(i) Derive the moment generating function of U .

[2]

(ii) Hence, derive the moment generating function of the random variable V with probability function:

$$P(V = v) = \binom{v-1}{k-1} p^k q^{v-k} \quad v = k, k+1, k+2, \dots \quad \text{where } p + q = 1$$

[2]

[Total 4]

Question 2.7

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. [7]

Question 2.8

The random variable V has a Poisson distribution with mean 5. For a given value of V , the 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]

Question 2.9

- (i) State the Central Limit Theorem. [2]
 - (ii) What is the approximate distribution of the arithmetic mean of a set of 100 independent observations from a $Gamma(2.5, 0.2)$ distribution? [3]
- [Total 5]

Question 2.10

Consider the following assertions relating to the random variable Y , which is the arithmetic mean of a fixed number n of identically distributed random variables X_1, X_2, \dots, X_n :

- I $E[Y] = E[X_1]$
- II $\text{var}(Y) = \frac{1}{n} \text{var}(X_1)$
- III Y has an approximate normal distribution.

State whether each of these assertions is necessarily true. [3]

Question 2.11

A fair coin is tossed repeatedly until 20 heads have been obtained. Calculate approximately the probability that this will require more than 50 tosses. [4]

Question 2.12

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 the:

- (i) sample mean of 5 claims is less than 15 mins [2]
- (ii) sample mean of 50 claims is greater than 22 mins [2]
- (iii) sample variance of 5 claims is greater than 6.65 mins [2]
- (iv) sample standard deviation of 30 claims is less than 7 mins [2]
- (v) both (i) and (iii) hold. [1]

[Total 9]

Question 2.13

The length of tarmac laid in a day by a motorway maintenance team is thought to be normally distributed with mean 400m and standard deviation 50m. Calculate the probability that the total length of tarmac laid in 5 days is between 1.9km and 2.4km.

[3]

Question 2.14

(i) Find:

(a) $P(F_{6,8} > 6.371)$

(b) $P(F_{7,12} > 0.3748)$. [2]

(ii) Find the value of c such that:

(a) $P(F_{2,15} < c) = 97.5\%$

(b) $P(F_{8,5} < c) = 5\%$. [2]

[Total 4]

Question 2.15(i) (a) State the definition of a t_k distribution.(b) Hence, using $\bar{X} \sim N(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, show that:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad [3]$$

(ii) (a) State the definition of an $F_{m,n}$ distribution.(b) Hence, using $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, show that for suitably defined samples:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{m-1, n-1} \quad [3]$$

(iii) Show that:

$$P(F_{m,n} > a) = b \Leftrightarrow P\left(F_{n,m} < \frac{1}{a}\right) = b \quad [2]$$

[Total 8]

Question 2.16

A random sample x_1, \dots, x_{10} is drawn from a $N(5, 2^2)$ distribution. Evaluate:

(i) $P\left[\sum X > 60\right]$ [3]

(ii) $P\left[\sum(X - \bar{X})^2 > 34\right]$ [2]

(iii) $P\left[\bar{X} > 4 \text{ and } \sqrt{\frac{1}{9}\sum(X - \bar{X})^2} < 2.6\right].$ [4]

[Total 9]

Question 2.17

The random variables X_1, \dots, X_n represent independent observations from a population with mean μ and variance σ^2 .

(i) State the mean and variance of the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$ [2]

(ii) Show that $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2.$ [2]

(iii) Determine the expected value of $\sum_{i=1}^n (X_i - \bar{X})^2.$ [3]

(iv) Hence show that $E(S^2) = \sigma^2$, where $S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right).$ [1]

(v) For the special case where the parent population is normally distributed and the values of the population parameters μ and σ^2 are known, state two results that specify the precise distributions of \bar{X} and $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2.$ [2]

[Total 10]

2 Exam-style questions

Question 2.18

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 $Z = X + Y$.

Calculate:

- (i) $\text{cov}(X, Z)$ [2]
 - (ii) $\text{var}(Z)$. [2]
- [Total 4]

Question 2.19

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]

Question 2.20

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]

Question 2.21

- (i) X_1, \dots, X_n are independent and identical $\text{Gamma}(\alpha, \lambda)$ distribution. Show, using moment generating functions, that \bar{X} has a $\text{Gamma}(n\alpha, n\lambda)$ distribution. [3]
 - (ii) If the random variable T , representing the total lifetime of an individual light bulb, has an $\text{Exp}(\lambda)$ distribution, where $1/\lambda = 2,000$ hours, calculate the probability that the average lifetime of 10 bulbs will exceed 4,000 hours. [3]
- [Total 6]

Question 2.22

The random variables X_1, \dots, X_n are independent and identically distributed with common moment generating function $M_X(t)$.

- (i) Obtain the moment generating function of Y , where $Y = \frac{1}{n} \sum_{i=1}^n X_i$. [3]
 - (ii) Use your result from part (i) to obtain the moment generating function of \bar{X} when $X \sim N(\mu, \sigma^2)$ and state its distribution. [2]
 - (iii) Comment briefly on your result to part (ii). [1]
- [Total 6]

Question 2.23

- (i) Given that X and Y are continuous random variables, prove from first principles that:

$$E(Y) = E[E(Y | X)] \quad [3]$$

- (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:

$$E(Y | X = x) = 3x + 1 \quad \text{var}(Y | X = x) = 2x^2 + 5$$

Calculate the unconditional mean and standard deviation of Y .

[5]

[Total 8]

Question 2.24

- (i) Two discrete random variables, X and Y , have the following joint probability function:

		X			
		1	2	3	4
Y	1	0.2	0	0.05	0.15
	2	0	0.3	0.1	0.2

Determine $\text{var}(X | Y = 2)$. [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 $E(U | V = v)$. [3]

[Total 6]

Question 2.25

The random variable S has a compound distribution so that $S = X_1 + \dots + X_N$ (and $S = 0$ if $N = 0$) where the X_i 's are independent, identically distributed (as a variable X) and are also independent of N .

- (i) Use the result $\text{var}[Y] = E[\text{var}(Y | X)] + \text{var}[E(Y | X)]$ to obtain an expression for the variance of S in terms of the means and variances of the random variables X and N . [4]
- (ii) N has a binomial distribution with parameters $n = 100$ and $p = 0.05$ and X has an exponential distribution with mean 50. Calculate the mean and standard deviation of S . [2]
- (iii) By using a suitable approximation, calculate the probability that S exceeds 500. [3]

[Total 9]

Question 2.26

The random variable Y is defined by:

$$Y = \sum_{i=1}^N X_i$$

where the X_i 's are independent $\text{Bin}(n, p)$ random variables and N is a Poisson random variable, independent of the X_i 's, with mean λ .

- (i) Prove from first principles that the moment generating function of Y is given by:

$$M_Y(t) = \exp\left\{\lambda\left[(q + pe^t)^n - 1\right]\right\} \quad [4]$$

- (ii) Hence derive expressions for the mean and variance of Y . [4]
[Total 8]

Question 2.27

A company issues questionnaires to clients to obtain feedback on the clarity of their brochure. It is thought that 5% of clients don't find the brochure helpful.

Calculate the approximate probability that in a sample of 1,000 responses, the number, N , of clients who don't find the brochure helpful satisfies $40 < N < 70$. [5]

Question 2.28

The random variable Y has a gamma distribution with parameters $\alpha (> 1)$ 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 $\alpha \rightarrow \infty$, but where λ is varied so that the mean $\frac{\alpha}{\lambda}$ has a constant value μ . [5]

- (ii) Given that $\alpha = 50$ and $\lambda = 0.2$, calculate the value of $P(Y > 350)$ using:

- (a) the gamma-chi square relationship

- (b) the Central Limit Theorem.

[5]

- (iii) Explain the reason for the difference between the answers obtained in part (ii) [1]
 [Total 11]

Question 2.29

A random sample of n observations is taken from a normal distribution with variance σ^2 . The sample variance is an observation of a random variable S^2 . Show that:

(i) $E(S^2) = \sigma^2$ [2]

(ii) $\text{var}(S^2) = \frac{2\sigma^4}{n-1}$. [2]

[Total 4]

Question 2.30

House prices in region X are normally distributed about a mean of £100,000 with a standard deviation of £10,000. House prices in region Y are normally distributed about a mean of £90,000 with 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 . Find the probability that:

- (i) (a) the region X sample mean is greater than the region Y sample mean
 (b) the difference between the sample means is less than £5,000 [6]
- (ii) (a) the region X sample variance is less than the region Y sample variance
 (b) the region X sample standard deviation is more than four times greater
 than the region Y sample standard deviation. [5]

[Total 11]

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.

Part 2 – Solutions

Solution 2.1

(i) $E(X + Y) = E(X) + E(Y)$

- A This is always true. [1]

(ii) $E(XY) = E(X)E(Y)$

- B This works if X and Y are independent. In fact X and Y only need to be uncorrelated, which is a slightly weaker condition. [1]

(iii) $E(cX) = cE(X)$

- A This is always true. [1]

(iv) $E[f(X)] = f(E[X])$

- D This is not true in general, but it may be true in particular cases eg if $f(X) = cX$ where c is a constant, then the equation reduces to identity (iii) above. [1]

(v) $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$

- B This is true if X and Y are independent, but again they need only be uncorrelated. [1]

(vi) $\text{var}(XY) = \text{var}(X)\text{var}(Y)$

- D This is only true in exceptional circumstances eg if X and Y are constants. [1]

(vii) $\text{var}(cX) = c\text{var}(X)$

- D This is only true if $c = 0$ or 1. It should be c^2 on the right hand side. [1]

(viii) $\text{var}(-X) = -\text{var}(X)$

- C Since $\text{var}(-X) = \text{var}(X)$, this can only be true if $\text{var}(X) = 0$ ie if X is not a random variable at all. [1]

Solution 2.2(i) **Mean**

$$E(X) = 1 \times 0.4 + 2 \times 0.2 + 3 \times 0.4 = 2$$

[1]

Or you could use the fact that the distribution of X is symmetrical about 2.

(ii) **Probability distribution of $Y | X=1$**

Using $P(Y = y | X = 1) = \frac{P(X = 1, Y = y)}{P(X = 1)}$ and $P(X = 1) = 0.4$ gives:

$Y = 1 X = 1$	$Y = 2 X = 1$	$Y = 3 X = 1$
0.5	0	0.5

[1]

(iii) **Correlated?**

To calculate the correlation coefficient, we first require the covariance.

$$E(X) = 2 \quad \text{from part (i)}$$

$$E(Y) = 1 \times 0.4 + 2 \times 0.2 + 3 \times 0.4 = 2$$

$$E(XY) = 1 \times 0.2 + 2 \times 0 + 3 \times 0.2 + \dots + 3 \times 0.2 + 6 \times 0 + 9 \times 0.2 = 4$$

$$\text{So } \text{cov}(X, Y) = E(XY) - E(X)E(Y) = 4 - 2 \times 2 = 0.$$

[1]

$$\text{Hence } \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}} = 0.$$

Therefore X and Y are uncorrelated.

[1]

(iv) **Independent?**

If X and Y are independent then $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x and y .

$$\text{Consider } P(X = 1, Y = 1) = 0.2 \neq 0.4 \times 0.4 = P(X = 1)P(Y = 1).$$

So X and Y are not independent.

[1]

Solution 2.3(i) **Marginal density**

$$f_X(x) = \int_{y=0}^1 \frac{4}{5} (3x^2 + xy) dy = \left[\frac{4}{5} \left(3x^2 y + \frac{1}{2} x y^2 \right) \right]_{y=0}^1 = \frac{4}{5} \left(3x^2 + \frac{1}{2} x \right) \quad [2]$$

(ii) **Conditional density**

$$f_{Y|X=x}(x, y) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{\frac{4}{5} (3x^2 + xy)}{\frac{4}{5} (3x^2 + \frac{1}{2} x)} = \frac{3x^2 + xy}{3x^2 + \frac{1}{2} x} = \frac{3x + y}{3x + \frac{1}{2}} \quad [1]$$

(iii) **Covariance**Using the marginal density function of X :

$$E(X) = \int_{x=0}^1 \frac{4}{5} \left(3x^3 + \frac{1}{2} x^2 \right) dx = \frac{4}{5} \left[\frac{3}{4} x^4 + \frac{1}{6} x^3 \right]_{x=0}^1 = \frac{11}{15} \quad [1]$$

Obtaining the marginal density function of Y :

$$\begin{aligned} f_Y(y) &= \int_{x=0}^1 \frac{4}{5} (3x^2 + xy) dx = \frac{4}{5} \left[x^3 + \frac{1}{2} x^2 y \right]_{x=0}^1 = \frac{4}{5} \left(1 + \frac{1}{2} y \right) \\ \Rightarrow E(Y) &= \int_{y=0}^1 \frac{4}{5} \left(y + \frac{1}{2} y^2 \right) dy = \frac{4}{5} \left[\frac{1}{2} y^2 + \frac{1}{6} y^3 \right]_{y=0}^1 = \frac{8}{15} \end{aligned} \quad [1]$$

Finally:

$$\begin{aligned}
 E(XY) &= \int_{x=0}^1 \int_{y=0}^1 \frac{4}{5} (3x^3y + x^2y^2) dx dy \\
 &= \int_{x=0}^1 \frac{4}{5} \left[\frac{3}{2}x^3y^2 + \frac{1}{3}x^2y^3 \right]_{y=0}^1 dx \\
 &= \int_{x=0}^1 \frac{4}{5} \left(\frac{3}{2}x^3 + \frac{1}{3}x^2 \right) dx \\
 &= \frac{4}{5} \left[\frac{3}{8}x^4 + \frac{1}{9}x^3 \right]_{x=0}^1 \\
 &= \frac{7}{18}
 \end{aligned} \tag{2}$$

Hence:

$$\text{cov}(X, Y) = \frac{7}{18} - \frac{11}{15} \times \frac{8}{15} = -\frac{1}{450} \tag{1}$$

Solution 2.4

Since the pdf must integrate to 1:

$$\int_{x=1}^{\infty} \int_{y=1}^{\infty} kx^{-\alpha} e^{-y/\beta} dx dy = 1$$

Integrating over the x values gives:

$$\int_{x=1}^{\infty} kx^{-\alpha} e^{-y/\beta} dx = ke^{-y/\beta} \left[\frac{x^{-\alpha+1}}{-\alpha+1} \right]_1^{\infty} = \frac{ke^{-y/\beta}}{\alpha-1} \tag{1}$$

Integrating this over the y values gives:

$$\int_{y=1}^{\infty} \frac{ke^{-y/\beta}}{\alpha-1} dy = \frac{k}{\alpha-1} \left[-\beta e^{-y/\beta} \right]_1^{\infty} = \frac{k\beta e^{-1/\beta}}{\alpha-1} \tag{1}$$

Equating this to 1:

$$\frac{k\beta e^{-1/\beta}}{\alpha - 1} = 1 \Rightarrow k = \frac{(\alpha - 1)e^{1/\beta}}{\beta} \quad [1]$$

Solution 2.5

Examining John's claim. Let $X_1 \sim Poi(\lambda)$, $X_2 \sim Poi(\mu)$ and $Y = X_1 - X_2$ then:

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E(e^{tX_1 - tX_2}) = E(e^{tX_1}e^{-tX_2}) \\ &= E(e^{tX_1})E(e^{-tX_2}) \quad \text{by independence} \\ &= M_{X_1}(t)M_{X_2}(-t) \\ &= e^{\lambda(e^t - 1)}e^{\mu(e^{-t} - 1)} \\ &= e^{\lambda e^t + \mu e^{-t} - (\lambda + \mu)} \end{aligned} \quad [1]$$

Comparing this with the MGF of a $Poi(\lambda - \mu)$ distribution, $e^{(\lambda - \mu)(e^t - 1)}$, we can see that John's claim is not true. [1]

Examining Jennie's claim. Let $Z = X_1 + X_2$, then:

$$\begin{aligned} M_Z(t) &= E(e^{tZ}) = E(e^{tX_1 + tX_2}) = E(e^{tX_1}e^{tX_2}) \\ &= E(e^{tX_1})E(e^{tX_2}) \quad \text{by independence} \\ &= M_{X_1}(t)M_{X_2}(t) \\ &= e^{\lambda(e^t - 1)}e^{\mu(e^t - 1)} \\ &= e^{(\lambda + \mu)(e^t - 1)} \end{aligned} \quad [1]$$

Comparing this with the MGF of a $Poi(\lambda + \mu)$ distribution, $e^{(\lambda + \mu)(e^t - 1)}$, we can see by the uniqueness property of MGFs that Jennie's claim is correct. [1]

Solution 2.6(i) **MGF of geometric**The MGF of U is:

$$\begin{aligned} M_U(t) &= E[e^{tU}] = \sum_{u=1}^{\infty} e^{tu} P(U=u) = \sum_{u=1}^{\infty} e^{tu} pq^{u-1} \\ &= pe^t + pqe^{2t} + pq^2e^{3t} + \dots \end{aligned} \quad [1]$$

This is an infinite geometric series with $a = pe^t$ and $r = qe^t$ so using the formula $S_{\infty} = \frac{a}{1-r}$ gives:

$$M_U(t) = \frac{pe^t}{1-qe^t} \quad [1]$$

(ii) **MGF of negative binomial** V has a negative binomial distribution with parameters k and p . Using the fact that a negative binomial with parameters (k, p) is the sum of k independent geometric random variables with parameter p and the formula:

$$M_{X_1+\dots+X_n}(t) = M_{X_1}(t) \cdots M_{X_n}(t) \quad [1]$$

We have:

$$M_V(t) = \left(\frac{pe^t}{1-qe^t} \right)^k \quad [1]$$

Solution 2.7

The chi square distribution is a continuous distribution that can take any positive value. So the PDF of the sum $Z = X + Y$ is given by the convolution formula:

$$\begin{aligned} f_Z(z) &= \int f_X(x)f_Y(z-x)dx \\ &= \int_0^z \frac{(\frac{1}{2})^{\frac{1}{2}m}}{\Gamma(\frac{1}{2}m)} x^{\frac{1}{2}m-1} e^{-\frac{1}{2}x} \frac{(\frac{1}{2})^{\frac{1}{2}n}}{\Gamma(\frac{1}{2}n)} (z-x)^{\frac{1}{2}n-1} e^{-\frac{1}{2}(z-x)} dx \\ &= \left(\frac{1}{2}\right)^{\frac{1}{2}(m+n)} e^{-\frac{1}{2}z} \int_0^z \frac{1}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} x^{\frac{1}{2}m-1} (z-x)^{\frac{1}{2}n-1} dx \end{aligned} \quad [2]$$

Using the substitution $t = x/z$ gives:

$$\begin{aligned} f_Z(z) &= (\frac{1}{2})^{\frac{1}{2}(m+n)} e^{-\frac{1}{2}z} \int_0^1 \frac{1}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} (zt)^{\frac{1}{2}m-1} (z-zt)^{\frac{1}{2}n-1} z dt \\ &= \frac{(\frac{1}{2})^{\frac{1}{2}(m+n)}}{\Gamma(\frac{1}{2}m + \frac{1}{2}n)} z^{\frac{1}{2}(m+n)-1} e^{-\frac{1}{2}z} \int_0^1 \frac{\Gamma(\frac{1}{2}m + \frac{1}{2}n)}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} t^{\frac{1}{2}m-1} (1-t)^{\frac{1}{2}n-1} dt \end{aligned} \quad [2]$$

Since the last integral represents the total probability for a $Beta(\frac{1}{2}m, \frac{1}{2}n)$ distribution, we get:

$$\begin{aligned} f_Z(z) &= \frac{(\frac{1}{2})^{\frac{1}{2}(m+n)}}{\Gamma(\frac{1}{2}m + \frac{1}{2}n)} z^{\frac{1}{2}(m+n)-1} e^{-\frac{1}{2}z} \times P[0 < Beta(\frac{1}{2}m, \frac{1}{2}n) < 1] \\ &= \frac{(\frac{1}{2})^{\frac{1}{2}(m+n)}}{\Gamma(\frac{1}{2}m + \frac{1}{2}n)} z^{\frac{1}{2}(m+n)-1} e^{-\frac{1}{2}z} \end{aligned} \quad [2]$$

Since this matches the PDF of a χ_{m+n}^2 distribution (and Z can take any positive value), Z has a χ_{m+n}^2 distribution. [1]

It is much easier to prove this result using MGFs.

Solution 2.8

We are given in the question that:

$$U | V = v \sim U(0, v) \quad V \sim Poi(5)$$

So:

$$E(V) = 5 \quad \text{var}(V) = 5 \quad [1/2]$$

and:

$$E(U | V) = \frac{1}{2}V \quad \text{var}(U | V) = \frac{1}{12}V^2 \quad [1/2]$$

Using the formulae on page 16 of the *Tables*, we have:

$$E[U] = E[E[U | V]] \quad \text{var}[U] = \text{var}[E[U | V]] + E[\text{var}[U | V]]$$

Therefore:

$$E[U] = E[E(U | V)] = E\left[\frac{1}{2}V\right] = \frac{1}{2}E[V] = 2\frac{1}{2} \quad [1]$$

$$\begin{aligned} \text{var}[U] &= \text{var}[E(U | V)] + E[\text{var}(U | V)] \\ &= \text{var}\left[\frac{1}{2}V\right] + E\left[\frac{1}{12}V^2\right] \\ &= \frac{1}{4}\text{var}[V] + \frac{1}{12}E[V^2] \end{aligned} \quad [1]$$

Since $E[V^2] = \text{var}[V] + E^2[V]$, we have:

$$\text{var}[U] = \frac{1}{4} \times 5 + \frac{1}{12}(5 + 5^2) = 3\frac{3}{4} \quad [1]$$

Solution 2.9(i) ***State the Central Limit Theorem***

If X_1, X_2, \dots, X_n is a sequence of IID random variables with finite mean μ and finite (nonzero) variance σ^2 , then the distribution of $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ approaches a standard normal distribution as n tends to ∞ . [2]

(ii) ***Approximating a gamma distribution***

The mean and variance of the $\text{Gamma}(2.5, 0.2)$ distribution are $\frac{2.5}{0.2} = 12.5$ and $\frac{2.5}{0.2^2} = 62.5$. [1]

So, by the Central Limit Theorem, the arithmetic mean of 100 independent observations has an approximate normal distribution with mean 12.5 and variance $\frac{62.5}{100} = 0.625$. [2]

Solution 2.10

I is true. [1]

II requires the X 's to be uncorrelated. [1]

III requires the X 's to be independent and n to be sufficiently large. [1]

Solution 2.11

We are counting the number of trials until the 20th success. So we have a Type I negative binomial with parameters $k = 20$ and $p = \frac{1}{2}$. [1]

The mean and variance of this distribution are:

$$\begin{aligned} E[X] &= \frac{k}{p} = \frac{20}{\frac{1}{2}} = 40 \\ \text{var}(X) &= \frac{kq}{p^2} = \frac{20 \times \frac{1}{2}}{\left(\frac{1}{2}\right)^2} = 40 \end{aligned} \quad [1]$$

Using a normal approximation with a continuity correction:

$$\begin{aligned} P(X > 50) &\approx P[N(40, 40) > 50\frac{1}{2}] \\ &= P[Z > (50\frac{1}{2} - 40) / \sqrt{40}] \\ &= 1 - \Phi(1.660) = 1 - 0.95154 = 0.04846 \end{aligned} \quad [2]$$

So the probability that more than 50 tosses will be required is approximately 4.8%.

Solution 2.12(i) **Probability of sample mean**

$\bar{X} \sim N(\mu, \sigma^2/n)$ 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.

$$\begin{aligned}
 P(\bar{X} < 15) &= P\left(Z < \frac{15 - 20}{\sqrt{5}}\right) \quad \text{since } \bar{X} \sim N(20, 5) \\
 &= P(Z < -2.236) \\
 &= 1 - \Phi(2.236) \\
 &= 0.0127
 \end{aligned} \tag{2}$$

(ii) **Probability of sample mean**

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.

$$\begin{aligned}
 P(\bar{X} > 22) &= P\left(Z > \frac{22 - 20}{\sqrt{0.5}}\right) \quad \text{since } \bar{X} \sim N(20, 0.5) \\
 &= P(Z > 2.828) \\
 &= 1 - \Phi(2.828) \\
 &= 0.00234
 \end{aligned} \tag{2}$$

(iii) **Probability of sample variance**

$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$ only holds for samples from a normal distribution. Therefore we require that we are sampling from a normal distribution.

$$\begin{aligned}
 P(S^2 > 6.65) &= P\left(\frac{4S^2}{\sigma^2} > \frac{4 \times 6.65}{5^2}\right) \\
 &= P(\chi_4^2 > 1.064) \\
 &= 0.9 \quad (\text{from page 168 of the Tables})
 \end{aligned} \tag{2}$$

(iv) ***Probability of sample standard deviation***

Again we require that we are sampling from a normal distribution.

$$\begin{aligned}
 P(S < 7) &= P\left(\frac{29S^2}{\sigma^2} < \frac{29 \times 7^2}{5^2}\right) \\
 &= P(\chi_{29}^2 < 56.84) \\
 &\simeq 0.998 \quad \text{by intepolation} \quad [2]
 \end{aligned}$$

(v) ***Probability of (i) and (iii) holding***

\bar{X} and S^2 are independent if we are sampling from a normal distribution. So assuming this (which is required anyway to calculate (i) and (ii)), we get:

$$\begin{aligned}
 P(\bar{X} < 15 \text{ and } S^2 > 6.65) &= P(\bar{X} < 15) \times P(S^2 > 6.65) \\
 &= 0.0127 \times 0.9 \\
 &= 0.0114 \quad [1]
 \end{aligned}$$

Solution 2.13

Using the result that $\sum X_i \stackrel{\text{d}}{\sim} N(n\mu, n\sigma^2) = N(2000, 12500)$: [1]

$$\begin{aligned}
 P(1,900 < \sum X_i < 2,400) &= P(\sum X_i < 2,400) - P(\sum X_i < 1,900) \\
 &= P\left(Z < \frac{2,400 - 2,000}{\sqrt{12,500}}\right) - P\left(Z < \frac{1,900 - 2,000}{\sqrt{12,500}}\right) \\
 &= \Phi(3.578) - [1 - \Phi(0.894)] \\
 &= 0.99983 - [1 - 0.81434] \\
 &= 0.814 \quad [2]
 \end{aligned}$$

Solution 2.14(i)(a) **Probability**

6.371 is greater than 1 so we simply use the upper critical values given:

$$P(F_{6,8} > 6.371) = 0.01 \quad [1]$$

(i)(b) **Probability**

Since this is a lower critical point we need to use the $\frac{1}{F_{m,n}}$ result:

$$P(F_{7,12} > 0.3748) = P\left(F_{12,7} < \frac{1}{0.3748}\right) = P(F_{12,7} < 2.688) = 0.9 \quad [1]$$

(ii)(a) **Find value**

Since 97.5% is below c it must be on the upper tail. So simply reading off the 2½% tables gives:

$$P(F_{2,15} < c) = 0.975 \Rightarrow P(F_{2,15} > c) = 0.025 \Rightarrow c = 4.765 \quad [1]$$

(ii)(b) **Find value**

Since only 5% is below c it must be on the lower tail. So we need to use the $\frac{1}{F_{m,n}}$ result:

$$P(F_{8,5} < c) = P\left(F_{5,8} > \frac{1}{c}\right) = 0.05 \Rightarrow \frac{1}{c} = 3.688 \Rightarrow c = 0.2711 \quad [1]$$

Solution 2.15(i)(a) ***Definition of t distribution***

$$t_k \equiv \frac{N(0,1)}{\sqrt{\chi_k^2/k}}, \text{ where } N(0,1) \text{ and } \chi_k^2 \text{ are independent.} \quad [1]$$

(i)(b) ***Show result is t distribution***

Standardising, we get:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad [\frac{1}{2}]$$

and also we have:

$$\frac{S^2}{\sigma^2} \sim \frac{\chi_{n-1}^2}{n-1} \quad [\frac{1}{2}]$$

Substituting these into the definition of the *t*-distribution:

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad [1]$$

(ii)(a) ***Definition of F distribution***

$$F_{m,n} \equiv \frac{\chi_m^2/m}{\chi_n^2/n}, \text{ where } \chi_m^2 \text{ and } \chi_n^2 \text{ are independent.} \quad [1]$$

(i)(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.

$$\frac{(m-1)S_1^2}{\sigma_1^2} \sim \chi_{m-1}^2 \Rightarrow \frac{S_1^2}{\sigma_1^2} \sim \frac{\chi_{m-1}^2}{m-1} \quad [\frac{1}{2}]$$

$$\frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi_{n-1}^2 \Rightarrow \frac{S_2^2}{\sigma_2^2} \sim \frac{\chi_{n-1}^2}{n-1} \quad [\frac{1}{2}]$$

Hence by the definition of the F distribution:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{m-1,n-1} \quad [1]$$

(iii) **Prove inverse F result**

By taking reciprocals, we obtain:

$$P(F_{m,n} > a) = b \Rightarrow P\left(\frac{1}{F_{m,n}} < \frac{1}{a}\right) = b$$

Now from the definition of the F distribution:

$$F_{m,n} \equiv \frac{\chi_m^2/m}{\chi_n^2/n} \Rightarrow \frac{1}{F_{m,n}} = \frac{\chi_n^2/n}{\chi_m^2/m} = F_{n,m} \quad [1]$$

Hence:

$$P(F_{m,n} > a) = b \Rightarrow P\left(\frac{1}{F_{m,n}} < \frac{1}{a}\right) = b \Rightarrow P\left(F_{n,m} < \frac{1}{a}\right) = b \quad [1]$$

Solution 2.16(i) **Probability of sum**

Using the result that $\sum X_i \sim N(n\mu, n\sigma^2) = N(50, 40)$ we obtain: [1]

$$P(\sum X_i > 60) = P\left(Z > \frac{60 - 50}{\sqrt{40}}\right) = P(Z > 1.581) = 1 - \Phi(1.581) = 0.0569 \quad [2]$$

(ii) **Probability of central moment**

Since $\sum (X_i - \bar{X})^2 = (n-1)S^2$ and using $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$:

$$\begin{aligned} P\left[\sum (X_i - \bar{X})^2 > 34\right] &= P[9S^2 > 34] \\ &= P\left[\frac{9S^2}{2^2} > \frac{34}{2^2}\right] \\ &= P\left[\chi_9^2 > 8.5\right] \\ &= 1 - 0.5154 = 0.485 \end{aligned} \quad [2]$$

(iii) **Joint probability**

Since $S = \sqrt{\frac{1}{9} \sum (X_i - \bar{X})^2}$ and using the fact that \bar{X} and S^2 are independent if we are sampling from a normal distribution:

$$P[\bar{X} > 4 \text{ and } S < 2.6] = P[\bar{X} > 4]P[S < 2.6] \quad [1]$$

Now $\bar{X} \sim N(5, 0.4)$, so:

$$P[\bar{X} > 4] = P\left(Z > \frac{4 - 5}{\sqrt{0.4}}\right) = P(Z > -1.581) = \Phi(1.581) = 0.9431 \quad [1]$$

Also using $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$:

$$P[S < 2.6] = P\left[\frac{9S^2}{2^2} < \frac{9 \times 2.6^2}{2^2}\right] = P[\chi_9^2 < 15.21] = 0.9145 \quad [1]$$

Hence $P[\bar{X} > 4 \text{ and } S < 2.6] = 0.9431 \times 0.9145 = 0.862$. [1]

Solution 2.17

(i) **State the mean and variance**

The sample mean has mean μ and variance σ^2/n . [2]

(ii) **Show result**

Expanding the square gives:

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum (X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\ &= \sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2 \\ &= \sum X_i^2 - 2\bar{X}(n\bar{X}) + n\bar{X}^2 = \sum X_i^2 - n\bar{X}^2 \end{aligned} \quad [2]$$

(iii) **Obtain expected value**

Now:

$$E[X_i^2] = \text{var}(X_i) + E^2[X_i] = \sigma^2 + \mu^2 \quad [\frac{1}{2}]$$

$$E[\bar{X}^2] = \text{var}(\bar{X}) + E^2[\bar{X}] = \frac{\sigma^2}{n} + \mu^2 \quad [\frac{1}{2}]$$

Hence:

$$\begin{aligned}
 E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] &= E\left[\sum X_i^2 - n\bar{X}^2\right] \\
 &= \sum E[X_i^2] - nE[\bar{X}^2] \\
 &= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\sigma^2
 \end{aligned} \tag{2}$$

(iv) ***Prove expectation of sample mean is population mean***

Using the result from part (iii), the expected value of the sample variance is:

$$E[S^2] = \frac{1}{n-1} E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n-1} \times (n-1)\sigma^2 = \sigma^2 \tag{1}$$

(v) ***State precise distributions***

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ and } \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2. \tag{2}$$

Solution 2.18(i) ***Covariance***

We have

$$\begin{aligned}\text{cov}(X, Z) &= \text{cov}(X, X + Y) \\ &= \text{cov}(X, X) + \text{cov}(X, Y) \\ &= \text{var}(X) + \text{cov}(X, Y)\end{aligned}$$

Using the correlation coefficient between X and Y gives:

$$\begin{aligned}\text{corr}(X, Y) = -0.3 &= \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{ var}(Y)}} = \frac{\text{cov}(X, Y)}{\sqrt{4 \times 1}} \\ \Rightarrow \text{cov}(X, Y) &= -0.6\end{aligned}$$

Hence:

$$\text{cov}(X, Z) = 4 - 0.6 = 3.4 \quad [2]$$

(ii) ***Variance***

Using $\text{var}(Z) = \text{cov}(Z, Z)$:

$$\begin{aligned}\text{var}(Z) &= \text{cov}(X + Y, X + Y) \\ &= \text{cov}(X, X) + 2 \text{cov}(X, Y) + \text{cov}(Y, Y) \\ &= \text{var}(X) + 2 \text{cov}(X, Y) + \text{var}(Y) \\ &= 4 + 2 \times -0.6 + 1 \\ &= 3.8 \quad [2]\end{aligned}$$

Note: $\text{var}(Z) \neq \text{var}(X) + \text{var}(Y)$ as X and Y are not independent.

Solution 2.19

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:

$$\begin{aligned} \text{var}(X - 2Y) &= \text{var}(X) + 4\text{var}(Y) - 2 \times 2\text{cov}(X, Y) & [1] \\ &= 5 + 4 \times 10 - 4 \times (-12) = 93 & [1] \end{aligned}$$

Solution 2.20

Let X be the claim size on car policies, so that $X \sim N(1800, 300^2)$.

Let Y be the claim size on home policies, so that $Y \sim N(1200, 500^2)$.

We want:

$$P(X > 2Y) = P(X - 2Y > 0) \quad [1]$$

So we need the distribution of $X - 2Y$:

$$\begin{aligned} X - 2Y &\sim N(1800 - 2 \times 1200, 300^2 + 4 \times 500^2) \\ &\sim N(-600, 1090000) \quad [2] \end{aligned}$$

Standardising:

$$z = \frac{0 - (-600)}{\sqrt{1,090,000}} = 0.575$$

So:

$$\begin{aligned} P(X - 2Y > 0) &= P(Z > 0.575) = 1 - P(Z < 0.575) \\ &= 1 - 0.71735 = 0.28265 \end{aligned} \quad [1]$$

Solution 2.21

- (i) **Show mean has a gamma distribution**

We have:

$$\begin{aligned} M_{\bar{X}}(t) &= E\left(e^{t\bar{X}}\right) = E\left(e^{\frac{t}{n}(X_1 + \dots + X_n)}\right) = E\left(e^{\frac{t}{n}X_1} \dots e^{\frac{t}{n}X_n}\right) \\ &= M_{X_1}\left(\frac{t}{n}\right) \dots M_{X_n}\left(\frac{t}{n}\right) \quad \text{by independence } [1] \\ &= \left[M_X\left(\frac{t}{n}\right)\right]^n \quad X_i's \text{ identical} \\ &= \left(1 - \frac{t}{n\lambda}\right)^{-n\alpha} \quad [1] \end{aligned}$$

This is the MGF of a $Ga(n\alpha, n\lambda)$ distribution. Hence, by the uniqueness property of MGFs \bar{X} has a $Ga(n\alpha, n\lambda)$ distribution. [1]

- (ii) **Probability that average lifetime of 10 bulbs exceeds 4,000 hours**

The individual lifetimes T have an $Exp(\lambda)$ distribution, which is the same as the $Gamma(1, \lambda)$ distribution. So, using the result from part (i) we have:

$$\bar{T} \sim Gamma(10 \times 1, 10 \times \frac{1}{2,000}) \equiv Gamma(10, 0.005) \quad [1]$$

Using the result from page 12 of the *Tables*, the probability that the average lifetime \bar{T} will exceed 4,000 hours is:

$$P(\bar{T} > 4,000) = P(\chi^2_{20} > 2 \times 0.005 \times 4,000) = P(\chi^2_{20} > 40) \quad [1]$$

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%. [1]

Solution 2.22

- (i) **Obtain MGF of Y**

The MGF of Y is:

$$\begin{aligned}
 M_Y(t) &= E(e^{tY}) = E\left(e^{\frac{t}{n}(X_1 + \dots + X_n)}\right) = E\left(e^{\frac{t}{n}X_1} \cdots e^{\frac{t}{n}X_n}\right) \\
 &= E\left(e^{\frac{t}{n}X_1}\right) \cdots E\left(e^{\frac{t}{n}X_n}\right) \quad \text{by independence} \quad [1] \\
 &= M_{X_1}\left(\frac{t}{n}\right) \cdots M_{X_n}\left(\frac{t}{n}\right) \quad [1] \\
 &= \left[M_X\left(\frac{t}{n}\right)\right]^n \quad [1]
 \end{aligned}$$

- (ii) **Obtain MGF of mean**

The MGF of a $N(\mu, \sigma^2)$ distribution is $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. Therefore since $Y = \bar{X}$, we can use the result from part (i):

$$M_{\bar{X}}(t) = \left[e^{\mu \frac{t}{n} + \frac{1}{2}\sigma^2 \left(\frac{t}{n}\right)^2}\right]^n = e^{\mu t + \frac{1}{2}\frac{\sigma^2}{n}t^2} \quad [1]$$

This has the same MGF as a $N\left(\mu, \frac{\sigma^2}{n}\right)$. Therefore by the uniqueness property of MGFs \bar{X} has a $N\left(\mu, \frac{\sigma^2}{n}\right)$ distribution. [1]

- (iii) **Comment on part (ii)**

This is the Central Limit Theorem. Therefore we have shown that the Central Limit Theorem is an *exact* result for a normal distribution. [1]

Solution 2.23(i) **Proof**

$E(Y | X = x)$ is a function of x so using $E[g(x)] = \int_x g(x)f(x) dx$, we have:

$$E[E(Y | X)] = \int_x E(Y | x)f(x) dx \quad [1]$$

Using the definition of $E(Y | X = x) = \int_y yf(y | x) dy$ gives:

$$E[E(Y | X)] = \int_x \left(\int_y yf(y | x) dy \right) f(x) dx$$

Using the definition $f(y | x) = \frac{f(x,y)}{f(x)}$ gives:

$$\begin{aligned} E[E(Y | X)] &= \int_x \left(\int_y y \frac{f(x,y)}{f(x)} dy \right) f(x) dx \\ &= \int_x \int_y yf(x,y) dy dx \\ &= \int_y y \left(\int_x f(x,y) dx \right) dy \end{aligned} \quad [1]$$

Since integrating the joint density function, $f(x,y)$, over all values of x gives the marginal density function, $f(y)$, we have:

$$E[E(Y | X)] = \int_y yf(y) dy = E(Y) \quad [1]$$

(ii) ***Calculate the unconditional mean and variance***

The mean and variance of the gamma distribution are given by:

$$E(X) = \frac{\alpha}{\lambda} = \frac{3}{2} = 1.5 \quad \text{var}(X) = \frac{\alpha}{\lambda^2} = \frac{3}{4} = 0.75 \quad [1]$$

Using the result from part (i), ie $E(Y) = E[E(Y | X)]$:

$$E(Y) = E[3X + 1] = 3E[X] + 1 = 3 \times 1.5 + 1 = 5.5 \quad [1]$$

Using the result $\text{var}(Y) = \text{var}[E(Y | X)] + E[\text{var}(Y | X)]$ from page 16 of the *Tables*:

$$\begin{aligned} \text{var}(Y) &= E[2X^2 + 5] + \text{var}[3X + 1] \\ &= 2E[X^2] + 5 + 9\text{var}[X] \end{aligned} \quad [1]$$

$$\text{Using the fact that } E(X^2) = \text{var}(X) + E^2(X) = 0.75 + 1.5^2 = 3 : \quad [1]$$

$$\text{var}(Y) = 2 \times 3 + 5 + 9 \times 0.75 = 17.75$$

$$\text{So the standard deviation is } \sqrt{17.75} = 4.21. \quad [1]$$

Solution 2.24(i) **Conditional variance**

$$\text{var}(X | Y = 2) = E(X^2 | Y = 2) - E^2(X | Y = 2)$$

$$\begin{aligned} E(X | Y = 2) &= \sum x P(X = x | Y = 2) = \sum x \frac{P(X = x \cap Y = 2)}{P(Y = 2)} \\ &= 1 \times \frac{0}{0.6} + 2 \times \frac{0.3}{0.6} + 3 \times \frac{0.1}{0.6} + 4 \times \frac{0.2}{0.6} \\ &= 2 \frac{5}{6} \end{aligned} \quad [1]$$

$$\begin{aligned} E(X^2 | Y = 2) &= \sum x^2 P(X = x | Y = 2) = \sum x^2 \frac{P(X = x \cap Y = 2)}{P(Y = 2)} \\ &= 1^2 \times \frac{0}{0.6} + 2^2 \times \frac{0.3}{0.6} + 3^2 \times \frac{0.1}{0.6} + 4^2 \times \frac{0.2}{0.6} \\ &= 8 \frac{5}{6} \end{aligned} \quad [1]$$

$$\text{So } \text{var}(X | Y = 2) = 8 \frac{5}{6} - \left(2 \frac{5}{6} \right)^2 = \frac{29}{36} = 0.80556. \quad [1]$$

(ii) **Conditional expectation**

We require:

$$E(U | V = v) = \int_u u f(u | v) du$$

Now:

$$f(v) = \int_{u=0}^1 \frac{48}{67} (2uv - u^2) du = \frac{48}{67} \left[u^2 v - \frac{1}{3} u^3 \right]_{u=0}^1 = \frac{48}{67} \left[v - \frac{1}{3} \right] \quad [1]$$

$$\Rightarrow f(u | v) = \frac{f(u, v)}{f(v)} = \frac{\frac{48}{67} (2uv - u^2)}{\frac{48}{67} (v - \frac{1}{3})} = \frac{2uv - u^2}{v - \frac{1}{3}} \quad [1]$$

So:

$$E(U | V = v) = \int_{u=0}^1 \frac{2u^2v - u^3}{v - \frac{1}{3}} du = \left[\frac{\frac{2}{3}u^3v - \frac{1}{4}u^4}{v - \frac{1}{3}} \right]_{u=0}^1 = \frac{\frac{2}{3}v - \frac{1}{4}}{v - \frac{1}{3}} \quad [1]$$

Solution 2.25

(i) **Variance formula**

Using the given formula:

$$\text{var}(S) = E[\text{var}(S | N)] + \text{var}[E(S | N)]$$

So we require the conditional mean and variance of S:

$$\begin{aligned} E(S | N) &= E(X_1 + X_2 + \dots + X_N) = E(X_1) + E(X_2) + \dots + E(X_N) \\ &= NE(X) \end{aligned} \quad [1]$$

$$\begin{aligned} \text{var}(S | N) &= \text{var}(X_1 + X_2 + \dots + X_N) \\ &= \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_N) \text{ by independence} \\ &= N \text{var}(X) \end{aligned} \quad [1]$$

Substituting into the original formula gives:

$$\text{var}(S) = E[N \text{var}(X)] + \text{var}[NE(X)] = \text{var}(X)E[N] + E^2(X)\text{var}[N] \quad [2]$$

Where we have used the results $E[aX] = aE[X]$ and $\text{var}[aX] = a^2 \text{var}[X]$.

(ii) **Mean and standard deviation**

We have $N \sim \text{Bin}(100, 0.05)$. So:

$$E(N) = 5 \quad \text{var}(N) = 4.75$$

We also have $X \sim \text{Exp}(0.02)$, so:

$$E(X) = 50 \quad \text{var}(X) = 2,500$$

So using the formulae given on page 16 of the *Tables*:

$$E(S) = 5 \times 50 = 250 \quad [1]$$

$$\text{var}(S) = 5 \times 2,500 + 4.75 \times 50^2 = 24,375 \Rightarrow sd(S) = 156.12 \quad [1]$$

(iii) ***Probability***

Using the Central Limit Theorem, we have:

$$S \sim N(250, 24375) \quad [1]$$

So:

$$P(S > 500) = P\left(Z > \frac{500 - 250}{\sqrt{24,375}}\right) = P(Z > 1.601) = 0.0547 \quad [2]$$

Solution 2.26(i) **MGF**

The definition of the MGF of Y is:

$$M_Y(t) = E(e^{tY})$$

Using the result $E(Y) = E[E(Y | X)]$ given on page 16 of the *Tables* with $Y = e^{tY}$ and $X = N$ gives:

$$M_Y(t) = E\left[E\left(e^{tY} | N = n\right)\right] \quad [1/2]$$

Now:

$$\begin{aligned} E\left(e^{tY} | N = n\right) &= E\left(e^{t(X_1 + \dots + X_n)}\right) \\ &= E\left(e^{tX_1} \dots e^{tX_n}\right) \\ &= E\left(e^{tX_1}\right) \dots E\left(e^{tX_n}\right) \quad X_i \text{ independent} \\ &= M_{X_1}(t) \dots M_{X_n}(t) \\ &= (M_X(t))^n \quad X_i \text{ identical} \end{aligned} \quad [1]$$

Hence:

$$M_Y(t) = E\left[\left(M_X(t)\right)^N\right] = E\left[e^{\ln(M_X(t))^N}\right] = E\left[e^{N \ln(M_X(t))}\right] \quad [1]$$

This is the MGF of N , $E\left(e^{tN}\right)$, but instead of t we have $\ln(M_X(t))$. Hence:

$$M_Y(t) = M_N\left[\ln(M_X(t))\right] \quad [1/2]$$

Now we have $M_N(t) = e^{\lambda(e^t - 1)}$ and $M_X(t) = (q + pe^t)^n$, so:

$$\begin{aligned} M_Y(t) &= \exp\left\{\lambda\left(e^{\ln M_X(t)} - 1\right)\right\} \\ &= \exp\left\{\lambda\left(e^{\ln(q+pe^t)^n} - 1\right)\right\} \\ &= \exp\left\{\lambda\left((q+pe^t)^n - 1\right)\right\} \end{aligned} \quad [1]$$

(ii) ***Mean and variance***

The CGF is given by:

$$C_Y(t) = \lambda\left((q+pe^t)^n - 1\right)$$

Hence:

$$C'_Y(t) = \lambda n p e^t (q + pe^t)^{(n-1)} \Rightarrow E(Y) = C'_Y(0) = \lambda np \quad [2]$$

And using the product rule:

$$\begin{aligned} C''_Y(t) &= \lambda n p e^t (q + pe^t)^{(n-1)} + \lambda n(n-1)p^2 e^{2t} (q + pe^t)^{(n-2)} \\ \Rightarrow \text{var}(Y) &= C''_Y(0) = \lambda np + \lambda n(n-1)p^2 = \lambda np(q + np) \end{aligned} \quad [1] \quad [1]$$

We could also use MGFs but it really is very messy.

Solution 2.27

We have $N \sim \text{Bin}(1000, 0.05)$. Using a normal approximation:

$$\text{Bin}(1000, 0.05) \approx N(50, 47.5) \quad [2]$$

Using a continuity correction $P(40 < N < 70) \approx P(40.5 < N < 69.5)$. Hence: [1]

$$\begin{aligned} P(40 < N < 70) &\approx P(N < 69.5) - P(N < 40.5) \\ &= P(Z < 2.829) - P(Z < -1.378) \\ &= P(Z < 2.829) - [1 - P(Z < 1.378)] \\ &= 0.99766 - [1 - 0.9159] \\ &= 0.91356 \end{aligned} \quad [2]$$

Solution 2.28

(i)(a) **Mode**

The mode is the maximum of the PDF $f(y)$:

$$f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} \quad y > 0$$

Differentiating and setting it equal to zero gives:

$$\begin{aligned} \frac{d}{dy} f(y) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \left[(\alpha-1)y^{\alpha-2} e^{-\lambda y} - \lambda y^{\alpha-1} e^{-\lambda y} \right] \\ \Rightarrow y^{\alpha-2} e^{-\lambda y} [(\alpha-1) - \lambda y] &= 0 \end{aligned} \quad [1]$$

Alternatively we could have differentiated the log of the PDF.

This gives:

$$y = 0 \quad \text{or} \quad y = \frac{\alpha-1}{\lambda} \quad [1]$$

Since $f(y) \geq 0$ and $f(0) = 0$, the first solution of zero **must** be a minimum and therefore the second solution **must** be a maximum.

Alternatively, this can be shown to be a maximum by considering the second derivative, but this was not expected.

$$\frac{d^2}{dy^2} f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda y} \left[(\alpha-1)(\alpha-2)y^{\alpha-3} - 2\lambda(\alpha-1)y^{\alpha-2} + \lambda^2 y^{\alpha-1} \right]$$

Substituting $y = \frac{\alpha-1}{\lambda}$ gives:

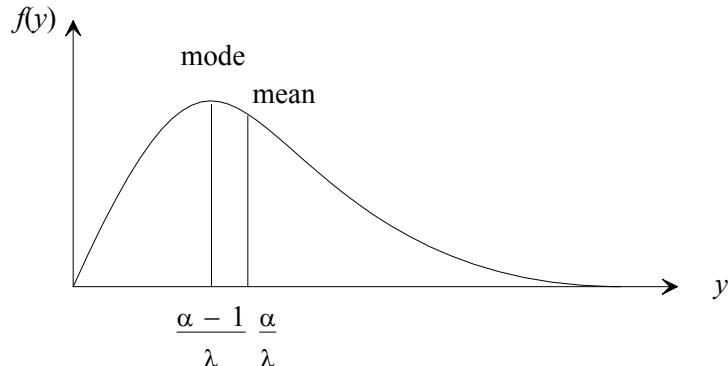
$$\begin{aligned} \frac{d^2}{dy^2} f(y) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-(\alpha-1)} \frac{(\alpha-1)^\alpha}{\lambda^\alpha} \left[\frac{\lambda^3(\alpha-2)}{(\alpha-1)^2} - \frac{\lambda^3}{(\alpha-1)} \right] \\ &= -\frac{1}{\Gamma(\alpha)} e^{-(\alpha-1)} (\alpha-1)^\alpha \frac{\lambda^3}{(\alpha-1)^2} \end{aligned}$$

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 $\alpha \rightarrow \infty$, but keeping μ constant. The mean is $\frac{\alpha}{\lambda}$, which will remain constant. The mode is $\frac{\alpha-1}{\lambda} = \frac{\alpha}{\lambda} - \frac{1}{\lambda} = \mu - \frac{\mu}{\alpha}$, which will be less than the mean μ , but will tend to μ as $\alpha \rightarrow \infty$.

So, for large α , the distribution will look like this:



The mean and mode are very close together. [2]

In fact, the distribution approaches a normal distribution in the limit.

(ii)(a) ***Probability using gamma-chi squared relationship***

$Y \sim \text{Gamma}(50, 0.2)$. Using the relationship $Y \sim \text{Gamma}(\alpha, \lambda) \Rightarrow 2\lambda Y \sim \chi^2_{2\alpha}$:

$$\begin{aligned} P(Y > 350) &= P(2\lambda Y > 2\lambda \times 350) \\ &= P(0.4Y > 140) \\ &= P(\chi^2_{100} > 140) \end{aligned} \quad [1]$$

Using the χ^2 tables on page 169 gives a value of approximately 0.5%. [1]

(ii)(b) ***Probability using normal approximation***

The mean and variance of the gamma distribution are:

$$E(Y) = \frac{\alpha}{\lambda} = \frac{50}{0.2} = 250 \quad \text{var}(Y) = \frac{\alpha}{\lambda^2} = \frac{50}{0.2^2} = 1,250$$

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:

$$Y \sim \text{Gamma}(50, 0.2) \approx N(250, 1250) \quad [1]$$

Hence:

$$\begin{aligned} P(Y > 350) &= P\left(Z > \frac{350 - 250}{\sqrt{1,250}}\right) \\ &= P(Z > 2.828) \quad [1] \\ &= 1 - 0.99766 = 0.234\% \quad [1] \end{aligned}$$

(iii) ***Explain the differences***

The gamma is only symmetrical when $\alpha \rightarrow \infty$. 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. [1]

Solution 2.29(i) ***Prove mean of sample variance***

Since we are sampling from a normal distribution, we have $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$.

Now $\chi_{n-1}^2 \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{1}{2}\right)$. Therefore:

$$\begin{aligned} E\left(\frac{(n-1)S^2}{\sigma^2}\right) &= E\left(\chi_{n-1}^2\right) = \frac{(n-1)/2}{1/2} = n-1 \\ \Rightarrow \frac{(n-1)}{\sigma^2} E(S^2) &= (n-1) \\ \Rightarrow E(S^2) &= \sigma^2 \end{aligned} \quad [2]$$

(ii) ***Prove variance of sample variance***

Similarly:

$$\begin{aligned} \text{var}\left(\frac{(n-1)S^2}{\sigma^2}\right) &= \text{var}\left(\chi_{n-1}^2\right) = \frac{(n-1)/2}{(1/2)^2} = 2(n-1) \\ \Rightarrow \frac{(n-1)^2}{\sigma^4} \text{var}(S^2) &= 2(n-1) \\ \Rightarrow \text{var}(S^2) &= \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1} \end{aligned} \quad [2]$$

Solution 2.30

(i)(a) **Probability mean of X greater than mean of Y**

We require $P(\bar{X} > \bar{Y}) = P(\bar{X} - \bar{Y} > 0)$, therefore we need the distribution of $\bar{X} - \bar{Y}$. Working in 1,000's, the distributions of the sample means are:

$$\bar{X} \sim N(100, 10) \quad \bar{Y} \sim N(90, 5)$$

Using the work from Chapter 6 gives:

$$\bar{X} - \bar{Y} \sim N(100 - 90, 10 + 5) = N(10, 15) \quad [1]$$

$$\begin{aligned} P(\bar{X} - \bar{Y} > 0) &= P\left(Z > \frac{0 - 10}{\sqrt{15}}\right) \\ &= P(Z > -2.582) \\ &= \Phi(2.582) \\ &= 0.995 \end{aligned} \quad [2]$$

(i)(b) **Probability difference between means is less than 5,000**

Using the distribution of $\bar{X} - \bar{Y}$ from before:

$$\begin{aligned} P(|\bar{X} - \bar{Y}| < 5) &= P(-5 < \bar{X} - \bar{Y} < 5) \\ &= P(\bar{X} - \bar{Y} < 5) - P(\bar{X} - \bar{Y} < -5) \quad [1] \\ &= P\left(Z < \frac{5 - 10}{\sqrt{15}}\right) - P\left(Z < \frac{-5 - 10}{\sqrt{15}}\right) \\ &= P(Z < -1.291) - P(Z < -3.873) \quad [1] \\ &= [1 - \Phi(1.291)] - [1 - \Phi(3.873)] \\ &= \Phi(3.873) - \Phi(1.291) \\ &= 0.0983 \quad [1] \end{aligned}$$

(ii)(a) **Probability sample variance of X less than sample variance of Y**

We require $P(S_X^2 < S_Y^2) = P\left(S_X^2 / S_Y^2 < 1\right)$. Using the definition of the F-distribution, we get:

$$\frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} = \frac{S_X^2 / S_Y^2}{\sigma_X^2 / \sigma_Y^2} = \frac{S_X^2 / S_Y^2}{10^2 / 5^2} = \frac{S_X^2 / S_Y^2}{4} \sim F_{9,4}$$

Hence:

$$P\left(S_X^2 / S_Y^2 < 1\right) = P\left(\frac{S_X^2 / S_Y^2}{4} < 0.25\right) = P(F_{9,4} < 0.25) \quad [1]$$

Since this is on the lower tail we need to use the $\frac{1}{F_{m,n}}$ result:

$$P(F_{9,4} < 0.25) = P\left(F_{4,9} > \frac{1}{0.25}\right) = P(F_{4,9} > 4) \quad [1]$$

This value is between 2½% and 5% (we could interpolate to get 4.2%). [1]

(ii)(b) **Probability sample s.d. of X is greater than four times sample s.d. of Y**

We require $P(S_X > 4S_Y) = P(S_X / S_Y > 4) = P\left(S_X^2 / S_Y^2 > 16\right)$. Using the result from (ii)(a) we get:

$$P\left(S_X^2 / S_Y^2 > 16\right) = P\left(\frac{S_X^2 / S_Y^2}{4} > 4\right) = P(F_{9,4} > 4) \approx 10\% \quad [2]$$

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.

Part 3 – Questions

Note that the split between development questions and exam-style questions is somewhat subjective. For example, there have been past CT3 exam questions that test knowledge of the Core Reading, and so are similar to what we've included here as development questions. The exam-style questions involve more application and a wider range of ideas and are typically the more challenging questions in the exam.

1 Development questions

Question 3.1

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 estimators for:

(a) λ from an $Exp(\lambda)$ distribution

(b) ν from a χ^2_ν distribution. [4]

(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. [6]

[Total 10]

Question 3.2

A random sample from a $Beta(\alpha, \beta)$ distribution has a mean of 0.8 and a standard deviation of 0.3. Derive the method of moments estimates of α and β . [5]

Question 3.3

Write down the equation(s) you would solve to estimate the parameters for the following distributions using the method of moments:

(i) $U(-\theta, \theta)$ [2]

(ii) $Beta(\alpha, \alpha)$. [2]
[Total 4]

Question 3.4

(i) A random sample $\underline{X} = (X_1, \dots, X_n)$ is taken. Derive from first principles the maximum likelihood estimators for:

(a) λ from an $Exp(\lambda)$ distribution

(b) p from a $Bin(m, p)$ distribution. [6]

(ii) Derive the $CRLB$ for each of the estimators in part (i). [4]
[Total 10]

Question 3.5

Write an expression for the likelihood of obtaining exactly 10 numbers in each of the intervals $(0, 0.1)$, $(0.1, 0.2)$, ..., $(0.9, 1)$ from a computer that generates 100 random numbers with a uniform distribution on the range $(0, 1)$. [3]

Question 3.6

The likelihood function for the parameter θ based on a random sample of n observations from a population with a continuous uniform distribution on the range $(-\theta/2, \theta/2)$ is:

- A $\prod_{i=1}^n x_i$ if $\max |x_i| < \theta/2$, and zero otherwise
- B $\prod_{i=1}^n \frac{x_i}{\theta}$ if $\max |x_i| < \theta/2$, and zero otherwise
- C $\frac{1}{\theta^n}$ if $\max |x_i| < \theta/2$, and zero otherwise
- D $\frac{1}{\theta^n}$ if $\max |x_i| < \theta/2$, and zero otherwise.

[3]

Question 3.7

Show that the likelihood that an observation from a $Poisson(\lambda)$ distribution takes an odd value (ie 1, 3, 5,...) is $\frac{1}{2}(1 - e^{-2\lambda})$. [4]

Question 3.8

A single fair die was thrown an unknown number of times. Ten sixes were observed. Determine the method of moments estimate of the total number of throws. [3]

Question 3.9

When 50 disk drives were tested in continuous use, 47 were still functioning perfectly after 200 days. If the lifetimes of the drives are assumed to have an exponential distribution, calculate the maximum likelihood estimate of the average lifetime. [6]

Question 3.10

If X denotes the number of successes in a series of n independent trials with constant success probability p , and $kX(n-X)$ is an unbiased estimator of $p(1-p)$, determine the value of k . [4]

Question 3.11

- (i) Explain briefly the importance of the Cramér-Rao lower bound. [3]
- (ii) Give a formula for calculating it, defining the symbols you use. [4]
- [Total 7]

Question 3.12

Determine the minimum possible variance of an unbiased estimator of the parameter p based on a total of x observed successes in a series of n independent Bernoulli trials. [4]

Question 3.13

State whether each of the following statements is true or false for IID observations from *any* population with finite variance σ^2 .

- (i) The second non-central sample moment is an unbiased estimator of the second non-central population moment. [2]
- (ii) The fourth central sample moment is an unbiased estimator of the fourth central population moment. [1]
- (iii) $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ is an unbiased estimator of σ^2 . [1]
- (iv) $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$ always has a chi square distribution. [1]
- (v) If the population is normal, the distribution of $\frac{1}{\sigma^2} \sum (X_i - \bar{X})^2 \sim \chi_{n-1}^2$ [1]
- (vi) The sample mean is an unbiased estimator of the population mean. [1]
- (vii) The sample mean is the maximum likelihood estimator of the population mean. [2]
- (viii) The sample mean always has variance σ^2/n . [1]
- (ix) The sample standard deviation S is an unbiased estimator of σ . [2]

- (x) If the population is normal, the sample standard deviation S is an unbiased estimator of σ . [1]
- (xi) The sample mean and sample variance are independent. [1]
- (xii) If the population is normal, the sample mean and sample variance are independent. [2]
- [Total 16]

Question 3.14

The mean square error of a biased estimator $\hat{\theta}$ of the parameter θ can be calculated as:

- A $\text{var}(\hat{\theta}) - [\text{Bias}(\hat{\theta})]^2$
- B $\text{var}(\hat{\theta})$
- C $E[(\hat{\theta} - \theta)^2]$
- D $E[\hat{\theta}^2] - \theta^2$. [2]

Question 3.15

Which of the following criteria is a necessary and sufficient condition for the sample variance, defined as $S^2 = \frac{1}{n-1}[\sum X_i^2 - n\bar{X}^2]$, of a set of n observations from a population to be an unbiased estimator of the population variance?

- A S^2 will always be unbiased.
- B S^2 will be unbiased if and only if the observations are independent.
- C S^2 will be unbiased if and only if the observations are normally distributed.
- D S^2 will be unbiased if and only if the observations are independent and normally distributed. [3]

Question 3.16

A random sample from a normal distribution gives the following values:

68.8, 79.4, 63.2, 67.1, 69.4, 78.2, 80.4, 69.6

- (i) Obtain a 95% confidence interval for the population mean, if the population standard deviation is:
 - (a) unknown
 - (b) 6.[5]

- (ii) Obtain a 90% confidence interval for the population standard deviation. [3]
[Total 8]

Question 3.17

The amounts of individual claims arising under a certain type of general insurance policy are 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. Find the lower and upper limits of a 95% confidence interval for the lognormal parameter μ . [3]

Question 3.18

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 [3]

- (ii) the population standard deviation is 8.4 cm. [3]
[Total 6]

2 Exam-style questions

Question 3.19

In a process, the number of failures before the first success has the following probability distribution:

$$P(X = x) = \theta(1 - \theta)^x \quad x = 0, 1, 2, \dots$$

Show that the method of moments and the method of maximum likelihood both lead to the same estimator for θ . [5]

Question 3.20

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 . [1]

A random sample of 30 observations gave respective frequencies of 7, 6 and 17.

- (ii) Calculate the method of moments estimator of α . [3]

- (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:

$$180\hat{\alpha}^2 + \frac{111}{8}\hat{\alpha} - \frac{91}{32} = 0 \quad [5]$$

- (iv) Hence determine the maximum likelihood estimate and explain why the second root is rejected as a possible estimate of α . [3]

[Total 12]

Question 3.21

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.

<i>No. of claims</i>	<i>No. of policies</i>
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]
 - (iv) Explain briefly why you would expect a negative binomial distribution to fit the above data better than a Poisson distribution. [2]
- [Total 15]

Question 3.22

Claims under a certain type of insurance policy occur as a Poisson process so that the number of claims arising from a policy in one calendar year is a Poisson random variable with mean μ .

In order to estimate μ a random sample of n such policies is examined and it is found that n_0 of these policies incurred no claims during the last full calendar year, n_1 of them incurred one claim and the remainder incurred more than one claim.

- (i) Write down the likelihood of these observations and show that the maximum likelihood estimator $\hat{\mu}$ is a solution of the equation:

$$n_1 - (n_0 + n_1)\hat{\mu} + e^{-\hat{\mu}}(n\hat{\mu}^2 + n_0\hat{\mu} - n_1) = 0$$

[You do not need to verify that a maximum is attained.] [10]

- (ii) A random sample of 20 such policies yields $n_0 = 8$ and $n_1 = 7$. By using the tables of the cumulative Poisson distribution, or otherwise, to find a good starting approximation for $\hat{\mu}$ and then using trial and error (or a more sophisticated method), determine $\hat{\mu}$ correct to two decimal places. [7]

[Total 17]

Question 3.23

The claim amounts for a certain insurance company follow a Pareto distribution with probability distribution function given by:

$$\alpha 250^\alpha (250 + x)^{-\alpha-1}$$

- (i) Show that the maximum likelihood estimator for α is given by:

$$\hat{\alpha} = \frac{n}{\sum \ln(250 + x_i) - n \ln 250} [4]$$

A random sample of 150 claims gives $\sum \ln(250 + x_i) = 869$.

- (ii) Given that the estimator is asymptotically unbiased, obtain a 95% confidence interval for the parameter α . [5]

[Total 9]

Question 3.24

The random variable X has a generalised Pareto distribution with density function, $f(x;\alpha)$ given by:

$$f(x;\alpha) = \alpha(\alpha+1)(1,000)^\alpha x(1,000+x)^{-\alpha-2}, \quad x > 0$$

A sample of 100 values of X gives the following information:

$$\sum_{i=1}^{100} \log_e(1,000 + x_i) = 750$$

- (i) Calculate the maximum likelihood estimate $\hat{\alpha}$ of α . [6]
 - (ii) Estimate the standard deviation of $\hat{\alpha}$. [3]
- [Total 9]

Question 3.25

- (i) Let X_1, X_2, \dots, X_n be a random sample of a Poisson variable with mean μ .
 - (a) Show that the maximum likelihood estimator of μ is $\bar{X} = \frac{1}{n} \sum X_i$.
 - (b) Use generating functions to derive the distribution of $\sum X_i$. [6]
 - (ii) An insurance company undertakes a study to estimate the mean annual claim rate for policies of a certain type. It assumes a Poisson distribution for the number of claims, with the same mean for each policy.
 - (a) A very small pilot study of 10 policies yields a total of 2 claims during the last year. Estimate the mean annual claim rate and determine a 95% confidence interval for this mean claim rate.
 - (b) The full study of 500 policies yields a total of 106 claims. Estimate the mean annual claim rate and determine a 95% confidence interval for this mean claim rate.
 - (c) Comment briefly on the comparison of the confidence intervals obtained in parts (ii)(a) and (ii)(b) above. [11]
- [Total 17]

Question 3.26

A random sample X_1, \dots, X_n is taken from a normal distribution with mean μ and variance σ^2 .

- (i) State the distribution of $\frac{\sum(X_i - \bar{X})^2}{\sigma^2}$. [1]

It is decided to estimate the variance, σ^2 , using the following estimator:

$$\hat{\sigma}^2 = \frac{1}{n+b} \sum (X_i - \bar{X})^2$$

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]

Question 3.27

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 $\sigma > k$. Calculate the value of k . [4]

Question 3.28

A sample of 51 life assurance policyholders has a mean sum assured of £180,000 and a standard deviation of £60,000. Eight of the policyholders had a sum assured greater than £250,000.

Obtain a symmetric 95% confidence interval for the:

- (i) true mean sum assured [3]
 - (ii) true percentage of policyholders with a sum assured greater than £250,000. [3]
- [Total 6]

Question 3.29

- (i) In an opinion poll, a random sample is to be asked whether they favour closer ties with Europe. Determine the minimum sample size required to ensure that 95% confidence limits for the underlying population proportion are of the form “ $\pm 5\%$ ”, justifying any approximations used. [4]
 - (ii) 1,000 people took the opinion poll in part (i). 30% said “Yes” to closer ties with Europe, 50% said “No” and 20% said “Don’t know”. Calculate 95% confidence intervals for the proportion of the whole population holding each opinion. [4]
 - (iii) After an extensive advertising campaign by the government another opinion poll of 800 people was taken. Of those questioned, 35% said “Yes”.
 - (a) Obtain a 90% confidence interval for the difference in proportions favouring closer ties with Europe before and after the campaign. [4]
 - (b) Comment on your answer. [4]
- [Total 12]

Question 3.30

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.

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) Find a 95% confidence interval for the difference between the mean claim amounts.
 (b) Comment on your answer. [6]

 - (ii) (a) Find 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]

Question 3.31

- (i) A sample value of 2 is obtained from a Poisson distribution with mean μ . Obtain an *exact* two-sided 90% confidence interval for μ . [5]

 - (ii) A sample of 30 values from the same Poisson distribution had a mean of 2. Use these data to obtain an approximate 90% confidence interval for μ . [3]
- [Total 8]

Question 3.32

The Chevalier de Méré, a seventeenth century gambler, thought that it paid to bet evens on the event:

A : you will get one or more sixes when four unbiased dice are thrown

In other words, the Chevalier thought that $P(A) > \frac{1}{2}$.

- (i) (a) Show that the probability of this event occurring is, in fact, 0.5177.
- (b) The experiment of throwing four unbiased dice is performed 10 times and results in the event A occurring 8 times.
 - (1) Write down an equation which must be satisfied by p_L , a lower 95% confidence limit for $p = P(A)$.
 - (2) Verify that $p_L = 0.493$ satisfies this equation.
 - (3) Comment on this value of p_L relative to the true value of p as specified in part (i)(a). [6]
- (ii) The experiment of throwing four unbiased dice is performed 1,000 times and results in the event A occurring Y times.
 - (a) Write down a large sample expression for p_L , a lower 95% confidence limit for $p = P(A)$.
 - (b) Determine how large Y would have to be for p_L to be greater than $\frac{1}{2}$.
 - (c) Using the true value of p as specified in part (i), calculate the probability that p_L will, in fact, be greater than $\frac{1}{2}$. [8]
- (iii) For the situation where the experiment is performed 10,000 times:
 - (a) repeat part (ii)(b)
 - (b) repeat part (ii)(c)

and comment briefly on any difference between your two answers. [6]
[Total 20]

Part 3 – Solutions

Solution 3.1

(i) ***Method of moments (one unknown)***

We have one unknown and so require only one equation:

$$E(X) = \frac{1}{n} \sum x_i = \bar{x}$$

For our data we have $\bar{x} = 2.95$.

[2]

(i)(a) ***Exponential***

$$\frac{1}{\hat{\lambda}} = 2.95 \Rightarrow \hat{\lambda} = 0.33898$$

[1]

(i)(b) ***Chi-square***

Since $\chi_v^2 = \text{Gamma}(\frac{v}{2}, \frac{1}{2})$, we get $E(X) = \frac{\alpha}{\lambda} = \frac{v/2}{1/2} = v$. Hence $\hat{v} = 2.95$.

[1]

(ii) ***Method of moments (two unknowns)***

We have two unknowns and so require two equations.

Either: $E(X) = \frac{1}{n} \sum x_i = \bar{x}$ and $E(X^2) = \frac{1}{n} \sum x_i^2$

or: $E(X) = \frac{1}{n} \sum x_i = \bar{x}$ and $\text{var}(X) = s^2$

For our data we have $\bar{x} = 2.95$, $\frac{1}{8} \sum x_i^2 = 13.635$ and $s^2 = 5.6371$.

[2]

(ii)(a) **Negative binomial**

Using the first method gives:

$$E(X) = \frac{\hat{k}(1 - \hat{p})}{\hat{p}} = 2.95$$

$$E(X^2) = \text{var}(X) + E^2(X) = \frac{\hat{k}(1 - \hat{p})}{\hat{p}^2} + \left(\frac{\hat{k}(1 - \hat{p})}{\hat{p}} \right)^2 = 13.635$$

Substituting the first equation into the second gives:

$$\frac{2.95}{\hat{p}} + 2.95^2 = 13.635 \Rightarrow \frac{2.95}{\hat{p}} = 4.9325 \Rightarrow \hat{p} = 0.59807$$

Hence, substituting this back into the first equation gives $\hat{k} = 4.3896$. [2]

Using the second method gives:

$$E(X) = \frac{\hat{k}(1 - \hat{p})}{\hat{p}} = 2.95 \quad \text{and} \quad \text{var}(X) = \frac{\hat{k}(1 - \hat{p})}{\hat{p}^2} = 5.6371$$

Substituting the first equation into the second gives:

$$\frac{2.95}{\hat{p}} = 5.6371 \Rightarrow \hat{p} = 0.52331$$

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:

$$e^{2(\hat{\mu} + \frac{1}{2}\hat{\sigma}^2)} e^{\hat{\sigma}^2} = 2.95^2 e^{\hat{\sigma}^2} = 13.635 \Rightarrow \hat{\sigma}^2 = 0.44903$$

Substituting this into the first equation gives $\hat{\mu} = 0.85729$. [2]

Using the second method gives:

$$E(X) = e^{\hat{\mu} + \frac{1}{2}\hat{\sigma}^2} = 2.95 \quad \text{and} \quad \text{var}(X) = e^{2\hat{\mu} + \hat{\sigma}^2} (e^{\hat{\sigma}^2} - 1) = 5.6371$$

Substituting the first equation into the second gives:

$$\text{var}(X) = 2.95^2 (e^{\hat{\sigma}^2} - 1) = 5.6371 \Rightarrow \hat{\sigma}^2 = 0.49942$$

Hence, substituting this into the first equation gives $\hat{\mu} = 0.83210$.

Solution 3.2

We have two unknowns, so we require two equations:

$$E(X) = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = 0.8 \quad \text{and} \quad \text{var}(X) = \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha} + \hat{\beta})^2(\hat{\alpha} + \hat{\beta} + 1)} = 0.3^2 \quad [2]$$

The first equation gives $\hat{\alpha} = 4\hat{\beta}$. Substituting this into the second equation gives:

$$\frac{4\hat{\beta}^2}{(5\hat{\beta})^2(5\hat{\beta} + 1)} = \frac{4}{25(5\hat{\beta} + 1)} = 0.09 \Rightarrow \hat{\beta} = 0.15 \quad [2]$$

Hence, we get $\hat{\alpha} = 4 \times 0.15 = 0.62$. [1]

Solution 3.3(i) **Method of moments (uniform)**

There is one parameter to estimate. However, since $E[X] = 0$, the first moments can't be used. So we must equate the second moments:

$$E[X^2] = \frac{1}{n} \sum x_i^2 \quad ie \quad \frac{\theta^2}{3} = \frac{1}{n} \sum x_i^2 \quad [2]$$

(ii) **Method of moments (beta)**

There is one parameter to estimate. However, $E[X] = \frac{\alpha}{\alpha + \alpha} = \frac{1}{2}$. So again the first moments can't be used. So we must equate the second moments:

$$E[X^2] = \frac{1}{n} \sum x_i^2 \quad ie \quad \frac{1}{4(2\alpha + 1)} + \frac{1}{4} = \frac{1}{n} \sum x_i^2 \quad [2]$$

Solution 3.4(i)(a) **MLE (Exponential)**

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i} \\ \Rightarrow \ln L(\lambda) &= n \ln \lambda - \lambda \sum x_i \\ \Rightarrow \frac{d}{d\lambda} \ln L(\lambda) &= \frac{n}{\lambda} - \sum x_i \end{aligned} \quad [1]$$

Setting the differential equal to zero to find the maximum:

$$\Rightarrow \frac{n}{\hat{\lambda}} - \sum x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}} \quad [1]$$

Checking that it gives a maximum:

$$\Rightarrow \frac{d^2}{d\lambda^2} \ln L(\lambda) = -\frac{n}{\lambda^2} < 0 \Rightarrow \max \quad [1]$$

So our estimator is $\hat{\lambda} = 1/\bar{X}$.

(i)(b) **MLE (Binomial)**

$$\begin{aligned} L(p) &= \prod_{i=1}^n \binom{m}{x_i} p^{x_i} (1-p)^{m-x_i} = \text{const} \times p^{\sum x_i} (1-p)^{mn - \sum x_i} \\ \Rightarrow \ln L(p) &= \text{const} + \sum x_i \ln p + (mn - \sum x_i) \ln(1-p) \\ \Rightarrow \frac{\partial}{\partial p} \ln L(p) &= \frac{\sum x_i}{p} - \frac{mn - \sum x_i}{1-p} \end{aligned}$$

Setting the differential equal to zero to find the maximum:

$$\begin{aligned} \Rightarrow \frac{\sum x_i}{\hat{p}} &= \frac{mn - \sum x_i}{1 - \hat{p}} \Rightarrow (1 - \hat{p}) \sum x_i = \hat{p}(mn - \sum x_i) \\ \Rightarrow \hat{p} &= \frac{\sum x_i}{mn} \end{aligned} \quad [2]$$

Checking that it gives a maximum:

$$\Rightarrow \frac{\partial^2}{\partial p^2} \ln L(p) = -\frac{\sum x_i}{p^2} - \frac{mn - \sum x_i}{(1-p)^2} < 0 \Rightarrow \max \quad [1]$$

Since $x_i = 0, 1, \dots, m \Rightarrow mn - \sum x_i \geq 0$.

So our estimator is $\hat{p} = \frac{\sum X_i}{mn}$.

(ii)(a) **CRLB (Exponential)**

We know that $\frac{d^2}{d\lambda^2} \ln L(\lambda) = -\frac{n}{\lambda^2}$, hence:

$$CRLB(\lambda) = -\frac{1}{E\left[\frac{d^2}{d\lambda^2} \ln L(\lambda)\right]} = -\frac{1}{E[-n/\lambda^2]} = \frac{1}{n/\lambda^2} = \frac{\lambda^2}{n} \quad [2]$$

(ii)(b) **CRLB (Binomial)**

We know that $\frac{\partial^2}{\partial p^2} \ln L(p) = -\frac{\sum X_i}{p^2} - \frac{mn - \sum X_i}{(1-p)^2}$, hence:

$$\begin{aligned} E\left[\frac{\partial^2}{\partial p^2} \ln L(p)\right] &= E\left[-\frac{\sum X_i}{p^2} - \frac{mn - \sum X_i}{(1-p)^2}\right] \\ &= -\frac{E\left[\sum X_i\right]}{p^2} - \frac{mn - E\left[\sum X_i\right]}{(1-p)^2} \\ &= -\frac{\sum E[X_i]}{p^2} - \frac{mn - \sum E[X_i]}{(1-p)^2} \end{aligned}$$

Now since $X_i \sim Bin(m, p)$, we have $E(X_i) = mp$. Hence:

$$\begin{aligned} E\left[\frac{\partial^2}{\partial p^2} \ln L(p)\right] &= -\frac{\sum mp}{p^2} - \frac{mn - \sum mp}{(1-p)^2} \\ &= -\frac{mnp}{p^2} - \frac{mn - mnp}{(1-p)^2} \\ &= -\frac{mn}{p} - \frac{mn}{1-p} \\ &= -\frac{mn}{p(1-p)} \end{aligned}$$

$$\Rightarrow CRLB(p) = -\frac{1}{E\left[\frac{\partial^2}{\partial p^2} \ln L(p)\right]} = \frac{p(1-p)}{mn} \quad [2]$$

Solution 3.5

Let $U \sim U(0,1)$, such that $f(u) = 1 \quad 0 \leq u \leq 1$. The likelihood is just the probability:

$$\begin{aligned} L &= [P(0 \leq u \leq 0.1)]^{10} \times [P(0.1 \leq u \leq 0.2)]^{10} \times \cdots \times [P(0.9 \leq u \leq 1)]^{10} \\ &= \text{const} \times 0.1^{10} \times 0.1^{10} \times \cdots \times 0.1^{10} \\ &= \text{const} \times 0.1^{100} \end{aligned}$$

where the constant is the combinatorial factor, $\frac{100!}{10! \dots 10!}$ due to the different possible ways of obtaining the ten results in each interval. [3]

Solution 3.6

The probability density function for each x_i is:

$$f_X(x) = \frac{1}{\theta} \text{ if } |x_i| < \frac{\theta}{2}, \text{ and zero otherwise}$$

So the joint likelihood function is:

$$L(\theta) = \frac{1}{\theta^n} \text{ if } |x_i| < \frac{\theta}{2} \text{ for each } x_i, \text{ and zero otherwise}$$

The answer is D. [3]

Solution 3.7

We have:

$$\begin{aligned} P[\text{Poisson}(\lambda) \text{ is odd}] &= P(X = 1) + P(X = 3) + P(X = 5) + \dots \\ &= e^{-\lambda} \left[\lambda + \frac{\lambda^3}{3!} + \frac{\lambda^5}{5!} + \dots \right] \end{aligned} \quad [1]$$

To sum the series in the square bracket, note that:

$$\begin{aligned} e^\lambda &= 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots & [1] \\ e^{-\lambda} &= 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \dots \end{aligned}$$

So $\frac{1}{2}(e^\lambda - e^{-\lambda}) = \lambda + \frac{\lambda^3}{3!} + \dots$, which is the required series. [1]

So the required probability is:

$$e^{-\lambda} \times \frac{1}{2}(e^\lambda - e^{-\lambda}) = \frac{1}{2}(1 - e^{-2\lambda}) \quad [1]$$

Solution 3.8

We are estimating the unknown parameter n based on a single observation from a $\text{Bin}(n, 1/6)$ distribution. [1]

Equating the observed value to the population mean gives $\hat{n}/6 = 10$. [1]

So the method of moments estimate is $\hat{n} = 60$. [1]

Solution 3.9

The likelihood that 47 will still be functioning after 200 days is the binomial probability:

$$L(p) = \binom{50}{47} p^{47} (1-p)^3$$

So the log-likelihood is:

$$\log L(p) = 47 \log p + 3 \log(1-p) + \text{constant} \quad [1]$$

Differentiating with respect to p to maximise this:

$$\frac{d}{dp} \log L(p) = \frac{47}{p} - \frac{3}{1-p} \quad [1]$$

This is zero when:

$$47(1-p) = 3p$$

So the MLE of p is:

$$\hat{p} = \frac{47}{50} = 0.94 \quad [2]$$

Expressed in terms of the exponential parameter λ , p is:

$$p = \int_{200}^{\infty} \lambda e^{-\lambda t} dt = e^{-200\lambda}$$

So, using the invariance property of MLEs, the MLE of λ satisfies:

$$e^{-200\hat{\lambda}} = 0.94 \quad [1]$$

So the MLE of the average lifetime $\mu = 1/\lambda$ is:

$$\hat{\mu} = 1/\hat{\lambda} = -1/\frac{1}{200} \log 0.94 = 3,232 \quad [1]$$

i.e 3,232 days or 8.85 years.

Note how we have used the invariance property of MLEs here. We actually wanted to find the MLE of μ which equals $\frac{1}{\lambda}$. We did this by first finding the MLE of p which equals $e^{-200\lambda}$. The invariance property tells us that the MLEs are related in the same way, ie $\hat{\mu} = \frac{1}{\hat{\lambda}}$ and $\hat{p} = e^{-200\hat{\lambda}}$, so we can deduce $\hat{\mu}$ from the value of \hat{p} .

Solution 3.10

The expectation of the estimator is:

$$E[kX(n-X)] = k(nE[X] - E[X^2]) \quad [1]$$

$$= k[n(np) - (n^2 p^2 + npq)] \quad [1]$$

$$= k[n^2 p(1-p) - np(1-p)] = kn(n-1)p(1-p) \quad [1]$$

This will equal pq if $k = \frac{1}{n(n-1)}$. [1]

Solution 3.11

(i) **Importance of CRLB**

When estimating a parameter (or a function of a parameter) of a parent population, it is desirable to select an estimator that has a small variance, since it will tend to produce estimates close to the true value. [1]

The *CRLB* provides a formula for expressing the minimum possible variance of an unbiased estimator as a function of the true parameter value. It can be used to compare the efficiency of different estimators. [1]

It also provides an approximate value for the variance of the MLE of a parameter when the sample size is large. [1]

(ii) ***CRLB formula***

One formula for calculating the *CRLB* is:

$$CRLB(\theta) = -\frac{1}{E\left[\frac{\partial^2}{\partial \theta^2} \log L(\theta)\right]} \quad [1]$$

where:

- θ is the quantity being estimated [1]
- $L(\theta)$ is the (joint) likelihood of obtaining the observed results [1]
- the expectation is calculated with respect to the probability distribution of the parent population. [1]

Solution 3.12

The minimum variance of an unbiased estimator is given by the *CRLB*. [½]

The likelihood of obtaining the value x , which has a $Bin(n, p)$ distribution is:

$$L = \binom{n}{x} p^x (1-p)^{n-x} \quad [½]$$

Taking logs:

$$\log L = constant + x \log p + (n-x) \log(1-p) \quad [½]$$

Differentiating with respect to p (the parameter we are estimating):

$$\frac{\partial}{\partial p} \log L = \frac{x}{p} - \frac{n-x}{1-p} \quad [½]$$

Differentiating again:

$$\frac{\partial^2}{\partial p^2} \log L = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} \quad [½]$$

Finding the expectation of this (with respect to the random variable X):

$$\begin{aligned}
 E\left[\frac{\partial^2}{\partial p^2} \log L\right] &= -\frac{1}{p^2} E(X) - \frac{1}{(1-p)^2} [n - E(X)] \\
 &= -\frac{np}{p^2} - \frac{n-np}{(1-p)^2} = -\frac{n}{p} - \frac{n}{(1-p)} \\
 &= -n \left[\frac{(1-p)+p}{p(1-p)} \right] = -\frac{n}{p(1-p)}
 \end{aligned} \tag{1}$$

So the Cramér-Rao lower bound is:

$$CRLB = -1 \left/ E\left[\frac{\partial^2}{\partial p^2} \log L\right]\right. = \frac{p(1-p)}{n} \tag{1/2}$$

Solution 3.13

- (i) True.

$$\begin{aligned}
 E\left[\frac{1}{n} \sum X_i^2\right] &= \frac{1}{n} \sum E(X_i^2) = \frac{1}{n} \sum [\text{var}(X_i) + \mu^2] \\
 &= \frac{1}{n} \times n[\text{var}(X) + \mu^2] = \text{var}(X) + \mu^2 = E[X^2]
 \end{aligned} \tag{2}$$

- (ii) False. There is no reason why (in general) the fourth central sample moment, ie $\frac{1}{n} \sum (X_i - \bar{X})^4$ should be an unbiased estimator of the fourth central population moment ie $E[(X - \mu)^4]$. [1]

- (iii) True. The sample variance is always an unbiased estimator of the population variance, whatever the distribution of the parent population. [1]

- (iv) False. This is only true if the parent population has a normal distribution. [1]

- (v) True. This is the same as $\frac{(n-1)S^2}{\sigma^2}$. [1]

- (vi) True. The sample mean is always an unbiased estimator of the population mean, whatever the distribution of the parent population. [1]

- (vii) False. For many of the common distributions, the maximum likelihood estimator of the population mean *is* the sample mean, but this is not true in general. For example, the MLE of θ from the $U(0, 2\theta)$ distribution is $\hat{\theta} = \max x_i$ (based on a random sample x_1, x_2, \dots, x_n). [2]
- (viii) True. The variance of the sample mean always equals the population variance divided by the sample size. [1]
- (ix) False. The sample *variance* is always unbiased, but the sample *standard deviation* is not. In fact, for a normal distribution, $E[S] = \frac{\sqrt{2}}{\sqrt{n-1}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \sigma$. [2]
- (x) False. The sample *variance* is always unbiased, but the sample *standard deviation* is not. [1]
- (xi) False. In general, the sample mean and sample variance are *not* independent. [1]
- (xii) True. If the population is normal, the sample mean and sample variance *are* independent. This is a special property of normal populations. [2]

Solution 3.14

The minus sign in A should be a plus sign.

B would only be correct for an unbiased estimator.

C is the definition of the mean square error.

D would only be correct for an unbiased estimator.

The only correct answer is C. [2]

Solution 3.15

- B The population doesn't have to be normal, but the observations do have to be independent. [3]

For example, if you made the observations statistically dependent by discarding observations if they are closer to the mean than the previous observation, the observations would contain a disproportionate number of extreme values and S^2 would tend to overestimate the population variance.

Solution 3.16

- (i)(a) **Mean confidence interval (variance unknown)**

Using the result:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

a 95% confidence interval is given by:

$$\bar{X} \pm t_{7;0.025} \frac{\sigma}{\sqrt{n}} \quad [1]$$

From the *Tables*, we know that $P(t_7 > 2.365) = 0.025$, and from the question we have:

$$\bar{x} = \frac{576.1}{8} = 72.0125 \quad [\frac{1}{2}]$$

$$s^2 = \frac{1}{7} \left\{ 41,774.37 - 8 \times 72.0125^2 \right\} \approx 41.138 \quad [\frac{1}{2}]$$

So we get the confidence interval to be:

$$72.0125 \pm 2.365 \frac{\sqrt{41.138}}{\sqrt{8}} = (66.6, 77.4) \quad [1]$$

(i)(b) ***Mean confidence interval (variance known)***

Using the result:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

a 95% confidence interval is given by:

$$\bar{X} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} \quad [1]$$

From the *Tables*, we know that $P(Z > 1.960) = 0.025$, and from the question we have:

$$\bar{x} = \frac{576.1}{8} = 72.0125$$

So we get the confidence interval to be:

$$72.0125 \pm 1.960 \frac{6}{\sqrt{8}} = (67.9, 76.2) \quad [1]$$

(ii) ***Standard deviation confidence interval***

Using the result:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

a 90% confidence interval for the variance is given by:

$$\left(\frac{7 \times 41.138}{\chi^2_{7;0.05}}, \frac{7 \times 41.138}{\chi^2_{7;0.95}} \right) = \left(\frac{287.97}{14.07}, \frac{287.97}{2.167} \right) = (20.47, 132.9) \quad [2]$$

So a 90% confidence for the standard deviation is $(4.52, 11.5)$. [1]

Solution 3.17

The formulae for the mean and variance of a lognormal distribution are:

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

Since the standard deviation equals 1.8 times the mean, we know that:

$$e^{\mu + \frac{1}{2}\sigma^2} (e^{\sigma^2} - 1)^{\frac{1}{2}} = 1.8e^{\mu + \frac{1}{2}\sigma^2} \quad [1]$$

So:

$$(e^{\sigma^2} - 1)^{\frac{1}{2}} = 1.8 \quad \Rightarrow \quad \sigma^2 = 1.4446 \quad [1]$$

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:

$$7.632 < \mu < 7.744 \quad [1]$$

Solution 3.18(i) ***Sample size needed (unknown variance)***

Using the pivotal quantity $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$, gives a 95% confidence interval of:

$$\bar{x} \pm t_{n-1;0.025} \frac{s}{\sqrt{n}}$$

The width of this confidence interval is $2 \times t_{n-1;0.025} \frac{s}{\sqrt{n}}$, so we require:

$$2 \times t_{n-1;0.025} \frac{8.4}{\sqrt{n}} < 10 \Rightarrow \frac{t_{n-1;0.025}}{\sqrt{n}} < 0.5952$$

Using trial and improvement, we get:

$$\frac{t_{13;0.05}}{\sqrt{14}} = \frac{2.160}{\sqrt{14}} = 0.5773$$

Therefore we need a sample size of at least 14 individuals.

(ii) ***Sample size needed (known variance)***

Using the pivotal quantity of $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, gives a 95% confidence interval of:

$$\mu = \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

The width of this confidence interval is $2 \times 1.96 \frac{\sigma}{\sqrt{n}}$, so we require:

$$2 \times 1.96 \frac{8.4}{\sqrt{n}} < 10 \Rightarrow 3.29 < \sqrt{n} \Rightarrow n > 10.8$$

Therefore we need a sample size of at least 11 individuals.

Solution 3.19

This is a Type 2 geometric distribution, which has $E(X) = \frac{1-\theta}{\theta}$.

The method of moments estimator is found by:

$$E(X) = \frac{1-\hat{\theta}}{\hat{\theta}} = \bar{X} \Rightarrow 1-\hat{\theta} = \hat{\theta}\bar{X} \Rightarrow \hat{\theta}(1+\bar{X}) = 1 \Rightarrow \hat{\theta} = \frac{1}{1+\bar{X}} \quad [2]$$

Note that since it's an estimator (rather than an estimate) we use capital letters.

The likelihood function is:

$$L(\theta) = \prod_{i=1}^n \theta(1-\theta)^{x_i} = \theta^n (1-\theta)^{\sum x} \quad [1]$$

Taking logs and differentiating gives:

$$\ln L(\theta) = n \ln \theta + \left(\sum x \right) \ln(1-\theta)$$

$$\Rightarrow \frac{d}{d\theta} \ln L(\theta) = \frac{n}{\theta} - \frac{\sum x}{1-\theta}$$

Setting this equal to zero in order to find the maximum gives:

$$\frac{n}{\hat{\theta}} - \frac{\sum x}{1-\hat{\theta}} = 0 \Rightarrow \frac{n}{\hat{\theta}} = \frac{\sum x}{1-\hat{\theta}} \Rightarrow n - n\hat{\theta} = \hat{\theta}\sum x \Rightarrow 1 = \hat{\theta}(1+\bar{X})$$

And hence $\hat{\theta} = \frac{1}{1+\bar{X}}$ as before. [2]

Solution 3.20(i) **Range of values**

Since $0 \leq P(X = x) \leq 1$, using this for each of the probabilities gives lower bounds for α of $-\frac{1}{16}, -\frac{1}{6}$ and $-\frac{3}{8}$. Hence, $\alpha \geq -\frac{1}{16}$. We also obtain upper bounds for α of $\frac{7}{16}, \frac{1}{6}$ and $\frac{5}{8}$. Hence, $\alpha \leq \frac{1}{6}$. [1]

(ii) **Method of moments estimator**

We have one unknown, so we will use $E(X) = \bar{x}$.

$$E(X) = 2\left(\frac{1}{8} + 2\alpha\right) + 4\left(\frac{1}{2} - 3\alpha\right) + 5\left(\frac{3}{8} + \alpha\right) = \frac{33}{8} - 3\alpha \quad [1]$$

From the data, we have:

$$\bar{x} = \frac{7 \times 2 + 6 \times 4 + 17 \times 5}{30} = \frac{123}{30} = 4.1 \quad [1]$$

Therefore:

$$\frac{33}{8} - 3\hat{\alpha} = 4.1 \Rightarrow \hat{\alpha} = 0.008\dot{3} \quad [1]$$

Note that this value lies between the limits derived in part (i).

(iii) **Maximum likelihood**

The likelihood of obtaining the observed results is:

$$L(\alpha) = \text{constant} \times \left(\frac{1}{8} + 2\alpha\right)^7 \times \left(\frac{1}{2} - 3\alpha\right)^6 \times \left(\frac{3}{8} + \alpha\right)^{17} \quad [1]$$

Takes logs and differentiating gives:

$$\begin{aligned} \Rightarrow \ln L(\alpha) &= \text{constant} + 7 \ln\left(\frac{1}{8} + 2\alpha\right) + 6 \ln\left(\frac{1}{2} - 3\alpha\right) + 17 \ln\left(\frac{3}{8} + \alpha\right) \\ \Rightarrow \frac{d}{d\theta} \ln L(\alpha) &= \frac{14}{\frac{1}{8} + 2\alpha} - \frac{18}{\frac{1}{2} - 3\alpha} + \frac{17}{\frac{3}{8} + \alpha} \end{aligned} \quad [2]$$

Equating this to zero to find the maximum value of θ gives:

$$\begin{aligned} \frac{14}{\frac{1}{8}+2\hat{\alpha}} - \frac{18}{\frac{1}{2}-3\hat{\alpha}} + \frac{17}{\frac{3}{8}+\hat{\alpha}} &= 0 \\ \Rightarrow 14\left(\frac{1}{2}-3\hat{\alpha}\right)\left(\frac{3}{8}+\hat{\alpha}\right) - 18\left(\frac{1}{8}+2\hat{\alpha}\right)\left(\frac{3}{8}+\hat{\alpha}\right) + 17\left(\frac{1}{8}+2\hat{\alpha}\right)\left(\frac{1}{2}-3\hat{\alpha}\right) &= 0 \\ \Rightarrow 14\left(\frac{3}{16}-\frac{5}{8}\hat{\alpha}-3\hat{\alpha}^2\right) - 18\left(\frac{3}{64}+\frac{7}{8}\hat{\alpha}+2\hat{\alpha}^2\right) + 17\left(\frac{1}{16}+\frac{5}{8}\hat{\alpha}-6\hat{\alpha}^2\right) &= 0 \\ \Rightarrow 180\hat{\alpha}^2 + \frac{111}{8}\hat{\alpha} - \frac{91}{32} &= 0 \end{aligned} \quad [2]$$

(iv) **MLE**

Solving the quadratic equation gives:

$$\hat{\alpha} = \frac{-\frac{111}{8} \pm \sqrt{\left(\frac{111}{8}\right)^2 - 4 \times 180 \times -\frac{91}{32}}}{360} = -0.170, 0.0929 \quad [1]$$

The maximum likelihood estimate is 0.0929. [1]

The other solution of -0.170 does not lie between the bounds calculated in (i). It is not feasible as it gives a probability of $P(X = 2) = -0.215$. [1]

Solution 3.21

(i) ***Expected results using method of moments (Poisson)***

The sample mean for the data is:

$$\frac{1}{100,000} \times (87,889 \times 0 + 11,000 \times 1 + 1,000 \times 2 + \dots) = 0.13345$$

The mean of a Poisson distribution with parameter λ is λ .

So the method of moments estimator of λ is 0.13345. [1]

The expected numbers, based on this estimate, are (using the iterative formula):

$$\begin{aligned}
 x = 0 : & 100,000 e^{-0.13345} = 87,507 \\
 x = 1 : & 0.13345 \times 87,507 = 11,678 \\
 x = 2 : & 0.13345 / 2 \times 11,678 = 779 \\
 x = 3 : & 0.13345 / 3 \times 779 = 35 \\
 x = 4 : & 0.13345 / 4 \times 35 = 1 \\
 x = 5 : & 0.13345 / 5 \times 1 = 0 \\
 x \geq 6 : & 100,000 - 87,507 - 11,678 - 779 - 35 - 1 - 0 = 0
 \end{aligned} \tag{2}$$

(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:

$$\begin{aligned}
 L(\lambda) &= \frac{n!}{n_0! n_1! n_2! \dots} (e^{-\lambda})^{n_0} (\lambda e^{-\lambda})^{n_1} (\lambda^2 e^{-\lambda})^{n_2} \dots \\
 &= \text{constant} \times \lambda^{n_1 + 2n_2 + 3n_3 + \dots} e^{-\lambda(n_0 + n_1 + n_2 + \dots)} \\
 &= \text{constant} \times \lambda^{13,345} e^{-100,000\lambda}
 \end{aligned} \tag{1}$$

So the log likelihood is:

$$\log L(\lambda) = 13,345 \log \lambda - 100,000\lambda + \text{constant} \tag{1}$$

Differentiating with respect to λ to maximise this:

$$\frac{d}{d\lambda} \log L(\lambda) = \frac{13,345}{\lambda} - 100,000 \tag{1}$$

This is zero when:

$$\lambda = 13,345 / 100,000 = 0.13345 \tag{1}$$

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:

$$\frac{1}{100,000} \times (87,889 \times 0^2 + 11,000 \times 1^2 + 1,000 \times 2^2 + \dots) = 0.16085$$

The mean and second non-central moment of the negative binomial distribution with parameters k and p are $\frac{kq}{p}$ and $\frac{kq}{p^2} + \left(\frac{kq}{p}\right)^2$.

So the method of moments estimators of k and p satisfy the equations:

$$\begin{aligned} \frac{kq}{p} &= 0.13345 \\ \text{and } \frac{kq}{p^2} + \left(\frac{kq}{p}\right)^2 &= 0.16085 \end{aligned} \quad [2]$$

From the second equation:

$$\frac{kq}{p^2} = 0.16085 - \left(\frac{kq}{p}\right)^2 = 0.16085 - (0.13345)^2 = 0.14304 \quad [\frac{1}{2}]$$

Dividing this into the first equation gives:

$$\begin{aligned} p &= 0.13345 / 0.14304 = 0.93295 & [\frac{1}{2}] \\ q &= 1 - p = 1 - 0.93295 = 0.06705 & [\frac{1}{2}] \\ \text{and } k &= 0.13345 \times 0.93295 / 0.06705 = 1.8569 & [\frac{1}{2}] \end{aligned}$$

The expected numbers, based on these estimates, are:

$$\begin{aligned} x = 0 : \quad 100,000(0.93295)^{1.8569} &= 87,909 \\ x = 1 : \quad 1.8569 / 1 \times 0.06705 \times 87,909 &= 10,945 \\ x = 2 : \quad 2.8569 / 2 \times 0.06705 \times 10,945 &= 1,048 \\ x = 3 : \quad 3.8569 / 3 \times 0.06705 \times 1,048 &= 90 \\ x = 4 : \quad 4.8569 / 4 \times 0.06705 \times 90 &= 7 \\ x = 5 : \quad 5.8569 / 5 \times 0.06705 \times 7 &= 1 \\ x \geq 6 : \quad 100,000 - 87,909 - 10,945 - 1,048 - 90 - 7 - 1 &= 0 \end{aligned} \quad [2]$$

(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.

The negative binomial distribution has more flexibility and can accommodate different values for the mean and variance (provided the variance exceeds the mean). [2]

Solution 3.22(i) ***MLE***

For a $Poi(\mu)$ distribution:

$$P(X = 0) = e^{-\mu}$$

$$P(X = 1) = \mu e^{-\mu}$$

$$\therefore P(X \geq 2) = 1 - \mu e^{-\mu} - e^{-\mu} \quad [1]$$

So the likelihood function for these observations is:

$$\begin{aligned} L(\mu) &= (e^{-\mu})^{n_0} \times (\mu e^{-\mu})^{n_1} \times (1 - \mu e^{-\mu} - e^{-\mu})^{n-n_0-n_1} \times const \\ &= \mu^{n_1} e^{-\mu(n_0+n_1)} \times (1 - \mu e^{-\mu} - e^{-\mu})^{n-n_0-n_1} \times const \end{aligned} \quad [2]$$

Taking logs:

$$\log L = n_1 \log \mu - (n_0 + n_1)\mu + (n - n_0 - n_1) \log(1 - \mu e^{-\mu} - e^{-\mu}) + const$$

Differentiating this expression with respect to μ :

$$\frac{d}{d\mu} \log L = \frac{n_1}{\mu} - (n_0 + n_1) + (n - n_0 - n_1) \times \frac{\mu e^{-\mu}}{1 - \mu e^{-\mu} - e^{-\mu}} \quad [2]$$

Setting this equal to zero, and multiplying through by $\mu(1 - \mu e^{-\mu} - e^{-\mu})$:

$$\begin{aligned} n_1(1 - \mu e^{-\mu} - e^{-\mu}) - (n_0 + n_1)\mu(1 - \mu e^{-\mu} - e^{-\mu}) \\ + \mu^2 e^{-\mu}(n - n_0 - n_1) = 0 \end{aligned} \quad [2]$$

Multiplying out brackets and cancelling terms:

$$n_1 - n_1 e^{-\mu} - n_0 \mu + n_0 \mu e^{-\mu} - n_1 \mu + n \mu^2 e^{-\mu} = 0 \quad [2]$$

Simplifying gives the equation satisfied by the MLE $\hat{\mu}$:

$$n_1 - (n_0 + n_1)\hat{\mu} + e^{-\hat{\mu}}(n\hat{\mu}^2 + n_0\hat{\mu} - n_1) = 0 \quad [1]$$

(ii) **Determine estimate**

We found that 15 out of 20 values in our sample gave a value of 0 or 1. So we can look in the *Tables* to see what parameter value gives $P(X \leq 1) = 0.75$. This suggests a value of 1 for the parameter (which gives $P(X \leq 1) = 0.73576$). [2]

Alternatively, to get an initial approximation for $\hat{\mu}$, we can equate the theoretical probability of no claims occurring *ie* $e^{-\mu}$ with the observed proportion $\frac{8}{20}$. This implies that $\hat{\mu}$ is approximately 0.916.

Using trial and error, starting with $\hat{\mu} = 1$, to find a value of $\hat{\mu}$ which makes the LHS equal to zero, we find:

μ	$f(\mu)$
1	-0.27453
0.9	0.16774
0.95	-0.03728
0.945	-0.01533
0.935	0.02761

where $f(\mu) = 7 - 15\mu + e^{-\mu}(20\mu^2 + 8\mu - 7)$. [3]

So the value of $\hat{\mu}$ lies between 0.935 and 0.945, *ie* $\hat{\mu} = 0.94$ to two decimal places. [2]

Solution 3.23(i) **MLE**

To find the maximum likelihood estimate we need to differentiate the log-likelihood function and set it equal to zero:

$$\begin{aligned}
 L(\alpha) &= \prod_{i=1}^n \alpha 250^\alpha (250 + x_i)^{-\alpha-1} \\
 &= \alpha^n 250^{n\alpha} \prod_{i=1}^n (250 + x_i)^{-\alpha-1} \\
 &= \alpha^n 250^{n\alpha} \left[\prod_{i=1}^n (250 + x_i) \right]^{-\alpha-1} \quad [1]
 \end{aligned}$$

Taking logs:

$$\begin{aligned}
 \ln L(\alpha) &= n \ln \alpha + n \alpha \ln 250 + (-\alpha - 1) \ln \left[\prod_{i=1}^n (250 + x_i) \right] \\
 &= n \ln \alpha + n \alpha \ln 250 + (-\alpha - 1) \left[\sum_{i=1}^n \ln(250 + x_i) \right] \quad [1]
 \end{aligned}$$

Differentiating:

$$\frac{d}{d\alpha} \ln L(\alpha) = \frac{n}{\alpha} + n \ln 250 - \sum_{i=1}^n \ln(250 + x_i) \quad [1]$$

Setting this equal to zero in order to find the maximum gives:

$$\begin{aligned}
 0 &= \frac{n}{\hat{\alpha}} + n \ln 250 - \sum \ln(250 + x_i) \\
 \frac{n}{\hat{\alpha}} &= \sum \ln(250 + x_i) - n \ln 250 \\
 \hat{\alpha} &= \frac{n}{\sum \ln(250 + x_i) - n \ln 250} \quad [1]
 \end{aligned}$$

(ii) ***Asymptotic confidence interval***

The value of $\hat{\alpha}$ for this sample is:

$$\hat{\alpha} = \frac{150}{869 - 150 \ln 250} = 3.68 \quad [1]$$

Asymptotically:

$$\hat{\alpha} \stackrel{d}{\sim} N(\alpha, CRLB)$$

Hence:

$$\frac{\hat{\alpha} - \alpha}{\sqrt{CRLB}} \stackrel{d}{\sim} N(0,1)$$

Now, the $CRLB$ is given by the equation:

$$CRLB = \frac{1}{-E\left[\frac{d^2}{d\alpha^2} \ln L\right]} \quad [1]$$

Now:

$$\frac{d^2}{d\alpha^2} \ln L(\alpha) = -\frac{150}{\alpha^2}$$

So:

$$CRLB = \frac{1}{E\left[\frac{150}{\alpha^2}\right]} = \frac{1}{\frac{150}{\hat{\alpha}^2}} = \frac{\hat{\alpha}^2}{150} = 0.090194 \quad [1]$$

So the confidence interval is given by:

$$\hat{\alpha} \pm 1.96\sqrt{CRLB} = 3.68 \pm 1.96\sqrt{0.090194} = (3.09, 4.27) \quad [2]$$

Solution 3.24(i) **MLE**

The likelihood function has the form:

$$\begin{aligned} L(\alpha) &= f(x_1; \alpha)f(x_2; \alpha)\dots f(x_{100}; \alpha) \\ &= \alpha^{100}(\alpha+1)^{100}1,000^{100\alpha} \prod x_i \prod (1,000+x_i)^{-\alpha-2} \end{aligned} \quad [1]$$

Taking logs, we obtain:

$$\begin{aligned} \log L &= 100\log\alpha + 100\log(\alpha+1) + 100\alpha\log 1,000 \\ &\quad + \sum \log x_i - (\alpha+2)\sum \log(1,000+x_i) \end{aligned} \quad [1]$$

Differentiating with respect to α , we obtain:

$$\frac{d}{d\alpha} \log L = \frac{100}{\alpha} + \frac{100}{\alpha+1} + 100\log 1,000 - \sum \log(1,000+x_i) \quad [1]$$

Setting this to zero and substituting the observed values:

$$\frac{100}{\alpha} + \frac{100}{\alpha+1} + 100\log 1,000 - 750 = 0 \quad [\frac{1}{2}]$$

Multiplying through by $\alpha(\alpha+1)$, dividing by 100 and collecting terms gives the quadratic equation:

$$0.59224\alpha^2 - 1.40776\alpha - 1 = 0 \quad [\frac{1}{2}]$$

Solving this using the quadratic equation formula gives $\alpha = 2.949$ or $\alpha = -0.572$.

Since α must be positive, the maximum likelihood estimate must correspond to the solution $\hat{\alpha} = 2.949$. [1]

Checking that this does give a maximum:

$$\frac{d^2}{d\alpha^2} \log L = -\frac{100}{\alpha^2} - \frac{100}{(\alpha+1)^2} \quad [1]$$

which is always negative. So this is a maximum value for the likelihood function.

(ii) ***Asymptotic standard deviation***

Asymptotically, MLEs attain the Cramér-Rao lower bound. So here, with a sample size of 100, we should obtain a good approximation to the variance of the estimator.

The *CRLB* is the negative of the reciprocal of the expected value of the second derivative of the log-likelihood:

$$CRLB = \frac{1}{100 \left[\frac{1}{\alpha^2} + \frac{1}{(\alpha+1)^2} \right]} = 0.05583 \approx 0.2363^2 \quad [2]$$

using the MLE $\hat{\alpha} = 2.949$ as an approximation for α .

So the standard deviation of $\hat{\alpha}$ is approximately 0.236. [1]

Solution 3.25(i)(a) ***MLE (Poisson)***

The likelihood function is:

$$L(\mu) = \frac{e^{-\mu} \mu^{X_1}}{X_1!} \times \cdots \times \frac{e^{-\mu} \mu^{X_n}}{X_n!} = \frac{e^{-n\mu} \mu^{\sum X_i}}{\prod X_i!} \quad [1]$$

Taking logs:

$$\log L = -n\mu + \sum X_i \log \mu - \text{constant} \quad [\frac{1}{2}]$$

Differentiating this with respect to μ :

$$\frac{d}{d\mu} (\log L) = -n + \frac{\sum X_i}{\mu} \quad [\frac{1}{2}]$$

Setting this equal to zero and rearranging, we get $\hat{\mu} = \frac{\sum x_i}{n} = \bar{X}$ as required. [1]

Checking the second derivative:

$$\frac{d^2}{d\mu^2}(\log L) = -\frac{\sum X_i}{\mu^2} < 0 \Rightarrow \max [1]$$

(i)(b) ***Derive distribution of sum***

The MGF for each individual X_i is $e^{\mu(e^t-1)}$. Since the X values form a random sample, and hence are independent, we can just multiply the MGFs together:

$$M_{\sum X_i}(t) = e^{\mu(e^t-1)} \times \dots \times e^{\mu(e^t-1)} = e^{n\mu(e^t-1)} [1]$$

We recognise this as the MGF of a *Poisson*($n\mu$) distribution. Hence, using the uniqueness of MGF's, $\sum X_i$ has this distribution. [1]

(ii)(a) ***Exact confidence interval***

The mean annual claim rate is $\hat{\mu} = 0.2$. The number of claims in a sample of 10 policies has a *Poisson*(10μ) distribution. Setting $\nu = 10\mu$, we want the value of ν such that:

- (1) the probability of getting 2 or more claims is only 0.025. This will give us the lower end of the confidence interval
- (2) the probability of getting 2 or fewer claims is only 0.025. This will give us the upper end of the confidence interval.

But the probability of getting 2 or more claims in a *Poisson*(ν) distribution is $1 - e^{-\nu} - \nu e^{-\nu} = 1 - (\nu + 1)e^{-\nu}$, and we want this to be equal to 0.025. So solving the equation $(\nu + 1)e^{-\nu} = 0.975$ by trial and error (or a more sophisticated method), we obtain $\nu = 0.242$, which gives $\mu = 0.0242$. [3]

Similarly for the upper end of our confidence interval, the probability of getting 2 or fewer claims is $e^{-\nu} (1 + \nu + \frac{1}{2}\nu^2)$, and setting this equal to 0.025 and solving, we obtain $\nu = 7.22$, which gives $\mu = 0.722$. So the 95% confidence interval for μ is (0.0242, 0.722). [3]

(ii)(b) ***Approximate confidence interval***

The estimate for the mean annual claim rate is now:

$$\hat{\mu} = \frac{106}{500} = 0.212 \quad [1]$$

Now using the normal approximation, we have that $\frac{\bar{X} - \mu}{\sqrt{\mu/n}}$ has a $N(0,1)$ distribution.

So:

$$0.95 = P\left(-1.96 < \frac{\bar{X} - \mu}{\sqrt{\mu/n}} < 1.96\right) \quad [1]$$

If we set $\hat{\mu} = \bar{X} = 0.212$ in the denominator, we have a confidence interval of $\bar{X} \pm 1.96\sqrt{\frac{0.212}{500}} = (0.172, 0.252)$. [1]

(ii)(c) ***Compare confidence intervals***

The large sample gives a much narrower confidence interval, as might be expected, since with a large sample we can predict the value of μ with greater certainty. [2]

Solution 3.26(i) **Distribution**

Using the result given on page 22 of the *Tables*:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2 \quad [1]$$

(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:

$$E\left[\frac{\sum(X_i - \bar{X})^2}{\sigma^2}\right] = E\left[\chi_{n-1}^2\right] = (n-1) \quad [\frac{1}{2}]$$

Since $(n+b)\hat{\sigma}^2 = \sum(X_i - \bar{X})^2$, we have:

$$\begin{aligned} E\left[\frac{(n+b)\hat{\sigma}^2}{\sigma^2}\right] &= (n-1) \\ \frac{(n+b)}{\sigma^2} E[\hat{\sigma}^2] &= (n-1) \\ E[\hat{\sigma}^2] &= \frac{(n-1)}{(n+b)} \sigma^2 \end{aligned} \quad [1]$$

Therefore the bias is given by:

$$bias(\hat{\sigma}^2) = \frac{(n-1)}{(n+b)} \sigma^2 - \sigma^2 = -\frac{(1+b)}{(n+b)} \sigma^2 \quad [\frac{1}{2}]$$

(ii)(b) **Unbiased**

Substituting $b = -1$ into the bias gives:

$$bias(\hat{\sigma}^2) = -\frac{(1-1)}{(n-1)} \sigma^2 = 0 \quad [1]$$

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 $MSE(\hat{\sigma}^2) = \text{var}(\hat{\sigma}^2) + \text{bias}^2(\hat{\sigma}^2)$. From part (i) we have:

$$\text{var}\left[\frac{\sum(X_i - \bar{X})^2}{\sigma^2}\right] = \text{var}\left[\chi_{n-1}^2\right] = 2(n-1) \quad [1]$$

Since $(n+b)\hat{\sigma}^2 = \sum(X_i - \bar{X})^2$, we have:

$$\begin{aligned} \text{var}\left[\frac{(n+b)\hat{\sigma}^2}{\sigma^2}\right] &= 2(n-1) \\ \frac{(n+b)^2}{\sigma^4} \text{var}[\hat{\sigma}^2] &= 2(n-1) \\ \text{var}[\hat{\sigma}^2] &= \frac{2(n-1)}{(n+b)^2} \sigma^4 \end{aligned} \quad [1]$$

Using this and the bias from (ii)(a), the mean square error is given by:

$$\begin{aligned} MSE(\hat{\sigma}^2) &= \frac{2(n-1)}{(n+b)^2} \sigma^4 + \frac{(1+b)^2}{(n+b)^2} \sigma^4 \\ &= \frac{2(n-1) + (1+b)^2}{(n+b)^2} \sigma^4 \end{aligned} \quad [1]$$

(iii)(b) **Consistent**

As $n \rightarrow \infty$, the mean square error becomes:

$$MSE(\hat{\sigma}^2) \rightarrow \frac{2}{n} \sigma^4 \rightarrow 0$$

So $\hat{\sigma}^2$ is consistent. [1]

(iii)(c) **Minimum mean square error**

Differentiating with respect to b using the quotient rule gives:

$$\frac{d}{db} MSE(\hat{\sigma}^2) = \frac{2(1+b)(n+b)^2 - [2(n-1)+(1+b)^2] \times 2(n+b)}{(n+b)^4} \sigma^4 \quad [2]$$

Substituting $b=1$ into this expression gives:

$$\begin{aligned} \left. \frac{d}{db} MSE(\hat{\sigma}^2) \right|_{b=1} &= \frac{2 \times 2(n+1)^2 - [2(n-1)+4] \times 2(n+1)}{(n+1)^4} \sigma^4 \\ &= \frac{4(n+1)^2 - 4(n+1)^2}{(n+1)^4} \sigma^4 \\ &= 0 \end{aligned} \quad [1]$$

So the MSE is minimised when $b=1$.

Alternatively, students may attempt to find the value of b that makes this zero as follows:

$$\begin{aligned} 2(1+b)(n+b)^2 &= [2(n-1)+(1+b)^2] \times 2(n+b) \\ (1+b)(n+b) &= [2(n-1)+(1+b)^2] \\ n+b+bn+b^2 &= 2n-1+2b+b^2 \\ b(n-1) &= n-1 \\ b &= 1 \end{aligned}$$

(iv) **Best estimator**

All values of b give consistent estimators. However, when $b=-1$, the estimator $\hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ is unbiased. Whereas when $b=1$, the estimator $\hat{\sigma}^2 = \frac{1}{n+1} \sum (X_i - \bar{X})^2$ 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]

Solution 3.27

Using the result:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

We have:

$$\bar{x} = \frac{104.2}{7} \simeq 14.88571 \quad [1/2]$$

$$s^2 = \frac{1}{6} \left\{ 1,581.98 - 7 \times 14.88571^2 \right\} \simeq 5.148 \quad [1/2]$$

So a 95% one-sided confidence interval for the variance is given by:

$$\left(\frac{6 \times 5.148}{\chi_{6;0.05}^2}, \infty \right) = \left(\frac{30.888}{12.59}, \infty \right) = (2.45, \infty) \quad [2]$$

So a 95% one-sided confidence interval for the standard deviation is $(1.57, \infty)$. [1]

Solution 3.28

(i) ***Confidence interval for mean sum assured***

We know that $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ has a t_{n-1} distribution. So using a t_{50} distribution, a symmetrical 95% confidence interval for μ is given by:

$$0.95 = P \left(-2.009 < \frac{\bar{X} - \mu}{S/\sqrt{51}} < 2.009 \right) \quad [1]$$

Rearranging the inequality gives:

$$180,000 - 2.009 \frac{60,000}{\sqrt{51}} < \mu < 180,000 + 2.009 \frac{60,000}{\sqrt{51}} \quad [1]$$

So the CI for μ (to 3 SF) is $(163000, 197000)$. [1]

(ii) ***Confidence interval for percentage***

Let X be the number of policyholders with a sum assured greater than £250,000.

$$X \sim Bin(51, p) \doteq N(51p, 51p(1-p)) \quad \text{or} \quad \hat{p} \doteq N\left(p, \frac{p(1-p)}{51}\right)$$

Estimating p in the variance as $\hat{p} = \frac{8}{51}$, we have:

$$Z = \frac{8 - 51p}{\sqrt{51 \times \frac{8}{51} \times \frac{43}{51}}} \quad \text{or} \quad Z = \frac{\frac{8}{51} - p}{\sqrt{\left(\frac{8}{51} \times \frac{43}{51}\right)/51}} \quad [1]$$

A symmetrical 95% confidence interval for p is given by:

$$0.95 = P(-1.960 < Z < 1.960)$$

So:

$$\frac{8 - 1.960\sqrt{51 \times \frac{8}{51} \times \frac{43}{51}}}{51} < p < \frac{8 + 1.960\sqrt{51 \times \frac{8}{51} \times \frac{43}{51}}}{51}$$

$$\text{or } \frac{8}{51} - 1.960\sqrt{\left(\frac{8}{51} \times \frac{43}{51}\right)/51} < p < \frac{8}{51} + 1.960\sqrt{\left(\frac{8}{51} \times \frac{43}{51}\right)/51} \quad [1]$$

Hence, our confidence interval for p is (0.0571, 0.257). [1]

Solution 3.29(i) **Sample size for given width**

If n is large enough to use a normal approximation, then the pivotal quantity is:

$$\frac{X - np}{\sqrt{np(1-p)}} \stackrel{\text{d}}{\sim} N(0,1) \quad \text{or} \quad \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \stackrel{\text{d}}{\sim} N(0,1)$$

where X is the number of success out of a sample size of n . Hence a 95% confidence interval can be obtained for p from:

$$\frac{X \pm 1.96\sqrt{np(1-p)}}{n} \quad \text{or} \quad \frac{X}{n} \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

where $\hat{p} = \frac{x}{n}$ with x being the number of observed successes in the sample of size n .

Hence, we require:

$$\frac{1.96\sqrt{np(1-p)}}{n} \leq 0.05 \quad \text{or} \quad 1.96\sqrt{\frac{p(1-p)}{n}} \leq 0.05 \quad [1]$$

Since n is constant it is the $\hat{p}(1-\hat{p})$ which determines the width. The greatest value this can take is when $\hat{p} = 0.5$. [1]

This can be shown by differentiation:

$$\begin{aligned} \text{Let } f(\hat{p}) &= \hat{p}(1-\hat{p}) \quad \Rightarrow \quad \frac{d}{d\hat{p}} f(\hat{p}) = 1 - 2\hat{p} = 0 \quad \Rightarrow \quad \hat{p} = 0.5 \\ &\Rightarrow \quad \frac{d^2}{d\hat{p}^2} f(\hat{p}) = -2 < 0 \quad \Rightarrow \quad \text{max} \end{aligned}$$

Substituting $\hat{p} = 0.5$ in, we get:

$$1.96\sqrt{\frac{0.5 \times 0.5}{n}} \leq 0.05 \quad \Rightarrow \quad \sqrt{n} \geq 19.6 \quad \Rightarrow \quad n \geq 384.16 \quad [1]$$

Hence, the minimum sample size required is 385. [1]

(ii) ***Proportion confidence interval***

Recall from part (i) that 95% confidence intervals were given by:

$$\frac{X \pm 1.96\sqrt{n\hat{p}(1-\hat{p})}}{n} \quad \text{or} \quad \frac{X}{n} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For the “yes” responses we have $\hat{p}_Y = 0.3$, $X_Y = 300$ and $n = 1,000$. Hence the confidence interval for the proportion of “Yes” answers is:

$$\frac{300 \pm 1.96\sqrt{1,000 \times 0.3 \times 0.7}}{1,000} \quad \text{or} \quad \frac{300}{1,000} \pm 1.96\sqrt{\frac{0.3 \times 0.7}{1,000}}$$

which gives:

$$(0.272, 0.328)$$

Similarly, for the “No” responses we have $\hat{p}_N = 0.5$, $X_N = 500$ and $n = 1,000$ which gives a confidence interval of:

$$(0.469, 0.531)$$

Finally, for the “Don’t know” responses we have $\hat{p}_D = 0.2$, $X_D = 200$ and $n = 1,000$ which gives a confidence interval of:

$$(0.175, 0.225)$$

(iii)(a) ***Difference between proportions confidence interval***

Let p_1 be the proportion responding “yes” before the campaign and let p_2 be the proportion responding “yes” afterwards.

The pivotal quantity is:

$$\Rightarrow \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \sim N(0,1) \quad [1]$$

So a 90% confidence interval for $p_2 - p_1$ is:

$$(\hat{p}_2 - \hat{p}_1) \pm 1.6449 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

We have $\hat{p}_1 = 0.3$, $\hat{p}_2 = 0.35$, $n_1 = 1,000$ and $n_2 = 800$. This gives:

$$(0.35 - 0.3) \pm 1.6449 \sqrt{\frac{0.3 \times 0.7}{1,000} + \frac{0.35 \times 0.65}{800}} = (0.0134, 0.0866) \quad [2]$$

Alternatively, the confidence interval for $p_1 - p_2$ is $(-0.0866, -0.0134)$.

(iii)(b) **Comment**

The interval does not contain zero, so there is a significant difference between those favouring closer ties before and after the campaign. [1]

Solution 3.30(i)(a) **Mean difference confidence interval**

Using the subscript 1 to refer to “without screening”. The pivotal quantity is:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \quad [1]$$

Calculating the required values:

$$\bar{x}_1 = \frac{257.4}{10} = 25.74 \quad \bar{x}_2 = \frac{200.7}{10} = 20.07 \quad [1]$$

$$s_1^2 = \frac{1}{9} \left\{ 6,951.16 - 10 \times 25.74^2 \right\} = 36.1871 \quad [\frac{1}{2}]$$

$$s_2^2 = \frac{1}{9} \left\{ 4,553.97 - 10 \times 20.07^2 \right\} = 58.4357 \quad [\frac{1}{2}]$$

The pooled variance is given by:

$$s_P^2 = \frac{1}{18} (9 \times 36.1871 + 9 \times 58.4357) = 47.3114 \quad [1]$$

Hence, a 95% confidence interval is given by:

$$(25.74 - 20.07) \pm 2.101 \sqrt{47.3114} \sqrt{\frac{2}{10}} = (-0.793, 12.1)$$

So the confidence interval is $(-0.793, 12.1)$. [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, we cannot say that the new screening programme significantly reduces the mean claim amount. [1]

(ii)(a) ***Ratio of variances confidence interval***

The pivotal quantity is:

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1, n_2-1} \quad [1]$$

Hence, a 95% confidence interval is given by:

$$\frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; 0.025}} < \sigma_1^2/\sigma_2^2 < \frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; 0.975}}$$

Which gives:

$$\frac{0.6193}{4.026} < \sigma_A^2/\sigma_B^2 < \frac{0.6193}{1/4.026} \Rightarrow (0.154, 2.49) \quad [2]$$

Alternatively, the confidence interval for σ_B^2/σ_A^2 is (0.401, 6.50).

(ii)(b) ***Comment***

Since the confidence interval contains 1, this means we are reasonably confident that the population variances are the same. [1]

(iii) ***Sample size***

The width of the confidence interval is:

$$2 \times t_{2n-2; 2.5\%} \sqrt{\frac{2}{n} \sqrt{\frac{36.1871(n-1) + 58.4357(n-1)}{2n-2}}} = \frac{19.455 t_{2n-2; 2.5\%}}{\sqrt{n}} \quad [1]$$

This must be less than 10, so using trial and improvement gives:

$$n = 16 \Rightarrow \frac{19.455 \times 2.042}{\sqrt{16}} = 9.93 < 10$$

The minimum sample size is 16. [2]

Solution 3.31(i) ***Exact confidence interval***

We require:

$$P(X \geq 2) = 0.05 \text{ under } Poi(\mu_1) \quad [\frac{1}{2}]$$

$$P(X \leq 2) = 0.05 \text{ under } Poi(\mu_2) \quad [\frac{1}{2}]$$

From the first equation:

$$0.95 = P(X = 0) + P(X = 1) = e^{-\mu_1} + \mu_1 e^{-\mu_1} \quad [1]$$

Solving this numerically we obtain $\mu_1 = 0.36$. [1]

From the second equation:

$$0.05 = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= e^{-\mu_2} + \mu_2 e^{-\mu_2} + \frac{\mu_2^2}{2} e^{-\mu_2} \quad [1]$$

Solving this numerically we obtain $\mu_2 = 6.3$. [1]

So the confidence interval is $0.36 < \mu < 6.3$.

(ii) ***Approximate confidence interval***

Since n is large enough to use a normal approximation, then the pivotal quantity is:

$$\frac{\sum X - n\lambda}{\sqrt{n\lambda}} \stackrel{d}{\sim} N(0,1) \quad \text{or} \quad \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\hat{\lambda}}{n}}} \stackrel{d}{\sim} N(0,1)$$

where $\hat{\lambda} = \bar{X}$. Hence, a 90% confidence interval can be obtained for λ from:

$$\frac{\sum X \pm 1.6449\sqrt{n\hat{\lambda}}}{n} \quad \text{or} \quad \hat{\lambda} \pm 1.6449\sqrt{\frac{\hat{\lambda}}{n}} \quad [1]$$

Substituting in $n = 30$, $\sum x_i = 60$ and $\hat{\lambda} = \bar{x} = 2$ gives:

$$\frac{60 \pm 1.6449\sqrt{30 \times 2}}{30} \quad \text{or} \quad 2 \pm 1.6449\sqrt{\frac{2}{30}} \quad [1]$$

which gives:

$$(1.58, 2.42) \quad [1]$$

Solution 3.32

(i)(a) **Probability**

The probability of this event is:

$$1 - \left(\frac{5}{6}\right)^4 = 0.5177 \quad [1]$$

(i)(b)(1) **Equation for lower confidence limit**

If the probability of A occurring is p , then the number of occurrences in 10 throws will have a $\text{Bin}(10, p)$ distribution. For a lower 95% confidence interval, we require:

$$0.05 = P(\text{Bin}(10, p_L) \geq 8)$$

You might recall from the notes that the lower tail value comes from the higher probabilities – the reason for this is that we must have a smaller probability of success for there to be a small chance of lots of successes.

So the required equation for p_L is:

$$0.05 = \binom{10}{8} p_L^8 (1-p_L)^2 + \binom{10}{9} p_L^9 (1-p_L) + p_L^{10} \quad [2]$$

(i)(b)(2) **Verify solution**

Substituting $p_L = 0.493$ into the RHS of this equation gives 0.049935, which is very close to the value on the LHS. [1]

(i)(b)(3) ***Comment***

The lower limit of the confidence interval, 0.493, is very close to 0.5 so it appears that there is a 95% chance that $p > 0.5$. [2]

(ii)(a) ***Approximate value for lower limit***

Using a normal approximation, the number of times that A will occur has a $N(1000p, 1000p(1-p))$ distribution. So the lower 95% confidence interval for p satisfies:

$$0.05 = P(N[1000p_L, 1000p_L(1-p_L)] \geq Y)$$

which is equivalent to:

$$0.95 = P\left[N(0, 1) < \frac{Y - 1,000p_L}{\sqrt{1,000p_L(1-p_L)}}\right]$$

From the tables of the normal distribution, this requires:

$$\frac{Y - 1,000p_L}{\sqrt{1,000p_L(1-p_L)}} = 1.645 \quad [2]$$

If we replace p_L by $\frac{Y}{1,000}$ (in the denominator only) and rearrange, we obtain:

$$p_L \approx \frac{Y}{1,000} - \frac{1.645}{1,000} \sqrt{1,000 \times \frac{Y}{1,000} \times \left(1 - \frac{Y}{1,000}\right)} \quad [1]$$

which is a large sample expression for p_L .

(ii)(b) ***Determine Y***

If $p_L = \frac{1}{2}$, then we would expect Y to be close to $\frac{1}{2}$. So:

$$\frac{1}{2} \approx \frac{Y}{1,000} - \frac{1.645}{1,000} \sqrt{1,000 \times \frac{1}{2} \times \frac{1}{2}}$$

which corresponds to $Y = 526.01$. [2]

Trying values close to $Y = 526$ in the expression in (ii)(a) we find that 526 gives 0.500025, whereas 525 gives 0.499023. So Y would have to be at least 526. [1]

(ii)(c) ***Calculate probability***

p_L will be greater than $\frac{1}{2}$ if (and only if) $Y \geq 526$. This probability is approximately:

$$P[N(517.7, 249.69) \geq 525.5] = P[N(0,1) \geq 0.4936] = 0.311 \quad [2]$$

(iii)(a) ***Repeat (ii)(b) for 10,000 simulations***

For p_L to be greater than $\frac{1}{2}$ now, the critical value of Y satisfies:

$$\frac{1}{2} = \frac{Y}{10,000} - \frac{1.645}{10,000} \sqrt{10,000 \times \frac{1}{2} \times \frac{1}{2}} \quad [1]$$

which gives $Y = 5,082.25$. So Y would have to be at least 5,083. [2]

(iii)(b) ***Repeat (ii)(c) for 10,000 simulations***

We now have:

$$\begin{aligned} P(Y \geq 5083) &= P[N(5177, 2496.87) \geq 5082.5] \\ &= P[Z \geq -1.8912] = 0.971 \end{aligned} \quad [1]$$

For the experiment with 10,000 trials, the lower limit is far more likely to exceed $\frac{1}{2}$ than for the experiment with only 1,000 trials. As the number of trials is increased the confidence interval becomes narrower, and the probability that the CI includes $\frac{1}{2}$ decreases. Since the true value of p is close to $\frac{1}{2}$, the decision is marginal and we need a large number of trials before this happens. [2]

Part 4 – Questions

Note that the split between development questions and exam-style questions is somewhat subjective. For example, there have been past CT3 exam questions that test knowledge of the Core Reading, and so are similar to what we've included here as development questions. The exam-style questions involve more application and a wider range of ideas and are typically the more challenging questions in the exam.

1 Development questions

Question 4.1

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%, what is the conclusion for a test of size:

- (i) 10% [1]
 - (ii) 5%. [1]
- [Total 2]

Question 4.2

A random sample, x_1, \dots, x_{10} , from a normal population gives the following values:

$$9.5 \quad 18.2 \quad 4.69 \quad 3.76 \quad 14.2 \quad 17.13 \quad 15.69 \quad 13.9 \quad 15.7 \quad 7.42$$

$$\sum x_i = 120.19 \quad \sum x_i^2 = 1,693.6331$$

- (i) Test at the 5% level whether the mean of the whole population is 15 if the variance is:
 - (a) unknown
 - (b) 20. [5]
 - (ii) Test at the 5% level whether the population variance is 20. [3]
- [Total 8]

Question 4.3

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. Do the results of your experiment support the gambler’s claims about the probabilities when a coin is spun? [3]

Question 4.4

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. [4]

Question 4.5

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. [3]

Question 4.6

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 . [5]

Question 4.7

An analysis using the simple linear regression model based on 19 data points gave:

$$s_{xx} = 12.2 \quad s_{yy} = 10.6 \quad s_{xy} = 8.1$$

- (i) (a) Calculate $\hat{\beta}$.
 - (b) Test whether β is significantly different from zero. [4]
 - (ii) (a) Calculate r .
 - (b) Test whether ρ is significantly different from zero. [4]
 - (iii) Comment on the results to your tests from part (i) and (ii). [2]
- [Total 10]

Question 4.8

The sums of the squares of the errors in a regression analysis are found to be:

$$\begin{aligned} SS_{REG} &= \sum(\hat{y}_i - \bar{y})^2 = 6.4 & SS_{RES} &= \sum(y_i - \hat{y}_i)^2 = 3.6 \\ SS_{TOT} &= \sum(y_i - \bar{y})^2 = 10.0 \end{aligned}$$

Find the coefficient of determination and explain what this represents. [2]

Question 4.9

How would you transform the following models to linear form:

$$(i) \quad y_i = a + bx_i^2 + e_i \quad [2]$$

$$(ii) \quad y_i = ae^{bx_i} \quad [2]$$

[Total 4]

2 Exam-style questions

Question 4.10

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:

$$H_0 : \mu = 7\text{cm} \quad vs \quad H_1 : \mu \neq 7\text{cm}$$

Assuming that the lengths of worms are normally distributed, find the probability-value of these sample results. [3]

Question 4.11

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.

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]

Question 4.12

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.

- (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]

Question 4.13

The total claim amounts (in £m) for home and car insurance over a year for similar sized companies are collected by an independent advisor:

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]

Question 4.14

A random variable X is believed to have probability density function, $f(x)$, where:

$$f(x) = 3\lambda^3(\lambda + x)^{-4} \quad x > 0$$

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]

Question 4.15

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.

- (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]

Question 4.16

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]

Question 4.17

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. [5]

[Total 13]

Question 4.18

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]

Question 4.19

Insurance claims (in £) arriving at an office over the last month have been analysed. The results are as follows:

Claim size, c	$0 \leq c < 500$	$500 \leq c < 1000$	$1000 \leq c < 2500$	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	$0 \leq c < 500$	$500 \leq c < 1000$	$1000 \leq c < 2500$	over 2,500
No. of claims	23	14	7	3

Postcode 2:

Claim size, c	$0 \leq c < 500$	$500 \leq c < 1000$	$1000 \leq c < 2500$	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]

Question 4.20

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]

Question 4.21

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:

	<i>Red</i>	<i>White</i>	<i>Pink</i>
<i>Plain</i>	97	42	77
<i>Variegated</i>	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:

	<i>Red</i>	<i>White</i>	<i>Pink</i>
<i>Plain</i>	q	$q/2$	$(1-3q)/2$
<i>Variegated</i>	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]

Question 4.22

- (i) An analysis using the bivariate correlation model resulted in a sample correlation coefficient of $r = 0.8$ based on a sample of 20 data pairs.

Test whether the population correlation coefficient, ρ , is greater than 0.5. [3]

- (ii) A sample of 10 pairs of values from a *different* bivariate normal distribution resulted in a sample correlation coefficient of $r = 0.62$.

Obtain a 95% confidence interval for the underlying population correlation coefficient. [3]

[Total 6]

Question 4.23

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

$$\sum(x_i - \bar{x})^2 = 60 \quad \sum(y_i - \bar{y})^2 = 925,262 \quad \sum(x_i - \bar{x})(y_i - \bar{y}) = 7,087$$

An actuary fits the following simple linear regression model to the data:

$$y_i = \alpha + \beta x_i + e_i \quad i = 0, 1, \dots, 8$$

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) Obtain 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]

Question 4.24

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) Calculate the correlation coefficient and carry out a test to establish whether or not the data agrees with the claim that class size does not affect GCSE results. [6]
 - (iii) Following his investigation, the teacher concludes, “bigger class sizes improve GCSE results”. Comment on this statement. [2]
 - (iv) 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 15]

Question 4.25

An actuary is fitting the following linear regression model through the origin:

$$Y_i = \beta x_i + e_i \quad e_i \sim N(0, \sigma^2) \quad i = 1, 2, \dots, n$$

- (i) Show that the least squares estimator of β is given by:

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} \quad [3]$$

- (ii) Derive the bias and mean square error of $\hat{\beta}$ under this model. [4]
- [Total 7]

Question 4.26

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:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{where } \varepsilon_i \sim N(0, \sigma^2) \text{ are independently distributed}$$

The summary results for eight ages were as follows:

Age, x	30	32	34	36	38	40	42	44
Force of mortality, $\mu_x (\times 10^{-4})$	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 . [4]
- (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 your transformation in part (a). [4]
- (ii) Use the data to obtain 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 them 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]

Question 4.27

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:

<i>Day (i)</i>	8	29	57	92	141	148
<i>Price (P_i)</i>	15	17	22	51	88	95
$\ln P_i$	2.7081	2.8332	3.0910	3.9318	4.4773	4.5539

$$\sum i = 475 \quad \sum i^2 = 54,403 \quad \sum \ln P_i = 21.5953 \quad \sum (\ln P_i)^2 = 81.1584$$

$$\sum i \ln P_i = 1,947.020$$

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) Obtain 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]

Question 4.28

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

$$\begin{aligned}\sum y &= 469.7 & \sum x_1 &= 235 & \sum x_2 &= 207.8 & \sum x_1^2 &= 11,202.68 & \sum x_2^2 &= 12,886.42 \\ \sum yx_1 &= 22,028.78 & \sum yx_2 &= 19,870.22 & \sum x_1 x_2 &= 8,985.96\end{aligned}$$

- (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, find their numerical values. [7]

- (ii) Predict the percentage effectiveness for a tablet containing 51.3 mg of drug x_1 and 18.3 mg of drug x_2 . [2]

[Total 9]

Question 4.29

Three insurance companies, A, B and C, write the same class of insurance business in each of four separate regions I, II, III, and IV. The table below shows their mean claim amounts, y , in 1998, per £100 sum insured.

<i>Region</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>Total</i>
<i>I</i>	89	68	62	219
<i>II</i>	78	59	61	198
<i>III</i>	114	85	83	282
<i>IV</i>	79	61	82	222
<i>Total</i>	360	273	288	921

$$\sum y^2 = 73,471$$

- (i) Ignoring possible regional effects, carry out a one-way analysis of variance to establish whether the data provide evidence of differences in mean claim amounts between companies. [5]
 - (ii) Obtain a 95% confidence interval for the true mean claim amount (per £100 sum assured) of Company B. [3]
- [Total 8]

Question 4.30

Three different drugs are being compared for their effectiveness in treating a certain illness. The mean number of days before the patient is discharged from hospital under each treatment is summarised below, together with the sample size and the sum of squares of the observations:

<i>Treatment</i>	<i>Sample size</i>	<i>Sample mean</i>	<i>Sum of squares</i>
A	10	5	264
B	6	7	310
C	8	3	84

- (i) For these three treatments, calculate estimates for the:
 - (a) overall mean
 - (b) common underlying variance. [4]
 - (ii) Perform an analysis of variance to show that real differences exist among the three treatments at the 1% level. [3]
 - (iii) Show that the mean number of days before discharge under treatment A is significantly better than under treatment B. [3]
 - (iv) The cost per day for treatments A, B and C are £7.50, £5.85 and £14.95 respectively. Given that it can also be shown that there are significant differences between each pair of treatments, briefly advise the hospital on which treatment it should use. [2]
- [Total 12]

Question 4.31

A company examining its employee retention rates considers the number of complete years a new employee works for a particular division before leaving. The results are shown in the table below:

<i>Division A</i>	4	7	6	8	6
<i>Division B</i>	3	9	5	8	
<i>Division C</i>	8	10	9	8	5

- (i) Perform a one-way analysis of variance at the 5% level to compare the retention rates for the three divisions. [5]
- (ii) Present the data in a simple diagram and hence comment briefly on the validity of the assumptions required for the analysis of variance. [2]
- (iii) Concern is expressed over the variability of the results in Division B. It is thought that this might be significantly different from the assumed common underlying variance.
- (a) Write down an estimate for the underlying common variance for all three divisions.
- (b) Calculate an unbiased estimate for the variance in Division B based on the data for Division B only.
- (c) Test at the 5% level whether the variance in Division B is significantly different from your estimate in (iii)(a). [5]
- [Total 12]

Question 4.32

The amounts (in £100 units) for claims arriving at 5 insurance companies, in a particular hour, were as follows:

<i>Company</i>	<i>Amounts</i>	<i>Sum</i>	<i>Sum of squares</i>
<i>A</i>	7.8, 11.6, 3.7, 5.2, 6.3	34.6	275.82
<i>B</i>	10.1, 12.7, 11.6, 14.7, 12.9, 18.3	80.3	1,115.25
<i>C</i>	8.6, 7.9, 15.2, 11.4	43.1	497.37
<i>D</i>	18.6, 17.6, 11.8, 9.2	57.2	879.6
<i>E</i>	16.3, 18.1, 19.8, 15.2, 9.4	78.8	1,304.74

- (i) Show that there are significant differences in the mean claim sizes between companies at the 1% level. [5]
 - (ii)
 - (a) Obtain estimates for the mean claim amounts for each company.
 - (b) Analyse the companies' mean claim sizes using a least significant difference approach. [8]
- [Total 13]

Part 4 – Solutions

Solution 4.1

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. [1]

(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. [1]

Solution 4.2

(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 $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$. From the data, we have:

$$\begin{aligned} \bar{x} &= \frac{120.19}{10} = 12.019 \\ s^2 &= \frac{1}{9}(1,693.6331 - 10 \times 12.019^2) = 27.674 \end{aligned} \quad [1]$$

This gives a statistic of:

$$t = \frac{12.019 - 15}{\sqrt{27.674} / \sqrt{10}} = -1.792 \quad [1]$$

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 $\mu = 15$. [1]

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 $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$. This gives:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{12.019 - 15}{\sqrt{20} / \sqrt{10}} = -2.108 \quad [1]$$

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 $\mu \neq 15$. [1]

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:

$$H_0: \sigma^2 = 20 \quad vs \quad H_1: \sigma^2 \neq 20 \quad [1/2]$$

We know that $\frac{(n-1)S^2}{\sigma^2}$ has a χ_{n-1}^2 distribution. [½]

The observed value of the test statistic is:

$$\frac{9 \times 27.674}{20} = 12.45 \quad [1]$$

The critical values of χ_9^2 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 $\sigma^2 = 20$. [1]

Solution 4.3

To test whether tails is more than twice as likely, we use the hypotheses:

$$H_0 : p = \frac{2}{3} \quad vs \quad H_1 : p > \frac{2}{3} \quad [½]$$

Let X be the number of tails obtained in the experiment, then:

$$X \sim Bin(25, p) \approx N(25p, 25pq) \Rightarrow \frac{X - 25p}{\sqrt{25pq}} \approx N(0,1) \quad [½]$$

Under H_0 , the statistic with continuity correction is:

$$z = \frac{17\frac{1}{2} - 16\frac{2}{3}}{\sqrt{5\frac{5}{9}}} = 0.354 \quad [1]$$

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 [1]

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.

Solution 4.4

We are testing:

$$H_0: \mu_A = \mu_B \quad vs \quad H_1: \mu_A \neq \mu_B \quad [1/2]$$

The test statistic is:

$$\frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{S_P \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_A+n_B-2} \text{ where } S_P^2 = \frac{(n_A-1)S_1^2 + (n_B-1)S_2^2}{n_A+n_B-2} \quad [1/2]$$

Now the observed value of the pooled variance is:

$$s_P^2 = \frac{9 \times 12.4 + 4 \times 25.8}{13} = 16.52 \quad [1]$$

So the value of the test statistic is:

$$\frac{4.5 - 0}{\sqrt{16.52} \sqrt{\frac{1}{10} + \frac{1}{5}}} = 2.021 \quad [1]$$

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 $\mu_A = \mu_B$. [1]

Alternatively, using probability values, we have $P(t_{13} > 2.021) \approx 0.034$. 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.

Solution 4.5

We are testing:

$$H_0: \text{The populations both have normal distributions}$$

vs $H_1: \text{At least one of the populations does not have a normal distribution. } [1/2]$

If H_0 is true, we know that the statistic $\frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2}$ has an $F_{4,9}$ distribution. [1/2]

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 $47.0 / 12.6 = 3.730$. [1]

The 5% critical values for an $F_{4,9}$ distribution are 0.1123 and 4.718. Since 3.633 lies between these, we have insufficient evidence at the 5% level to reject H_0 . Therefore we conclude that the populations *are* both normal. [1]

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.

Solution 4.6

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:

$$\frac{\text{Likelihood under } H_0}{\text{Likelihood under } H_1} < \text{critical value} \quad [1]$$

Since the populations are normal, this is:

$$\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_0}{\sigma}\right)^2} \Bigg/ \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2} < \text{constant}$$

Cancelling the constants reduces this to:

$$e^{-\frac{1}{2\sigma^2} \sum (x_i - \mu_0)^2} \Bigg/ e^{-\frac{1}{2\sigma^2} \sum (x_i - \mu_1)^2} < \text{constant} \quad [1]$$

Taking logs:

$$-\frac{1}{2\sigma^2} \sum (x_i - \mu_0)^2 + \frac{1}{2\sigma^2} \sum (x_i - \mu_1)^2 < \text{constant} \quad [1]$$

Multiplying through by $2\sigma^2$ and expanding the squares:

$$-\sum (x_i^2 - 2\mu_0 x_i + \mu_0^2) + \sum (x_i^2 - 2\mu_1 x_i + \mu_1^2) < \text{constant}$$

Simplifying this gives $(\mu_0 - \mu_1) \sum x_i < \text{constant}$. [1]

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:

$$\bar{x} > \text{constant}$$

So the best test requires us to reject H_0 if the sample mean exceeds a specified critical value. [1]

Solution 4.7(i)(a) ***Calculate slope parameter estimate***

Using the formula given on page 24 of the *Tables*:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{8.1}{12.2} = 0.66393 \quad [1]$$

(i)(b) ***Test whether slope parameter is significantly different from zero***

We are testing:

$$H_0: \beta = 0 \quad \text{vs} \quad H_1: \beta \neq 0$$

Under H_0 , $\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$ has a t_{n-2} distribution. Now:

$$\hat{\sigma}^2 = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = \frac{1}{17} \left(10.6 - \frac{8.1^2}{12.2} \right) = 0.30718 \quad [1]$$

So the observed value of the test statistic is:

$$\frac{0.66393 - 0}{\sqrt{0.30718/12.2}} = 4.184 \quad [1]$$

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 $\beta \neq 0$. [1]

(ii)(a) ***Calculate the correlation coefficient***

Using the formula on page 25 of the *Tables*:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{8.1}{\sqrt{12.2 \times 10.6}} = 0.71228 \quad [1]$$

(ii)(b) ***Test whether correlation coefficient is significantly different from zero***

We are testing:

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0 \quad [1/2]$$

Under H_0 , $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ has a t_{n-2} distribution. [1/2]

So the observed value of the test statistic is:

$$\frac{0.71228\sqrt{17}}{\sqrt{1-0.71228^2}} = 4.184 \quad [1]$$

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 $\rho \neq 0$. [1]

(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). [2]

Solution 4.8

The coefficient of determination is given by:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{6.4}{10.0} = 0.64 \quad [1]$$

This explains the proportion of the total variance “explained” by the model. So 64% of the variance can be “explained” by the model. [1]

Solution 4.9

- (i)
- Transform quadratic to linear form**

Let $Y_i = y_i$ and $X_i = x_i^2$. [1]

Then the model becomes $Y_i = a + bX_i + e_i$. [1]

- (ii)
- Transform exponential to linear form**

Taking logs gives:

$$\ln y_i = \ln a + bx_i \quad [\frac{1}{2}]$$

Let $Y_i = \ln y_i$ and $X_i = x_i$. [½]

Then the model becomes $Y_i = \alpha + \beta X_i$ where $\alpha = \ln a$ and $\beta = b$. [1]

Solution 4.10

We are testing:

$$H_0 : \mu = 7\text{cm} \quad vs \quad H_1 : \mu \neq 7\text{cm} \quad (\sigma^2 \text{ unknown})$$

Under H_0 , the statistic $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ has a t_{11} distribution.

So the value of our test statistic is:

$$\frac{8.54 - 7}{2.97/\sqrt{12}} = 1.796 \quad [1]$$

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]

Solution 4.11

- (i) **Test whether new screening programme reduces mean claim amount**

We are testing:

$$H_0: \mu_1 = \mu_2 \quad vs \quad H_1: \mu_2 < \mu_1 \quad [1/2]$$

where subscript 1 refers to without screening.

The test statistic is:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ where } S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \quad [1/2]$$

Calculating the observed values:

$$\bar{x}_1 = \frac{267.4}{10} = 26.74, \quad \bar{x}_2 = \frac{200.7}{10} = 20.07 \quad [1/2]$$

$$s_1^2 = \frac{1}{9} (7,755.16 - 10 \times 26.74^2) = 67.2093 \quad [1/2]$$

$$s_2^2 = \frac{1}{9} (4,553.97 - 10 \times 20.07^2) = 58.4357 \quad [1/2]$$

$$s_P^2 = \frac{9 \times 67.2093 + 9 \times 58.4357}{18} = 62.8225 \quad [1/2]$$

So the value of the test statistic is:

$$\frac{(26.74 - 20.07) - 0}{\sqrt{62.8225} \sqrt{\frac{2}{10}}} = 1.882 \quad [1]$$

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 $P(t_{18} > 1.882) \approx 0.04$ 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:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \quad [1/2]$$

The test statistic is:

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1} \quad [1/2]$$

Under H_0 , the value of the test statistic is:

$$\frac{67.2093}{58.4357} = 1.150 \quad [1]$$

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 for part (i) seems valid). [1]

Solution 4.12(i) ***Test if chemicals are present***

We are testing the proportion p of defective animals using the hypotheses:

$$H_0 : p = 0.005 \quad vs \quad H_1 : p > 0.005 \quad [1/2]$$

Let X be the number of deformed animals obtained, then:

$$X \sim Bin(1000, p) \div N(1000p, 1000pq) \Rightarrow \frac{X - 1000p}{\sqrt{1000pq}} \div N(0, 1) \quad [1/2]$$

Under H_0 , the statistic with continuity correction is:

$$\frac{14.5 - 5}{\sqrt{4.975}} = 4.26 \quad [2]$$

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.

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:

$$H_0 : p_1 = p_2 \quad vs \quad H_1 : p_2 < p_1 \quad [1/2]$$

where the subscript 1 refers to “before” and 2 refers to “after”.

The test statistic is:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \sim N(0,1) \quad [1/2]$$

Here we have:

$$\hat{p}_1 = \frac{15}{1,000} = 0.015 \quad \hat{p}_2 = \frac{10}{800} = 0.0125 \quad \hat{p} = \frac{25}{1,800} = 0.0138 \quad [1/2]$$

which gives us a value of 0.450 for our test statistic. [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). [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.

Solution 4.13

- (i) **Test whether mean home and car claims are equal**

We are testing:

$$H_0: \mu_H = \mu_C \quad vs \quad H_1: \mu_H \neq \mu_C \quad [\frac{1}{2}]$$

The test statistic is:

$$\frac{(\bar{X}_H - \bar{X}_C) - (\mu_H - \mu_C)}{S_P \sqrt{\frac{1}{n_H} + \frac{1}{n_C}}} \sim t_{n_H+n_C-2}$$

$$\text{where } S_P^2 = \frac{(n_H - 1)S_H^2 + (n_C - 1)S_C^2}{n_H + n_C - 2} \quad [\frac{1}{2}]$$

Calculating the observed values:

$$\bar{x}_H = \frac{78.8}{5} = 15.76, \quad \bar{x}_C = \frac{88.3}{5} = 17.66 \quad [\frac{1}{2}]$$

$$s_1^2 = \frac{1}{4} (1,271.34 - 5 \times 15.76^2) = 7.363 \quad [\frac{1}{2}]$$

$$s_2^2 = \frac{1}{4} (1,631.93 - 5 \times 17.66^2) = 18.138 \quad [\frac{1}{2}]$$

$$s_P^2 = \frac{4 \times 7.363 + 4 \times 18.138}{8} = 12.7505 \quad [\frac{1}{2}]$$

The value of the test statistic is:

$$\frac{(15.76 - 17.66) - 0}{\sqrt{12.7505} \sqrt{\frac{2}{5}}} = -0.841 \quad [1]$$

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 $P(t_8 < -0.841) \approx 0.21$. 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:

$$\text{Sample 2} - \text{Sample 1}: \quad 1.0 \quad 1.8 \quad -0.1 \quad 1.6 \quad 5.2 \quad [1/2]$$

Now:

$$\bar{x}_D = \frac{9.5}{5} = 1.9 \quad s_D^2 = \frac{1}{4}(33.85 - 5 \times 1.9^2) = 3.95 \quad [1/2]$$

So our test statistic is:

$$\frac{\bar{x}_D - \mu_D}{s_D / \sqrt{n}} = \frac{1.9 - 0}{\sqrt{3.95/5}} = 2.138 \quad [1]$$

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 $P(t_4 > 2.138) \approx 0.05$. 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.

Solution 4.14(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.

$$\begin{aligned}\alpha &= P(X > 93.5 \text{ when } \lambda = 50) & [1] \\ &= \int_{93.5}^{\infty} 3 \times 50^3 (50+x)^{-4} dx \\ &= \left[-50^3 (50+x)^{-3} \right]_{93.5}^{\infty} \\ &= 0.0423\end{aligned}$$

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.

$$\begin{aligned}1 - \beta &= P(X > 93.5 \text{ when } \lambda = 60) & [1] \\ &= \int_{93.5}^{\infty} 3 \times 60^3 (60+x)^{-4} dx \\ &= \left[-60^3 (60+x)^{-3} \right]_{93.5}^{\infty} \\ &= 0.0597\end{aligned}$$

The power of the test is 5.97%. [1]

Solution 4.15(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:

$$H_0 : p = 0.2 \quad vs \quad H_1 : p \neq 0.2 \quad [1]$$

Under H_0 , the number of correctly identified patterns has a $\text{Bin}(10, 0.2)$ 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 $\text{Bin}(10, 0.2)$ distribution (iteratively) shows that:

$$\begin{aligned} P[\text{Bin}(10, 0.2) \leq 4] &= 0.1074 + 0.2684 + 0.3020 + 0.2013 + 0.0881 \\ &= 0.9672 \end{aligned} \quad [1]$$

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]

Solution 4.16

- (i) **Maximum likelihood estimate of p**

The likelihood of observing the given sample is:

$$\begin{aligned} L &= C[(1-p)^3]^{60}[3p(1-p)^2]^{75}[3p^2(1-p)]^{16}[p^3]^2 \\ &= K(1-p)^{180}p^{75}(1-p)^{150}p^{32}(1-p)^{16}p^6 \\ &= K(1-p)^{346}p^{113} \end{aligned} \quad [1]$$

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 :

$$\frac{d \ln L}{dp} = -\frac{346}{1-p} + \frac{113}{p} \quad [1]$$

Setting this equal to zero gives:

$$\begin{aligned} 346\hat{p} &= 113(1-\hat{p}) \\ \Rightarrow 459\hat{p} &= 113 \\ \Rightarrow \hat{p} &= \frac{113}{459} = 0.246 \end{aligned} \quad [1]$$

Checking that we do have a maximum:

$$\frac{d^2 \ln L}{dp^2} = -\frac{346}{(1-p)^2} - \frac{113}{p^2} < 0 \Rightarrow \max \quad [1]$$

(ii) ***Goodness of fit test***

We are testing the following hypotheses using a χ^2 goodness of fit test:

$$H_0 : \text{the probabilities conform to a } Bin(3, p) \text{ distribution}$$

vs $H_1 : \text{the probabilities do not conform to a } Bin(3, p) \text{ distribution}$

Using $\hat{p} = \frac{113}{459}$ from part (i), the probabilities for this binomial distribution are:

$$P(X = 0) = (1 - p)^3 = 0.4283$$

$$P(X = 1) = 3p(1 - p)^2 = 0.4197$$

$$P(X = 2) = 3p^2(1 - p) = 0.1371$$

$$P(X = 3) = p^3 = 0.0149$$

[1]

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:

$$\begin{aligned} \sum \frac{(O_i - E_i)^2}{E_i} &= \frac{(60 - 65.54)^2}{65.54} + \frac{(75 - 64.21)^2}{64.21} + \frac{(18 - 23.25)^2}{23.25} \\ &= 0.4683 + 1.813 + 1.185 = 3.47 \end{aligned}$$

[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]

Solution 4.17(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:

$$\hat{\mu} = \frac{0 \times 40 + 1 \times 66 + 2 \times 93 + \dots + 8 \times 1}{400} = \frac{1,034}{400} = 2.585 \quad [1]$$

(i)(b) ***Goodness of fit test***

The hypotheses are:

$$\begin{aligned} H_0: & \text{ The observed numbers conform to a Poisson distribution} \\ \text{vs } H_1: & \text{ The observed numbers don't conform to a Poisson distribution.} \end{aligned} \quad [\frac{1}{2}]$$

We can use our estimate from part (i)(a) to calculate the expected numbers using the Poisson PF:

$$\text{eg } P(X=0) = e^{-2.585} = 0.07540 \Rightarrow 30.16 \text{ cells} \quad [\frac{1}{2}]$$

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:

$$\begin{aligned} \sum \frac{(O-E)^2}{E} &= \frac{(40-30.2)^2}{30.2} + \frac{(66-78.0)^2}{78.0} + \dots + \frac{(6-6.6)^2}{6.6} \\ &= 3.180 + 1.846 + 0.604 + 0.597 + 0.620 + 0.552 + 0.18 + 0.055 \\ &= 7.63 \end{aligned} \quad [3]$$

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:

$$H_0 : \mu = 3 \quad vs \quad H_1 : \mu < 3 \quad [1/2]$$

Let X be the count per cell, then:

$$\sum X \sim Poi(400\mu) \div N(400\mu, 400\mu) \Rightarrow \frac{X - 400\mu}{\sqrt{400\mu}} \div N(0,1) \quad [1/2]$$

Under H_0 , the statistic with continuity correction is:

$$z = \frac{1,034.5 - 1,200}{\sqrt{1,200}} = -4.78 \quad [1]$$

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 $P(Z < -4.78) < 0.000001$ which is highly significant!

Solution 4.18

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
<i>Total</i>	<i>122</i>	<i>100</i>	<i>327</i>	<i>549</i>

[1]

The value of the chi square test statistic is:

$$\sum \frac{(O_i - E_i)^2}{E_i} = \frac{(24 - 10.7)^2}{10.7} + \dots + \frac{(310 - 298.4)^2}{298.4} = 23.79 \quad [2]$$

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 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]

Solution 4.19(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 $\frac{1}{\lambda}$. 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:

$$\begin{aligned} P(0 < X < 500) &= e^0 - e^{-500\lambda} = 1 - e^{-500\lambda} = 0.4531 \\ P(500 < X < 1,000) &= e^{-500\lambda} - e^{-1,000\lambda} = 0.2478 \\ P(1,000 < X < 2,500) &= e^{-1,000\lambda} - e^{-2,500\lambda} = 0.2502 \\ P(2,500 < X < 10,000) &= e^{-2,500\lambda} - e^{-10,000\lambda} = 0.0489 \end{aligned} \quad [1]$$

Multiplying these figures by 153, we obtain the expected values 69.33, 37.91, 38.27 and 7.48 respectively. [1]

We then substitute into the test statistic $\sum \frac{(O_i - E_i)^2}{E_i}$:

$$\frac{(75 - 69.33)^2}{69.33} + \frac{(51 - 37.91)^2}{37.91} + \frac{(22 - 38.27)^2}{38.27} + \frac{(5 - 7.48)^2}{7.48} = 12.7 \quad [1]$$

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 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:

$$\begin{aligned} H_0 &: \text{the claim size is independent of postcode} \\ \text{vs } H_1 &: \text{the claim size is not independent of postcode} \end{aligned}$$

The observed values in each of the categories are:

Claim size, c	$0 \leq c < 500$	$500 \leq c < 1000$	$1000 \leq c < 2500$	$2500 \leq c < 10000$	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	$0 \leq c < 500$	$500 \leq c < 1000$	$1000 \leq c < 2500$	$2500 \leq c < 10000$
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	$0 \leq c < 500$	$500 \leq c < 1000$	$1000 \leq c < 10000$
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:

$$\chi^2 = \frac{(23 - 23.04)^2}{23.04} + \dots + \frac{(5 - 8.47)^2}{8.47} = 4.58 \quad [1]$$

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]

Solution 4.20

(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:

$$\begin{aligned}\sum \frac{(O-E)^2}{E} &= \frac{(100-85)^2}{85} + \frac{(300-315)^2}{315} + \frac{(240-255)^2}{255} + \frac{(960-945)^2}{945} \\ &= 4.482\end{aligned}\quad [2]$$

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^2_1 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]

Solution 4.21(i) **Test for association**

The test required is a χ^2 contingency table test. The hypotheses are:

$$\begin{aligned} H_0: & \text{ There is no association between flower colour and leaf type} \\ \text{vs } H_1: & \text{ There is some association between flower colour and leaf type.} \end{aligned} \quad [1]$$

The expected frequencies are:

	<i>Red</i>	<i>White</i>	<i>Pink</i>
<i>Plain</i>	87.3	82.1	46.7
<i>Variegated</i>	114.7	107.9	61.3

[2]

So the test statistic is:

$$\sum \frac{(O-E)^2}{E} = \frac{(97-87.3)^2}{87.3} + \dots + \frac{(31-61.3)^2}{61.3} = 71.0 \quad [2]$$

Comparing this with the figures in the *Tables* for the χ^2 distribution, we see that this figure is far larger than the 1% 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:

$$\begin{aligned} L(q) &= q^{97} \left(\frac{q}{2}\right)^{42} \left(\frac{1-3q}{2}\right)^{77} q^{105} \left(\frac{3q}{2}\right)^{148} \left(\frac{1-5q}{2}\right)^{31} \times \text{constant} \\ &= q^{392} (1-3q)^{77} (1-5q)^{31} \times \text{constant} \end{aligned} \quad [1]$$

Taking logs:

$$\log L = 392 \log q + 77 \log(1-3q) + 31 \log(1-5q) + \text{constant} \quad [1]$$

Differentiating with respect to q :

$$\frac{d}{dq} \log L = \frac{392}{q} - \frac{231}{1-3q} - \frac{155}{1-5q} \quad [1]$$

Setting this equal to zero, and multiplying through by $q(1-3q)(1-5q)$, we obtain:

$$392(1-3q)(1-5q) - 231q(1-5q) - 155q(1-3q) = 0 \quad [1]$$

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 $\frac{1-5q}{2}$ 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:

$$\frac{d^2}{dq^2} \log L = -\frac{392}{q^2} - \frac{693}{(1-3q)^2} - \frac{775}{(1-5q)^2} < 0 \Rightarrow \max \quad [1]$$

(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:

	<i>Red</i>	<i>White</i>	<i>Pink</i>
<i>Plain</i>	90.6	45.3	114.0
<i>Variegated</i>	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:

$$\sum \frac{(O - E)^2}{E} = \frac{(97 - 90.6)^2}{90.6} + \dots + \frac{(31 - 23.4)^2}{23.4} = 18.5 \quad [2]$$

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]

Solution 4.22

- (i) ***Test whether correlation coefficient greater than 0.5***

We are testing:

$$H_0 : \rho = 0.5 \quad \text{vs} \quad H_1 : \rho > 0.5$$

Under H_0 , $\frac{Z_r - Z_\rho}{\sqrt{1/(n-3)}}$ has a $N(0,1)$ distribution. The observed values are:

$$z_r = \tanh^{-1} 0.8 = 1.0986 \quad z_\rho = \tanh^{-1} 0.5 = 0.54931 \quad [1]$$

So the observed value of the test statistic is:

$$\frac{1.0986 - 0.54931}{\sqrt{1/17}} = 2.265 \quad [1]$$

Since this is greater than 2.2571, the upper 1.2% point of the $N(0,1)$ distribution, we have sufficient evidence to reject H_0 at the 1.2% level. Therefore it is reasonable to conclude that $\rho > 0.5$. [1]

- (ii) ***Confidence interval for correlation coefficient***

From page 25 of the *Tables*, we have:

$$\frac{Z_r - Z_\rho}{\sqrt{1/(n-3)}} \stackrel{d}{\sim} N(0,1)$$

Now $z_r = \tanh^{-1} 0.62 = 0.72501$, so the 95% confidence interval for ρ is given by:

$$-1.96 \leq \frac{0.72501 - z_\rho}{\sqrt{1/7}} \leq 1.96 \quad [1]$$

$$-0.015805 \leq Z_\rho \leq 1.465815 \quad [1]$$

Using the \tanh function gives:

$$-0.0158 \leq \rho \leq 0.899 \quad [1]$$

Solution 4.23

- (i) **Regression line**

We are given:

$$s_{xx} = 60 \quad s_{yy} = 925,262 \quad s_{xy} = 7,087$$

So:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{7,087}{60} = 118.117 \quad [1]$$

Since $\bar{x} = \frac{36}{9} = 4$ and $\bar{y} = \frac{3,960}{9} = 440$, we get:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 440 - 118.117 \times 4 = -32.47 \quad [1]$$

So the regression line is:

$$\hat{y} = -32.47 + 118.117x$$

Deduct ½ mark if regression line is not given.

- (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:

$$\hat{\sigma}^2 = \frac{1}{7} \left(925,262 - \frac{7,087^2}{60} \right) = 12,595.6 \quad [1]$$

So the 99% confidence interval is given by:

$$118.117 \pm 3.499 \sqrt{\frac{12,595.6}{60}} = 118.117 \pm 50.696 = (67.4, 169) \quad [2]$$

(ii)(b) ***Confidence interval for variance***

The pivotal quantity is given by:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2 \quad [1]$$

A 99% confidence interval is given by:

$$0.90 = P\left(\chi_{n-2;0.995}^2 < \frac{(n-1)\hat{\sigma}^2}{\sigma^2} < \chi_{n-2;0.005}^2\right)$$

which gives a confidence interval of:

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi_{n-2;0.005}^2}, \frac{(n-2)\hat{\sigma}^2}{\chi_{n-2;0.995}^2} \right)$$

Substituting in, the confidence interval (to 3 SF) is:

$$\left(\frac{7 \times 12,595.6}{20.28}, \frac{7 \times 12,595.6}{0.9893} \right) = (4350, 89100) \quad [1]$$

(iii)(a) ***Partition***

The total sum of squares, $SS_{TOT} = \sum (y_i - \bar{y})^2$ is given by s_{yy} which is 925,262. [1]

The partition given on the bottom of page 25 in the *Tables* is:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$ie \quad SS_{TOT} = SS_{RES} + SS_{REG}$$

Now, modifying the $\hat{\sigma}^2$ formula on page 24 of the *Tables*, we have:

$$SS_{RES} = \sum (y_i - \hat{y}_i)^2 = \left(s_{yy} - \frac{s_{xy}^2}{s_{xx}} \right) = 925,262 - \frac{7,087^2}{60} = 88,169 \quad [1]$$

Alternatively, using $\hat{\sigma}^2$ from part (ii), we get $SS_{RES} = (n - 2)\hat{\sigma}^2 = 7 \times 12,595.6$.

Hence:

$$SS_{REG} = 925,262 - 88,169 = 837,093 \quad [1]$$

Alternatively, this could be calculated as $SS_{REG} = \frac{s_{xy}^2}{s_{xx}} = \frac{7,087^2}{60} = 837,093$.

(iii)(b) **Proportion of variability explained by the model**

This is the coefficient of determination, R^2 , which is given by:

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{837,093}{925,262} = 90.5\% \quad [1]$$

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:

$$x = 1 \Rightarrow \hat{y} = -32.47 + 118.117 \times 1 \approx 86 \Rightarrow \hat{e} = 131 - 86 \approx 45 \quad [1]$$

$$x = 4 \Rightarrow \hat{y} = -32.47 + 118.117 \times 4 \approx 440 \Rightarrow \hat{e} = 330 - 440 \approx -110 \quad [1]$$

$$x = 8 \Rightarrow \hat{y} = -32.47 + 118.117 \times 8 \approx 912 \Rightarrow \hat{e} = 1,095 - 912 \approx 183 \quad [1]$$

(iv)(b) ***Dotplot of residuals***

The dotplot is:

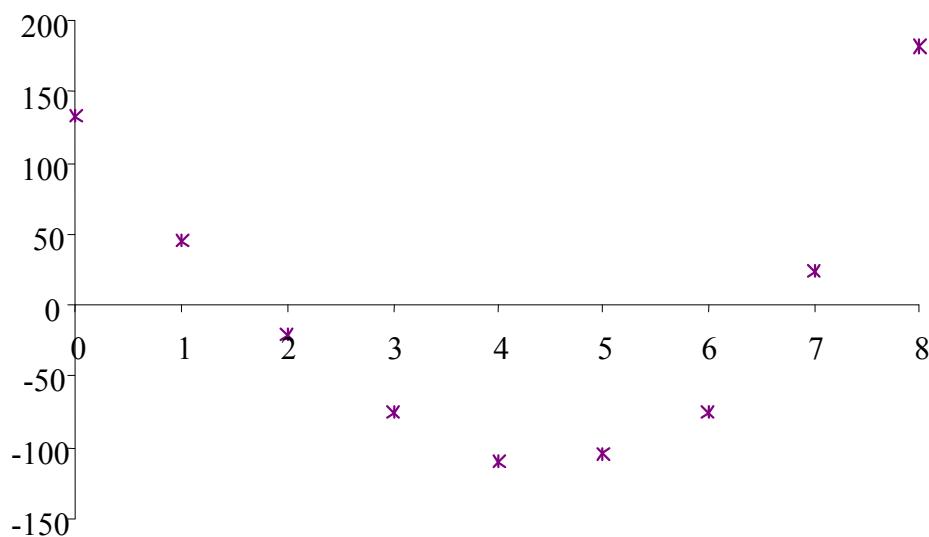


[1]

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***

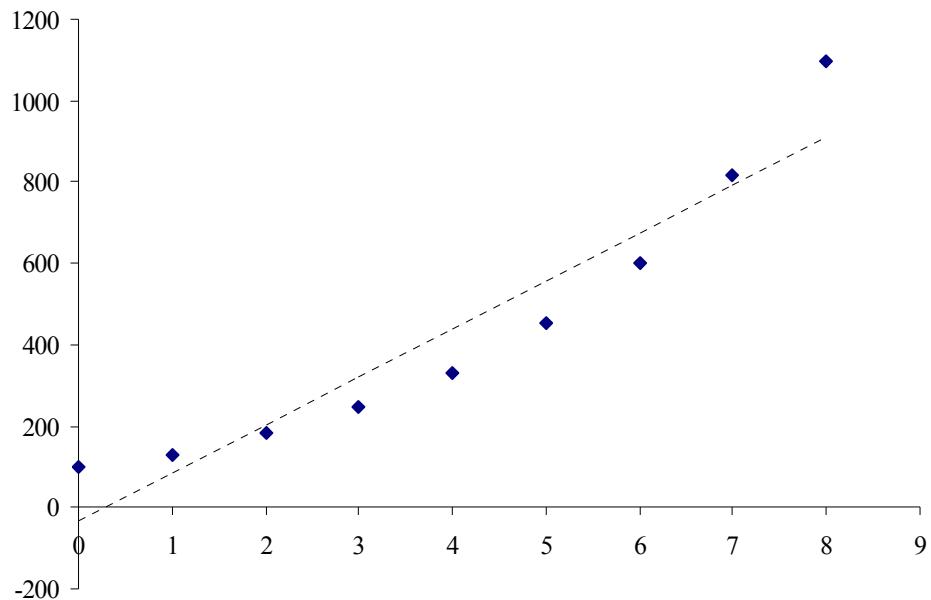
The graph is:



[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 the mistake:



The price increases in an exponential (rather than linear) way. We should have used the log of the price against time instead.

Solution 4.24

- (i) ***Obtain the fitted regression line***

The regression line for p on c is given by:

$$p = \hat{\alpha} + \hat{\beta}c$$

where $\hat{\beta} = \frac{S_{cp}}{S_{cc}}$ and $\hat{\alpha} = \bar{p} - \hat{\beta}\bar{c}$.

$$S_{cc} = \sum c^2 - \frac{(\sum c)^2}{n} = 6,884 - \frac{238^2}{9} = 590.2222$$

[1]

$$S_{cp} = \sum cp - \frac{(\sum c)(\sum p)}{n} = 983 - \frac{238 \times 33.4}{9} = 99.75556$$

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) ***Test the correlation coefficient***

The correlation coefficient is calculated using:

$$r = \frac{S_{cp}}{\sqrt{S_{cc}S_{pp}}}$$

We require:

$$S_{pp} = \sum p^2 - \frac{(\sum p)^2}{n} = 149.62 - \frac{33.4^2}{9} = 25.66889$$

So:

$$r = \frac{99.75556}{\sqrt{590.2222 \times 25.66889}} = 0.81045 \quad [2]$$

We are testing:

$$H_0: \rho = 0 \quad vs \quad H_1: \rho \neq 0$$

The statistic is:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \quad [1]$$

Under H_0 our t -statistic is $\frac{0.81045\sqrt{7}}{\sqrt{1-0.81045^2}} = 3.660$ which is greater than the 0.5%

critical value of 3.499 for 7 degrees of freedom. 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). [3]

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 $\rho = 0$.

(iii) ***Comment***

There is strong positive correlation between class size and GCSE results (*ie* bigger classes have better GCSE results). 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. [2]

(iv) ***Estimate the GCSE score and its standard error***

The estimate of the average GCSE point score is obtained from the regression line:

$$\hat{P} = -0.75836 + 0.16901 \times 15 = 1.78 \quad [1]$$

The standard error of this individual response is given by:

$$\sqrt{\left\{1 + \frac{1}{n} + \frac{(c_0 - \bar{c})^2}{S_{cc}}\right\}\hat{\sigma}^2} \quad [1]$$

$$\text{where } \hat{\sigma}^2 = \frac{1}{n-2} \left(S_{pp} - \frac{S_{cp}^2}{S_{cc}} \right) = \frac{1}{7} \left(25.66889 - \frac{99.75556^2}{590.2222} \right) = 1.25841. \quad [1]$$

Hence, the standard error is given by:

$$\begin{aligned} &\sqrt{\left\{1 + \frac{1}{9} + \frac{(15 - \frac{238}{9})^2}{590.2222}\right\}1.25841} \\ &= \sqrt{1.33302 \times 1.25841} \\ &= \sqrt{1.67748} \\ &= 1.29518 \end{aligned} \quad [1]$$

Solution 4.25(i) **Least squares estimate of slope parameter**

The least squares estimate minimises $\sum e_i^2$. Now:

$$\sum e_i^2 = \sum (Y_i - \beta x_i)^2 \quad [1/2]$$

Differentiating this gives:

$$\frac{d}{d\beta} \Rightarrow -2 \sum x_i(Y_i - \beta x_i) \quad [1]$$

Setting this equal to zero gives:

$$\begin{aligned} \sum x_i(Y_i - \hat{\beta}x_i) &= 0 \\ \sum x_i Y_i - \hat{\beta} \sum x_i^2 &= 0 \\ \hat{\beta} &= \frac{\sum x_i Y_i}{\sum x_i^2} \end{aligned} \quad [1]$$

The second derivative is $2 \sum x_i^2 > 0$, so we do have a minimum. [1/2]

(ii) **Bias and mean square error**

The expectation of $\hat{\beta}$ is:

$$E(\hat{\beta}) = E\left(\frac{\sum x_i Y_i}{\sum x_i^2}\right) = \frac{\sum x_i E(Y_i)}{\sum x_i^2} \quad [1/2]$$

Now $E(Y_i) = E(\beta x_i + e_i) = \beta x_i + 0 = \beta x_i$. So: [1/2]

$$E(\hat{\beta}) = \frac{\beta \sum x_i^2}{\sum x_i^2} = \beta \quad [1/2]$$

Hence:

$$bias(\hat{\beta}) = E(\hat{\beta}) - \beta = 0 \quad [1/2]$$

The variance of $\hat{\beta}$ is:

$$\text{var}(\hat{\beta}) = \text{var}\left(\frac{\sum x_i Y_i}{\sum x_i^2}\right) = \frac{\sum x_i^2 \text{var}(Y_i)}{\left(\sum x_i^2\right)^2} \quad [\frac{1}{2}]$$

Now $\text{var}(Y_i) = \text{var}(\beta x_i + e_i) = \text{var}(e_i) = \sigma^2$. So: [\frac{1}{2}]

$$\text{var}(\hat{\beta}) = \frac{\sigma^2 \sum x_i^2}{\left(\sum x_i^2\right)^2} = \frac{\sigma^2}{\sum x_i^2}$$

Hence:

$$MSE(\hat{\beta}) = \text{var}(\hat{\beta}) + \text{bias}^2(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2} \quad [1]$$

Solution 4.26(i)(a) *Expressions for parameters*

Taking logs of the original expression gives:

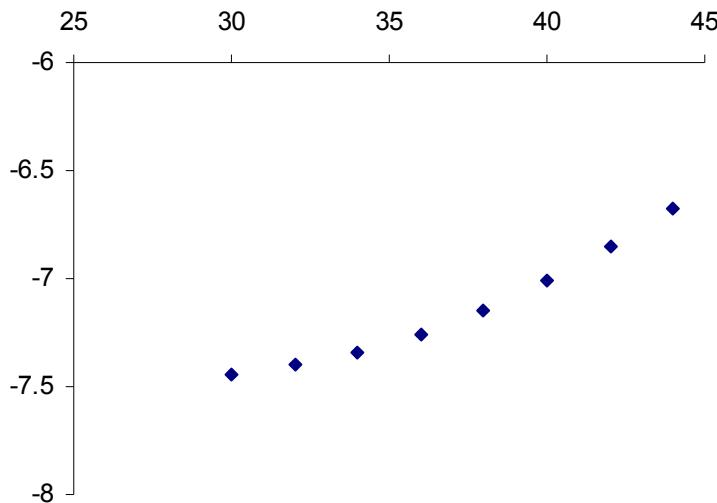
$$\ln \mu_x = \ln B + x \ln c \quad [1]$$

This expression is now linear in x . Comparing the expression with $Y = \alpha + \beta x$ gives:

$$Y = \ln \mu_x \quad \alpha = \ln B \quad \beta = \ln c \quad [1]$$

(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 to 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 $y = \ln \mu_x$:

$$s_{xx} = \sum x^2 - n\bar{x}^2 = 11,120 - 8\left(\frac{296}{8}\right)^2 = 168$$

$$s_{xy} = \sum xy - n\bar{x}\bar{y} = -2,104.5 - 8\left(\frac{296}{8}\right)\left(\frac{-57.129}{8}\right) = 9.273$$

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{9.273}{168} = 0.055196 \quad [1]$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{-57.129}{8} - 0.055196 \times \frac{296}{8} = -9.1834 \quad [1]$$

Therefore, we obtain:

$$B = e^\alpha = e^{-9.1834} = 0.000103 \quad [1]$$

$$c = e^\beta = e^{0.055196} = 1.06$$

(iii)(a) **Coefficient of determination and comment**

The coefficient of determination is given by:

$$R^2 = r^2 = \frac{s_{xy}^2}{s_{xx}s_{yy}} = \frac{9.273^2}{168 \times 0.53467} = 95.7\% \quad [1]$$

$$\text{where } s_{yy} = \sum y^2 - n\bar{y}^2 = 408.50 - 8\left(\frac{-57.129}{8}\right)^2 = 0.53467$$

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

$$\text{Age 32 yrs } (-7.40) - (-9.1834 + 0.055196 \times 32) = 0.02 \quad [1]$$

$$\text{Age 36 yrs } (-7.26) - (-9.1834 + 0.055196 \times 36) = -0.06 \quad [1]$$

$$\text{Age 40 yrs } (-7.01) - (-9.1834 + 0.055196 \times 40) = -0.03 \quad [1]$$

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:

$$\left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} \hat{\sigma}^2 = \left\{ \frac{1}{8} + \frac{(35 - 37)^2}{168} \right\} \times 0.0038056 = 0.0005663 \quad [1]$$

$$\text{where } \hat{\sigma}^2 = \frac{1}{6} \left(0.53467 - \frac{9.273^2}{168} \right) = 0.0038056 \quad [1]$$

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:

$$-7.251 \pm 2.447 \sqrt{0.0005663} = (-7.309, -7.193) \quad [1]$$

Hence the 95% confidence interval for μ_{35} is given by:

$$(0.000669, 0.000752) \quad [1]$$

Solution 4.27(i) **Estimate parameters**

Now using x for i and y for $\ln P_i$, we get:

$$\begin{aligned}s_{xx} &= \sum x^2 - n\bar{x}^2 = 16,799 \\ s_{xy} &= \sum xy - n\bar{x}\bar{y} = 237.39 \\ s_{yy} &= \sum y^2 - n\bar{y}^2 = 3.4322\end{aligned}\quad [2]$$

So the estimates for a , b and σ^2 are:

$$\hat{b} = \frac{s_{xy}}{s_{xx}} = \frac{237.39}{16,799} = 0.01413 \quad [1]$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{21.5953}{6} - 0.01413\left(\frac{475}{6}\right) = 2.4805 \quad [1]$$

$$\hat{\sigma}^2 = \frac{1}{n-2}(s_{yy} - \frac{s_{xy}^2}{s_{xx}}) = \frac{1}{4}(3.4322 - \frac{237.39^2}{16,799}) = 0.01940 \quad [1]$$

(ii) **Correlation coefficient**

The correlation coefficient is:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{237.39}{\sqrt{16,799 \times 3.4322}} = 0.989 \quad [1]$$

(iii) **Confidence interval for slope parameter**

Using the result given on page 24 of the *Tables*, we have:

$$\hat{b} \pm t_{4;0.005} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = 0.01413 \pm 4.604 \sqrt{\frac{0.01940}{16,799}} \quad [1]$$

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:

$$\hat{y}_{365} = 2.4805 + 0.01413 \times 365 = 7.638 \quad [1]$$

The distribution of $\frac{y_{365} - \hat{y}_{365}}{s_{365}}$ is t_4 , where:

$$s_{365}^2 = \left[\frac{1}{n} + \frac{(365 - \bar{x})^2}{S_{xx}} \right] \hat{\sigma}^2 = \left[\frac{1}{6} + \frac{[365 - (475/6)]^2}{16799} \right] \times 0.01940 = 0.09758 \quad [1]$$

So a symmetrical 95% confidence interval for y_{365} is:

$$y_{365} = 7.638 \pm 2.776\sqrt{0.09758} = 7.638 \pm 0.867 \quad [1]$$

and the corresponding confidence interval for P_{365} is:

$$(e^{6.771}, e^{8.505}) = (870, 4940) \quad [1]$$

(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 $\frac{y_{365}^* - \hat{y}_{365}}{s_{365}^*}$ has a t_4 distribution, where:

$$s_{365}^{*2} = \left[1 + \frac{1}{n} + \frac{(365 - \bar{x})^2}{S_{xx}} \right] \hat{\sigma}^2 = s_{365}^2 + \hat{\sigma}^2 = 0.09758 + 0.01940 = 0.11698 \quad [1]$$

This gives a confidence interval of:

$$y_{365}^* = 7.638 \pm 2.776\sqrt{0.11698} = 7.638 \pm 0.949 \quad [1]$$

So the confidence interval for P_{365}^* is:

$$(e^{6.689}, e^{8.587}) = (800, 5360) \quad [1]$$

Solution 4.28(i)(a) ***Least squares estimates equations***

We need to minimise the expression $R = \sum (y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2}))^2$. [1]

To do this, we need to differentiate the expression with respect to the parameters and set the expressions equal to zero:

$$\begin{aligned}\frac{\partial R}{\partial \alpha} &= -2 \sum (y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})) = 0 \\ \Rightarrow \sum y_i &= n\alpha + \beta_1 \sum x_{i1} + \beta_2 \sum x_{i2}\end{aligned}\quad \text{eqn (1)}$$

$$\begin{aligned}\frac{\partial R}{\partial \beta_1} &= -2 \sum x_{i1} (y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})) = 0 \\ \Rightarrow \sum y_i x_{i1} &= \alpha \sum x_{i1} + \beta_1 \sum x_{i1}^2 + \beta_2 \sum x_{i2} x_{i1}\end{aligned}\quad \text{eqn (2)}$$

$$\begin{aligned}\frac{\partial R}{\partial \beta_2} &= -2 \sum x_{i2} (y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})) = 0 \\ \Rightarrow \sum y_i x_{i2} &= \alpha \sum x_{i2} + \beta_1 \sum x_{i1} x_{i2} + \beta_2 \sum x_{i2}^2\end{aligned}\quad \text{eqn (3)}\quad [3]$$

(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:

$$47 \times (1) - (2) \Rightarrow 47.12 = -157.68\beta_1 + 780.64\beta_2 \quad \text{eqn (4)}$$

$$5 \times (3) - 207.8 \times (1) \Rightarrow 1,747.44 = -3,903.2\beta_1 + 21,251.26\beta_2 \quad \text{eqn (5)}$$

$$157.68 \times (5) - 3,903.2 \times (4) \Rightarrow \beta_2 = 0.301468 \quad [1]$$

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) ***Predict the percentage effectiveness***

Substituting in the values given:

$$\hat{y} = 25.31 + (1.194 \times 51.3) + (0.3015 \times 18.3) = 92.1\% \quad [1]$$

Solution 4.29

(i) ***ANOVA***

We are testing the hypotheses:

H_0 : There is *no* difference between the mean claim amounts (per £100 sum assured) of each company.

H_1 : There *is* a difference between the mean claim amounts (per £100 sum assured) of each company.

Using the formulae given on page 26 of the *Tables*, we have:

$$SS_T = 73,471 - \frac{921^2}{12} = 2,784.25 \quad [1]$$

$$SS_B = \left(\frac{360^2}{4} + \frac{273^2}{4} + \frac{288^2}{4} \right) - \frac{921^2}{12} = 1,081.5 \quad [1]$$

$$SS_R = 2,784.25 - 1,081.5 = 1,702.75$$

The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between companies	2	1081.5	540.75
Residual	9	1,702.75	189.19
Total	11	2,784.25	

[1]

The variance ratio is $F = \frac{540.75}{189.19} = 2.858$. [1]

Under H_0 , this has a $F_{2,9}$ distribution.

The 5% critical point is 4.256, so we have insufficient evidence to reject H_0 at the 5% level. Therefore we conclude that there is *no* difference in mean claim amounts between companies. [1]

(ii) ***Confidence interval for mean of Company B***

Using:

$$\frac{\bar{Y}_{B\bullet} - \mu_B}{\hat{\sigma}/\sqrt{n_B}} \sim t_{n-k} \quad [\frac{1}{2}]$$

we have:

$$\hat{\sigma}^2 = \frac{SS_R}{n-k} = \frac{1,702.75}{9} = 189.19 \quad [\frac{1}{2}]$$

and $n_B = 4$ and $\bar{y}_{B\bullet} = 273/4 = 68.25$. [½]

Hence, a 95% confidence interval for μ_B is given by:

$$68.25 \pm 2.262 \sqrt{\frac{189.19}{4}} \quad [\frac{1}{2}]$$

This gives a confidence interval of (52.7, 83.8). [1]

Solution 4.30(i)(a) *Estimate overall mean*

$$\hat{\mu} = \bar{y}_{..} = \frac{1}{n} \sum_i \sum_j y_{ij} \Rightarrow \hat{\mu} = \frac{10 \times 5 + 6 \times 7 + 8 \times 3}{24} = \frac{116}{24} = 4.8\dot{3} \quad [1]$$

(i)(b) *Estimate common underlying variance*

Now from the *Tables* on page 26, we have $\hat{\sigma}^2 = \frac{1}{n-k} SS_R$.

$$SS_T = \sum \sum y_{ij}^2 - \frac{1}{n} y_{..}^2 = (264 + 310 + 84) - \frac{116^2}{24} = 97.\dot{3} \quad [1]$$

$$SS_B = \sum \frac{1}{n_i} y_{i..}^2 - \frac{1}{n} y_{..}^2 = \left(\frac{50^2}{10} + \frac{42^2}{6} + \frac{24^2}{8} \right) - \frac{116^2}{24} = 55.\dot{3} \quad [1]$$

$$SS_R = SS_T - SS_B = 97.\dot{3} - 55.\dot{3} = 42$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n-k} SS_R = \frac{42}{24-3} = 2 \quad [1]$$

(ii) *ANOVA*

We are testing the hypotheses:

H_0 : Each drug has the same mean number of days before discharge

H_1 : There are differences between the mean number of days before discharge using different treatments

The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between treatments	2	55. <dot{3}< td=""><td>27.<dot{6}< td=""></dot{6}<></td></dot{3}<>	27. <dot{6}< td=""></dot{6}<>
Residual	21	42	2
Total	23	97. <dot{3}< td=""><td></td></dot{3}<>	

[1]

The variance ratio is $F = \frac{27.6}{2} = 13.8\dot{3}$. [1]

Under H_0 , this has an $F_{2,21}$ distribution.

The 1% critical point is 5.780, so we have sufficient evidence to reject H_0 at the 1% level. We conclude that there *are* differences between the mean number of days before discharge using the different treatments. [1]

(iii) ***Test between treatments A and B***

We are testing:

$$H_0: \mu_A = \mu_B \quad \text{vs} \quad H_1: \mu_A < \mu_B$$

Under H_0 , the statistic $\frac{(\bar{Y}_{A\bullet} - \bar{Y}_{B\bullet}) - (\mu_A - \mu_B)}{\hat{\sigma} \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$ has a t_{n-k} distribution. [1]

The observed value of the test statistic is:

$$\frac{(5 - 7) - 0}{\sqrt{2} \sqrt{\frac{1}{10} + \frac{1}{6}}} = -2.739 \quad [1]$$

The 1% critical value for a t_{21} distribution is -2.518 .

Since $-2.739 < -2.518$, we have sufficient evidence at the 1% level to reject H_0 . Therefore it is reasonable to conclude that the mean number of days before discharge under treatment A is significantly better than under treatment B. [1]

(iv) ***Advice to the hospital on the treatments***

The significant differences between pairs of treatments means that treatment C is significantly better than treatment A which is significantly better than treatment B. So if the hospital wishes to use the best treatment it should use treatment C. [1]

However, looking at the costing, treatment C will cost $3 \times £14.95 = £44.85$ before the patient is recovered, whereas treatment A will cost $5 \times £7.50 = £37.50$ and treatment B will cost $7 \times £5.85 = £40.95$. This means that treatment A is likely to be significantly cheaper in treating patients than treatment C. So if the hospital wants to minimise costs it should use treatment A. [1]

Solution 4.31(i) ***ANOVA***

We are testing the hypotheses:

H_0 : There is *no* difference between the mean number of years worked in each division before leaving

H_1 : There *is* a difference between the mean number of years worked in each division before leaving

Using the formulae given on page 26 of the *Tables*, we have:

$$SS_{TT} = 763 - \frac{103^2}{15} = 55.733 \quad [1]$$

$$SS_B = \left(\frac{31^2}{5} + \frac{25^2}{4} + \frac{47^2}{6} \right) - \frac{103^2}{15} = 9.35 \quad [1]$$

$$\Rightarrow SS_R = 55.73 - 9.35 = 46.383$$

The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between divisions	2	9.35	4.675
Residual	12	46.383	3.865
Total	14	55.733	

[1]

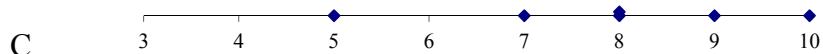
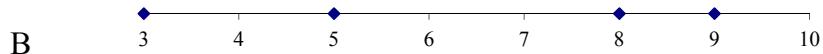
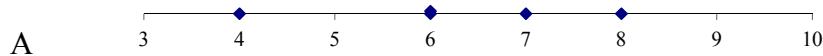
The variance ratio is $F = \frac{4.675}{3.8653} = 1.209$. [1]

Under H_0 this has an $F_{2,12}$ distribution.

This is less than the 5% critical point of 3.885, so we have insufficient evidence to reject H_0 at the 5% level, and conclude that there is no difference between the mean number of years worked in each division before leaving. [1]

(ii) ***Comment on the validity of the ANOVA test***

Line plots for the three divisions look like this:



[1]

A one-way analysis of variance requires that each of the populations is normally distributed with the same variance. Whilst Divisions A and C appear normally distributed with the same variance, Division B does not appear to have the same variance, although it is difficult to draw a definite conclusion when we have such a small amount of data. Also, the normal distribution can only be an approximation here since the data consists of whole numbers and so actually has a discrete distribution. [1]

(iii)(a) ***Estimate for underlying common variance***

This estimate is found directly from the ANOVA table. It is the figure of 3.865, estimated from the residual variance.

Alternatively, from the Tables on page 26, we have:

$$\hat{\sigma}^2 = \frac{1}{n-k} SS_R = \frac{46.383}{12} = 3.8653 \quad [1]$$

(iii)(b) ***Estimate for variance in Division B only***

The sample variance of Division B is:

$$s^2 = \frac{1}{3} [179 - 4 \times 6.25^2] = 7.5833 \quad [1]$$

(iii)(c) ***Test whether variance in Division B is significantly different***

We are testing:

$$H_0 : \sigma_2^2 = 3.8653 \quad \text{vs} \quad H_1 : \sigma_2^2 \neq 3.8653$$

Under H_0 , the statistic $\frac{(n-1)S^2}{\sigma^2}$ has a χ_3^2 distribution. The value of the test statistic is $\frac{3 \times 7.5833}{3.8653} = 5.886$. [1]

The 5% critical values for a χ_3^2 distribution are 9.348 and 0.2158. [1]

Since our test statistic lies between these we have insufficient evidence at the 5% level to reject H_0 . Therefore it is reasonable to conclude that $\sigma_2^2 = 3.8653$ (*ie* the variance in Division B is not significantly different from the underlying common variance). [1]

Solution 4.32(i) ***ANOVA***

We are testing the hypotheses:

H_0 : There is *no* difference between the mean claim sizes of each company.

H_1 : There *is* a difference between the mean claim sizes of each company.

Using the formulae given on page 26 of the *Tables*, we have:

$$SS_T = 4,072.78 - \frac{294^2}{24} = 471.28 \quad [1]$$

$$SS_B = \left(\frac{34.6^2}{5} + \frac{80.3^2}{6} + \frac{43.1^2}{4} + \frac{57.2^2}{4} + \frac{78.8^2}{5} \right) - \frac{294^2}{24} = 236.864 \quad [1]$$

$$SS_R = 471.28 - 236.864 = 234.416$$

The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between treatments	4	236.864	59.216
Residual	19	234.416	12.338
Total	23	471.28	

[1]

$$\text{The variance ratio is } F = \frac{59.216}{12.338} = 4.80.$$

[1]

Under H_0 , this has a $F_{4,19}$ distribution.

The 1% critical point is 4.5, so we have very strong evidence to reject H_0 , and we conclude that the mean claim sizes in the companies are different. [1]

(ii)(a) ***Estimates of mean claim amounts***

Simply using the treatment sample means, we get:

$$\bar{y}_{A\bullet} = 6.92 \quad \bar{y}_{B\bullet} = 13.383 \quad \bar{y}_{C\bullet} = 10.775 \quad \bar{y}_{D\bullet} = 14.3 \quad \bar{y}_{E\bullet} = 15.76 \quad [2]$$

(ii)(b) ***Analyse the mean claim sizes using a least significant difference approach***

Writing the sample means in order:

$$\bar{y}_{A\bullet} < \bar{y}_{C\bullet} < \bar{y}_{B\bullet} < \bar{y}_{D\bullet} < \bar{y}_{E\bullet}$$

$$\text{Now } \hat{\sigma}^2 = \frac{SS_R}{n-k} = \frac{234.416}{19} = 12.338.$$

The least significant difference between $\bar{y}_{A\bullet}$ and $\bar{y}_{C\bullet}$ is:

$$t_{19,0.025} \hat{\sigma} \sqrt{\frac{1}{n_A} + \frac{1}{n_C}} = 2.093 \times \sqrt{12.338} \times \sqrt{\frac{1}{5} + \frac{1}{4}} = 4.932 \quad [1]$$

Since $\bar{y}_{C\bullet} - \bar{y}_{A\bullet} = 10.775 - 6.92 = 3.855$, there is *no* significant difference between Companies A and C. Hence:

$$\underline{\bar{y}_{A\bullet} < \bar{y}_{C\bullet} < \bar{y}_{B\bullet} < \bar{y}_{D\bullet} < \bar{y}_{E\bullet}} \quad [1]$$

For the next pair, $\bar{y}_{C\bullet}$ and $\bar{y}_{B\bullet}$, the least significant difference is:

$$2.093 \times \sqrt{12.338} \times \sqrt{\frac{1}{4} + \frac{1}{6}} = 4.746 \quad [1/2]$$

Since $\bar{y}_{B\bullet} - \bar{y}_{C\bullet} = 13.383 - 10.775 = 2.608$, there is *no* significant difference between Companies C and B. Hence:

$$\underline{\bar{y}_{A\bullet}} < \underline{\bar{y}_{C\bullet}} < \underline{\bar{y}_{B\bullet}} < \underline{\bar{y}_{D\bullet}} < \bar{y}_E. \quad [1/2]$$

Now we will see if there is any significant difference between $\bar{y}_{A\bullet}$ and $\bar{y}_{B\bullet}$ (and whether we can combine A, B and C into one group).

The least significant difference for $\bar{y}_{A\bullet}$ and $\bar{y}_{B\bullet}$ is 4.452. Since $\bar{y}_{B\bullet} - \bar{y}_{A\bullet} = 13.383 - 6.92 = 6.463$ the diagram remains as before. [1]

Examining the next pair, $\bar{y}_{B\bullet}$ and $\bar{y}_{D\bullet}$, the least significant difference is 4.746. Since $\bar{y}_{D\bullet} - \bar{y}_{B\bullet} = 14.3 - 13.383 = 0.917$, there is *no* significant difference between Companies B and D. Hence:

$$\underline{\bar{y}_{A\bullet}} < \underline{\bar{y}_{C\bullet}} < \underline{\bar{y}_{B\bullet}} < \underline{\bar{y}_{D\bullet}} < \bar{y}_E. \quad [1/2]$$

Now we will see if we can combine C, B and D into one group. The least significant difference for $\bar{y}_{D\bullet}$ and $\bar{y}_{C\bullet}$ is 5.1982. Since $\bar{y}_{D\bullet} - \bar{y}_{C\bullet} = 14.3 - 10.775 = 3.525$ we can combine C, B and D into one group:

$$\underline{\bar{y}_{A\bullet}} < \underline{\bar{y}_{C\bullet}} < \underline{\bar{y}_{B\bullet}} < \underline{\bar{y}_{D\bullet}} < \bar{y}_E. \quad [1/2]$$

For the next pair, $\bar{y}_{D\bullet}$ and $\bar{y}_{E\bullet}$, the least significant difference is 4.932. Since $\bar{y}_{E\bullet} - \bar{y}_{D\bullet} = 15.76 - 14.3 = 1.46$, there is *no* significant difference between Companies D and E. Hence:

$$\underline{\bar{y}_{A\bullet}} < \underline{\bar{y}_{C\bullet}} < \underline{\bar{y}_{B\bullet}} < \underline{\bar{y}_{D\bullet}} < \bar{y}_E. \quad [1/2]$$

Finally, we shall see if we can combine C, B, D and E into one group. The least significant difference for $\bar{y}_{E\bullet}$ and $\bar{y}_{C\bullet}$ is 4.932. Since $\bar{y}_{E\bullet} - \bar{y}_{C\bullet} = 4.985$ the diagram remains as before. [1/2]

Part 5 – Revision Questions

This part contains 100 marks of questions testing the material from the whole course. You may like to try these questions under exam conditions as a mock exam.

Question 5.1

Show that S^2 , the sample variance of a random sample, is an unbiased estimator of the population variance σ^2 . [3]

Question 5.2

- (i) In a simple linear regression model the correlation coefficient has been calculated to be -0.28 . Describe what this tells you about the relationship between the response and the explanatory variable. [1]
 - (ii) Calculate the coefficient of determination in this case. Explain what the coefficient of determination measures. [2]
- [Total 3]

Question 5.3

The random variable N has a Poisson distribution with mean λ . Y is a continuous random variable such that the conditional distribution of $Y | N$ is uniform on the range $(-N, 2N)$. Find expressions for $E[Y]$ and $\text{var}[Y]$. [3]

Question 5.4

A small financial consultancy with 120 clients has estimated the probability that any given existing client will take out a new policy in any given year to be 42%. Calculate the approximate probability that in the next year the consultancy will issue more than 62 new policies to its existing clients, stating any assumptions you make. [4]

Question 5.5

An insurance company wishes to estimate the mean age of a particular group of its pension policyholders. Past experience has suggested the standard deviation of their ages is 5.28. The company wants to sample policyholders in order to construct a 90% confidence interval for the mean age of maximum width 2.5. Calculate the minimum sample size required to achieve this. [3]

Questions 6 to 9 relate to a customer survey carried out by a television company, in which a random sample of 50 men and 50 women were asked six questions. The questionnaire and a summary of the results are given on the data sheet on pages 7/8.

Question 5.6

- (i) Show that, if X and Y are random variables and c is an arbitrary constant, then $\text{var}(X - c) = \text{var}(X)$ and $\text{cov}(X - c, Y - c) = \text{cov}(X, Y)$. [4]
 - (ii) Calculate the correlation coefficient for the replies to the two opinion questions and interpret your answer. (You are not required to carry out a statistical test.) [4]
- [Total 8]

Question 5.7

- (i) Stating clearly any assumptions you make, carry out statistical tests to determine whether, for the population being studied:
 - (a) individuals under 40 watch more television on average than those over 40
 - (b) the average number of hours of viewing amongst the over 40s is less than the national average of 17.5 hours. [5]
 - (ii) State clearly one reason why the assumption(s) you made in (i) may not be valid. [1]
- [Total 6]

Question 5.8

- (i) Sketch a graph of the cumulative distribution function for the salary dataset (both sexes combined), assuming that all salaries lie in the range (£0, £100,000) and that they are distributed uniformly within each range. [2]
- (ii) Hence estimate the median and the interquartile range for the salaries. [3]
- (iii) Estimate the mean salary using the same assumptions and comment on your answer. [3]
- [Total 8]

Question 5.9

Carry out a statistical test to determine whether the salary levels can be considered to be independent of sex. [4]

Question 5.10

The sizes of claims made to an insurance company have a lognormal distribution with mean £1,800 and standard deviation £280.

- (i) Calculate the probability that a claim is more than £2,000. [4]
- An actuary goes through a pile of claims noting those which are in excess of £2,000.
- (ii) Calculate the probability that she examines eight claims before she finds three that are in excess of £2,000. [2]
- [Total 6]

Question 5.11

A random variable X has a Gamma distribution with parameters α and λ .

- (i) Show that the moment generating function of X , $M_X(t)$, is given by:

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-\alpha}, \quad t < \lambda \quad [3]$$

- (ii) Use this moment generating function to show that the mean and variance of X are given by $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$, respectively. [2]

- (iii) In the case when 2α is an integer, prove that $2\lambda X$ is a $\chi_{2\alpha}^2$ random variable. [3]
[Total 8]

Question 5.12

- (i) Define the size and power of a statistical test. [2]

An actuary is comparing a set of observed mortality rates with the values predicted by a certain formula. He has calculated the “standardised residuals” for 25 different ages. If the formula is satisfactory these residuals should have a standard normal distribution.

He is assuming that the residuals X_1, \dots, X_{25} are independent identically distributed random variables from a $N(\mu, 1)$ distribution and he wishes to carry out a test of the hypotheses $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$.

- (ii) Write down the test criterion the actuary should use for a symmetrical test of size 5%. [2]

- (iii) Calculate the power of this test if the true value of μ is:

- (a) 0 (b) 0.5 (c) -0.2 [3]

- (iv) Hence sketch a graph of the power function for this test as a function of μ . [2]
[Total 9]

Question 5.13

The number of new customers generated per month by different branches of a small building society is being monitored for employee bonus purposes. Head Office has collated the figures sent in by four branches over recent months, which are as follows:

Branch 1: 11, 5, 4, 9, 3, 0

Branch 2: 9, 7, 6, 8, 12

Branch 3: 5, 4, 5, 6, 0, 8, 6

Branch 4: 7, 8, 12, 0, 1, 15, 6

There are different numbers of figures because of incomplete data being sent to Head Office.

- (i) You have been asked to investigate whether there is any difference between the mean number of new customers generated in each branch. Carry out an appropriate test and explain your conclusion. [6]
 - (ii) State the assumptions you made in this analysis. [2]
- [Total 8]

Question 5.14

An actuarial student runs a stall at a children's charity fair every year. He wants to decide an appropriate fee to charge for next year. Each child is given three quoits, which they have to throw and try to "hoop" a peg. Any child who gets three hoops over the peg wins a prize. The frequencies of successful throws last year (which the student diligently recorded) were as follows:

<i>Number of successful throws</i>	0	1	2	3
<i>Frequency</i>	12	25	39	22

The probability of any particular throw being successful can be considered to have a constant value p for all children, and the outcome of different throws can be considered to be statistically independent.

- (i) Derive the maximum likelihood estimate of p . [6]
 - (ii) Find the standard error of p . [2]
 - (iii) Find a symmetrical 90% confidence interval for p . [2]
- [Total 10]

Question 5.15

A revolutionary study programme has been developed to help students memorise facts before exams. The authors of the programme state that if you study their material for several days, your performance in standardised tests will improve.

A publishing company wants you to investigate the authors' claims, and has provided you with the following information, which shows the average scores obtained by eight random samples of five students, with each sample being given a different number of days to study the materials:

Number of days spent studying (x)	1	2	3	4	5	6	7	8
Average score (%) in test (y)	25	28	35	48	50	61	71	79

You are given the following information:

$$\sum x = 36, \sum y = 397, \sum x^2 = 204, \sum y^2 = 22,441, \sum xy = 2,123$$

- (i) Plot a scatter diagram of the data. Comment on the appropriateness of using a linear regression model here. [3]
 - (ii) One of the authors has claimed that the correlation coefficient is more than 0.9. Test this claim at the 5% level. [5]
 - (iii) Find the equation of the least squares linear regression line. [2]
 - (iv) Find a 95% confidence interval for the slope parameter β . [3]
 - (v) Find a 99% confidence interval for the mean score you could expect in the test if you studied the material for 4 days. [3]
 - (vi) Comment on the claim “Study this material for a fortnight and you'll perform really well in tests”. [1]
- [Total 17]

DATA SHEET**Questionnaire***Opinions*

Q1: Do you agree with the statement “Advertising is a good thing.” (Statement 1)?

- 1 = Disagree
- 2 = Don’t Know
- 3 = Agree

Q2: Do you agree with the statement “I am prepared to pay to watch television” (Statement 2)?

- 1 = Disagree
- 2 = Don’t Know
- 3 = Agree

TV viewing habits

Q3 How many hours of television do you watch in an average week?

Personal details

Q4: Please indicate your sex.

- Male
- Female

Q5: Please indicate your age group.

- Under 40
- 40 or over

Q6: Please indicate the range in which your annual salary (before tax) falls.

- Under £10,000
- £10,001 to £25,000
- £25,001 to £50,000
- Over £50,000

Summary of results

Opinions

		<i>Statement 2</i>		
		<i>Disagree (=1)</i>	<i>Don't Know (=2)</i>	<i>Agree (=3)</i>
<i>Statement 1</i>	<i>Disagree (=1)</i>	0	10	10
	<i>Don't Know (=2)</i>	15	30	5
	<i>Agree (=3)</i>	25	0	5

Analysis of TV viewing

<i>TV viewing</i>	<i>Hours per week</i>		
	<i><40</i>	<i>>40</i>	<i>All ages</i>
<i>Number</i>	40	60	100
<i>Mean</i>	25	15	19
<i>Standard deviation</i>	15	10	12

Analysis of salaries

<i>Annual salary, x</i>	<i>Number</i>		
	<i>M</i>	<i>F</i>	<i>Total</i>
$x < £10,000$	6	14	20
$£10,000 \leq x < £25,000$	25	25	50
$£25,000 \leq x < £50,000$	15	10	25
$£50,000 \leq x$	4	1	5

Part 5 – Revision Solutions

Solution 5.1

The sample variance $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ or $\frac{1}{n-1} (\sum X_i^2 - n\bar{X}^2)$ is an unbiased estimator of σ^2 if $E[S^2] = \sigma^2$. So let us consider $E[S^2]$:

$$\begin{aligned} E[S^2] &= E\left[\frac{1}{n-1} (\sum X_i^2 - n\bar{X}^2)\right] \\ &= \frac{1}{n-1} \left(\sum E[X_i^2] - nE[\bar{X}^2] \right) \end{aligned} \quad [1]$$

Now $E[X^2] = \text{var}[X] + E^2[X] = \sigma^2 + \mu^2$. Similarly, $E[\bar{X}^2] = \frac{\sigma^2}{n} + \mu^2$. [1]

$$\begin{aligned} E[S^2] &= \frac{1}{n-1} \left[\sum (\mu^2 + \sigma^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) \right] \\ &= \frac{1}{n-1} \left[n(\mu^2 + \sigma^2) - \sigma^2 - n\mu^2 \right] \\ &= \frac{1}{n-1} (n-1)\sigma^2 = \sigma^2 \end{aligned}$$

Therefore S^2 is unbiased. [1]

Solution 5.2(i) ***Relationship***

This value indicates a negative correlation between the response and the explanatory variable, *ie* when a linear relationship is assumed, the response tends to decrease when the explanatory variable increases, and conversely. [1]

(ii) ***Coefficient of determination***

For the simple linear regression model, the coefficient of determination R^2 is given by $R^2 = r^2$.

The coefficient of determination is therefore $(-0.28)^2 = 0.0784$. [1]

The coefficient of determination quantifies the proportion of the total variability (in the responses) that is explained by the model. [1]

Solution 5.3

We can apply the identities:

$$E[Y] = E(E[Y|N]) \text{ and } \text{var}[Y] = E[\text{var}(Y|N)] + \text{var}(E[Y|N])$$

We are given that:

$$N \sim \text{Poisson}(\lambda) \text{ and } Y|N \sim U(-N, 2N)$$

So:

$$E[Y|N] = \frac{1}{2}(2N - N) = \frac{1}{2}N$$

and:

$$\text{var}[Y|N] = \frac{1}{12}(2N + N)^2 = \frac{3}{4}N^2 \quad [1]$$

Using the formulae for the mean and variance of the $U(a,b)$ from the Tables.

Therefore:

$$E[Y] = E\left(\frac{1}{2}N\right) = \frac{1}{2}E[N] = \frac{1}{2}\lambda \quad [1]$$

and:

$$\begin{aligned} \text{var}[Y] &= E\left[\frac{3}{4}N^2\right] + \text{var}\left[\frac{1}{2}N\right] \\ &= \frac{3}{4}E[N^2] + \frac{1}{4}\text{var}[N] = \frac{3}{4}(\lambda + \lambda^2) + \frac{1}{4}\lambda = \lambda + \frac{3}{4}\lambda^2 \end{aligned} \quad [1]$$

Solution 5.4

The number of new policies (X) taken out in the next year by current clients is distributed binomially with parameters $n = 120$ and $p = 0.42$.

This assumes that clients behave independently of each other and independently from one year to the next. [1]

We want $P(X > 62)$. Since n is large and p is close to 0.5, the normal approximation will provide a good estimate here.

$$X \sim N(50.4, 29.232) \quad [1]$$

We must use a continuity correction since we are using a continuous distribution to approximate a discrete one.

The probability is calculated as:

$$P(X > 62.5) = P\left(Z > \frac{62.5 - 50.4}{\sqrt{29.232}}\right) = P(Z > 2.24) = 1 - 0.98745 = 0.01255 \quad [2]$$

Solution 5.5

The confidence interval will be of the form $\bar{x} \pm 1.6449 \frac{\sigma}{\sqrt{n}}$ where $\sigma = 5.28$.

The width is $2 \times 1.6449 \times \frac{\sigma}{\sqrt{n}}$, so: [1]

$$2 \times 1.6449 \times \frac{\sigma}{\sqrt{n}} = 2.5 \Rightarrow n = \left(\frac{2 \times 1.6449 \times 5.28}{2.5} \right)^2 = 48.28 \quad [1]$$

So the sample should consist of at least 49 policyholders. [1]

Solution 5.6(i) **Proof**

There are two ways of proving this result:

Method 1

If we start from the most basic definition of variance and covariance, we have:

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

$$\begin{aligned}\text{cov}(X - c, Y - c) &= E[\{(X - c) - E(X - c)\} \{(Y - c) - E(Y - c)\}] \\ &= E[\{X - c - [E(X) - c]\} \{Y - c - [E(Y) - c]\}] \\ &= E[\{X - E(X)\} \{Y - E(Y)\}] = \text{cov}(X, Y)\end{aligned}\quad [2]$$

Method 2

Alternatively, we can start from the formula usually used for finding the variance and covariance:

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

$$\begin{aligned}\text{cov}(X - c, Y - c) &= E[(X - c)(Y - c)] - [E(X - c)][E(Y - c)] \\ &= E(XY - cX - cY + c^2) - [E(X) - c][E(Y) - c] \\ &= E(XY) - cE(X) - cE(Y) + c^2 \\ &\quad - \{E(X)E(Y) - cE(X) - cE(Y) + c^2\} \\ &= E(XY) - E(X)E(Y) = \text{cov}(X, Y)\end{aligned}\quad [2]$$

(ii) ***Correlation coefficient***

We can use the formulae proved in part (i) to simplify the calculations. If we subtract 2 from all the values, most will become zero and will drop out of the calculation.

If we refer to the responses for Statements 1 and 2 as X and Y respectively, the table becomes:

		Y		
		-1	0	+1
X	-1	0	10	10
	0	15	30	5
	1	25	0	5

We can then calculate:

$$\Sigma x = -1 \times 20 + 1 \times 30 = 10 \quad \Sigma y = -1 \times 40 + 1 \times 20 = -20$$

$$\Sigma x^2 = (-1)^2 \times 20 + 1^2 \times 30 = 50 \Rightarrow s_{xx} = 50 - \frac{10^2}{100} = 49$$

$$\Sigma y^2 = -1^2 \times 40 + 1^2 \times 20 = 60 \Rightarrow s_{yy} = 60 - \frac{(-20)^2}{100} = 56$$

$$\Sigma xy = -1 \times 1 \times 10 + 1 \times -1 \times 25 + 1 \times 1 \times 5 = -30$$

$$\Rightarrow s_{xy} = -30 - \frac{10 \times (-20)}{100} = -28$$

and the correlation coefficient is:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{-28}{\sqrt{49 \times 56}} = -0.535 \quad [3]$$

If you don't apply the "subtract 2" trick, the figures work out as:

$$\Sigma x = 210, \Sigma y = 180, \Sigma x^2 = 490, \Sigma y^2 = 380, \Sigma xy = 350$$

Note that in this question we were calculating the sample correlation coefficient (r), not the correlation coefficient (ρ) for an entire population where the joint distribution was known. So you should follow the approach given above (involving s_{xy} etc), rather than calculating $\text{cov}(X, Y)$ etc. (In fact, in this example, both approaches would give the same answer.)

This shows that the responses are negatively correlated, ie people tend to have opposite opinions on the two statements. [1]

Solution 5.7

(i)(a) **Test difference between 2 means**

Here we need to apply a two-sample t test to the hypotheses:

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_1 : \mu_X > \mu_Y$$

where X and Y denote the under 40s and over 40s, respectively.

This test assumes that the populations are normal with a common variance.

The test statistic is $t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_p^2 \left(\frac{1}{m} + \frac{1}{n} \right)}}$, has a t_{m+n-2} distribution under H_0 .

Here the observed value of the pooled sample variance is:

$$s_p^2 = \frac{39 \times 15^2 + 59 \times 10^2}{39 + 59} = 149.74 \quad [1]$$

So the observed value of the test statistic is:

$$t = \frac{25 - 15}{\sqrt{149.74 \left(\frac{1}{40} + \frac{1}{60} \right)}} = 4.00 \quad [1]$$

For a one-sided test this corresponds to a p value of 0.013%. (The t_{98} distribution is very similar to the standard normal distribution.). So we can confidently reject the null hypothesis and conclude that the under 40s watch more TV than the over 40s. [1]

(i)(b) ***Test mean***

Here we need to apply a one-sample t test to the hypotheses:

$$H_0: \mu_Y = 17.5 \quad vs \quad H_1: \mu_Y < 17.5$$

This test assumes that the population is normal.

The test statistic is $t = \frac{\bar{Y} - \mu_Y}{S_Y / \sqrt{n}}$, which has a t_{n-1} distribution under H_0 .

So here the observed value of the test statistic is:

$$t = \frac{15 - 17.5}{10 / \sqrt{60}} = -1.94$$

For a one-sided test this corresponds to a p value just over 2½%. (From the *Tables* the 2½% point of the t_{59} distribution would be just over 2.000.) We can reject the null hypothesis at the 5% level and conclude that the over 40s watch less TV than the national average. [2]

(ii) ***Assumptions not valid***

There may be a proportion of the population who don't watch TV at all (eg if they don't own a TV). Having a non-zero probability associated with the value 0 would be inconsistent with a normal distribution.

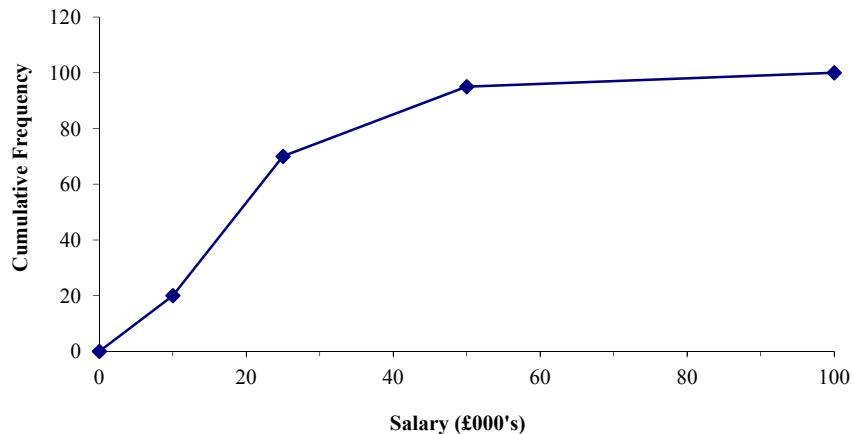
There may be some "TV-aholics" who watch TV for a disproportionate number of hours, implying a positively skewed distribution that would be inconsistent with a normal assumption.

If the hours of viewing followed a normal distribution this would imply a positive probability associated with a negative number of hours!

[1 mark for any of these]

Solution 5.8(i) **Cumulative distribution graph**

The empirical (observed) distribution function looks like this:



[2]

(ii) **Median and IQR**

The median value splits the distribution into two equal halves, so that there are 50 observations below the median and 50 observations above the median. The median lies in the £10,000 ≤ x < £25,000 group. Using linear interpolation gives:

$$\text{£}10,000 + \frac{30}{50} \times (\text{£}25,000 - \text{£}10,000) = \text{£}19,000 \quad [1]$$

Alternatively, using the $\frac{1}{2}(n+1)$ formula we get:

$$\text{£}10,000 + \frac{30.5}{50} \times (\text{£}25,000 - \text{£}10,000) = \text{£}19,150$$

Similarly, the quartiles split the distribution into quarters. The lower quartile lies in the £10,000 ≤ x < £25,000 group and is approximately:

$$\text{£}10,000 + \frac{5}{50} \times (\text{£}25,000 - \text{£}10,000) = \text{£}11,500 \quad [\frac{1}{4}]$$

And the upper quartile lies in the $\text{£}25,000 \leq x < \text{£}50,000$ group and is approximately:

$$\text{£}25,000 + \frac{5}{25} \times (\text{£}50,000 - \text{£}25,000) = \text{£}30,000 \quad [1\frac{1}{2}]$$

So the interquartile range is $\text{£}30,000 - \text{£}11,500 = \text{£}18,500$. [1]

Alternatively, using the $\frac{1}{4}n + \frac{1}{2}$ and $\frac{3}{4}n + \frac{1}{2}$ formulae we get an interquartile range of:

$$\text{£}30,500 - \text{£}11,650 = \text{£}18,850. \quad [1]$$

Finally, using the $\frac{1}{4}n + \frac{1}{4}$ and $\frac{3}{4}n + \frac{3}{4}$ formulae we get an interquartile range of:

$$\text{£}30,750 - \text{£}11,575 = \text{£}19,175.$$

(iii) **Mean**

If the values are distributed uniformly within each band, the mean of each band will equal the midpoint value. So the overall mean is:

$$\frac{(20 \times 5,000) + (50 \times 17,500) + (25 \times 37,500) + (5 \times 75,000)}{100} = \text{£}22,875 \quad [2]$$

The mean is significantly higher than the median because the distribution of salaries is positively skewed. [1]

Solution 5.9

We can use a chi squared (contingency table) test to test the hypotheses:

$$H_0 : \text{Salaries are independent of sex}$$

vs $H_1 : \text{Salaries are not independent of sex.}$

The test statistic $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ has a $\chi^2_{(r-1)(c-1)}$ distribution under H_0 .

The expected numbers can be calculated from the row and column totals. (In fact the expected numbers for each sex are just half of the corresponding row totals.)

	Annual salary	Actual (Expected) Number		
		M	F	Total
	$x < £10,000$	6 (10)	14 (10)	20
	$£10,000 \leq x < £25,000$	25 (25)	25 (25)	50
	$£25,000 \leq x < £50,000$	15 (12.5)	10 (12.5)	25
	$£50,000 \leq x$	4 (2.5)	1 (2.5)	5
	<i>Total</i>	50	50	100

[1]

We must first combine the “small” groups (in the bottom row) so that the expected numbers will all exceed 5:

	Annual salary	Actual (Expected) Number		
		M	F	Total
	$x < £10,000$	6 (10)	14 (10)	20
	$£10,000 \leq x < £25,000$	25 (25)	25 (25)	50
	$£25,000 \leq x$	19 (15)	11 (15)	30
	<i>Total</i>	50	50	100

[1]

The chi squared statistic is then calculated as:

$$\chi^2 = \frac{(6-10)^2}{10} + \dots + \frac{(11-15)^2}{15} = 5.33 \quad [1]$$

Under H_0 the distribution of this test statistic has $(3-1) \times (2-1) = 2$ degrees of freedom. Since the observed value of 5.33 is below 5.991, the 95th percentile of this distribution, we cannot reject H_0 , ie we don't have sufficient evidence to show that salaries are dependent on sex. [1]

Solution 5.10

- (i) **Probability claim is more than £2,000**

Let X be the claim size, such that $X \sim \log N(\mu, \sigma^2)$.

$$\text{mean} = e^{\mu + \frac{1}{2}\sigma^2} = 1,800 \quad (1)$$

$$\text{variance} = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1) = 280^2 \quad [1]$$

Substituting equation (1) into equation (2) gives:

$$\begin{aligned} 1800^2 (e^{\sigma^2} - 1) &= 280^2 \\ e^{\sigma^2} &= 1.0242 \\ \sigma^2 &= 0.023909 \end{aligned} \quad [1]$$

Now substituting this value back into equation (1) gives:

$$\mu = \ln 1800 - \frac{1}{2} \times 0.023909 = 7.4836 \quad [1]$$

Hence, $X \sim \log N(7.4836, 0.023909)$, so:

$$P(X > 2,000) = P(\ln X > \ln 2,000)$$

Standardising, $z = \frac{\ln 2000 - 7.4836}{\sqrt{0.023909}} = 0.759$, so:

$$\begin{aligned} P(X > 2,000) &= P(Z > 0.759) \\ &= 1 - P(Z < 0.759) \\ &= 1 - 0.77607 = 0.224 \end{aligned} \quad [1]$$

- (ii) **Probability examine eight claims before three in excess of £2,000**

This is a negative binomial distribution with parameters $k = 3$ and $p = 0.224$. Hence:

$$P(X = 8) = {}^7C_2 \times 0.224^3 \times 0.776^5 = 0.0664 \quad [2]$$

Solution 5.11(i) **MGF**

Considering the definition of the MGF:

$$\begin{aligned}
 M_X(t) &= E(e^{tX}) = \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \\
 &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x} dx
 \end{aligned} \tag{1}$$

The integral looks like the PDF of a $Ga(\alpha, \lambda - t)$, so putting in the appropriate constants:

$$\begin{aligned}
 M_X(t) &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \int_0^\infty \frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} dx \\
 &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \quad \text{provided } t < \lambda \\
 &= \left(\frac{\lambda}{\lambda-t} \right)^\alpha = \left(\frac{\lambda-t}{\lambda} \right)^{-\alpha} = \left(1 - \frac{t}{\lambda} \right)^{-\alpha}
 \end{aligned} \tag{1}$$

Since the integral of a Gamma pdf over the whole range is 1.

(ii) **Mean and variance**Using the results $E(X) = M'_X(0)$ and $E(X^2) = M''_X(0)$:

$$M'_X(t) = \frac{\alpha}{\lambda} \left(1 - \frac{t}{\lambda} \right)^{-\alpha-1} \Rightarrow E(X) = M'_X(0) = \frac{\alpha}{\lambda} \tag{1}$$

$$M''_X(t) = \frac{\alpha(\alpha+1)}{\lambda^2} \left(1 - \frac{t}{\lambda} \right)^{-\alpha-2} \Rightarrow E(X^2) = M''_X(0) = \frac{\alpha(\alpha+1)}{\lambda^2}$$

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

(iii) ***Prove gamma-chi square relationship***

Let $Y = 2\lambda X$, then:

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E(e^{2\lambda tX}) = M_X(2\lambda t) \\ &= \left(1 - \frac{2\lambda t}{\lambda}\right)^{-\alpha} \\ &= (1 - 2t)^{-\alpha} \end{aligned} \quad [1]$$

Now $\chi_n^2 \sim Ga\left(\frac{n}{2}, \frac{1}{2}\right)$, so its MGF is given by:

$$M(t) = \left(1 - \frac{t}{\frac{1}{2}}\right)^{-\frac{n}{2}} = (1 - 2t)^{-\frac{n}{2}} \quad [1]$$

Comparing MGFs, $n = 2\alpha$ and so by the uniqueness property of MGFs $2\lambda X \sim \chi_{2\alpha}^2$ [1]

Solution 5.12(i) ***Size and power***

The size of a test is the probability of a Type I error, *ie* the probability of rejecting the null hypothesis when it is true. [1]

The power of a test is 1 minus the probability of a Type II error, *ie* the probability of rejecting the null hypothesis when it is false. Usually this is a function of the parameter whose value is being tested. [1]

(ii) ***Test criterion***

With this statistical model, the sample mean \bar{X} has a $N(\mu, \frac{1}{25})$ distribution. So under H_0 , $5\bar{X}$ has a $N(0,1)$ distribution. So, for a symmetrical two-sided test, we would reject H_0 if $5\bar{X}$ falls outside the interval $(-1.96, 1.96)$, *ie* if $|\bar{X}| > 0.392$. [2]

(iii) **Power**

We know that \bar{X} has a $N(\mu, \frac{1}{25})$ distribution, whatever the value of μ . So the power function, $\pi(\mu)$ say, is:

$$\begin{aligned}\pi(\mu) &= P(|\bar{X}| > 0.392) = P\left|N(\mu, \frac{1}{25})\right| > 0.392 \\ &= 1 - P[-0.392 < N(\mu, \frac{1}{25}) < 0.392] \\ &= 1 - P[5(-0.392 - \mu) < N(0,1) < 5(0.392 - \mu)] \\ &= 1 - [\Phi(1.96 - 5\mu) - \Phi(-1.96 - 5\mu)]\end{aligned}$$

Evaluating this function for the values of μ given, we get:

$$\pi(0) = 1 - [\Phi(1.96) - \Phi(-1.96)] = 1 - [0.97500 - 0.02500] = 0.05 \quad [1]$$

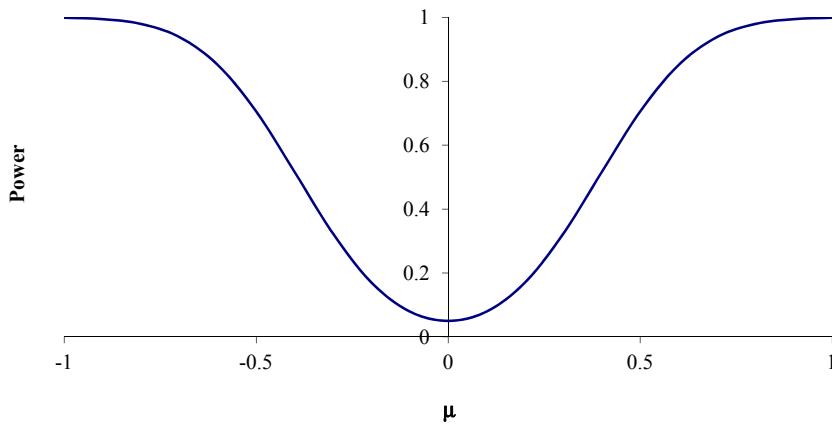
$$\pi(0.5) = 1 - [\Phi(-0.540) - \Phi(-4.460)] = 1 - [0.29460 - 0] = 0.705 \quad [1]$$

$$\pi(-0.2) = 1 - [\Phi(2.96) - \Phi(-0.96)] = 1 - [0.99846 - 0.16853] = 0.170 \quad [1]$$

It may seem paradoxical working out the power of the test (ie the probability of rejecting H_0) when $\mu = 0$, because this value of μ would mean that H_0 was correct, and hence we shouldn't be rejecting it! What we're actually doing is working out the probability of rejecting H_0 when μ is close to 0.

(iv) **Power function sketch**

The graph of the power function will approach 1 for values of μ a long way from zero and will look like this:



[2]

Solution 5.13(i) ***ANOVA***

We need to carry out an analysis of variance to test:

- H_0 : Each branch has the same mean number of new customers per month
 vs H_1 : There are differences between the mean numbers of new customers obtained by the different branches.

For this data set, we have:

$$\text{Number of treatments } k = 4$$

$$\text{Sample sizes } n_1 = 6 \quad n_2 = 5 \quad n_3 = 7 \quad n_4 = 7 \quad n = 25$$

Sample totals

$$y_{1\bullet} = 32 \quad y_{2\bullet} = 42 \quad y_{3\bullet} = 34 \quad y_{4\bullet} = 49 \quad y_{\bullet\bullet} = 157 \quad \sum \sum y_{ij}^2 = 1347 \quad [1]$$

$$SS_T = 1347 - \frac{157^2}{25} = 361.04 \quad [1]$$

$$SS_B = \left(\frac{32^2}{6} + \frac{42^2}{5} + \frac{34^2}{7} + \frac{49^2}{7} \right) - \frac{157^2}{25} = 45.650 \quad [1]$$

$$SS_R = 361.04 - 45.650 = 315.390$$

The ANOVA table is:

Source of variation	DF	Sum of squares	Mean squares
Between treatments	3	45.650	15.217
Residual	21	315.390	15.019
Total	24	361.04	

[1]

$$\text{The variance ratio is } F = \frac{15.217}{15.019} = 1.013. \quad [1]$$

Under H_0 , this has an $F_{3,21}$ distribution.

The 5% critical point is 3.072, so we cannot reject H_0 , and we conclude that the average number of new customers doesn't differ between branches. [1]

(ii) ***Assumptions***

The assumptions are that the underlying distribution is normal, there is a common variance between branches and that the samples have been drawn randomly and independently of each other. [2]

Since we have incomplete data, a particular concern here would be that some of the branches might have only submitted figures for the “better” months.

Solution 5.14(i) ***MLE***

The number of successful throws will have a binomial distribution with parameters $n = 3$ and p .

For each child, the probabilities of getting 0, 1, 2, and 3 successful throws are therefore $(1-p)^3$, $3p(1-p)^2$, $3p^2(1-p)$ and p^3 respectively.

The likelihood of the obtaining these results is therefore:

$$L = C[(1-p)^3]^{12} [3p(1-p)^2]^{25} [3p^2(1-p)]^{39} [p^3]^{22} \quad [1]$$

where C is a constant of proportionality.

We can simplify this and then take logs in order to make the differentiation easier:

$$L = D(1-p)^{3 \times 12 + 2 \times 25 + 39} p^{25 + 2 \times 39 + 3 \times 22} = D(1-p)^{125} p^{169} \quad [1]$$

where D is another constant.

$$\Rightarrow \ln L = \text{constant} + 125 \ln(1-p) + 169 \ln p$$

Differentiating:

$$\frac{d \ln L}{dp} = -\frac{125}{1-p} + \frac{169}{p} \quad [1]$$

Setting this equal to zero:

$$\begin{aligned} -\frac{125}{1-\hat{p}} + \frac{169}{\hat{p}} &= 0 \Rightarrow 169(1-\hat{p}) = 125\hat{p} \\ \Rightarrow 294\hat{p} &= 169 \Rightarrow \hat{p} = \frac{169}{294} \end{aligned} \quad [2]$$

We need to check that this is a maximum. Differentiating again:

$$\frac{d^2 \ln L}{dp^2} = -\frac{125}{(1-p)^2} - \frac{169}{p^2} < 0$$

So this value of p is a maximum. [1]

Alternatively, we can note that the likelihood function is positive (which it always must be) and that it equals zero if $p = 0$ or $p = 1$. Since there is only one turning point, it must be a maximum.

(ii) Standard error

The standard error is the square root of the Cramér-Rao Lower Bound:

$$CRLB = 1 \left/ -E \left[\frac{d^2 \ln L}{dp^2} \right] \right. = \left(\frac{125}{(1-0.5748)^2} + \frac{169}{0.5748^2} \right)^{-1} = 0.000831$$

So the standard error is $\sqrt{0.000831} = 0.0288$. [2]

(iii) Confidence interval

We know for a large sample that $\hat{p} \sim N(p, CRLB)$. So our confidence interval can be derived from the probability equation:

$$0.90 = P \left(-1.645 < \frac{\hat{p} - p}{\sqrt{CRLB}} < 1.645 \right)$$

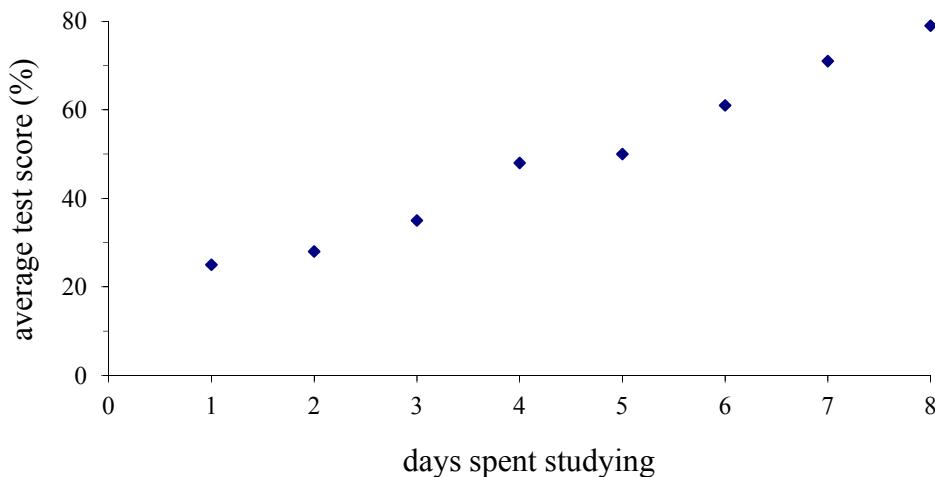
The 90% confidence interval has limits:

$$\hat{p} \pm 1.645\sqrt{CRLB} = 0.5748 \pm 1.645 \times 0.0288 = (0.527, 0.622) \quad [2]$$

Solution 5.15(i) **Scatter diagram**

The scatter diagram looks like this:

[2]



Yes. The points appear to follow a straight line and so a linear regression model appears to be appropriate.

[1]

(ii) **Test correlation coefficient**

We need to calculate the sample correlation coefficient:

$$s_{xy} = \sum xy - n \bar{x} \bar{y} = 2123 - 8 \times \frac{36}{8} \times \frac{397}{8} = 336.5$$

$$s_{xx} = \sum x^2 - n \bar{x}^2 = 204 - 8 \times \left(\frac{36}{8} \right)^2 = 42$$

$$s_{yy} = \sum y^2 - n \bar{y}^2 = 22,441 - 8 \times \left(\frac{397}{8} \right)^2 = 2,739.875$$

Substituting these values into the formula for the correlation coefficient:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{336.5}{\sqrt{42 \times 2,739.875}} = 0.992$$

[2]

We are testing:

$$H_0: \rho = 0.9 \quad \text{versus} \quad H_1: \rho > 0.9$$

We know that the distribution of the Fisher Z statistic is $Z \sim N\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$,

$$\text{where } Z = \frac{1}{2}\ln\frac{1+r}{1-r}.$$

Alternatively, you can calculate the transformed valued as $\tanh^{-1} \rho$ and $\tanh^{-1} r$.

The standardised value of our test statistic is:

$$\frac{\frac{1}{2}\ln\frac{1.992}{0.008} - \frac{1}{2}\ln\frac{1.9}{0.1}}{\sqrt{\frac{1}{5}}} = \frac{2.756 - 1.472}{\sqrt{0.2}} = 2.871 \quad [2]$$

Since this is greater than 1.645, we have sufficient evidence to reject H_0 at the 5% level and we conclude that the correlation coefficient is greater than 0.9. [1]

(iii) ***Regression line equation***

We need to calculate the values of $\hat{\alpha}$ and $\hat{\beta}$:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{336.5}{42} = 8.012 \quad [1]$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = \frac{397}{8} - 8.012 \times \frac{36}{8} = 13.571 \quad [1]$$

So the equation of the regression line is $y = 8.012x + 13.571$.

(iv) ***Slope parameter confidence interval***

We know that:

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$$

where $\hat{\sigma}^2 = \frac{1}{n-2} \left(s_{yy} - \frac{s_{xy}^2}{S_{xx}} \right) = \frac{1}{6} \left(2739.875 - \frac{336.5^2}{42} \right) = 7.312$. [1]

So our confidence interval can be derived from the probability equation:

$$0.95 = P \left(-2.447 < \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} < 2.447 \right)$$

This gives the confidence interval:

$$\hat{\beta} \pm 2.447 \sqrt{\hat{\sigma}^2 / S_{xx}} = (6.99, 9.03) [2]$$

(v) ***Mean score confidence interval***

The variance of the mean predicted response is:

$$\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \hat{\sigma}^2$$

The confidence interval is therefore given by:

$$8.012 \times 4 + 13.571 \pm 3.707 \sqrt{\frac{1}{8} + \frac{(4 - 4.5)^2}{42}} \sqrt{7.312} = (41.99, 49.25) [3]$$

(vi) ***Comment***

Trying to work out the score in the test after a fortnight's study, *ie* 14 days, relies upon extrapolation. It is dangerous to extrapolate because we cannot guarantee that the same linear relationship holds beyond the range of the data given in the question and any error in the estimated values of the parameters will be exaggerated. In this particular case we find that if we extrapolate to 11 days we will get a predicted score exceeding 100%, which is impossible! [1]

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 CT3: Assignment X1

2014 Examinations

Time allowed: 2½ hours

Instructions to the candidate

1. *Please note that assignment marking is not included in the price of the Course Materials. Please purchase Series Marking or Marking Vouchers before submitting your script.*
2. *We only accept the current version of assignments for marking, ie you can only submit this assignment in the sessions leading to the 2014 exams.*
3. *Attempt all of the questions, leaving space in the margin and beginning your answer to each question on a new page.*
4. *Write in black ink using a medium-sized nib because we will be unable to mark illegible scripts.*
5. ***Leave at least 2cm margin on all borders.***
6. *Attempt the questions as far as possible under exam conditions.*
7. *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 in the Study Guide for the 2014 exams, 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. ActEd will not be responsible for scripts lost or damaged in the post or for scripts received after the deadline date. 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.

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

There are three methods for you to submit your script, namely by *email*, by *fax* or by *post*.

If you are submitting by **email**:

- complete the cover sheet, including the checklist
- scan your script (and Marking Voucher if applicable) to a pdf document, then email it to: **ActEdMarking@bpp.com**.

Please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* “CT3 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 3 MB in size. **The file size cannot exceed 4 MB.**
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

If you are submitting by **fax**:

- only write on one side of the paper when completing the assignment
- complete the cover sheet, including the checklist
- fax your script (including cover sheet and Marking Voucher if applicable) to **0844 583 4501**.

In addition:

- We recommend that you stay by the fax machine until the fax has been sent so that you can deal with any problems immediately. (If an error occurs, please re-fax the whole script.)
- An email will be sent by the end of the next working day to confirm that we have processed your script. Please do not phone to check progress before then. If the fax was sent without error then it's very unlikely that there will be a problem.

We will **not** accept:

- scripts submitted to other ActEd fax numbers – please use **0844 583 4501**
- scripts that have been split over a number of faxes. (If an error occurs, please re-fax the whole script.)
- more than one script per fax
- jumbled scripts – please fax the pages in the correct order.

If you are submitting by **post**:

- complete the cover sheet, including the checklist.
- we recommend that you photocopy your script before posting, in case your script is lost in the post.
- post your script to: **First Floor, Kimber House, 1 Kimber Road, Abingdon, Oxfordshire, OX14 1BZ**
- please staple the cover sheet (and Marking Voucher if applicable) to the front of your assignment
- please do not staple more than one assignment together.

Subject CT3: Assignment X1

2014 Examinations

Please complete the following information:

Name:

Address:

ActEd Student Number (see Note below):

--	--	--	--	--

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting this number will help us to process your scripts quickly. If you do not complete this box, your script may be delayed. If you do not know your ActEd Student Number, please email ActEd@bpp.com. **Your ActEd Student Number is not the same as your Faculty/Institute 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:

Time to do assignment (see Note below): _____ hrs _____ mins

Under exam conditions (delete as applicable): yes / nearly / no

Note: If you spend more than 2½ hours on the assignment, you should indicate on the assignment how much you completed within this time so that the marker can provide useful feedback on your chances of success in the exam.

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
2	3	4	4	5	7	5	6	6	6	5	8	8	11	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, eg 2014 for the sessions leading to the 2014 exams?
- [] Written your full name and postal address 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?
- [] Attached your Marking Voucher or ordered Series X Marking?

Please follow the instructions on the previous page when submitting your script for marking.

***This page has been left blank in case you wish to submit your
script by fax.***

Feedback from marker

Notes on marker's section

The marker's main objective is to give you advice on how to improve your answers. The marker will also assess your script quantitatively and qualitatively. The percentage score gives you a quantitative assessment. The grade is a qualitative assessment of how your script might be classified in the exam. The grades are as follows:

A = Clear Pass B = Probable Pass C = Borderline D = Probable Fail E = Clear Fail

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

or when you submit your next script.

***This page has been left blank in case you wish to submit your
script by fax.***

Question 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 $\frac{1}{2}$
- (iii) the gamma distribution with $\alpha = 1$ and $\lambda = \frac{1}{2}$.

State with reasons whether the student is correct.

[2]

Question X1.2

Medical research has been carried out into the diagnosis of a certain disease. Doctors test the blood of a patient and look for high concentrations of any one of four chemicals thought to be indicators of the disease. By studying the general population, it is known that the probabilities of having high concentrations of the four chemicals are 0.4, 0.2, 0.1 and 0.3 respectively and that only one chemical is found in any one person.

It has been found that the probability of a patient having the disease if they have high concentrations of the first chemical is 0.2. The probabilities for the other chemicals are 0.4, 0.8 and 0.1 respectively.

If a patient has the disease, what is the probability that they have high concentrations of the third chemical in their blood?

[3]

Question X1.3

Suppose that the continuous random variable X has a uniform distribution on the interval (1,6).

- (i) Determine the probability density function of the random variable Y , where:

$$Y = \frac{X}{6-X} \quad [3]$$

- (ii) State the range of y values for which the density function of Y is valid. [1]

[Total 4]

Question X1.4

Pearson's measure of skewness is defined as:

$$\kappa = \frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

Find the value of κ for a gamma distribution with parameters $\alpha = 2.5$ and $\lambda = 0.4$. [4]

Question X1.5

The random variable X has a mixed distribution with the probability density function:

$$\begin{aligned} P(X = 0) &= \alpha && \text{at } x = 0 \\ f_X(x) &= \beta x^2(1-x) && \text{for } 0 < x < 1 \end{aligned}$$

- (i) Show that the total probability is given by:

$$\alpha + \frac{\beta}{12} \quad [1]$$

- (ii) Obtain an expression for the mean value of X . [1]

- (iii) Given that $E(X) = 0.4$, calculate α , β and the standard deviation of X . [3]

[Total 5]

Question X1.6

A random variable X has probability density function:

$$f(x) = \frac{2 \times 5^2}{(5+x)^3}, \quad x > 0$$

- (i) Obtain an expression for the distribution function, $F(x)$. [1]

- (ii) Hence, calculate the $P(X > 5 | X > 3)$. [3]

- (iii) Simulate two observations from the distribution using the random numbers 0.656 and 0.285 selected from the $U(0,1)$ distribution. [3]

[Total 7]

Question X1.7

A large life office has 1,000 policyholders, each with 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]

Question X1.8

The table below shows the approximate monetary equivalent of the annual pay and benefit packages (in £000s) of two independent samples of actuarial students, one for male students and one for female students, all of whom have passed 2 exams.

Male students	21	18	24	25	21	32	23	21
Female students	22	22	20	19	24	28	27	22

Draw boxplot diagrams for the pay packages of the male and female students. Use these to compare and contrast the two distributions.

[6]

Question X1.9

A company has a portfolio of 50 high-risk car insurance policies. The number of claims per policy in a 3-month period has a Poisson distribution with mean 0.5. It is assumed that all of the policies in the portfolio are independent.

- (i) Calculate the probability that:
 - (a) there are a total of 30 claims in a 3-month period for the whole portfolio
 - (b) there is a wait of more than $\frac{1}{2}$ month before a claim is made across the whole portfolio. [4]

- (ii) Simulate the number of claims made in a 3-month period by a single policy using the random number 0.975 taken from the $U(0,1)$ distribution. [2]

[Total 6]

Question X1.10

- (i) The records of cash payments received in a week by two clerks are summarised below:

Clerk	No. of payments	Sample mean	Sample standard deviation
A	1,000	£250	£30
B	1,500	£240	£25

Calculate the overall mean and standard deviation of the 2,500 payments. [4]

- (ii) The marks (%) of a sample of 20 students from a large class in a recent examination had a sample mean of 43 and a sample standard deviation of 6. The marks were subsequently adjusted – each mark was multiplied by 1.3 and the result was then increased by 10. Calculate the sample standard deviation of the adjusted marks. [2]

[Total 6]

Question X1.11

The random variable X has a beta distribution with parameters $\alpha = 1$ and $\beta = 4$.

- (i) State $E(X)$. [1]
 - (ii) Obtain the median of X . [3]
 - (iii) Hence comment on the shape of this distribution. [1]
- [Total 5]

Question X1.12

- (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]

Question X1.13

- (i) Derive, from first principles, the cumulant generating function of a gamma distribution and show that it can be written as:

$$C_X(t) = -\alpha \ln\left(1 - \frac{t}{\lambda}\right) \quad t < \lambda \quad [4]$$

- (ii) Hence derive an expression for the coefficient of skewness of a gamma distribution. [4]
- [Total 8]

Question X1.14

(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) A random variable X has PDF given by:

$$f(x) = \frac{1}{2}e^{-|x|} \quad -\infty < x < \infty$$

(a) Show that the MGF of X is given by:

$$\frac{1}{1-t^2} \quad -1 < t < 1$$

(b) Obtain the standard deviation of X .

(c) By using a series expansion or otherwise obtain $E(X^6)$. [8]

[Total 11]

END OF PAPER

Subject CT3: Assignment X2

2014 Examinations

Time allowed: 2½ hours

Instructions to the candidate

1. *Please 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 2014 exams.*
2. *Attempt all of the questions, leaving space in the margin and beginning your answer to each question on a new page.*
3. *Write in black ink using a medium-sized nib because we will be unable to mark illegible scripts.*
4. ***Leave at least 2cm margin on all borders.***
5. *Attempt the questions as far as possible under exam conditions.*
6. *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 in the Study Guide for the 2014 exams, 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. ActEd will not be responsible for scripts lost or damaged in the post or for scripts received after the deadline date. 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.

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

There are three methods for you to submit your script, namely by *email*, by *fax* or by *post*.

If you are submitting by **email**:

- complete the cover sheet, including the checklist
- scan your script (and Marking Voucher if applicable) to a pdf document, then email it to: **ActEdMarking@bpp.com**.

Please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* “CT3 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 3 MB in size. **The file size cannot exceed 4 MB.**
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

If you are submitting by **fax**:

- only write on one side of the paper when completing the assignment
- complete the cover sheet, including the checklist
- fax your script (including cover sheet and Marking Voucher if applicable) to **0844 583 4501**.

In addition:

- We recommend that you stay by the fax machine until the fax has been sent so that you can deal with any problems immediately. (If an error occurs, please re-fax the whole script.)
- An email will be sent by the end of the next working day to confirm that we have processed your script. Please do not phone to check progress before then. If the fax was sent without error then it's very unlikely that there will be a problem.

We will **not** accept:

- scripts submitted to other ActEd fax numbers – please use **0844 583 4501**
- scripts that have been split over a number of faxes. (If an error occurs, please re-fax the whole script.)
- more than one script per fax
- jumbled scripts – please fax the pages in the correct order.

If you are submitting by **post**:

- complete the cover sheet, including the checklist.
- we recommend that you photocopy your script before posting, in case your script is lost in the post.
- post your script to: **First Floor, Kimber House, 1 Kimber Road, Abingdon, Oxfordshire, OX14 1BZ**
- please staple the cover sheet (and Marking Voucher if applicable) to the front of your assignment
- please do not staple more than one assignment together.

Subject CT3: Assignment X2

2014 Examinations

Please complete the following information:

Name:

Address:

ActEd Student Number (see Note below):

--	--	--	--	--

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting this number will help us to process your scripts quickly. If you do not complete this box, your script may be delayed. If you do not know your ActEd Student Number, please email ActEd@bpp.com. **Your ActEd Student Number is not the same as your Faculty/Institute 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:

Time to do assignment (see Note below): _____ hrs _____ mins

Under exam conditions (delete as applicable): yes / nearly / no

Note: If you spend more than 2½ hours on the assignment, you should indicate on the assignment how much you completed within this time so that the marker can provide useful feedback on your chances of success in the exam.

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	Q15	Total
3	3	3	3	2	4	4	5	3	6	6	7	7	7	17	80

Grade: A B C D E **Marker's initials:** _____

Please grade your Assignment X1 marker by ticking the appropriate box.

- [] **Excellent** – the marker's comments were thorough and very helpful
- [] **Good** – the marker's comments were generally helpful
- [] **Acceptable** – please explain below how the marker could have been more helpful
- [] **Poor** – the marker's comments were generally unhelpful; please give details below

Please give any additional comments here (especially if you rate the marker less than good):

Note: Giving feedback on your marker helps us to improve the quality of marking.

Please follow the instructions on the previous page when submitting your script for marking.

***This page has been left blank in case you wish to submit your
script by fax.***

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, eg 2014 for the sessions leading to the 2014 exams?
- Written your full name and postal address in the appropriate box?
- Completed your ActEd Student Number in the appropriate box?
- 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?
- Attached your Marking Voucher or ordered Series X Marking?
- Rated your Assignment X1 marker?

Feedback from marker

Notes on marker's section

The marker's main objective is to give you advice on how to improve your answers. The marker will also assess your script quantitatively and qualitatively. The percentage score gives you a quantitative assessment. The grade is a qualitative assessment of how your script might be classified in the exam. The grades are as follows:

A = Clear Pass B = Probable Pass C = Borderline D = Probable Fail E = Clear Fail

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

or when you submit your next script.

***This page has been left blank in case you wish to submit your
script by fax.***

Question 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]

Question X2.2

Derive, from first principles, the moment generating function of a Type I negative binomial distribution, with parameters k and p . [3]

Question X2.3

For a random sample of size 10 from a normal distribution with mean 40 and variance 20, which of the following statements are true? Explain your answers.

(i) The variance of the sample mean is 2. [1]

(ii) The mean of the sample variance is 20. [1]

(iii) The probability that the sample variance exceeds 30 is less than 5%. [1]

[Total 3]

Question X2.4

Suppose that a random sample of nine observations is taken from a normal distribution with mean $\mu = 0$. Let \bar{X} and S^2 denote the sample mean and variance respectively.

Determine (to 2 decimal places) the probability that the value of \bar{X} exceeds that of S , ie determine $P(\bar{X} > S)$. [3]

Question X2.5

A random sample of size 5 is drawn from a normal distribution with mean 50 and variance 20, and a random sample of size 20 is drawn from a normal distribution with mean 60 and variance 40. Calculate the probability that the variance of the first sample will be more than $2\frac{1}{4}$ times that of the second sample. [2]

Question X2.6

A general insurance portfolio contains 100 policies. The probability that any policy in the portfolio makes one or more claims is 0.2, independent of any other policy.

The number of claims arising on a policy is modelled as a Type 1 negative binomial with $k = 1$ and $p = 0.6$, independently for each policy and independent of the number of policies giving rise to claims.

Calculate the mean and standard deviation of the total number of claims for this portfolio. [4]

Question X2.7

The time to complete a tricky pension review is normally distributed with mean 8 hours and standard deviation 2 hours. Calculate the probability that the times taken for two randomly selected tricky reviews differ by no more than 3 hours. [4]

Question X2.8

- (i) For a pair of jointly distributed random variables X and Y , derive the result:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y) \quad [2]$$

- (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]

Question X2.9

Let X have a Poisson distribution with mean 6, and let Y be a related variable with conditional mean and variance given by:

$$E(Y | X = x) = 3x - 5, \quad \text{var}(Y | X = x) = \frac{x^2}{4}$$

Calculate the unconditional standard deviation of Y .

[3]

Question X2.10

- (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]

Question X2.11

Two children play an “incy-wincy” spider game. They take it in turns to each roll two dice 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]

Question X2.12

The number of claims N which arise under a general insurance portfolio, in a certain period, is modelled as a discrete random variable. The claim amounts X_i , $i = 1, 2, \dots, N$, are modelled as independent, identically distributed random variables and they are also independent of N .

Let Y be the total aggregate claim amount arising from this portfolio.

- (i) Show that the moment generation function of Y is given by:

$$M_Y(t) = M_N[\ln M_X(t)] \quad [3]$$

The number of claims is modelled as a Poisson random variable with mean 12 and the claim amounts are modelled as gamma variables with mean 150 and variance 50.

- (ii) By using the cumulant generating function of Y , or otherwise, obtain the coefficient of skewness of the aggregate claim amount of this portfolio. [4]
- [Total 7]

Question X2.13

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]

Question X2.14

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. [2]

(ii) A random sample X_1, \dots, X_n is taken from a $Gamma(\alpha, \lambda)$ 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 \bar{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 \bar{X} exceeds 40. [2]

[Total 7]

Question X2.15

X and Y are discrete random variables. The only possible combinations of these two variables have the following probabilities:

		X			
		0	1	2	
Y		0	$\frac{1}{2}$	0	$\frac{1}{16}$
		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 $E(X) = E[E(X | Y)]$ holds. [1]
 - (iii) Calculate:
 - (a) $E(X + Y | X = 1)$
 - (b) $E(X | Y = 2)$
 - (c) $\text{var}(X | Y = 2)$. [7]
 - (iv) Obtain the values of the function $E[Y^2 | X]$ and hence calculate $E[E(Y^2 | X)]$. [3]
 - (v) Calculate $E[Y^2]$ and comment on your answer. [2]
- [Total 17]

END OF PAPER

Subject CT3: Assignment X3

2014 Examinations

Time allowed: 3 hours

Instructions to the candidate

1. *Please 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 2014 exams.*
2. *Attempt all of the questions, leaving space in the margin and beginning your answer to each question on a new page.*
3. *Write in black ink using a medium-sized nib because we will be unable to mark illegible scripts.*
4. ***Leave at least 2cm margin on all borders.***
5. *Attempt the questions as far as possible under exam conditions.*
6. *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 in the Study Guide for the 2014 exams, 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. ActEd will not be responsible for scripts lost or damaged in the post or for scripts received after the deadline date. 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.

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

There are three methods for you to submit your script, namely by *email*, by *fax* or by *post*.

If you are submitting by **email**:

- complete the cover sheet, including the checklist
- scan your script (and Marking Voucher if applicable) to a pdf document, then email it to: **ActEdMarking@bpp.com**.

Please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* “CT3 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 3 MB in size. **The file size cannot exceed 4 MB.**
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

If you are submitting by **fax**:

- only write on one side of the paper when completing the assignment
- complete the cover sheet, including the checklist
- fax your script (including cover sheet and Marking Voucher if applicable) to **0844 583 4501**.

In addition:

- We recommend that you stay by the fax machine until the fax has been sent so that you can deal with any problems immediately. (If an error occurs, please re-fax the whole script.)
- An email will be sent by the end of the next working day to confirm that we have processed your script. Please do not phone to check progress before then. If the fax was sent without error then it's very unlikely that there will be a problem.

We will **not** accept:

- scripts submitted to other ActEd fax numbers – please use **0844 583 4501**
- scripts that have been split over a number of faxes. (If an error occurs, please re-fax the whole script.)
- more than one script per fax
- jumbled scripts – please fax the pages in the correct order.

If you are submitting by **post**:

- complete the cover sheet, including the checklist.
- we recommend that you photocopy your script before posting, in case your script is lost in the post.
- post your script to: **First Floor, Kimber House, 1 Kimber Road, Abingdon, Oxfordshire, OX14 1BZ**
- please staple the cover sheet (and Marking Voucher if applicable) to the front of your assignment
- please do not staple more than one assignment together.

Subject CT3: Assignment X3

2014 Examinations

Please complete the following information:

Name:

Address:

ActEd Student Number (see Note below):

--	--	--	--	--

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting this number will help us to process your scripts quickly. If you do not complete this box, your script may be delayed. If you do not know your ActEd Student Number, please email ActEd@bpp.com. **Your ActEd Student Number is not the same as your Faculty/Institute 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:

Time to do assignment (see Note below): _____ hrs _____ mins

Under exam conditions (delete as applicable): yes / nearly / no

Note: If you spend more than 3 hours on the assignment, you should indicate on the assignment how much you completed within this time so that the marker can provide useful feedback on your chances of success in the exam.

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
3	3	4	3	6	5	7	7	6	5	15	18	18	100 = ____ %

Grade: A B C D E

Marker's initials: _____

Please grade your Assignment X2 marker by ticking the appropriate box.

- [] **Excellent** – the marker's comments were thorough and very helpful
- [] **Good** – the marker's comments were generally helpful
- [] **Acceptable** – please explain below how the marker could have been more helpful
- [] **Poor** – the marker's comments were generally unhelpful; please give details below

Please give any additional comments here (especially if you rate the marker less than good):

Note: Giving feedback on your marker helps us to improve the quality of marking.

Please follow the instructions on the previous page when submitting your script for marking.

***This page has been left blank in case you wish to submit your
script by fax.***

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, eg 2014 for the sessions leading to the 2014 exams?
- Written your full name and postal address in the appropriate box?
- Completed your ActEd Student Number in the appropriate box?
- 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?
- Attached your Marking Voucher or ordered Series X Marking?
- Rated your Assignment X2 marker?

Feedback from marker

Notes on marker's section

The marker's main objective is to give you advice on how to improve your answers. The marker will also assess your script quantitatively and qualitatively. The percentage score gives you a quantitative assessment. The grade is a qualitative assessment of how your script might be classified in the exam. The grades are as follows:

A = Clear Pass B = Probable Pass C = Borderline D = Probable Fail E = Clear Fail

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

or when you submit your next script.

***This page has been left blank in case you wish to submit your
script by fax.***

Question X3.1

A random sample of 16 observations (x_1, \dots, x_{16}) from a normal distribution gives:

$$\sum_{i=1}^{16} x_i = 128 \quad \sum_{i=1}^{16} x_i^2 = 1,168$$

Obtain a 90% confidence interval for the population standard deviation. [3]

Question X3.2

A student is examining the numbers of claims from a sample of motor policies, all of which have experienced at least one claim in the last year. He obtains these results:

Number of claims	1	2	3	4	5
Number of policies	58	14	3	1	1

He wishes to model these figures using a negative binomial distribution with PF:

$$P(X = x) = \binom{x-1}{k-1} p^k q^{x-k}, \quad x = k, k+1, \dots \quad k = \text{positive integer}, \quad q = 1 - p$$

- (i) State the appropriate value for the parameter k . [1]
 - (ii) Hence, estimate the value of the parameter p using the method of moments. [2]
- [Total 3]

Question X3.3

The following sample was taken from a normal distribution with mean μ and variance 20:

56, 32, 49, 57, 44

- (i) Obtain a symmetrical 95% confidence interval for μ . [2]
 - (ii) Repeat part (i) for the situation where the population variance is unknown. [2]
- [Total 4]

Question X3.4

A random sample of size $2n$ is taken from a geometric distribution for which:

$$P(X = x) = pq^{x-1} \quad x = 1, 2, \dots$$

Give an expression for the likelihood that the sample contains equal numbers of odd and even values of X . [3]

Question X3.5

A sample of 50 independent and identically distributed observations from an $\text{Exp}(\lambda)$ distribution gave:

Range	$0 < x < 1$	$1 < x < 2$	$x > 2$
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]

Question X3.6

A research agency is testing the effectiveness of a new surface to air missile which is believed to be more effective than existing defence systems at hitting its target under poor lighting conditions. The agency has carried out a series of 100 trials in which the missile was required to respond to an enemy fighter plane following a randomly selected trajectory. The missile successfully hit its target in 72 of these trials.

- (i) Obtain a 90% confidence interval for the probability that one of the new missiles will hit its target. [3]
- (ii) The new missiles are to be deployed in banks of three, which will be launched together, separated by 2-second intervals, at the same target. Use part (i) to obtain a 90% confidence interval for the probability that the target will be hit when a missile bank is fired. [2]
[Total 5]

Question X3.7

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:

$$\hat{\mu} = \bar{X} \quad [3]$$

- (ii) Obtain 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]

Question X3.8

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 $Poi(\lambda)$ 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 obtain 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]

Question X3.9

A random variable X has probability density function:

$$2e^{-2(x-\theta)} \quad x \geq \theta$$

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 $E(X) = \theta + \frac{1}{2}$ and hence calculate the method of moments estimate of θ . [2]
 - (iii) Comment briefly on your results. [1]
- [Total 6]

Question X3.10

The percentage return on an investment of a particular type over a period of one year is modelled as a normal random variable X with mean μ and standard deviation 0.5%. An investor wants to estimate the probability that the return on such an investment will exceed 3%.

A sample of 8 such returns have values:

2.1%, 2.6%, 0.8%, 3.2%, 0.6%, 1.5%, 3.8%, 2.9%

- (i) Derive a formula for the maximum likelihood estimator of μ . [3]
 - (ii) Hence calculate the maximum likelihood estimate of $\theta = P(X > 3\%)$ based on the given data. [2]
- [Total 5]

Question X3.11

Following archaeological excavations at a site in Egypt, ten samples of wood were carbon-dated and their ages x (years) estimated as:

$$\begin{array}{ccccc} 4,900 & 4,750 & 4,820 & 4,710 & 4,760 \\ 4,570 & 4,300 & 4,680 & 4,800 & 4,670 \end{array}$$

$$\sum x = 46,960 \quad \sum x^2 = 220,772,800$$

- (i) Calculate a 95% confidence interval for the true mean age of the wood found at this site. [3]
- (ii) Present these data 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, using the variance of the existing sample as an estimate of the variance for all such wood samples. [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]

Question X3.12

- (i) Show, using moment generating functions, that if X_1, \dots, 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, \dots, X_n . [3]

A random sample of 10 pet insurance claims had an average size of £680. It is believed that claim amounts are exponentially distributed.

- (ii) Using your result from part (i) obtain an exact 90% confidence interval for the *mean* pet insurance claim size. [3]
- (iii) 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]
- (iv) (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, obtain the estimated asymptotic standard error of the mean. [3]
- (v) (a) Use your results from (iv) 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 (ii) and (v)(a). [4]
 [Total 18]

Question X3.13

Let X be a gamma random variable with parameters 2 and $\frac{1}{\theta}$, such that:

$$f(x) = \frac{1}{\theta^2} x e^{-x/\theta}, \quad x > 0$$

It is required to estimate θ based on X_1, X_2, \dots, X_n , a random sample of n observations of X , with mean \bar{X} .

- (i) Write down the mean and variance of X . [2]
 - (ii) Show that the maximum likelihood estimator $\hat{\theta}$ is $\frac{1}{2}\bar{X}$ and obtain an expression for its mean square error (MSE). [7]
 - (iii) Consider the set of estimators of the form $a\bar{X}$, where a is a constant.
 - (a) Show that the mean square error of $a\bar{X}$ is:

$$\frac{\theta^2}{n} [2a^2 + n(2a - 1)^2]$$
 - (b) Hence, show that the value of a that minimises the MSE of $a\bar{X}$ is $a = n/(2n+1)$. [5]
 - (iv) Compare the MSE of the maximum likelihood estimator found in part (ii) with that of the optimal estimator found in part (iii)(b) for the three sample sizes $n = 1, 5$ and 100 , and comment briefly on the bias and the MSE of the two estimators. [4]
- [Total 18]

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 CT3: Assignment X4

2014 Examinations

Time allowed: 3 hours

Instructions to the candidate

1. *Please 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 2014 exams.*
2. *Attempt all of the questions, leaving space in the margin and beginning your answer to each question on a new page.*
3. *Write in black ink using a medium-sized nib because we will be unable to mark illegible scripts.*
4. ***Leave at least 2cm margin on all borders.***
5. *Attempt the questions as far as possible under exam conditions.*
6. *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 in the Study Guide for the 2014 exams, 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. ActEd will not be responsible for scripts lost or damaged in the post or for scripts received after the deadline date. 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.

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

There are three methods for you to submit your script, namely by *email*, by *fax* or by *post*.

If you are submitting by **email**:

- complete the cover sheet, including the checklist
- scan your script (and Marking Voucher if applicable) to a pdf document, then email it to: **ActEdMarking@bpp.com**.

Please note the following:

- Please title the email to ensure that the subject and assignment are clear *eg* “CT3 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 3 MB in size. **The file size cannot exceed 4 MB.**
- Before emailing to ActEd, please check that your scanned assignment includes all pages and conforms to the above.

If you are submitting by **fax**:

- only write on one side of the paper when completing the assignment
- complete the cover sheet, including the checklist
- fax your script (including cover sheet and Marking Voucher if applicable) to **0844 583 4501**.

In addition:

- We recommend that you stay by the fax machine until the fax has been sent so that you can deal with any problems immediately. (If an error occurs, please re-fax the whole script.)
- An email will be sent by the end of the next working day to confirm that we have processed your script. Please do not phone to check progress before then. If the fax was sent without error then it's very unlikely that there will be a problem.

We will **not** accept:

- scripts submitted to other ActEd fax numbers – please use **0844 583 4501**
- scripts that have been split over a number of faxes. (If an error occurs, please re-fax the whole script.)
- more than one script per fax
- jumbled scripts – please fax the pages in the correct order.

If you are submitting by **post**:

- complete the cover sheet, including the checklist.
- we recommend that you photocopy your script before posting, in case your script is lost in the post.
- post your script to: **First Floor, Kimber House, 1 Kimber Road, Abingdon, Oxfordshire, OX14 1BZ**
- please staple the cover sheet (and Marking Voucher if applicable) to the front of your assignment
- please do not staple more than one assignment together.

Subject CT3: Assignment X4

2014 Examinations

Please complete the following information:

Name:

Address:

ActEd Student Number (see Note below):

--	--	--	--	--

Note: Your ActEd Student Number is printed on all personal correspondence from ActEd. Quoting this number will help us to process your scripts quickly. If you do not complete this box, your script may be delayed. If you do not know your ActEd Student Number, please email ActEd@bpp.com. **Your ActEd Student Number is not the same as your Faculty/Institute 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:

Time to do assignment (see Note below): _____ hrs _____ mins

Under exam conditions (delete as applicable): yes / nearly / no

Note: If you spend more than 3 hours on the assignment, you should indicate on the assignment how much you completed within this time so that the marker can provide useful feedback on your chances of success in the exam.

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
2	2	3	6	3	4	6	6	10	15	16	19	8	100 = _____ %

Grade: A B C D E

Marker's initials: _____

Please grade your Assignment X3 marker by ticking the appropriate box.

- [] **Excellent** – the marker's comments were thorough and very helpful
- [] **Good** – the marker's comments were generally helpful
- [] **Acceptable** – please explain below how the marker could have been more helpful
- [] **Poor** – the marker's comments were generally unhelpful; please give details below

Please give any additional comments here (especially if you rate the marker less than good):

Note: Giving feedback on your marker helps us to improve the quality of marking.

Please follow the instructions on the previous page when submitting your script for marking.

***This page has been left blank in case you wish to submit your
script by fax.***

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, eg 2014 for the sessions leading to the 2014 exams?
- Written your full name and postal address in the appropriate box?
- Completed your ActEd Student Number in the appropriate box?
- 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?
- Attached your Marking Voucher or ordered Series X Marking?
- Rated your Assignment X3 marker?

Feedback from marker

Notes on marker's section

The marker's main objective is to give you advice on how to improve your answers. The marker will also assess your script quantitatively and qualitatively. The percentage score gives you a quantitative assessment. The grade is a qualitative assessment of how your script might be classified in the exam. The grades are as follows:

A = Clear Pass B = Probable Pass C = Borderline D = Probable Fail E = Clear Fail

Please note that you can provide feedback on the marking of this assignment at:

www.ActEd.co.uk/marking

***This page has been left blank in case you wish to submit your
script by fax.***

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:

$$Y_i = a + bx_i + e_i \text{ where } e_i \sim N(0, \sigma^2)$$

Question X4.1

Calculate the least squares estimate for the slope parameter b . [2]

Question X4.2

Calculate an unbiased estimate of the variance parameter. [2]

Question X4.3

Give a 95% confidence interval for b based on the sample data. [3]

Question X4.4

(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]

Question X4.5

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]

Question X4.6

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]

Question X4.7

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]

Question X4.8

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]

Question X4.9

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	$\frac{1-q}{2}$	$\frac{1-q^2}{2}$
colour-blind	$\frac{q}{2}$	$\frac{q^2}{2}$

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]

Question X4.10

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, ..., 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]

Question X4.11

A research worker employed by a large estate agency conducted a survey into the difference (D) between the initial asking price (IAP) and the eventual selling price (ESP) of private houses sold in 1998 in Glasgow. She categorised houses into two types, A (divided into flats) and B (others), and into two areas of the city, I and II, based on distance from the city centre. The table below gives the results ($D = \text{ESP} - \text{IAP}$, in units of £100) for a random sample of seven sales in each of the four joint categories. The last two rows give the sum and sum of squares for the respective columns.

House type A		House type B	
Area I	Area II	Area I	Area II
61	25	22	3
43	56	-21	25
112	7	12	-40
30	39	39	-8
49	68	-31	15
22	-4	28	-64
77	30	-5	11
$\sum y$	394	221	44
$\sum y^2$	27828	10871	4360
			6740

- (i) Display the data in a simple diagram that illustrates their main features. Comment briefly on any general conclusions that you can draw about the importance of the factors “house type” and “city area”. [4]
 - (ii) Perform a one-way analysis of variance to establish whether the underlying average price difference differs among the four joint categories. [5]
 - (iii) Calculate a 95% confidence interval for the underlying common standard deviation, σ , of the price difference in the four categories. [3]
 - (iv) Carry out an analysis of the mean differences using a least significant difference approach with a significance level of 5%. [4]
- [Total 16]

Question X4.12

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

$$\sum(x - \bar{x})^2 = 42 \quad \sum(y - \bar{y})^2 = 16,492 \quad \sum(x - \bar{x})(y - \bar{y}) = 826$$

The following model is used to represent the data:

$$Y_i = a + bx_i + e_i$$

where Y_i ($i = 1, 2, \dots, 8$) are the costs, x_i ($i = 1, 2, \dots, 8$) are the fixed times and e_i ($i = 1, 2, \dots, 8$) 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. [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]

Question X4.13

Consider a one-way analysis of variance for comparing k treatments with n_i responses for the i th treatment. The model is:

$$Y_{ij} = \mu + \tau_i + e_{ij} \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, n_i$$

where e_{ij} are independent $N(0, \sigma^2)$ random variables and $\sum_i n_i \tau_i = 0$.

Consider the least squares estimators for this model:

$$\hat{\mu} = \bar{Y}_{\bullet\bullet}, \quad \hat{\tau}_i = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}, \quad \hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$$

Show that:

(i) $\hat{\mu}$ is an unbiased estimator of μ [1]

(ii) $\hat{\tau}_i$ is an unbiased estimator of τ_i [1]

(iii) $\hat{\sigma}^2$ can be rewritten as $\frac{1}{n-k} \left[\sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^k \frac{1}{n_i} \bar{Y}_{i\bullet}^2 \right]$ [2]

(iv) $\hat{\sigma}^2$ is an unbiased estimator of σ^2 . [4]

[Total 8]

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.

For the session leading to the April 2014 exams – CT Subjects

Marking vouchers

Subjects	Assignments	Mocks
CT1, CT8	26 March 2014	7 April 2014
CT2, CT3, CT4, CT5, CT6, CT7	2 April 2014	14 April 2014

Series X Assignments

Subjects	Assignment	Recommended submission date	Final deadline date
CT1, CT8	X1	13 November 2013	22 January 2014
CT2, CT3, CT4, CT5, CT6, CT7		20 November 2013	29 January 2014
CT1, CT8	X2	27 November 2013	12 February 2014
CT2, CT3, CT4, CT5, CT6, CT7		4 December 2013	19 February 2014
CT1, CT8	X3	29 January 2014	5 March 2014
CT2, CT3, CT4, CT5, CT6, CT7		5 February 2014	12 March 2014
CT1, CT8	X4	19 February 2014	19 March 2014
CT2, CT3, CT4, CT5, CT6, CT7		26 February 2014	26 March 2014

Mock Exams

Subjects	Recommended submission date	Final deadline date
CT1, CT8	26 March 2014	7 April 2014
CT2, CT3, CT4, CT5, CT6, CT7	2 April 2014	14 April 2014

We encourage you to work to the recommended submission dates where possible.

We strongly recommend that you submit your mock exam electronically, by email or fax, in order for us to return your marked script to you in plenty of time before your exam. If you submit your mock by post to arrive with us on the final deadline date, you are likely to receive your script back less than a week before your exam.

In general, the turnaround of all scripts is likely to be quicker if you submit it electronically and well before the final deadline date.

For the session leading to the September/October 2014 exams – CT Subjects**Marking vouchers**

Subjects	Assignments	Mocks
CT1, CT2, CT4, CT6, CT7	27 August 2014	8 September 2014
CT3, CT5, CT8	3 September 2014	15 September 2014

Series X Assignments

Subjects	Assignment	Recommended submission date	Final deadline date
CT1, CT2, CT4, CT6, CT7	X1	11 June 2014	2 July 2014
CT3, CT5, CT8		18 June 2014	9 July 2014
CT1, CT2, CT4, CT6, CT7	X2	2 July 2014	23 July 2014
CT3, CT5, CT8		9 July 2014	30 July 2014
CT1, CT2, CT4, CT6, CT7	X3	23 July 2014	6 August 2014
CT3, CT5, CT8		30 July 2014	13 August 2014
CT1, CT2, CT4, CT6, CT7	X4	6 August 2014	20 August 2014
CT3, CT5, CT8		13 August 2014	27 August 2014

Mock Exams

Subjects	Recommended submission date	Final deadline date
CT1, CT2, CT4, CT6, CT7	20 August 2014	8 September 2014
CT3, CT5, CT8	27 August 2014	15 September 2014

We encourage you to work to the recommended submission dates where possible.

We strongly recommend that you submit your mock exam electronically, by email or fax, in order for us to return your marked script to you in plenty of time before your exam. If you submit your mock by post to arrive with us on the final deadline date, you are likely to receive your script back less than a week before your exam.

In general, the turnaround of all scripts is likely to be quicker if you submit it electronically and well before the final deadline date.