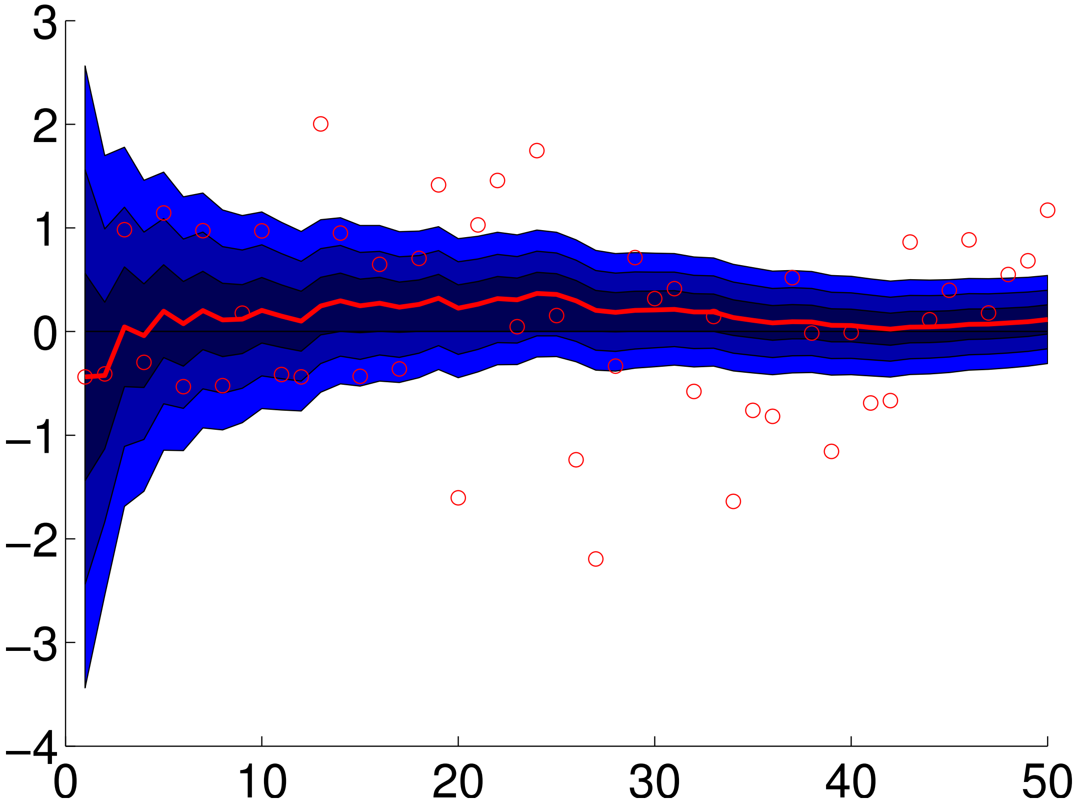
Adam Johansen
Department of Statistics
University of Warwick, UK

ST911 Fundamentals of Statistics (part 2)



Warwick
Statistics

Table of Contents

Part I. Background Reading Background

1. **Probability** 9
   - 1.1 Motivation 9
   - 1.2 Sets and Suchlike Background 9
   - 1.3 Outcomes and Events 11
   - 1.4 Probability Functions / Measures 12
     - 1.4.1 Properties of $\mathbb{P}[\cdot]$ 12
   - 1.5 Conditional Probability and Independence 14

2. **Random Variables** 17
   - 2.1 Random Variables and cumulative distribution functions 17
   - 2.2 Density Functions 18
   - 2.3 Expectations and Moments 20
   - 2.4 Expectation of a Function of a Random Variable 21
   - 2.5 Two important inequality results 21
   - 2.6 Moments and Moment Generating Functions 23
   - 2.7 Other Distribution Summaries 24

3. **Special Univariate Distributions** 25
   - 3.1 Discrete Distributions 25
     - 3.1.1 Discrete Uniform Distribution 25
     - 3.1.2 Bernoulli Distribution 25
     - 3.1.3 Binomial Distribution 26
     - 3.1.4 Hypergeometric Distribution 26
     - 3.1.5 Poisson Distribution 27
     - 3.1.6 Geometric and Negative Binomial Distribution 28
   - 3.2 Continuous Distributions 28
     - 3.2.1 Uniform Distribution 28
     - 3.2.2 Normal/Gaussian Distribution 29
     - 3.2.3 Exponential and Gamma Distribution 30
     - 3.2.4 Beta Distribution 30
   - 3.3 Exercises 31

4. **Joint and Conditional Distributions** 33
   - 4.1 Joint Distributions 33
   - 4.2 Special Multivariate Distributions 34
     - 4.2.1 Multinomial Distribution 34
     - 4.2.2 Bivariate Normal Distribution 35
   - 4.3 Conditional Distributions and Densities 35
   - 4.4 Conditional Expectation 36
   - 4.5 Conditional Expectations of Functions of random variables 37
   - 4.6 Independence of Random Variables 38
   - 4.7 Covariance and Correlation 38
   - 4.8 Sums of Random Variables 39
   - 4.9 Transformation of Random Variables: $Y = g(X)$ 39
   - 4.10 Moment-Generating-Function Technique 40
   - 4.11 Exercises 41

5. Inference 43
5.1 Sample Statistics 43
5.2 Sampling Distributions 43
5.3 Point Estimation 50
5.4 Interval Estimation 53
5.5 Exercises 58

Part II. Core Material

6. Maximum Likelihood Estimation 63
6.1 Likelihood Function and ML estimator 63
6.2 MLE and Exponential Families of Distributions 64
6.3 The Cramer-Rao Inequality and Lower Bound 67
6.4 Properties of MLEs 68
6.5 MLE and properties of MLE for the multi-parameter Case 70
6.6 Computation of MLEs 72
6.6.1 The Newton-Raphson Method 72
6.6.2 Fisher's Method of Scoring 73
6.6.3 Newton-Raphson and the Exponential Family of Distributions 73
6.6.4 The Expectation-Maximisation (EM) Algorithm 73
6.7 Exercises 74

7. Hypothesis Testing 75
7.1 Introduction 75
7.2 Simple Hypothesis Tests 76
7.3 Simple Null, Composite Alternative 78
7.4 Composite Hypothesis Tests 79
7.5 Exercises 82
7.6 The Multinomial Distribution and χ^2 Tests 83
7.6.1 Chi-Squared Tests 83
7.7 Exercises 85

8. Simulation for Inference 87
8.1 Background and History 87
8.1.1 A Simple Example 88
8.1.2 History 89
8.2 Basic Monte Carlo Inference 90
8.2.1 Monte Carlo Integration 91
8.2.2 Inversion Sampling 93
8.2.3 Rejection Sampling 93
8.2.4 Importance Sampling 96
8.3 Bootstrap Procedures 97
8.4 Notes 100

9. Elements of Bayesian Inference 101
9.1 Introduction 101
9.2 Parameter Estimation 102
9.3 Conjugate Prior Distributions 104
9.4 Uninformative Prior Distributions 106
9.5 Hierarchical Models 107
9.6 Inference using Markov Chain Monte Carlo (MCMC) Methods 108
9.6.1 Foundations 109

9.6.2 Gibbs Sampling 109
9.6.3 Metropolis Hastings Algorithm 111
9.6.4 Extensions NE 112

10. Linear Statistical Models 115
10.1 Regression 115
10.2 The Linear Model 115
10.2.1 Selected Extensions NE 117
10.3 Maximum Likelihood Estimation for NLMs 117
10.3.1 ML estimation for σ^2 118
10.4 Confidence Intervals 120
10.5 Hypothesis Tests for the Regression Coefficients 121
10.6 Prediction 122
10.7 Test for Significance of a Regression 123
10.7.1 Analysis of Variance for a Regression 124
10.8 The Coefficient of Multiple Determination (R^2) 124
10.9 Residual Analysis 125

A. Solutions to Exercises 127

Preface

These lecture notes serve as a primary reference for two modules, ST903 and ST911 (MASDOC S1). Two slightly different versions are produced from common source files; the differences are small but do provide some specific additional information for each of the modules. This version of the notes is intended to be used by students studying ST911.

This course aims to provide a grounding in fundamental aspects of modern statistics. The first part of the course covers Monte Carlo methods which lie at the heart of much modern statistics. The second part, covered by the second half of these notes, is concerned with statistical rather than computational aspects of modern statistics.

The notes are divided into two parts. The first contains background reading with which you are expected to be familiar by the time that you start attending lectures on the second part of the course. The second part of these lecture notes is intended to support the lectures occuring in weeks 6-10.

Beware: although there are some parts of this course which you might feel you know already, such as chapter 8 you should make sure that you really are comfortable with everything that these parts of the notes contain (including, say, section 8.3) as well as those parts which you may have seen before.

Books

There are a great many books which cover the topics studied in this course. A number of these are recommended, but you should spend some time in the library deciding which books you find useful.

Some which are recommended include:

- **Statistical Inference**, G. Casella and R. L. Berger, Duxbury, 2001 (2nd ed.).
- **All of Statistics: A Concise Course in Statistical Inference**, L. Wasserman, Springer Verlag, 2004.

These Notes

Although these lecture notes are reasonably comprehensive, you are strongly encouraged to consult these or other books if there are any areas which you find unclear; or just to see a different presentation of the material. The lecture notes are based (in some places rather closely) on a previous version prepared by Dr. Bärbel Finkelstadt. Any mistakes in the present version were undoubtedly introduced by me; any errors should be reported to me (`a.m.johansen@warwick.ac.uk`).

Throughout the notes, exercises and examples are indicated in different colours in sans serif font for easy identification.

For the avoidance of doubt some parts of the lecture notes are marked NE in order to indicate that they are *not examinable* (or, more accurately, will not be examined as a part of this course). Other parts, maked, *Background* are background material provide to refresh your memory; in the event that you haven't seen this material before it's your responsibility to familiarise yourself with it: don't expect it to be covered in detail in lectures.

Some of the notation used may not be familiar to you. Whilst every attempt has been made to make these notes self contained, you may find `http://en.wikipedia.org/wiki/Table_of_mathematical_symbols` a useful first reference if there is any symbolic notation which you're not comfortable with in these notes.

Exercises

Exercises are placed at appropriate points throughout these lecture notes. They aren't intended to be decorative: do them. This is by far the most effective way to learn mathematics. You're encouraged to answer the questions as soon as possible. The course contains a lot of material and you are strongly encouraged to read ahead of the lectures and think about the exercises in advance. Some of the exercises will be discussed in lectures — please be prepared to contribute your own thoughts to these discussions. Once the material has been covered in lectures, try to make sure that you understand how to answer these questions.

The relatively modest number of exercises present within the lecture notes provide a good starting point, but it would be to your advantage to answer additional exercises. The recommended textbooks (and many others that you'll find in the library) contain appropriate exercises — it's always important to attempt the exercises in textbooks.

Assessment

Assessment of ST911 has a number of components.

Coursework: the coursework from the first half of the course on Monte Carlo Methods will make up 30% of the module mark. An additional 20% will be provided by an assignment from this part of the module:

Assignment	**Released**	**Deadline**
Assignment	23/11/2010	9/12/2010

The remainder of the marks will come from an oral examination to be conducted at the start of term 2. Details will be made available later.

Office Hours and Contact Details

	Dr. Adam Johansen
email	`a.m.johansen@warwick.ac.uk`
Office	C0.20 (Maths and Statistics)
Office hours	Tuesday 9:30–10:30
	Friday 10:30 – 11:30
Telephone	024 761 - 50919

Part I

Background Reading Background

1. Probability

1.1 Motivation

Probability theory

- provides *probabilistic models* that try to explain the patterns observed in the data, and to provide an answer to the question “Where did the data come from?”
- allows a mathematical analysis of many aspects of statistical inference *under the assumption that the model is correct.*

Caveat:

- Every model is an idealised simplification of reality. We often assume that our model is adequate for practical purposes.
- Always check that the model does not disagree violently with the actually observed data.

With this course, and these notes, we begin from the beginning. It is not assumed that you have seen any formal probability or statistics before. However, if you really haven’t seen anything in this course before it will be rather intense and you will find that there are areas in which you need to do some background reading. It’s likely that ST911 students will already have seen much of the early material before; if you’ve already had a course in measure theoretic probability then keep that in mind and don’t worry about the less formal definitions given in this chapter.

1.2 Sets and Suchlike Background

This section is intended to serve as a reminder of a few basic mathematical concepts that will be important throughout the course. If you don’t remember having seen them before then you might find it helpful to consult a relevant book, such as “Set & Groups: A First Course in Algebra”, J. A. Green, Routledge 1965. Reprinted: Chapman Hall (1995).

In mathematics a set is a collection of objects. The order of those objects is irrelevant and those objects themselves may be sets of some sort.

The key concept when dealing with sets is **membership**. Object a is a member of a set A, written $a \in A$ if one of the objects contained in A is exactly a. This may equivalently be written as $A \ni a$ indicating that set A contains object a.

Specifying Sets. It’s important that we are precise when we specify which objects are in a particular set. All sorts of problems can arise if we allow imprecise definitions (the *Russell set paradox* is such an example: the set of all sets which do not contain themselves is not sufficiently precise because we don’t have a full definition of the set of all sets). There are broadly two ways in which we can specify the membership of a set:

- Semantically (intensively) by specifying the defining property of members of the set, for example $A = \{\text{odd integers}\}$ or $B = \{\text{rooms in the Zeeman building}\}$.
- Extensively, by listing all members of the set, for example, $A = \{1, 3, 5\}$ or $B = \{1, 2, 3, 4, \dots, 107\}$ or $C = \{2, 4, 6, 8, 10, 12, \dots\}$.

We will often find it useful to employ a formal form of semantic definition, specifying that we are interested in all objects of a certain type which satisfy a particular condition, for example, the set off all odd integers may be written as:

$$A = \{2k - 1 : k \in \mathbb{N}\}$$

Whilst the set of all points in a plane within a distance of 1 of the origin may be written as:

$$B = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$

One special set is the **empty set**, $\emptyset = \{\}$ which contains no points. Nothing is a member of $\emptyset$.

Comparing Sets. Given two sets, A and B, it's possible that all of the members of one also belong to the other. If all of the members of A are also members of B then we may write $A \subset B$ indicating that A is a **subset** of B. Similarly, if the converse were true and every member of B were a member of A we might write $A \supset B$ indicating that A is a **superset** of B. If $A \subset B$ and $B \subset A$ then we say that the two sets are equal: $A = B$. Notice that none of these ideas have involved the order of the members of A or B. Indeed, it's meaningless to talk about the ordering of members of a set and $\{1, 2, 3\} = \{3, 1, 2\} = \{2, 1, 3\}$.

Set Operations. There are a number of basic operations which its possible to perform when dealing with sets and these will prove to be important:

- Those elements which are members of two sets. The **intersection** of A and B, written $A \cap B$ is exactly the collection of objects which are members of both A and B:

  $$A \cap B = \{x : x \in A, x \in B\}.$$

- Those elements which are members of at least one of a collection of sets. The **union** of A and B, written $A \cup B$ is the collection of objects which are members of either A or B (including those points which are in both A and B):

  $$A \cup B = \{x : x \in A \text{ or } B\}.$$

- The **complement** of a set is the collection of objects which are not members of that set. We write $a \notin A$ to indicate that a point a is not a member of set A. The complement:

  $$\bar{A} = \{x : x \notin A\}.$$

  Some care is required here. The complement must always be taken with respect to the **universe** of interest that is the set containing all points that we might be interested in. This is often specified implicitly and will often be something like $\mathbb{R}$. For the purposes of the current course we will use the symbol Ω to denote this object (in some settings $\mathbf{X}$ is more commonly used). With this convention:

  $$\bar{A} = \{x \in \Omega : x \notin A\}.$$

- The **difference** between set A and B is the collection of points which are members of A but not of B:

  $$A \setminus B = \{x \in A : x \notin B\}.$$

  Notice that $A \setminus B \neq B \setminus A$ in general: it's an asymmetric definition. The complement of a set may be written as the difference between the universe and the set of interest $\bar{A} = \Omega \setminus A$.
- The **symmetric difference** makes symmetric the notion of the difference between two sets:

  $$\begin{aligned} A \Delta B =& (A \setminus B) \cup (B \setminus A) = \{x \in A \cup B : x \notin A \cap B\} \\ =& [A \cup B] \setminus [A \cap B]. \end{aligned}$$

De Morgan's Laws. There is a useful result which relates complements, intersections and unions in a way which is often useful when manipulating algebraic expressions.

Theorem 1.1 (De Morgan's Laws). *For sets A and B:*

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

Singletons, Subsets and Membership. A **singleton** is a set containing a single object. Mathematically, a set $A = \{a\}$ which contains one member, a, is different to the member itself: $A \neq \{a\}$. If we have a larger set $B = \{a, b, c, d, e, f, \dots, z\}$ then it is true to say that $a \in B$ and that $A \subset B$. However, it is not true that $A \in B$ or that $a \subset B$. This seemingly pedantic point is actually rather important.

This has close parallels in the real world. A box containing a cake is different to a cake and a box containing four different sorts of cake need not contain four boxes which each contain a different cake. Sets are really just a formal type of mathematical container.

1.3 Outcomes and Events

It's convenient to begin with some definitions. We consider *experiments*, which comprise: a collection of distinguishable outcomes, which are termed elementary events, and typically denoted Ω and a collection of sets of possible outcomes to which we might wish to assign probabilities, $\mathcal{A}$, the *event space.*

It's important to be careful when considering sets of sets. Notice that if $A \in \mathcal{A}$ then $A \subset \Omega$ — it is *not* true that $A \in \Omega$ or that for $\omega \in \Omega$ we can expect $\omega \in \mathcal{A}$ although it is possible that $\{\omega\} \in \mathcal{A}$. There is a difference between ω, a point in the space Ω and $\{\omega\}$ a set containing a single point, ω.

Example 1.1. Consider the experiment consisting of the tossing of a single die. $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let $A = \{\text{even number}\}$. A is an event. It is a subset of Ω. $A = \{2, 4, 6\}$. Let $A_i = \{i \text{ showing}\}; i = 1, 2, ..., 6$. Each A_i is an elementary event. If we are interested in all possible subsets of Ω, then there are $2^6 = 64$ events, of which only 6 are elementary, in $\mathcal{A}$ (including both the empty set and Ω). ◁

Exercise 1.3.1. Prove that a set of size M has 2^M subsets. (Hint: either use binomial coefficients or think of different representations of a subset of a set with M members).

In order to obtain a sensible theory of probability, we require that our collection of events $\mathcal{A}$ is an *algebra* over Ω, *i.e.* it must possess the following properties

(i) $\Omega \in \mathcal{A}$
(ii) If A in $\mathcal{A}$, then $\overline{A} \in \mathcal{A}$
(iii) If A_1 and $A_2 \in \mathcal{A}$, then $A_1 \cup A_2 \in \mathcal{A}$.

In the case of finite Ω, we might note that the collection of all subsets of Ω necessarily satisfies the above properties and by using this default choice of algebra, we can assign probabilities to any possible combination of elementary events.

Several results follow from the above properties

Proposition 1.1. *If $\mathcal{A}$ is an algebra, then $\emptyset \in \mathcal{A}$.*

Proof. By property (i) $\Omega \in \mathcal{A}$; by (ii) $\overline{\Omega} \in \mathcal{A}$; but $\overline{\Omega} = \emptyset$; so $\emptyset \in \mathcal{A}$ □

Proposition 1.2. *If A_1 and $A_2 \in \mathcal{A}$, then $A_1 \cap A_2 \in \mathcal{A}$ for any algebra $\mathcal{A}$.*

Proof. $\overline{A}_1$ and $\overline{A}_2 \in \mathcal{A}$; hence $\overline{A}_1 \cup \overline{A}_2$, and $\overline{(\overline{A}_1 \cup \overline{A}_2)} \in \mathcal{A}$; but $\overline{(\overline{A}_1 \cup \overline{A}_2)} = \overline{\overline{A}}_1 \cap \overline{\overline{A}}_2 = A_1 \cap A_2$ by De Morgan's law (theorem 1.1: "union and intersection interchange under complementation"). □

Proposition 1.3. *If $\mathcal{A}$ is an algebra and $A_1, A_2, ..., A_n \in \mathcal{A}$, then $\bigcup_{i=1}^n A_i$ and $\bigcap_{i=1}^n A_i \in \mathcal{A}$.*

Proof. This follows from property (iii) by induction for any *finite n*. □

Aside 1.1. Two problems arise, if we try to deal with general[1] Ω:

- It is necessary to replace the third defining property of an algebra with a slightly stronger one:
(iii') If $A_1, A_2, \ldots$ are a countable sequence of members of $\mathcal{A}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.
Of course, (iii')⇒(iii). If $\mathcal{A}$ satisfies (i)–(ii) and (iii') then it is termed a σ-algebra.
- In such advanced settings, the set of all subsets (or power set) of Ω is simply too big and it isn't possible to construct probability distributions which can assign probabilities consistently to all members of the power set. This is why σ-algebras are used.

The present course is not concerned with the technical details of probability but you may subsequently discover a need to familiarise yourself with it, depending upon the type of statistics in which you are interested. A. N. Shiryaev's book, "Probability", published by Springer Verlag provides one of many good references on the subject. J. Rosenthal's "A First Look at Rigorous Probability" provides a shorter introduction to the area which is very readable. The MA911 module will also discuss probability at a rather higher level.

1.4 Probability Functions / Measures

Let Ω denote the sample space and $\mathcal{A}$ denote a collection of events assumed to be a σ-algebra (an algebra will suffice if $|\Omega| < \infty$) of events that we shall consider for some random experiment.

Definition 1.1 (Probability function). *A probability function $\mathbb{P}[\cdot]$ is a set function with domain $\mathcal{A}$ (a σ-algebra of events) and range [0,1], i.e., $\mathbb{P} : \mathcal{A} \to [0, 1]$, which satisfies the following axioms*

(i) $\mathbb{P}[A] \geq 0$ *for every* $A \in \mathcal{A}$
(ii) $\mathbb{P}[\Omega] = 1$
(iii) If $A_1, A_2, \ldots$ is a sequence of mutually exclusive events(i.e. $A_i \cap A_j = \emptyset$ for any $i \neq j$) in $\mathcal{A}$ and if $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{A}$, then

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$$

Axioms (i) to (iii) are called Kolmogorov's axioms or simply the axioms of probability.

Aside 1.2 (Measures). Some references refer to $\mathbb{P}[\cdot]$ as a probability measure referring to the fact that its a particular example of a class of set-functions termed measures which assign a "size" to the sets upon which they operate. This is outside the scope of the present course, but it's always useful to at least be familiar with alternative terminology.

1.4.1 Properties of $\mathbb{P}[\cdot]$

A remarkably rich theory emerges from these three axioms (together, of course, with those of set theory). Indeed, all formal probability follows as a logical consequence of these axioms. Some of the most important simple results are summarised here. Throughout this section, assume that Ω is our collection of possible outcomes, $\mathcal{A}$ is a σ-algebra over Ω and $\mathbb{P}[\cdot]$ is an associated probability distribution.

Many of these results simply demonstrate that things which we would intuitively want to be true of probabilities do, indeed, arise as logical consequences of this simple axiomatic framework.

[1] Such complex objects as $\mathbb{R}$ are uncountable and complicated enough to require this more sophisticated theory.

Proposition 1.4. $\mathbb{P}[\emptyset] = 0$.

Proof. Take $A_1 = \emptyset, A_2 = \emptyset, ...$; then by axiom (iii)

$$\mathbb{P}[\emptyset] = \mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i] = \sum_{i=1}^{\infty} \mathbb{P}[\emptyset]$$

which, as $\forall A : 0 \leq \mathbb{P}[A] \leq 1$ (eliminating infinite solutions), can only hold if $\mathbb{P}[\emptyset] = 0$. □

Proposition 1.5. *If $A_1, A_2, \ldots, A_n$ are pairwise disjoint elements of $\mathcal{A}$, corresponding to mutually exclusive outcomes in our experiment, then*

$$\mathbb{P}[A_1 \cup A_2 \cup \ldots \cup A_n] = \sum_{i=1}^{n} \mathbb{P}[A_i]$$

Proof. Let $A_{n+1} = \emptyset, A_{n+2} = \emptyset, ...$ then the statement follows immediately from axiom (iii). □

Proposition 1.6. *If $A \in \mathcal{A}$ then*

$$\mathbb{P}[\bar{A}] = 1 - \mathbb{P}[A]$$

Proof. $A \cup \bar{A} = \Omega$, and $A \cap \bar{A} = \emptyset$; so $\mathbb{P}[\Omega] = \mathbb{P}[A \cup \bar{A}] = \mathbb{P}[A] + \mathbb{P}[\bar{A}]$ by proposition 1.5. But $\mathbb{P}[\Omega] = 1$ by axiom (ii); the result follows. □

Proposition 1.7. *For any two events $A, B \in \mathcal{A}$*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$

Proof. $A \cup B = A \cup (\bar{A} \cap B)$ and $A \cap \bar{A} \cap B = \emptyset$; so $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[\bar{A} \cap B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$. □

Proposition 1.8. *If $A, B \in \mathcal{A}$ and $A \subset B$, then*

$$\mathbb{P}[A] \leq \mathbb{P}[B]$$

Proof. $B = (B \cap A) \cup (B \cap \bar{A})$, and $B \cap A = A$; so $B = A \cup (B \cap \bar{A})$, and $A \cap (B \cap \bar{A}) = \emptyset$; hence $\mathbb{P}[B] = \mathbb{P}[A] + \mathbb{P}[B \cap \bar{A}]$. The statement follows by noting that $\mathbb{P}[B \cap \bar{A}] \geq 0$. □

Proposition 1.9 (Boole's inequality). *If $A_1, ..., A_n \in \mathcal{A}$, then*

$$\mathbb{P}[A_1 \cup A_2 \cup ... \cup A_n] \leq \mathbb{P}[A_1] + \mathbb{P}[A_2] + ... + \mathbb{P}[A_n].$$

Proof. $\mathbb{P}[A_1 \cup A_2] = \mathbb{P}[A_1] + \mathbb{P}[A_2] - \mathbb{P}[A_1 \cap A_2] \leq \mathbb{P}[A_1] + \mathbb{P}[A_2]$. The proof is completed by using mathematical induction. □

We are now ready to define arguably the most important object in probability:

Definition 1.2 (Probability Space). *A* **probability space** *is the triple $(\Omega, \mathcal{A}, \mathbb{P}[\cdot])$, where Ω is a sample space, $\mathcal{A}$ is a σ-algebra over Ω, and $\mathbb{P}[\cdot]$ is a probability function with domain $\mathcal{A}$.*

Thus, a probability space is a single entity that gives us a convenient and compact way to specify all of the components which we need to make use of probability.

1.5 Conditional Probability and Independence

Sometimes it's possible to observe that one event has occurred. In this situation, we wish to have a model for the behaviour of the probability that other events compatible with B. Conditional probability is the appropriate language.

Definition 1.3 (Conditional Probability). *Let A and B be events in $\mathcal{A}$ of the given probability space $(\Omega, \mathcal{A}, \mathbb{P}[\cdot])$. The conditional probability of event A given event B, denoted by $\mathbb{P}[A|B]$, is defined as*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \text{ if } \mathbb{P}[B] > 0,$$

and is left undefined when $\mathbb{P}[B] = 0$.

Exercise 1.5.1. Consider the experiment of tossing two coins, $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$, and assume that each point is equally likely. Find

(i) the probability of two heads given a head on the first coin.

(ii) the probability of two heads given at least one head.

When speaking of conditional probabilities we are conditioning on some event B, *i.e.* we are assuming that the experiment has (already) resulted in some outcome B. B, in effect, becomes our new sample space. The following result allows us to break a probability up into smaller, perhaps more manageable pieces.

Theorem 1.2 (Law of Total Probability). *For a given probability space $(\Omega, \mathcal{A}, \mathbb{P}[\cdot])$, if $B_1, ..., B_n$ is a collection of mutually disjoint events in $\mathcal{A}$ satisfying*

$$\Omega = \bigcup_{j=1}^{n} B_j,$$

i.e. *$B_1, \ldots, B_n$ partition Ω and $\mathbb{P}[B_j] > 0$, $j = 1, \ldots, n$, then for every $A \in \mathcal{A}$,*

$$\mathbb{P}[A] = \sum_{j=1}^{n} \mathbb{P}[A \cap B_j].$$

Proof. Note that $A = \bigcup_{j=1}^{n} A \cap B_j$ and the $A \cap B_j$ are mutually disjoint, hence

$$\mathbb{P}[A] = \mathbb{P}\left[\bigcup_{j=1}^{n} A \cap B_j\right] = \sum_{j=1}^{n} \mathbb{P}[A \cap B_j].$$

□

Conditional probability has a number of useful properties. The following elementary result is surprisingly important and has some far-reaching consequences.

Theorem 1.3 (Bayes' Formula). *For a given probability space $(\Omega, \mathcal{A}, \mathbb{P}[\cdot])$, if $A, B \in \mathcal{A}$ are such that $\mathbb{P}[A] > 0, \mathbb{P}[B] > 0$, then:*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

Proof.

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

by using both the definition of conditional probability twice (in particular, note that $\mathbb{P}[B|A] = \mathbb{P}[A \cap B]/\mathbb{P}[A]$). □

Theorem 1.4 (Partition Formula). *If $B_1, \dots, B_n \in \mathcal{A}$ partition Ω, then for any $A \in \mathcal{A}$:*

$$P(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Proof: By the law of total probability:

$$P(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i)$$

and $\mathbb{P}(A \cap B_i) = \mathbb{P}(A|B_i)\mathbb{P}(B_i)$ by definition of $\mathbb{P}(A|B_i)$.

Combining Bayes theorem with the partition formula provides the following useful result:

$$\mathbb{P}[B_k|A] = \frac{\mathbb{P}[A|B_k]\mathbb{P}[B_k]}{\sum_{j=1}^{n} \mathbb{P}[A|B_j]\mathbb{P}[B_j]}$$

for any partition, $B_1, \dots, B_n \in \mathcal{A}$ of Ω.

Theorem 1.5 (Multiplication Rule). *For a given probability space $(\Omega, \mathcal{A}, \mathbb{P}[\cdot])$, let $A_1, \dots, A_n$ be events belonging to $\mathcal{A}$ for which $\mathbb{P}[A_1, \dots, A_{n-1}] > 0$, then*

$$\mathbb{P}[A_1, A_2, \dots, A_n] = \mathbb{P}[A_1]\mathbb{P}[A_2|A_1] \dots \mathbb{P}[A_n|A_1 \dots A_{n-1}].$$

Proof. The proof follows from the definition of conditional probability by mathematical induction. □

Exercise 1.5.2. There are 5 urns, numbered 1 to 5. Each urn contains 10 balls. Urn i has i defective balls, i = 1, . . . , 5. Consider the following experiment: First an urn is selected uniformly (i.e. each urn is selected with the same probability) at random and then a ball is selected uniformly at random from the selected urn. The experimenter does not know which urn was selected.

(i) What is the probability that a defective ball will be selected?
(ii) If we have already selected the ball and noted that it is defective, what is the probability that it came from urn 5? Generalise to urn $k; k = 1, \dots, 5$.

Definition 1.4 (Independent Events). *For a given probability space $(\Omega, \mathcal{A}, \mathbb{P}[\cdot])$, let A and B be two events in $\mathcal{A}$. Events A and B are defined to be independent iff one of the following conditions is satisfied*

(i) $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$
(ii) $\mathbb{P}[A|B] = \mathbb{P}[A]$ *if* $\mathbb{P}[B] > 0$
(iii) $\mathbb{P}[B|A] = \mathbb{P}[B]$ *if* $\mathbb{P}[A] > 0$.

Remark: to demonstrate the equivalence of (i) to (iii) it suffices to show that (i) implies (ii), (ii) implies (iii), (iii) implies (i).

Exercise 1.5.3. Consider the experiment of rolling two dice. Let $A = \{$total is odd$\}$, $B = \{$6 on the first die$\}$, $C = \{$total is seven$\}$.

(i) Are A and B independent?
(ii) Are A and C independent?
(iii) Are B and C independent?

Definition 1.5 (Independence of Several Events). *For a given probability space $(\Omega, \mathcal{A}, \mathbb{P}[\cdot])$, let $A_1, \dots, A_n$ be events in $\mathcal{A}$. Events $A_1, \dots, A_n$ are defined to be independent iff*

1. $\mathbb{P}[A_i \cap A_j] = \mathbb{P}[A_i]\mathbb{P}[A_j], i \neq j$

2. $\mathbb{P}[A_i \cap A_j \cap A_k] = \mathbb{P}[A_i]\mathbb{P}[A_j]\mathbb{P}[A_k], i \neq j, j \neq k, i \neq k$

$\vdots$

n. $\mathbb{P}\left[\bigcap_{i=1}^{n} A_i\right] = \prod_{i=1}^{n} \mathbb{P}[A_i]$

Exercise 1.5.4. Show that pairwise independence (the first condition in definition 1.5) does not imply independence using the following events in the random experiment of rolling two unbiased dice:

(i) $A_1 = \{$odd face on first die$\}$

(ii) $A_2 = \{$odd face on second die$\}$

(iii) $A_3 = \{$odd total$\}$

2. Random Variables

In statistics and other areas which deal routinely with uncertainty, it is useful to think of experiments in which the outcome is more subtle than one of a collection of possible events occurring. In order to deal with such experiments we do not need to depart far from the event-based probability introduced in the previous chapter.

This chapter provides an introduction to the language of distribution theory. The principal rôle of this chapter is to introduce some concepts and definitions that will be important throughout the rest of the course. The notion of a random variable will be used to relate quantities of interest to events and a distribution function will be used to give the probabilities of certain events defined in terms of random variables.

2.1 Random Variables and cumulative distribution functions

We considered random events in the previous chapter: experimental outcomes which either do or do not occur. In general we cannot predict whether or not a random event will or will not occur before we observe the outcome of the associated experiment — although if we know enough about the experiment we may be able to make good probabilistic predictions. The natural generalisation of a random event is a random variable: an object which can take values in the set of real numbers (rather than simply happening or not happening) for which the precise value which it takes is not known before the experiment is observed.

The following definition may seem a little surprising if you've seen probability only outside of measure-theoretic settings in the past. In particular, random variables are deterministic functions: neither random nor variable in themselves. This definition is rather convenient; all randomness stems from the underlying probability space and it is clear that random variables and random events are closely related. This definition also makes it straightforward to define multiple random variables related to a single experiment and to investigate and model the relationships between them.

Definition 2.1 (Random Variable). *Given a probability space $(\Omega, \mathcal{A}, \mathbb{P}[.])$, a* ***random variable****, X, is a function with domain Ω and codomain $\mathbb{R}$ (the real line) (i.e. $X : \Omega \to \mathbb{R}$).*

We generally use capital letters to denote random variables and the corresponding lower case letter to denote a realisation (or specific value) of the random variable. Some authors use the term random variable a little more liberally and allow it to include functions with domain Ω and *any* codomain whilst others refer to members of this more general class of objects as *random elements*.

Example 2.1. Roll 2 dice $\Omega = \{(i, j); i, j = 1, ..., 6\}$. Several random variables can be defined, for example $X((i, j)) = i + j$, also $Y((i, j)) = |i - j|$. Both, X and Y are random variables. X can take values $2, 3, ..., 12$ and Y can take values $0, 1, ..., 5$. ◁

Definition 2.2 (Distribution Function). *The* ***distribution function*** *of a random variable X, denoted by $F_X(\cdot)$ is defined to be the function $F_X : \mathbb{R} \to [0,1]$ which assigns*

$$F_X(x) = \mathbb{P}[X \le x] = \mathbb{P}[\{\omega : X(\omega) \le x\}]$$

for every $x \in \mathbb{R}$.

The distribution function connects random variables to random events. It allows us to characterise a random variable by talking about the probability of a class of events defined in terms of the lower level sets of the random variable. Conveniently, this class of events is general enough that it allows us to completely characterise the random variable as we shall see later.

Example 2.2. Toss a coin. Let X =number of heads. Then

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \le x < 1 \\ 1 & \text{if } x \ge 1. \end{cases}$$

◁

Exercise 2.1.1. Consider the experiment in example 2.1. Sketch F_Y.

Properties of $F_X(\cdot)$

1. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$
2. $F_X(a) \le F_X(b)$ for $a < b$ (monotone and non-decreasing)
3. $F_X(\cdot)$ is continuous from the right

$$\lim_{h \downarrow 0} F_X(x+h) = F_X(x)$$

Any function $F(\cdot)$ with $F : \mathbb{R} \to [0,1]$ satisfying the above properties is a distribution function for some random variable.

2.2 Density Functions

For two distinct classes of random variables, the distribution of values can be described more simply by using density functions. These classes are termed ‘discrete’ and ‘continuous’.

Definition 2.3 (Discrete random variable). *A random variable X is* ***discrete*** *if the range of X is countable (*i.e. *it is finite or isomorphic to the natural numbers).*

Definition 2.4 (Discrete density function). *If X is a discrete random variable with distinct values $x_1, x_2, \ldots, x_n, \ldots$, then the* ***discrete density function*** *of X is defined by*

$$f_X(x) = \begin{cases} \mathbb{P}[X = x_j] & \text{if } x = x_j, j = 1, 2, \ldots, n, \ldots \\ 0 & \text{if } x \ne x_j \end{cases}$$

Remark: the x_i are often called atoms, mass points or points of concentration. Other terms are sometimes used instead of discrete density function, including probability function or probability mass function.

Theorem 2.1. *Let X be a discrete random variable. $F_X(\cdot)$ can be obtained from $f_X(\cdot)$ and vice versa.*

Proof. Suppose $f_X(\cdot)$ is given. Then

$$F_X(x) = \sum_{j:x_j \leq x} f_X(x_j).$$

Suppose $F_X(\cdot)$ is given. Then

$$f_X(x_j) = F_X(x_j) - \lim_{h \downarrow 0} F_X(x_j - h).$$

□

Exercise 2.2.1. Consider the experiment of tossing two dice. Let $X = \{\text{total of upturned faces}\}$ and $Y = \{\text{absolute difference of upturned faces}\}$.

1. Give the probability function f_X and sketch it.
2. Give f_Y.

Definition 2.5. *Any function $f : \mathbb{R} \to [0, 1]$ is defined to be a discrete density function if for some countable set $x_1, x_2, \dots, x_n, \dots$*

1. $f(x_j) \geq 0 \quad j = 1, 2, \dots$
2. $f(x) = 0$ *for* $x \neq x_j; j = 1, 2, \dots$
3. $\sum_j f(x_j) = 1$ *where summation is over* $x_1, x_2, \dots, x_n, \dots$

Definition 2.6 (Continuous random variable). *A random variable X is called **continuous** if there exists a function $f_X(\cdot)$ such that*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u \quad \textit{for every } x \in \mathbb{R}.$$

Definition 2.7 (Probability density function of a continuous random variable). *If X is a continuous random variable, the function f_X in $F_X(x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u$ is called the **probability density function** of X.*

Theorem 2.2. *Let X be a continuous random variable. Then $F_X(\cdot)$ can be obtained from $f_X(\cdot)$, and vice versa.*

Proof. Suppose $f_X(\cdot)$ is given, then

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u.$$

Suppose $F_X(\cdot)$ is given, then

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}.$$

□

Note: For discrete random variables $f_X(x) = \mathbb{P}[X = x]$. However, this is not true for continuous random variables. In the continuous case:

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x - \Delta x)}{2\Delta x},$$

hence, for sufficiently small Δx,

$$\begin{aligned} f_X(x)2\Delta x &\approx F_X(x + \Delta x) - F_X(x - \Delta x) \\ &= \mathbb{P}[x - \Delta x < X \leq x + \Delta x]. \end{aligned}$$

Definition 2.8. *A function $f : \mathbb{R} \to [0, \infty)$ is a probability density function iff*

1. $f(x) \geq 0 \quad \forall x$
2. $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1.$

Notice that the second of these requirements includes, implicitly, the existence of this integral.

2.3 Expectations and Moments

Definition 2.9 (Expectation, Mean). *Let X be a random variable. The **mean** of X, denoted by μ_X or $\mathbb{E}[X]$, is defined by*

(i) $\mathbb{E}[X] = \sum_j x_j f_X(x_j)$ *if X is discrete with mass points $x_1, x_2, \ldots, x_j, \ldots$*

(ii) $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$ *if X is continuous with density $f_X(x)$.*

Intuitively, $\mathbb{E}[X]$ is the centre of gravity of the unit mass that is specified by the density function. Expectation in a formal mathematical sense does not necessarily coincide with what we would *expect* in the ordinary sense of the word. It's quite possible that the *exepcted value* of a random variable is a value that the random variable can never take.

Exercise 2.3.1. Consider the experiment of rolling two dice. Let X denote the total of two dice and Y their absolute difference. Compute $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

Exercise 2.3.2. Let X be a continuous random variable with density

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } 0 \leq x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Compute $\mathbb{E}[X]$ and $F_X(x)$.

Definition 2.10 (Variance). *Let X be a random variable, and let $\mu_X = \mathbb{E}[X]$. The **variance** of X, denoted by σ_X^2 or $\mathbb{V}\mathrm{ar}[X]$ is defined by*

1. $\mathbb{V}\mathrm{ar}[X] = \sum_j (x_j - \mu_X)^2 f_X(x_j)$ *if X is discrete with mass points $x_1, x_2, \ldots, x_j, \ldots$*
2. $\mathbb{V}\mathrm{ar}[X] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\,\mathrm{d}x$ *for continuous X with density $f_X(x)$.*

Variance is a measure of spread or dispersion. If the values of a random variable X tend to be far from their mean, the variance of X will be larger than the variance of a comparable random variable Y whose values are typically nearer to the mean. It is clear from (i) and (ii) that variance is always non-negative (and is strictly positive for any random variable which has positive probability of differing from its mean).

Definition 2.11 (Standard deviation). *If X is a random variable, the **standard deviation** of X, denoted by σ_X, is defined as $+\sqrt{\mathbb{V}\mathrm{ar}[X]}$.*

Standard deviation is useful because it gives a quantification of the spread of a random variable on the same scale as the random variable itself. Despite its convenient mathematical properties, variance is essentially a squared distance and can be difficult to interpret.

2.4 Expectation of a Function of a Random Variable

Definition 2.12 (Expectation of a Function of a Random Variable). *Let X be a random variable and $g(\cdot)$ a function $g(\cdot) : \mathbb{R} \to \mathbb{R}$. The expectation or expected value of the function $g(\cdot)$ of the random variable X, denoted by $\mathbb{E}[g(X)]$ is defined by*

1. $\mathbb{E}[g(X)] = \sum_j g(x_j) f_X(x_j)$ *if X is discrete with mass points $x_1, x_2, \ldots, x_j, \ldots$ and provided the series is absolutely convergent*
2. $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, \mathrm{d}x$, *for continuous X with density $f_X(x)$, provided the integral exists.*

If $g(X) = X$, then $\mathbb{E}[g(X)] = \mathbb{E}[X] \equiv \mu_X$. If $g(X) = (X - \mu_X)^2$, then $\mathbb{E}[g(X)] = \sigma_X^2$. Furthermore, as $X : \Omega \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, we can consider $g \circ X(\omega) := g(X(\omega)) : \Omega \to \mathbb{R}$ to be a random variable in its own right. That is, a real-valued function of a random variable defines a new random variable.

Proposition 2.1 (Properties of Expectations). *Expectations have a number of useful properties which are very often useful:*

1. $\mathbb{E}[c] = c$ *for a constant c*
2. $\mathbb{E}[cg(X)] = c\mathbb{E}[g(X)]$ *for a constant c*
3. $\mathbb{E}[c_1 g_1(X) + c_2 g_2(X)] = c_1 \mathbb{E}[g_1(X)] + c_2 \mathbb{E}[g_2(X)]$
4. $\mathbb{E}[g_1(X)] \leq \mathbb{E}[g_2(X)]$ *if $g_1(x) \leq g_2(x)\ \forall\ x$*

The proof of these properties is left as an exercise (see lecture).

Proposition 2.2 (Variance in terms of expectations). *If X is a random variable, then*

$$\mathbb{V}\mathsf{ar}[X] = \mathbb{E}[(X - E(X)^2)] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Proof. See lecture. □

2.5 Two important inequality results

Theorem 2.3 (Chebyshev's Inequality). *Let X be a random variable and $g(\cdot)$ a non-negative function, then*

$$\mathbb{P}[g(X) \geq k] \leq \frac{\mathbb{E}[g(X)]}{k} \quad \forall \quad k > 0.$$

Proof. If X is a continuous random variable with density $f_X(\cdot)$, then

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_{-\infty}^{+\infty} g(x)f_X(x)\mathrm{d}x \\ &= \int_{x:g(x)\geq k} g(x)f_X(x)\mathrm{d}x + \int_{x:g(x)<k} g(x)f_X(x)\mathrm{d}x \\ &\geq \int_{x:g(x)\geq k} g(x)f_X(x)\mathrm{d}x \\ &\geq \int_{x:g(x)\geq k} kf_X(x)\mathrm{d}x \\ &= k\mathbb{P}[g(X)\geq k].\end{aligned}$$

Where the expression on the second line is an example of commonly-used notation indicating that we integrate the argument of an integral over a set specified as the subscript of the integral. If you're more used to seeing integrals over intervals in the form $\int_a^b f(x)\mathrm{d}x$ then this is only a slight generalisation and $\int_a^b f(x)\mathrm{d}x \equiv \int_{[a,b]} f(x)\mathrm{d}x$.

The result follows immediately. The discrete case is analogous (convince yourself that this is true). □

Corollary 2.1. *If X is a random variable with finite variance, σ_X^2, then:*

$$\mathbb{P}[|X-\mu_X|\geq r\sigma_X] = \mathbb{P}[(X-\mu_X)^2 \geq r^2\sigma_X^2] \leq \frac{1}{r^2}.$$

Proof. Take $g(X) = (X-\mu_X)^2$ and $k = r^2\sigma_X^2$ in theorem 2.3. □

Note, that the last statement can also be written as

$$\mathbb{P}[|X-\mu_X|\leq r\sigma_X] \geq 1 - \frac{1}{r^2}.$$

Thus the probability that X falls within $r\sigma$ units of μ_X is greater than or equal to $1-\frac{1}{r^2}$. For $r=2$ one gets

$$\mathbb{P}[\mu_X - 2\sigma_X < X < \mu_X + 2\sigma_X] \geq \frac{3}{4}$$

or, for each random variable X having a finite variance, at least 3/4 of the mass of X falls within two standard deviations of its mean. Chebyshev's inequality gives a bound which does not depend on the distribution of X. Despite its apparent simplicity, it is useful in a very wide range of settings, for example, it plays an important role in proving the law of large numbers in these notes. Perhaps even more widely used is the following result:

Theorem 2.4 (Jensen's Inequality). *For a RV X with mean $\mathbb{E}[X]$ and $g(\cdot)$ a convex continuous function,*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]).$$

Note, a convex function, g, is any function which lies beneath any of its chords. That is, given points a and b, with $a<b$, for any $\lambda \in [0,1]$:

$$g(a+\lambda(b-a)) = g((1-\lambda)a+\lambda b) \leq (1-\lambda)g(a) + \lambda g(b).$$

This may alternatively be expressed in terms of continuity, and the second derivative of g, which must satisfy:

$$\frac{d^2g}{dx^2}(x) \geq 0$$

everywhere where it is defined if g is convex.

Proof. See lecture. □

2.6 Moments and Moment Generating Functions

Definition 2.13 (Moments and central moments). *For a RV X, the r^{th}* ***moment*** *of X is given by*

$$\mu'_r = \mathbb{E}[X^r]$$

if the expectation exists. For a RV X, the r^{th} ***central moment*** *of X (the r^{th} moment about μ'_1) is given by*

$$\mu_r = \mathbb{E}[(X - \mu'_1)^r] = \mathbb{E}\left[\sum_{i=1}^{r} \binom{r}{i} (\mu'_1)^{r-i} X^i\right] = \sum_{i=0}^{r} \binom{r}{i} [\mu'_1]^{r-i} \mathbb{E}[X^i]$$

where the right hand side is simply a binomial expansion. Thus the moments of a random variable are the expectation of the powers of the random variable. Note that $\mu'_1 = \mathbb{E}[X] = \mu_X$, the mean of X.

The First Four Central Moments.

Zero $\mu_1 = \mathbb{E}[(X - \mu_x)] = 0$
Variance $\mu_2 = \mathbb{E}[(X - \mu_x)^2]$
Skewness $\frac{\mu_3}{\sigma_X^3} = \frac{\mathbb{E}[(X-\mu_x)^3]}{\sigma_X^3}$
Kurtosis $\frac{\mu_4}{\sigma_X^4} = \frac{\mathbb{E}[(X-\mu_x)^4]}{\sigma_X^4}$

Note that all odd central moments of X about μ_X are 0 if the density function is symmetrical about μ_X. The skewness is used to indicate whether a density is skewed to the left (value for skewness is negative) or skewed to the right (value for skewness is positive). The kurtosis is sometimes used to indicate that a density is more peaked (and heavier tailed) around its centre than the normal density. The kurtosis of a normally distributed random variable (introduced next chapter) can be shown to be 3. The normal distirbution is very widely used in statistics and is often taken as a default assumption; it's consequently common to compare properties of distributions with those of the normal distribution and thus the quantity $\frac{\mu_4}{\sigma_X^4} - 3$ is known as the excess kurtosis.

The moments of a density function play an important rôle in statistics. If we believe that a distribution must lie in a particular parametric family of distributions then it may be possible to specify those parameters exactly if we know all (or in some cases just a subset) of the moments of the distribution. The moment generating function gives us a function from which all the moments can be reconstructed via differentiation.

Definition 2.14 (Moment Generating Function (MGF)). *Let X be a RV with density $f_X(\cdot)$. We may define it's* ***moment generating function****, $m(t)$, as:*
Discrete

$$m(t) = \mathbb{E}[e^{tX}] = \sum_x e^{tx} f_X(x)$$

Continuous

$$m(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, \mathrm{d}x$$

The MGF has the property that

$$\left.\frac{d^r m(t)}{dt^r}\right|_{t=0} = \mu'_r$$

The following useful result is stated without proof here. Although the proof of this result is beyond the scope of this module (if you're familiar with the Laplace transform then you might notice that the moment generating function is essentially the Laplace transform of the density function and the general argument holds) the result itself is of sufficient usefulness to justify its inclusion here.

Theorem 2.5 (Equality of distributions). *Let X and Y be two random variables with densities $f_X(\cdot)$ and $f_Y(\cdot)$, respectively. Suppose that $m_X(t)$ and $m_Y(t)$ both exist and are equal for all t. Then the two distribution functions $F_X(\cdot)$ and $F_Y(\cdot)$ are equal.*

Exercise 2.6.1. Find the MGF of the binomial distribution

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, \ldots, n$$

Use it to show that the mean is np and the variance is $np(1-p)$

2.7 Other Distribution Summaries

Many other quantities are used as "summary statistics", numbers which provide as much information as possible about a distribution in as compact a form as possible. This section lists a few of the more common such statistics:

Definition 2.15 (Quantile). *For a RV X, the q^{th} **quantile** η_q is the smallest number η satisfying*

$$F_X(\eta) \geq q$$

*Certain quantiles are often used in statistics and are given particular names: **Median** $\eta_{0.5}$, a measure of the location of the distribution's centre. **Lower Quartile** $\eta_{0.25}$ and **Upper Quartile** $\eta_{0.75}$ can be combined to give a measure of the spread of the distribution termed the **Interquartile Range** $\eta_{0.75} - \eta_{0.25}$ and together with the median provide some idea of the skewness of the distribution.*

Another measure of the centre of a distribution is its **mode**: the point at which $f_X(\cdot)$ obtains its maximum.

3. Special Univariate Distributions

This chapter provides an overview of a number of parametric families of univariate density functions. These distributions are often used to model particular physical systems and have standard names. We consider discrete distributions and continuous distributions.

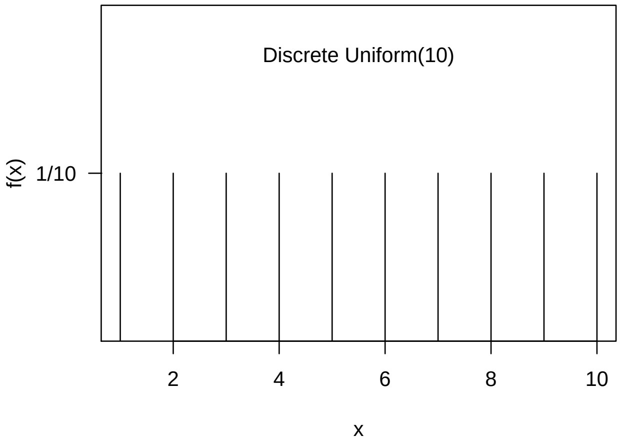
You might view this chapter as a taxonomy of useful building blocks from which sophisticated statistical models may be built. In addition to their rôle in statistical models, these distributions will arise directly in the probabilistic description of methods and situations of interest to the statistician.

3.1 Discrete Distributions

3.1.1 Discrete Uniform Distribution

$$f(x) = \mathsf{U}\left(x; \{1, \ldots, N\}\right) = \begin{cases} \frac{1}{N} & x = 1, 2, 3, \ldots, N \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = \frac{N+1}{2} \qquad \mathbb{V}\mathsf{ar}[X] = \frac{N^2 - 1}{12}$$

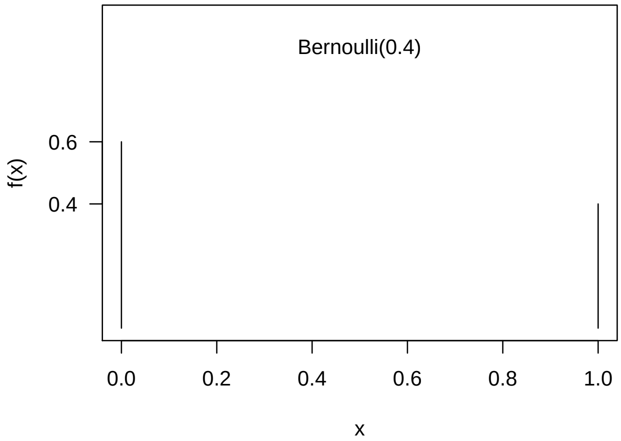


The discrete uniform distribution is an appropriate model when there are finitely many possible outcomes which are equally probable. If X has a discrete uniform distribution, we write $X \sim \mathsf{U}\{1, \ldots, N\}$.

3.1.2 Bernoulli Distribution

$$f(x) = \mathsf{Ber}\left(x; p\right) = \begin{cases} p^x(1-p)^{1-x} & x = 0, 1 \quad 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = p \qquad \mathbb{V}\mathsf{ar}[X] = p(1-p) = pq$$

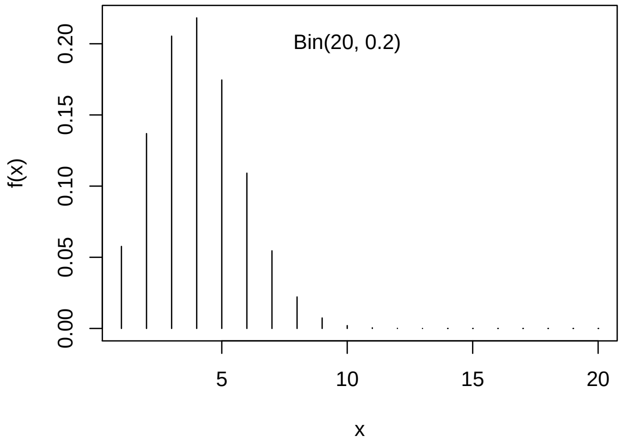


The Bernoulli distribution can be used to model random experiments with 2 possible outcomes (usually termed success and failure). Written: $X \sim \mathsf{Ber}\,(p)$.

3.1.3 Binomial Distribution

$$f(x) = \mathsf{Bin}\,(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \quad 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = np \qquad \mathbb{V}\mathsf{ar}[X] = np(1-p)$$

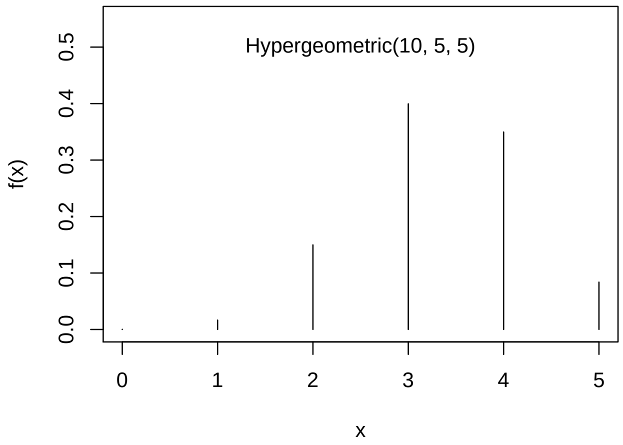


The binomial distribution corresponds to a model for the number of "successes" which occur when n independent Bernoulli random variables of probability p are generated. It also corresponds to the distribution of the number of black balls drawn from an urn which contains balls of which a proportion p are black if n balls are *sampled with replacement* (i.e. balls are removed from the urn one at a time, inspected, and returned before another ball is drawn) from that urn. Written: $X \sim \mathsf{Bin}\,(n, p)$.

3.1.4 Hypergeometric Distribution

$$f(x; a, b, n) = \begin{cases} \dfrac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}} & x = \max(0, n-b), 1, \ldots, \min(n, a) \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = na/(a+b) \qquad \mathbb{V}\mathsf{ar}[X] = \frac{nab}{(a+b)^2}\frac{a+b-n}{a+b-1}$$

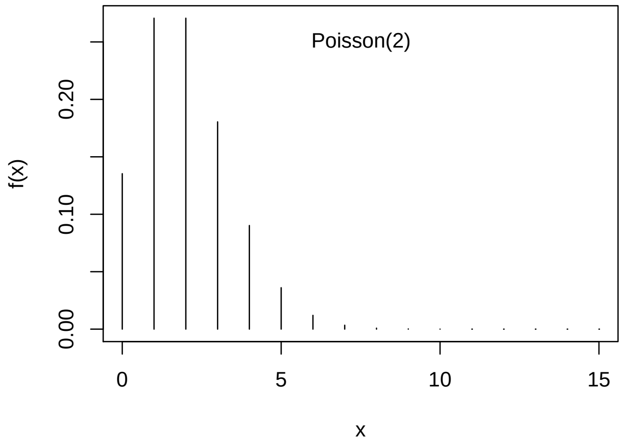


The hypergeometric distribution corresponds to the distribution of the number of black balls which are drawn from an urn which contains $a + b$ balls of which a are black if they are *sampled without replacement* (i.e. n balls are removed from the urn and not returned).

3.1.5 Poisson Distribution

$$f(x;\lambda) = \mathsf{Poi}\,(x;\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0, 1, \ldots \quad \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = \mathbb{V}\mathsf{ar}[X] = \lambda$$



Exercise 3.1.1. Prove that $\mathbb{E}[X] = \mathbb{V}\mathsf{ar}[X] = \lambda$ using the MGF of the Poisson Distribution

Uses of the Poisson Distribution

- For large n, and small p, $X \sim Bin(n,p)$ is approximately distributed as $\mathsf{Poi}\,(np)$. this is sometimes termed the “law of small numbers”.
- A **Poisson Process** with rate λ per unit time is such that
  1. X, the number of occurrences of an event in any given time interval of length t is $\mathsf{Poi}\,(\lambda t)$.
  2. The number of events in non-overlapping time intervals are independent random variables (see later).

Exercise 3.1.2. The number X of insect larvæfound on a cm^2 on a petri plate is assumed to follow a Poisson distribution with $\lambda = 3$. Find

- $P(X \leq 3)$
- $P(X > 1)$
- $P(2 \leq X \leq 4)$

– $F(4.2)$

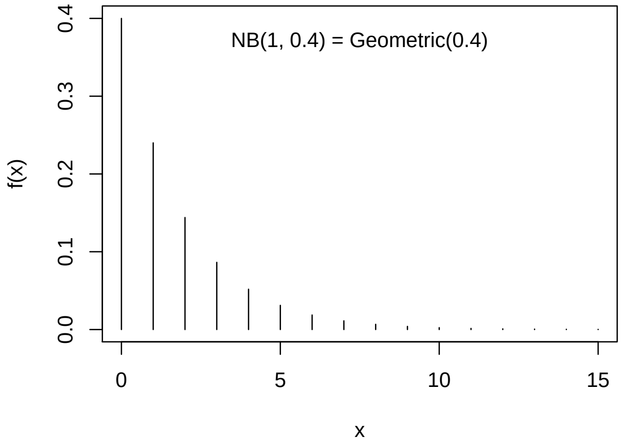
Is the assumption of a Poisson distribution likely to be reasonable for the entire plate?

3.1.6 Geometric and Negative Binomial Distribution

$$f(x;r,p) = \mathsf{NB}\,(x;r,p) = \begin{cases} \binom{x+r-1}{r-1} p^r (1-p)^x & x = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}$$

where $0 \le p \le 1, q = 1-p$

$$\mathbb{E}[X] = \frac{rq}{p} \qquad \mathbb{V}\mathsf{ar}[X] = \frac{rq}{p^2}$$



This, negative binomial distribution, corresponds to that of the number of failures which occur before the r^{th} success occurs in a sequence of independent Bernoulli trials with common probability p. In the case $r = 1$, the distribution is sometimes referred to as a geometric distribution and corresponds to the distribution of the number of failures which occur before the first success.

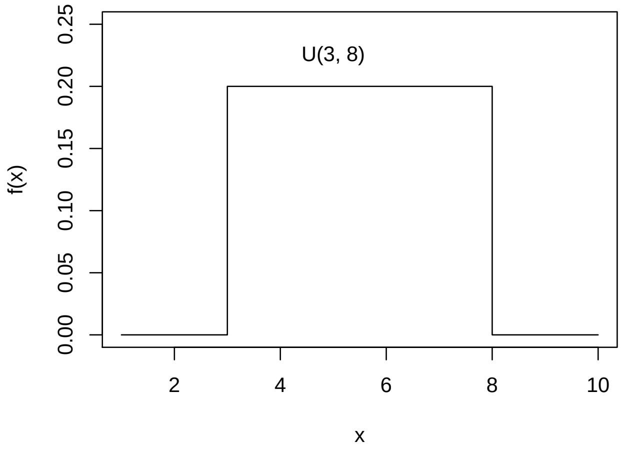
Exercise 3.1.3. Let $X \sim \mathsf{NB}\,(r,p)$. Find its MGF and use it to derive $\mathbb{E}[X]$ and $\mathbb{V}\mathsf{ar}[X]$.

3.2 Continuous Distributions

3.2.1 Uniform Distribution

$$f(x;a,b) = \mathsf{U}\,(x;[a,b]) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathbb{V}\mathsf{ar}[X] = \frac{(b-a)^2}{12}$$

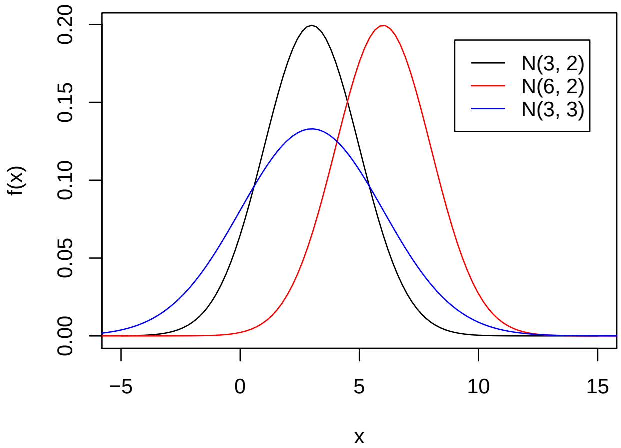


If any outcome within some region of the continuum is equally probable then a uniform distribution is appropriate: it corresponds to assigning probabilities to sub-intervals which depend only upon their length. Written $X \sim \mathsf{U}[a, b]$.

3.2.2 Normal/Gaussian Distribution

$$f(x; \mu, \sigma) = \mathsf{N}\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \qquad x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0$$

$$\mathbb{E}[X] = \mu \qquad\qquad \mathbb{V}\mathsf{ar}[X] = \sigma^2$$



The normal distribution is ubiquitous throughout statistics (and many other fields). In some instances it arises as a consequence of physical considerations; in others it can be justified by mathematical arguments. It arises as a consequence of theoretical arguments such as the central limit theorem.

Standard Normal Distribution A standard normal random variable is simply a special case of a normal random variable: $X \sim \mathsf{N}(0, 1)$, which has distribution function $\Phi(x) = F(x; 0, 1)$.

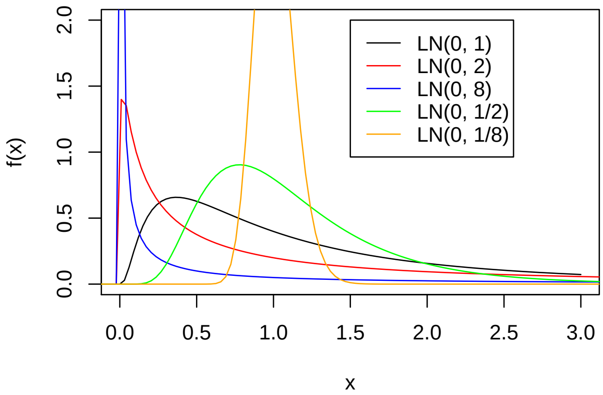
For $Z \sim \mathsf{N}\left(\mu, \sigma^2\right)$:

$$\mathbb{P}[a < Z < b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

This is useful because statistical tables and approximations for the standard normal random variable are readily available and via this representation it is possible to deal with any normal random variable by representing it as a rescaled and shifted standard normal random variable.

Lognormal Distribution If $\log(X) \sim \mathsf{N}\left(\mu, \sigma^2\right)$, then

$$f(x; \mu, \sigma^2) = \mathsf{logN}\left(x; \mu, \sigma^2\right) \frac{1}{(2\pi)^{\frac{1}{2}}\sigma x} \exp\left\{-\frac{1}{2\sigma^2}(\log x - \mu)^2\right\}, x > 0, \mu \in \mathbb{R}, \sigma > 0$$



The lognormal distribution is widely used. It is particularly prevalent in settings in which one knows that a random variable must be positive and would expect behaviour on a logarithmic scale which can be adequately modelled by a normal random variable.

Exercise 3.2.1. Let $X \sim \mathsf{N}(0,1)$. Show that

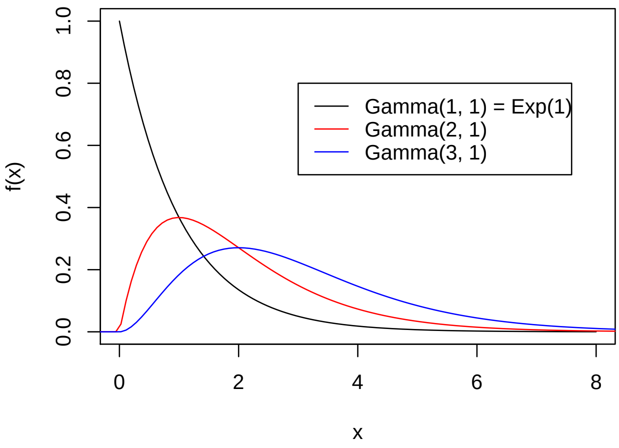
$$\beta_2 = \frac{\mu_4}{\sigma^4} = 3$$

(*i.e.* the kurtosis of a standard normal is 3).

3.2.3 Exponential and Gamma Distribution

$$f(x;r,\lambda) = \mathsf{Gamma}(x;r,\lambda) = \begin{cases} \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1}e^{-\lambda x} & \text{for } 0 \leq x < \infty, r > 0, \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = \frac{r}{\lambda} \qquad \mathbb{V}\mathsf{ar}[X] = \frac{r}{\lambda^2}$$

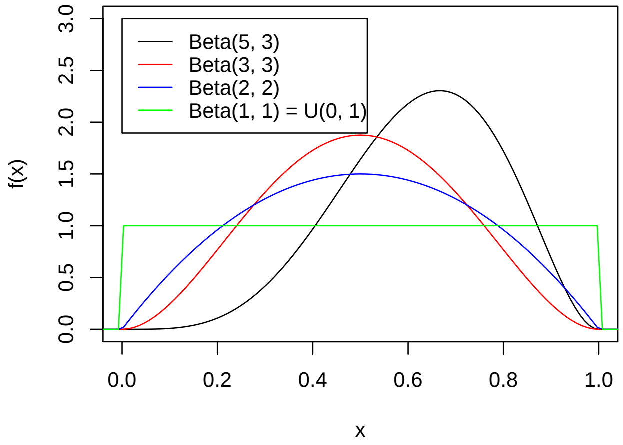


The exponential distribution of rate λ has density $\mathsf{Exp}(x;\lambda) = \mathsf{Gamma}(x;1,\lambda)$. It corresponds to a "memoryless distribution" and has connections with Poisson processes. The general gamma distribution can be interpreted, for integer r, as the distribution of the sum of r exponential random variables.

3.2.4 Beta Distribution

$$f(x;a,b) = \mathsf{Beta}(x;a,b) = \begin{cases} \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1} & \text{for } 0 \leq x < 1, a > 0, b > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = \frac{a}{a+b} \qquad \mathbb{V}\mathsf{ar}[X] = \frac{ab}{(a+b+1)(a+b)^2}$$



The beta distribution provides a flexible parametric family of distributions over the unit interval. These can be asymmetric (consider the mean) and may have mass concentrated around the mean (when a and b are large) or pushed towards the edges of the interval (when a and b are both small).

3.3 Exercises

Exercise 3.3.1. Consider the shopping preferences of a person buying a bar of chocolate. Suppose on the first purchase they are equally likely to choose a Mars or a Yorkie bar. They like a bit of variety, so for each subsequent purchase the probability that they will switch brands is $\frac{2}{3}$ whilst the probability of purchasing the same brand as the preceding purchase is $\frac{1}{3}$.

1. What is the probability that the shopper buys a Mars bar on their first and second purchases and a Yorkie on their third and fourth purchases?
2. Suppose now that the initial probability of purchasing a Mars bar is $\frac{3}{4}$. What is the probability that the shopper's second purchase will be a Yorkie bar?

Exercise 3.3.2. In a certain factory, four-packs of fruit-flavoured yoghurt are packaged on two production lines. Records show that a small percentage of the yoghurt packs are not packaged properly for sale: 1% from the first production line and 2.5% from the second.

1. If the percentages of total output that have come from the production lines are 55% from the first and 45% from the second, what is the probability that a yoghurt pack chosen at random from the whole output is faulty?
2. If we find a box which is faulty, what is the probability that it came from the second production line?

Exercise 3.3.3. A firm is running 500 vending machines of canned drinks in a town. Every vending machine (independently of the other vending machines) is defective within one week with probability $\frac{1}{50}$. In that case the firm has to send out a mechanic. In order to decide whether or not a mechanic should be employed permanently it is of interest to know the probability that the number of defective vending machines X is between 5 and 15 during a week.

1. What is the distribution of X? Determine also the expected value and the variance of X.
2. Determine $Pr[5 \leq X \leq 15]$ approximately with the Poisson distribution. Fix the parameter λ such that the expected value of the Poisson distribution corresponds to the expected value of X.
3. Give a lower bound of $Pr[5 \leq X \leq 15]$ by using Chebyshev's inequality.

Exercise 3.3.4. A mutual fund has an annual rate of return that is assumed to be normally distributed with mean 10% and standard deviation 4%.

1. Find the probability that the one-year return is negative.

2. Find the probability that the one-year return exceeds 15%.
3. If the managers of the mutual fund modify the composition of its portfolio, they can raise its mean annual return to 12% but will also raise the standard deviation to 5%. What would the probabilities in parts (a) and (b) be for the modified portfolio? Would you advise the managers to make the change?

4. Joint and Conditional Distributions

In this chapter we review the concept of k-dimensional distribution functions (for $k > 1$), conditional distributions, and independence of random variables.

4.1 Joint Distributions

Definition 4.1 (Joint Distribution Function). *For random variables $X_1, \dots X_k$ all defined on the same $(\Omega, \mathcal{A}, \mathbb{P}[.])$, the function $F : \mathbb{R}^k \to [0, 1]$*

$$F_{X_1,\dots,X_k}(x_1, \dots, x_k) = \mathbb{P}[X_1 \leq x_1; \dots ; X_k \leq x_k] \quad \forall (x_1, \dots, x_k)$$

is called the joint distribution function.

Definition 4.2 (Marginal Distribution Function). *For $F_{X_1,\dots,X_k}$ and $X_{i1}, \dots, X_{in}$ a strict subset of $X_1, \dots, X_k$, the function $F_{X_{i1},\dots X_{in}}$ is called a marginal distribution function.*

Definition 4.3 (Joint Discrete Density Function). *For a k-dimensional discrete random variable $(X_1, \dots X_k)$, the function*

$$f_{X_1,\dots,X_k}(x_1, \dots, x_k) = \mathbb{P}[X_1 = x_1; \dots ; X_k = x_k] \quad \forall (x_1, \dots, x_k)$$

is called the joint discrete density function, joint density function, or joint probability function.

Definition 4.4 (Marginal Discrete Density Function). *For the joint discrete density function $f_{X_1,\dots,X_k}$ and $X_{i1}, \dots, X_{in}$ a strict subset of $X_1, \dots, X_k$, the function $f_{X_{i1},\dots X_{in}}$ is called a marginal discrete density function.*

Given a collection of random variables, their joint distribution encodes information about the full set of random variables simultaneously. If we are interested in only a subset of a collection of random variables, then we can make use of the marginal distribution of that subset to consider the properties of interest in a compact way.

Exercise 4.1.1. Consider the tossing of 3 different coins. Let X_1: number of heads on the first and second coin and X_2: number of heads on the second and third coin.

- What is the joint density function of X_1 and X_2? Present it in a table.
- Find
  - $F(0.4, 1.3)$
  - $F(0, 0)$
  - $F(1.4, 2.1)$
  - $F(-1, 2)$

– $P(X_1 = 1, X_2 \geq 1)$

Definition 4.5 (Joint Continuous Density Function). *For a k-dimensional random variable $(X_1, \dots, X_k)$, the function $f_{X_1,\dots,X_k}(x_1, \dots, x_k) \geq 0$ such that*

$$F_{X_1,\dots,X_k}(x_1, \dots, x_k) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f_{X_1,\dots,X_k}(u_1, \dots, u_k)\, du_1 \dots du_k$$

for all $(x_1, \dots, x_k)$ is called a joint probability density function.

Definition 4.6 (Marginal Density Function). *For the joint continuous density function $f_{X_1,\dots,X_k}$ and $X_{i1}, \dots, X_{in}$ a strict subset of $X_1, \dots, X_k$, the function $f_{X_{i1},\dots,X_{in}}$ is called a marginal density function.*

Exercise 4.1.2. A random variable (X_1, X_2) has joint density

$$f(x_1, x_2) = \begin{cases} \frac{1}{8}(6 - X_1 - X_2) & \text{for } 0 \leq X_1 \leq 2,\, 2 \leq X_2 \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

– Show that f is a density function
– Find
1. F(1, 3)
2. F(0, 1)
3. F(3, 5)

Note that multivariate distribution functions and densities are exactly analogous to their univariate equivalents. Distribution functions now correspond to the probability that all elements of a vector are inferior to their argument whilst densities are integrated over areas/volumes/hypervolumes rather than intervals but nothing fundamental has changed.

4.2 Special Multivariate Distributions

There are many ways to construct multivariate distributions and a great many named multivariate distributions exist. This section provides a brief summary of a small number of the most ubiquitous multivariate distributions.

4.2.1 Multinomial Distribution

The multivariate distribution is a generalisation of the binomial distribution. It models the case of iidrepeated trials with $k+1$ distinct possible outcomes for each trial:

$$f_{X_1,\dots,X_k}(x_1, \dots, x_k) = \mathsf{Mult}\left(\mathbf{x}; n, p_1, \dots, p_{k+1}\right) = \frac{n!}{\prod_{i=1}^{k+1} x_i!} \prod_{i=1}^{k+1} p_i^{x_i} \tag{4.1}$$

where $x_i = 0, \dots, n$ and $\sum_{i=1}^{k+1} x_i = n$. (n is fixed, so the value of x_{k+1} is determined by values of $x_1, \dots, x_k$).

If one has n iid trials, each of which produces outcome i with probability p_i for any $i \in \{1, \dots, k+1\}$ then the $k+1$-dimensional random vector, $\mathbf{x} = x_1, \dots, x_{k+1}$, whose i^{th} element corresponds to the number of times that outcome i was observed has a multinomial distribution.

Note that the i^{th} element of $\mathbf{X}$, X_i, has a $\mathsf{Bin}\left(n, p_i\right)$ marginal distribution.

4.2.2 Bivariate Normal Distribution

The normal distribution can be straightforwardly extended to a bivariate random variable by considering to random variables which are correlated and which each is marginally normal:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \right.\right.$$
$$\left.\left. \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) \right] \right\}$$

for $-\infty < x_1, x_2, \mu_1, \mu_2 < \infty, \sigma_1, \sigma_2 > 0, -1 < \rho < 1$.

- ρ is the correlation coefficient
- for $\rho = 0$ the bivariate normal density is the product of two univariate normal densities: this corresponds to X_1 and X_2 being independent Normal random variables.

Extension to Multivariate Normal Distribution
In fact, it's straightforward to extend the normal distribution to vectors or arbitrary length, the multivariate normal distribution has density:

$$f(\mathbf{x}; \mu, \Sigma) = \mathsf{N}(\mathbf{x}; \mu, \Sigma) = \left(\frac{1}{\sqrt{2\pi}}\right)^r |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) \right\}$$

where

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_r \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_r \end{pmatrix}, \qquad \Sigma = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$$

Note that $\mathbf{x}$ is a vector; it has mean μ which is itself a vector and Σ is the variance-covariance matrix. If $\mathbf{x}$ is k-dimensional then Σ is a $k \times k$ matrix.

The multivariate normal distribution is very important throughout statistics, probability and related areas. It has a number of appealing properties, including the fact that its marginal and conditional (see section 4.3) distributions are themselves all normal.

4.3 Conditional Distributions and Densities

Given several random variables how much information does knowing one provide about the others? The notion of conditional probability provides an explicit answer to this question.

Definition 4.7 (Conditional Discrete Density Function). *For discrete random variables with X and Y with probability mass points $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$,*

$$f_{Y|X}(y_j|x_i) = \frac{\mathbb{P}[X = x_i; Y = y_j]}{\mathbb{P}[X = x_i]} = \mathbb{P}[Y = y_j | X = x_i]$$

is called the conditional discrete density function of Y given $X = x$.

Definition 4.8 (Conditional Discrete Distribution). *For jointly discrete random variables X and Y,*

$$F_{Y|X}(y|x) = \mathbb{P}[Y \leq y | X = x] = \sum_{j: y_j \leq y} f_{Y|X}(y_j|x)$$

is called the conditional discrete distribution of Y given $X = x$.

Exercise 4.3.1. Let Y_1 and Y_2 be two random variables with joint density

	0	1	2
0	q^3	pq^2	0
1	pq^2	pq	p^2q
2	0	p^2q	p^3

– Find the marginal densities of Y_1 and Y_2
– Find the conditional density function of y_2 given y_1
– Find
1. $\mathbb{E}[Y_1 - Y_2]$
2. $\mathbb{E}[Y_1 + Y_2]$
3. $\mathbb{E}[Y_1]$

Definition 4.9 (Conditional Probability Density Function). *For continuous random variables X and Y with joint probability density function $f_{X,Y}(x, y)$,*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \ \textit{if}\ f_X(x) > 0$$

where $f_X(x)$ is the marginal density of X.
Conditional Distribution *For jointly continuous random variables X and Y,*

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(z|x)\,\mathrm{d}z \qquad \forall x \textit{ such that } f_X(x) > 0.$$

4.4 Conditional Expectation

We can also ask what the *expected* behaviour of one random variable is, given knowledge of the value of a second random variable and this gives rise to the idea of conditional expectation.

Definition 4.10 (Conditional Expectation). *The **conditional expectation** in discrete and continuous cases corresponds to an expectation with respect to the appropriate conditional probability distribution:*

Discrete

$$\mathbb{E}[Y|X = x] = \sum_{\textit{all } y} y P_{Y|X}(Y = y|X = x)$$

Continuous

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(Y|X = x)\,\mathrm{d}y.$$

Exercise 4.4.1. A random variable (X_1, X_2) has joint density

$$f(x_1, x_2) = \begin{cases} \frac{1}{8}(6 - X_1 - X_2) & \text{for } 0 \le X_1 \le 2, 2 \le X_2 \le 4 \\ 0 & \text{otherwise} \end{cases}$$

– Find $f_{X_1|X_2}$ and $f_{X_2|X_1}$
– Determine $F_{X_1|X_2}$ and $F_{X_2|X_1}$
– Find $\mathbb{E}[X_1|X_2 = x_2]$

Note that before X is known to take the value x, $\mathbb{E}[Y|X]$ is itself a random variable being a function of the random variable X. We might be interested in the distribution of the random variable $\mathbb{E}[Y|X]$, and comparing it with the unconditional expectation $\mathbb{E}[X_1]$. The following is an important result

Theorem 4.1 (Tower property of conditional expectation). *For any two random variables X_1 and X_2*

$$\mathbb{E}[\mathbb{E}[X_1|X_2]] = \mathbb{E}[X_1]$$

Exercise 4.4.2. Prove theorem 4.1 for continuous random variables X_1 and X_2

Exercise 4.4.3. Suppose that the random variable X has a uniform distribution, $X \sim \mathsf{U}[0,1]$, and that, once $X = x$ has been observed, the conditional distribution of Y is $[Y|X = x] \sim \mathsf{U}[x,1]$. Find $\mathbb{E}[Y|x]$ and hence, or otherwise, show that $\mathbb{E}[Y] = 3/4$.

Exercise 4.4.4. Suppose that $\Theta \sim \mathsf{U}[0,1]$ and $(X|\Theta) \sim \mathsf{Bin}\,(2,\Theta)$.
Find $\mathbb{E}[X|\Theta]$ and hence or otherwise show that $\mathbb{E}[X] = 1$.

4.5 Conditional Expectations of Functions of random variables

By extending the theorem on marginal expectations we can relate the conditional and marginal expectations of *functions* of random variables (in particular, their variances).

Theorem 4.2 (Marginal expectation of a transformed Random Variables). *For any random variables X_1 and X_2, and for any function $h(\cdot)$,*

$$\mathbb{E}\big[\mathbb{E}[h(X_1)|X_2]\big] = \mathbb{E}[h(X_1)]. \tag{4.2}$$

Exercise 4.5.1. Prove theorem 4.2 for discrete random variables X_1 and X_2.

Theorem 4.3 (Marginal variance). *For any random variables X_1 and X_2,*

$$\mathbb{V}\mathsf{ar}(X_1) = E\big[\mathbb{V}\mathsf{ar}(X_1|X_2)\big] + \mathbb{V}\mathsf{ar}\big(\mathbb{E}[X_1|X_2]\big). \tag{4.3}$$

Interpretation

"marginal variance = expectation of conditional variance + variance of conditional expectation". So: The uncertainty involved in predicting the value x_1 taken by a random variable X_1 can be decomposed into two parts. One component is the unavoidable uncertainty due to random variation in X_1, but the other can be reduced by observing quantities (here the value x_2 of X_2) related to X_1.

Exercise 4.5.2. Prove theorem 4.3 for continuous random variables X_1 and X_2.
Expand $E\big[\mathbb{V}\mathsf{ar}(X_1|X_2)\big]$ and $\mathbb{V}\mathsf{ar}\big(\mathbb{E}[X_1|X_2]\big)$. Hence show that $\mathbb{V}\mathsf{ar}(X_1) = E\big[\mathbb{V}\mathsf{ar}(X_1|X_2)\big] + \mathbb{V}\mathsf{ar}\big(\mathbb{E}[X_1|X_2]\big)$.

Exercise 4.5.3. Continuing Exercise 4.4.4, in which $\Theta \sim \mathsf{U}[0,1]$, $(X|\Theta) \sim \mathsf{Bin}\,(2,\Theta)$, and $\mathbb{E}[X|\Theta] = 2\Theta$, find $\mathbb{V}\mathsf{ar}\big(\mathbb{E}[X|\Theta]\big)$ and $E\big[\mathbb{V}\mathsf{ar}(X|\Theta)\big]$. Hence or otherwise show that $\mathbb{V}\mathsf{ar}[X] = 2/3$, and comment on the effect on the uncertainty in X of observing Θ.

4.6 Independence of Random Variables

Whilst the previous sections have been concerned with the information that one random variable carries about another, it would seem that there must be pairs of random variables which each provide no information whatsoever about the other. It is, for example, difficult to imagine that the value obtain when a die is rolled in Coventry will tell us much about the outcome of a coin toss taking place at the same time in Lancaster.

There are two equivalent statements of a property termed stochastic independence which capture precisely this idea. The following two definitions are equivalent for both discrete and continuous random variables.

Definition 4.11 (Stochastic Independence). ***Definition 1*** *Random variables $X_1, X_2, \ldots, X_n$ are* ***stochastically independent*** *iff*

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_k) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

Definition 2 *Random variables $X_1, X_2, \ldots, X_n$ are stochastically independent iff*

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_k) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

If X_1 and X_2 are independent then their conditional densities are equal to their marginal densities

Exercise 4.6.1. – Show that for the bivariate normal distribution

$$\rho = 0 \Rightarrow f(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

– Consider the two-dimensional exponential distribution with distribution function

$$F(x_1, x_2) = 1 - e^{-x_1} - e^{-x_2} + e^{-x_1 - x_2 - \rho x_1 x_2}, \quad x_1, x_2 > 0$$

Under what condition are X_1 and X_2 independent? What are the marginal distributions F_{x_1} and F_{x_2} under independence?

4.7 Covariance and Correlation

Having established that sometimes one random variable does convey information about another and in other cases knowing the value of a random variable tells us nothing useful about another random variable it is useful to have mechanisms for characterising the relationship between pairs (or larger groups) of random variables.

Definition 4.12 (Covariance and Correlation). ***Covariance:*** *For random variables X and Y defined on the same probability space*

$$\begin{aligned} \mathbb{C}\mathsf{ov}[X, Y] &= \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] \\ &= \mathbb{E}[XY] - \mu_X \mu_Y \end{aligned}$$

Correlation: *For random variables X and Y defined on the same probability space*

$$\rho[X, Y] = \frac{\mathbb{C}\mathsf{ov}[X, Y]}{\sigma_X \sigma_Y} = \frac{\mathbb{C}\mathsf{ov}[X, Y]}{\sqrt{\mathbb{V}\mathsf{ar}[X]\mathbb{V}\mathsf{ar}[Y]}}$$

provided that $\sigma_X > 0$ and $\sigma_Y > 0$.

Exercise 4.7.1. Let

$$f(x_1, x_2) = \begin{cases} X_1 + X_2 & \text{for } 0 < x_1, x_2 < 1, \\ 0 & \text{otherwise} \end{cases}$$

– Show that X_1 and X_2 are dependent
– Find $\mathbb{C}\mathsf{ov}[X_1, X_2]$ and $\rho[X_1, X_2]$

The following theorem is really a particular example of a result from analysis which finds rather wide application:

Theorem 4.4 (Cauchy-Schwarz Inequality). *Let X and Y have finite second moments. Then*

$$(\mathbb{E}[XY])^2 = |\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

with equality if and only if $\mathbb{P}[Y = cX] = 1$ for some constant c.

Exercise 4.7.2. By considering $\mathbb{E}\big((tX - Y)^2\big)$, or otherwise, prove the *Cauchy Schwarz inequality*. Hence or otherwise prove that the correlation $\rho_{X,Y}$ between X and Y satisfies $|\rho_{X,Y}| \leq 1$. Under what circumstances does $\rho_{X,Y} = 1$?

4.8 Sums of Random Variables

It's very often necessary to combine random variables in various ways, in particular by summing them or taking sample averages. It is possible to make a number of useful statements about the sums of random variables, in particular about their expectation and variance:

For random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ and constants $a_1, \ldots, a_n$, $b_1, \ldots, b_m$

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\ \mathbb{V}\mathsf{ar}\left[\sum_{i=1}^{n} X_i\right] &= \sum_{i=1}^{n} \mathbb{V}\mathsf{ar}[X_i] + 2\sum_{i}\sum_{j<i} \mathbb{C}\mathsf{ov}\left[X_i, X_j\right] \\ \mathbb{C}\mathsf{ov}\left[\sum_{i=1}^{n} a_i X_i \sum_{j=1}^{M} b_j Y_j\right] &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathbb{C}\mathsf{ov}[X_i, Y_j] \end{aligned}$$

Exercise 4.8.1. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Let

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Derive $\mathbb{E}\left[\overline{X}_n\right]$ and $\mathbb{V}\mathsf{ar}\left[\overline{X}_n\right]$.

4.9 Transformation of Random Variables: $Y = g(X)$

Theorem 4.5 (Distribution of a Function of a Random Variable). *Let X be a random variable and $Y = g(X)$ where g is injective (*i.e.* it maps at most one x to any value y). Then*

$$f_Y(y) = f_X(g^{-1}(y)) \left|\frac{dg^{-1}(y)}{dy}\right|$$

given that $(g^{-1}(y))'$ exists and $(g^{-1}(y))' > 0 \quad \forall y$ or $(g^{-1}(y))' < 0 \quad \forall y$. If g is not bijective (one-to-one) there may be values of y for which there exists no x such that $y = g(x)$. Such points clearly have density zero.

When the conditions of this theorem are not satisfied it is necessary to be a little more careful. The most general approach for finding the density of a transformed random variable is to explicitly construct the distribution function of the transformed random variable and then to use the standard approach to turn the distribution function into a density (this approach is discussed in Larry Wasserstein's "All of Statistics").

Exercise 4.9.1. Let X be distributed exponentially with parameter α, that is

$$f_X(x) = \begin{matrix} \alpha e^{-\alpha x} & x \geq 0 \\ 0 & x < 0 \end{matrix}$$

Find the density function of

1. $Y = g(X)$ with $g(X) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\alpha x} & \text{for } x \geq 0 \end{cases}$
2. $Y = X^{\frac{1}{\beta}}, \quad \beta > 0$
3. $Y = g(X)$ with $g(X) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$

Theorem 4.6 (Probability Integral Transformation). *If X is a random variable with continuous $F_X(x)$, then $U = F_X(X)$ is uniformly distributed over the interval $(0, 1)$.*
Conversely if U is uniform over $(0, 1)$, then $X = F_X^{-1}(U)$ has distribution function F_X.

Exercise 4.9.2. Suppose that you wish to generate two independent values from a distribution whose density function is

$$f(x) = x + \frac{1}{2} \quad \text{for } 0 < x < 1.$$

Show how such values can be obtained from uniform $[0, 1]$ random variables. Given the following values generated by a uniform pseudo-random number generator over the range $[0, 1]$

$$x_1 = 0.25, x_2 = 0.46$$

what values would your method provide as samples of the desired type?

4.10 Moment-Generating-Function Technique

The following technique is but one example of a situation in which the moment generating function proves invaluable. It's very often much simpler to deal with the moment generating function of the transformation of a random variable than to deal explicitly with the function itself and sums of random variables can be dealt with especially easy via the moment generating function:

Function of a Variable. For $Y = g(X)$ compute

$$m_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[e^{tg(X)}]$$

If the result is the MGF of a known distribution then it will follow that Y has that distribution (recall from Theorem 2.5 that random variables with equivalent moment generating functions have the same distributions).

Sums of Independent random variables. For $Y = \sum_i X_i$, where the X_i are independent random variables for which the MGF exists $\forall -h < t < h, h > 0$

$$m_Y(t) = \mathbb{E}[e^{\sum_i tX_i}] = \prod_i m_{X_i}(t) \quad \text{for } -h < t < h$$

Thus $\prod_i m_{X_i}(t)$ may be used to identify the distribution of Y as above.

Exercise 4.10.1. Let X_i, $i = 1, \ldots, n$ be independent Poisson random variables with parameter λ_i. Find the distribution of $\sum_i X_i$ using the MGF technique.

4.11 Exercises

Exercise 4.11.1. A small internet cafe has five computers available for use. For $i = 1, \ldots, 5$, let p_i denote the probability that exactly i computers will be in use on any Tuesday evening at 8pm. Suppose $p_0 = 0.05$, $p_1 = 0.10$, $p_2 = 0.20$, $p_3 = 0.30$, $p_4 = 0.25$, and $p_5 = 0.10$. Let X and Y denote the number of computers that will be in use on two independent Tuesday evenings. Determine

1. the joint density of X and Y,
2. $\mathbb{P}[X = Y]$,
3. $\mathbb{P}[X > Y]$.

Exercise 4.11.2. Let Y be the rate (calls per hour) at which calls are made to a customer helpline. Let X be the number of calls during a two-hour period and suppose

$$f(x, y) = \begin{array}{ll} \frac{(2y)^x}{x!} e^{-3y} & \text{if } y > 0 \text{ and } x = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{array}$$

1. Verify that f is a joint density function.
2. Find $\mathbb{P}[X = 0]$.

Exercise 4.11.3. Consider a set of medium-quality bonds, rated according to Standard & Poor's bond-rating scheme, where in terms of risk B > Bb > Bbb. If X is the bond rating and Y is the bond yield (%), suppose the joint density of X and Y is as follows

$f_{X,Y}(x, y)$	1 (Bbb)	2 (Bb)	3 (B)
8.5	0.26	0.10	0.00
11.5	0.04	0.28	0.04
17.5	0.00	0.02	0.26

1. Find $f_X(x)$ and $f_Y(y)$.
2. Find $E(X)$ and $E(Y)$.
3. Are the bond rating and bond yield independent?
4. Find $\mathbb{C}\text{ov}(X, Y)$ and $\rho(X, Y)$.
5. Find $\mathbb{E}[Y|X = 1]$.

Exercise 4.11.4. Suppose that the numbers X_1 and X_2 of male and female piglets in a litter follow independent Poisson distributions with means λ_1 & λ_2 respectively. Find the joint probability function.

1. Let N denote the total number of piglets in a litter. Now assume the model $N \sim \text{Poi}\,(\lambda)$ and $(X_1|N) \sim \text{Bin}\,(N, \theta)$. Again find the joint probability function.
2. Verify that the two models above give identical fitted values, and are therefore in practice indistinguishable.

Exercise 4.11.5. Let Z denote the rate at which customers are served at a certain bank (customers per minute). Assume that Z has the density function

$$f(z) = \begin{cases} 5e^{-5z} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases}$$

1. Find the density function of the average waiting time $T = \frac{1}{Z}$.
2. Find $\mathbb{P}[T > t]$ for $t > 0$. Hence find $\mathbb{P}[T > 5]$ and $\mathbb{P}[T > 10]$.

Exercise 4.11.6. Suppose that the random variables X and Y have a continuous joint distribution, with pdf $f(x, y)$, means μ_X & μ_Y respectively, variances σ_X^2 & σ_Y^2 respectively, and correlation ρ. Suppose also that

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$$

where β_0 and β_1 are constants.

Show that

1. $\int_{-\infty}^{\infty} y f(x, y) \mathrm{d}y = (\beta_0 + \beta_1 x) f_X(x)$,
2. $\mu_Y = \beta_0 + \beta_1 \mu_X$, and
3. $\rho \sigma_X \sigma_Y + \mu_X \mu_Y = \beta_0 \mu_X + \beta_1(\sigma_X^2 + \mu_X^2)$.
   (Hint: use the fact that $\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|X]]$).
4. Hence or otherwise express β_0 and β_1 in terms of μ_X, μ_Y, σ_X, σ_Y & ρ.

5. Inference

As noted in Chapter 3, probability theory provides the basis for statistical inference. Assuming a probability model for a population we can infer characteristics of that population from a random sample taken from it.

5.1 Sample Statistics

Suppose we select a sample of size n from a population of size N. For each i in $\{1, \dots, n\}$, let X_i be a random variable denoting the outcome of the i^{th} observation of a variable of interest. For example, X_i might be the height of the i^{th} person sampled. Under the assumptions of simple random sampling, the X_i are *independent and identically distributed (iid).*

Therefore, if the distribution of a single unit sampled from the population can be characterised by a distribution with density function f, the marginal density function of each X_i is also f and their joint density function g is a simple product of their marginal densities:

$$g(x_1, x_2, \dots, x_n) = f(x_1)f(x_2)\dots f(x_n)$$

In order to make inferences about a population *parameter*, we use sample data to form an *estimate* of the population parameter. We calculate our estimate using an *estimator* or *sample statistic*, which is a function of the X_i. We saw examples of sample statistics in the first lecture, for example the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where n is the size of the sample is an estimator of the population mean, e.g. for discrete X

$$\mu_X = \sum_{j=1}^{N} x_j \mathbb{P}[X = x_j]$$

where N is the number of distinct values which it is possible for an X_i to take.

5.2 Sampling Distributions

Since an estimator $\hat{\theta}$ is a function of random variables, it follows that $\hat{\theta}$ is itself a random variable and possesses its own distribution. The probability distribution of an estimator itself is called a *sampling distribution*. Note that the interpretation of this distribution is important and slightly subtle. The original sample of random variables produces a single value for an estimator; any other sample would also produce a value for the estimator. The sampling distribution of the estimator describes the distribution of these values.

Proposition 5.1 (Distribution of the sample mean). *Let $\overline{X}$ denote the sample mean of a random sample of size n from a normal distribution with mean μ and variance σ^2. Then*

$$\overline{X} \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Proof.

$$\begin{aligned}
m_{\overline{X}}(t) &= \mathbb{E}\left[e^{t\overline{X}}\right] = \mathbb{E}\left[\exp\left\{\frac{t\sum X_i}{n}\right\}\right] \\
&= \mathbb{E}\left[\prod_i \exp\left\{\frac{tX_i}{n}\right\}\right] = \prod_i \mathbb{E}\left[\exp\left\{\frac{tX_i}{n}\right\}\right] \\
&= \prod_i m_{X_i}\left(\frac{t}{n}\right) = \prod_i \exp\left\{\frac{\mu t}{n} + \frac{1}{2}\left(\frac{\sigma t}{n}\right)^2\right\} \\
&= \exp\left\{\mu t + \frac{\frac{1}{2}(\sigma t)^2}{n}\right\}
\end{aligned}$$

which is the MGF of a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. The result follows from theorem 2.5. □

Exercise 5.2.1. Let X denote the number of miles per gallon achieved by cars of a particular model. Suppose that $X \sim \mathsf{N}\left(20, 2^2\right)$. What is the probability that for a random sample of 25 cars, the average miles per gallon will be greater than 21?

The following distributional result is a simple example of a Central Limit Theorem, one of the most important classes of theorem to feature in statistics.

Theorem 5.1 (Central Limit Theorem). *Let f be a density function with mean μ and finite variance σ^2. Let $\overline{X}$ be the sample mean of a random sample of size n from f and let*

$$Z_n = \frac{\overline{X} - \mathbb{E}\left[\overline{X}\right]}{\sqrt{\mathbb{V}\mathsf{ar}\left[\overline{X}\right]}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Then the distribution of Z_n approaches the standard normal distribution as $n \to \infty$. This is often written as: $Z_n \xrightarrow{d} \mathsf{N}(0,1)$ with $\xrightarrow{d}$ denoting convergence in distribution*.*

Proof (for distributions for which MGF exists).

$$\begin{aligned}
m_{z_n(t)} &= \mathbb{E}\left[e^{tz_n}\right] = \mathbb{E}\left[\exp\left\{\frac{t(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{\frac{t}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\frac{\sigma}{\sqrt{n}}}\right\}\right] = \mathbb{E}\left[\prod_{i=1}^{n}\exp\left\{\frac{t}{n}\frac{X_i - \mu}{\frac{\sigma}{\sqrt{n}}}\right\}\right] \\
&= \prod_{i=1}^{n}\mathbb{E}\left[\exp\left\{\frac{t}{n}\frac{X_i - \mu}{\frac{\sigma}{\sqrt{n}}}\right\}\right] = \prod_{i=1}^{n}\mathbb{E}\left[\exp\left\{\frac{t}{\sqrt{n}}\frac{X_i - \mu}{\sigma}\right\}\right] \\
&= \prod_{i=1}^{n}\mathbb{E}\left[\exp\left\{\frac{t}{\sqrt{n}}Y_i\right\}\right] \quad \text{where } Y_i = \frac{X_i - \mu}{\sigma} \\
&= \prod_{i=1}^{n} m_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \prod_{i=1}^{n} m_Y\left(\frac{t}{\sqrt{n}}\right) \quad \text{since } Y_i \text{ are iid.} \\
&= \left[m_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n.
\end{aligned}$$

We may write

$$\begin{aligned} m_Y\left(\frac{t}{\sqrt{n}}\right) &= 1 + \frac{t}{\sqrt{n}} m'_Y(0) + \frac{\left(\frac{t}{\sqrt{n}}\right)^2}{2!} m''_Y(0) + \dots \\ &= 1 + \frac{1}{n}\left(\frac{1}{2}t^2 m''_Y(0) + \frac{1}{3!\sqrt{n}} t^3 m'''_Y(0) + \dots\right) \quad \text{since } m'_Y(0) = 0 \\ &=: 1 + \frac{u_n}{n} \end{aligned}$$

where the final equality defines u_n. Thus:

$$m_{Z_n} = \left[1 + \frac{u_n}{n}\right]^n.$$

Now if $\lim_{n\to\infty} a_n = b$

$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^b$$

and

$$\lim_{n\to\infty} u_n = \frac{t^2}{2}$$

so

$$\lim_{n\to\infty} m_{z_n(t)} = \lim_{n\to\infty}\left[1 + \frac{u_n}{n}\right]^n = e^{\frac{1}{2}t^2}$$

which can be recognised as the MGF of a standard normal. □

Thus, if the sample size is "large enough", the sample mean (not the sample! just having lots of samples from a distribution doesn't change the distribution itself) can be assumed to follow a normal distribution regardless of the population distribution. In practice, this assumption is often taken to be valid for a sample size $n > 30$ (although the size of sample required for it to be a good approximation depends to some extent upon the actual distribution of the sample).

Corollary 5.1. *If $X_1, \dots, X_n$ is a random sample from a density function f with mean μ and variance $\sigma^2 < \infty$, then the distribution of*

$$Z_n = \frac{n\overline{X} - n\mu}{n\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

approaches the standard normal distribution as $n \to \infty$.

Exercise 5.2.2. Suppose a fair coin is tossed 900 times. Let $X_i, i = 1, \dots, 900$ denote a RV that takes the value 1 if the i^{th} outcome is a head and 0 otherwise, so that $E(X_i) = \frac{1}{2}$ and $\mathbb{V}\text{ar}(X_i) = \frac{1}{4}$. What is the probability of obtaining more than 495 heads?

The Chi-Squared Distribution

The chi-squared distribution is a special case of the gamma distribution. We shall see that, up to a scale factor, the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$$

of a standard normal distribution is χ^2 with $n-1$ degrees of freedom.

Definition 5.1. *If X is a random variable with density*

$$f(x) = \chi_k(x; =) \begin{cases} \frac{1}{\Gamma(\frac{k}{2})} \left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1} e^{-\frac{1}{2}x} & x > 0 \\ 0 & \textit{otherwise} \end{cases}$$

then X is defined to have a χ^2 distribution with k degrees of freedom (χ^2_k) where k is a positive integer.

Thus the χ^2_k density is a gamma density with $r = \frac{k}{2}$, $\lambda = \frac{1}{2}$. Hence, if a RV X has a χ^2_k distribution then applying general results for the gamma distribution to these special cases:

$$\begin{aligned} \mathbb{E}[X] &= \frac{\frac{k}{2}}{\frac{1}{2}} = k \\ \mathbb{V}\mathsf{ar}[X] &= \frac{\frac{k}{2}}{\left(\frac{1}{2}\right)^2} = 2k \\ \text{and} \quad m_X(t) &= \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{k}{2}} = \left(\frac{1}{1-2t}\right)^{\frac{k}{2}}, \quad t < \frac{1}{2}. \end{aligned}$$

Exercise 5.2.3. Show that for $X_i \sim \chi^2_{r_i}$

$$\sum_i X_i \sim \chi^2_{\sum r_i}$$

using the MGF technique.

Theorem 5.2. *If the RVs $X_i, i = 1, \ldots, n$ are independently normally distributed with means μ_i and variances σ_i^2 then*

$$U = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

has a χ^2_n distribution.

Proof Let $Z_i = \frac{X_i - \mu}{\sigma_i}$, then $Z_i \sim \mathsf{N}(0,1)$ and $U = \sum_{i=1}^n z_i^2$ is the sum of n squared standard normal RVs.

We have already shown that

$$m_{Z_i^2}(t) = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}}.$$

Thus

$$m_U(t) = \prod_{i=1}^{n} m_{z_i^2}(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}$$

which is the MGF of a χ^2_n distribution.

Corollary 5.2. *If $X_1, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then*

$$U = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n.$$

Theorem 5.3. *If $X_1, \dots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2 then*

(i) $\overline{X}$ and $\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ are independent,
(ii) $\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2}$ has a χ^2_{n-1} distribution.

Proof. Proof of i) is cumbersome and will not be given here.

For part ii) note that

$$\sum_i X_i^2 = \sum_i (X_i - \overline{X})^2 + n\overline{X}^2$$

and given part i) $n\overline{X}$ and $\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ are independent and so we can decompose the MGF as

$$m_{\sum Z_i^2}(t) = m_{\sum_i (Z_i - \overline{Z})^2}(t) m_{n\overline{Z}^2}(t)$$

$$m_{\sum_i (Z_i - \overline{Z})^2} = \frac{m_{\sum Z_i^2}(t)}{m_{n\overline{Z}^2}(t)} = \frac{\left(\frac{1}{1-2t}\right)^{\frac{n}{2}}}{\left(\frac{1}{1-2t}\right)^{\frac{1}{2}}}$$

$$= \left(\frac{1}{1-2t}\right)^{\frac{n-1}{2}}, \quad t < \frac{1}{2}.$$

Thus

$$\sum_i (Z_i - \overline{Z})^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sigma^2}$$

has χ^2_{n-1} distribution. □

Corollary 5.3. *If*

$$s^2 = \frac{1}{n-1} \sum_i (X_i - \overline{X})^2$$

is the sample variance of a random sample of size n from a normal distribution with mean μ and variance σ^2, then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

N.B. The term "degrees of freedom" refers to the number of *independent* squares in the sum. The sum $\sum_{i=1}^{n}(X_i - \mu)^2$ has n independent summands. The sum $\sum_{i=1}^{n}(X_i - \overline{X})^2$ has $n-1$ independent summands because the relation $\sum_{i=1}^{n}(x_i - \overline{X}) = 0$ determines any one of the $X_i - \overline{X}$ given the other $n-1$ of them.

Exercise 5.2.4. Consider the sample of cars from Exercise 5.2.1. What is the probability of observing a sample variance greater than 8?

The t Distribution

The t distribution is closely related to the normal distribution and is needed for making inferences about the mean of a normal distribution when the variance is also unknown.

Definition 5.2. *If $Z \sim \mathsf{N}(0,1)$, $U \sim \chi^2_k$ and Z and U are independent of one other, then*

$$X = \frac{Z}{\sqrt{\frac{U}{k}}} \sim \mathsf{t}_k$$

where t_k denotes a t distribution with k degrees of freedom.

The density of the t_k distribution is:

$$f(x) = \mathsf{t}_k\,(x) = \frac{\Gamma\left[\frac{k+1}{2}\right]}{\Gamma\left[\frac{k}{2}\right]} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+\frac{x^2}{k})^{\frac{k+1}{2}}}.$$

for $-\infty < x < \infty$.

The MGF for the t distribution does not exist[1], but

$$\mathbb{E}[X] = 0$$

as the distribution is symmetric (actually the expectation only exists when $k > 1$ although an extension known as the Cauchy principle value can be defined more generally), and it can be shown that

$$\mathbb{V}\mathsf{ar}[X] = \frac{k}{k-2}, \quad \text{for } k > 2.$$

Theorem 5.4. *If $X \sim \mathsf{t}_k$ then*

$$f(x) = \frac{\Gamma\left[\frac{k+1}{2}\right]}{\Gamma\left[\frac{k}{2}\right]} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+\frac{x^2}{k})^{\frac{k+1}{2}}} \to \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

as $k \to \infty$. That is, as $k \to \infty$ the density approaches the density of a standard normal.

Proof (Outline). Since

$$\lim_{k\to\infty} \left(\left(1+\frac{x^2}{k}\right)^k\right)^{-\frac{1}{2}} = e^{-\frac{x^2}{2}}$$

and

$$\lim_{k\to\infty} \left(1+\frac{x^2}{k}\right)^{-\frac{1}{2}} = 1$$

it can be seen that

$$\frac{1}{(1+\frac{x^2}{k})^{\frac{k+1}{2}}} = \left(1+\frac{x^2}{k}\right)^{-\frac{k}{2}} \left(1+\frac{x^2}{k}\right)^{-\frac{1}{2}} \to e^{-\frac{x^2}{2}}$$

as $k \to \infty$. It can also be shown that

$$\frac{\sqrt{2}\Gamma\left[\frac{k+1}{2}\right]}{\sqrt{k}\Gamma\left[\frac{k}{2}\right]} \to 1$$

as $k \to \infty$, and the result follows. □

[1] This is because the t distribution decays at a polynomial, rather than exponential rate and so the expectation of any exponential of a t-distributed random variable is undefined.

Previously we saw that

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim \mathsf{N}(0,1)$$

and

$$U = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

when $\bar{X}$ is the sample mean and s^2 the sample variance of a sample from a normal population with mean μ and variance σ^2. Hence

$$\begin{aligned} T &= \frac{Z}{\sqrt{\frac{U}{n-1}}} \sim \mathsf{t}_{n-1} \\ &= \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \Bigg/ \sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}} \sim \mathsf{t}_{n-1} \\ &= \frac{\sqrt{n}(\overline{X} - \mu)}{s} \sim \mathsf{t}_{n-1}. \end{aligned}$$

Example 5.1. Suppose that X, the life of a candle, follows a normal distribution with a mean of 8.6 hours. In a random sample of 10 candles, the sample mean was 8.2 and the sample variance was 0.4. What is the probability of observing a mean of 8.2 or lower?

◁

Suppose we have two independent random samples

$$X_1, \ldots, X_{n_1} \sim \mathsf{N}\left(\mu_1, \sigma^2\right) \quad Y_1, \ldots, Y_{n_2} \sim \mathsf{N}\left(\mu_2, \sigma^2\right)$$

We can use the result that for $X_i \sim \mathsf{N}\left(\mu_i, \sigma_i^2\right)$

$$\sum a_i X_i \sim \mathsf{N}\left(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2\right)$$

to give

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)} \sim \mathsf{N}(0,1)$$

and the result that for $X_i \sim \chi^2_{r_i}$

$$\sum X_i \sim \chi^2_{\sum r_i}$$

to give

$$U = \frac{1}{\sigma^2}\left[(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2\right] \sim \chi^2_{n_1+n_2-2}$$

from which we can establish, by the same arguments as were used previously, that

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{s_p\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)} \sim \mathsf{t}_{n_1+n_2-2}$$

where s_p^2 is the *pooled variance*

$$s_p^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}.$$

Convince yourself that these steps all make sense.

Exercise 5.2.5. Consider two brands of washing powder. Suppose 15 random samples were taken and tested for each powder, giving the following results

	Mean Score (Variance)
Brand A	17.89 (3.43)
Brand B	16.97 (4.76)

Assuming the scores are normally distributed and that there is no difference in population means, what is the probability of observing a difference in sample means of 17.89 - 16.97 = 0.92 or greater?

The F Distribution

The distribution of the difference in sample means derived in the last section was based on the simplifying assumption that the variances of the two populations were equal. The F Distribution is useful for making inferences about the ratio of two *unknown* variances.

Definition 5.3. *Suppose U and V are independently distributed with $U \sim \chi^2_m$ and $v \sim \chi^2_n$. Then the random variable*

$$X = \frac{U}{m} \Big/ \frac{V}{n}$$

is distributed according to an F distribution with m and n degrees of freedom.

The density of X is given by

$$f(x) = \mathsf{F}_{m,n}(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{x^{\frac{m-2}{2}}}{\left[1+\frac{m}{n}x\right]^{\frac{m+n}{2}}}, \quad x > 0$$

The proof of this uses a multivariate version of Theorem 4.5 given in the previous chapter. Writing $X = g(U, V)$ it is possible to express the density of X in terms of those of U and V as well as the derivatives of $g(u, v) = u/v$.

Returning to the case of two independent random samples from normal populations of common variance, but differing means:

$$X_1, \ldots, X_{n_1} \sim \mathsf{N}\left(\mu_1, \sigma^2\right) \quad Y_1, \ldots, Y_{n_2} \sim \mathsf{N}\left(\mu_2, \sigma^2\right)$$

we have

$$\frac{(n_1-1)s_X^2}{\sigma^2} \sim \chi^2_{n_1-1} \quad \text{and} \quad \frac{(n_2-1)s_Y^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

so that

$$\frac{\frac{(n_1-1)s_X^2}{\sigma^2}/(n_1-1)}{\frac{(n_2-1)s_Y^2}{\sigma^2}/(n_2-1)} = \frac{s_X^2}{s_Y^2} \sim \mathsf{F}_{n_1-1,n_2-1}.$$

5.3 Point Estimation

The sample mean and variance are examples of *point estimators*, because the estimates they produce are single point values, rather than a range of values. For a given parameter there are an infinite number of possible estimators, hence the question arises: what makes a “good” estimator? This section describes two properties of “good” estimators.

Definition 5.4 (Unbiasedness). *Let X be a random variable with pdf $f(x;\theta)$, where $\theta \in \Omega \subset \mathbb{R}^p$ is some unknown parameter, $p \geq 1$. Let $X_1, \ldots, X_n$ be a random sample from the distribution of X and let $\hat{\theta}$ denote a statistic. $\hat{\theta}$ is an* unbiased estimator *of θ if*

$$\mathbb{E}[\hat{\theta}] = \theta \quad \forall \, \theta \in \Omega$$

where the expectation is with respect to $f(x;\theta)$.

If $\hat{\theta}$ is not unbiased, we say that $\hat{\theta}$ is a biased estimator of θ, with

$$\mathsf{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

If $\mathsf{Bias}(\hat{\theta}) \to 0$ when $n \to \infty$ then we say that $\hat{\theta}$ is asymptotically unbiased.

Exercise 5.3.1. Consider a random sample $X_1, \ldots, X_n$ from a population with unknown mean μ and variance σ^2. Consider the following estimators for μ

1. $\hat{\mu}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
2. $\hat{\mu}_2 = \frac{X_1 + X_2}{2}$

Determine whether or not these estimators are unbiased.

Example 5.2. Consider the following estimator of the population variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$$

We find that

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n X_i^2 - n\overline{X}^2\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^n \mathbb{E}[X_i^2] - n\mathbb{E}[\overline{X}^2]\right) \\
&= \frac{1}{n}\left(n\sigma^2 + n\mu^2 - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

So this estimator is biased with bias equal to $\frac{-\sigma^2}{n}$. As this decays to zero as $n \to \infty$ it is an asymptotically unbiased estimator. However, we can see that the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$$

is unbiased. ◁

Consistency

In order to define a consistent estimator, we first define **convergence in probability**.

Definition 5.5 (Convergence in probability). *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. We say that X_n converges in probability to X if $\forall\ \epsilon > 0$*

$$\lim_{n\to\infty} P\left[|X_n - X| \geq \epsilon\right] = 0, \textit{ or equivalently} \qquad \lim_{n\to\infty} P\left[|X_n - X| < \epsilon\right] = 1.$$

If so, it is common to write $X_n \xrightarrow{p} X$.

In statistics the limiting variable X is often a constant, say a (*i.e.* it is a degenerate random variable, one which places probability 1 on a single value). In this case we write $X_n \xrightarrow{p} a$.

Definition 5.6 (Consistent estimator). *Let $X_1, \dots, X_n$ be a sample from the distribution of X where X is a random variable with distribution function $F(x;\theta)$. Let $\hat{\theta}$ denote a statistic. $\hat{\theta}$ is a* ***consistent estimator*** *of θ if, whatever the value of θ,*

$$\hat{\theta} \xrightarrow{p} \theta.$$

A particular case of consistency is often used to justify using sample averages:

Theorem 5.5 (Weak Law of Large Numbers). *Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ with $X_1, \dots, X_n$ iid. Then*

$$\overline{X} \xrightarrow{p} \mu,$$

i.e. $\overline{X}$ *is a consistent estimator of μ.*

Proof. Recall $\mathbb{E}[\overline{X}] = \mu$; $\mathbb{V}\mathsf{ar}[\overline{X}] = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$. By Chebyshev's inequality (Theorem 2.3) we have for any $\epsilon > 0$

$$P\left[|\overline{X} - \mu| \geq \epsilon\right] = P\left[|\overline{X} - \mu| \geq \left(\frac{\epsilon}{\sigma_{\overline{X}}}\right)\sigma_{\overline{X}}\right] \leq \frac{\sigma^2_{\overline{X}}}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

□

Theorem 5.6. *Suppose $X_n \xrightarrow{p} a$ and the real function g is continuous at a. Then $g(X_n) \xrightarrow{p} g(a)$.*

Proof. Let $\epsilon > 0$. Then since g is continuous at a, there exists $\delta > 0$ such that if $|x - a| < \delta$, then $|g(x) - g(a)| \leq \epsilon$. Thus (using the *contrapositive*)

$$|g(x) - g(a)| \geq \epsilon \Rightarrow |x - a| \geq \delta.$$

Substituting X_n for x we obtain

$$\mathbb{P}[|g(X_n) - g(a)| \geq \epsilon] \leq \mathbb{P}[|X_n - a| \geq \delta.]$$

Since $X_n \xrightarrow{p} a$, the last term goes to 0 as $n \to \infty$. □

As a result of this theorem, if $X_n \xrightarrow{p} a$, then

$$X_n^2 \xrightarrow{p} a^2 \qquad \frac{1}{X_n} \xrightarrow{p} \frac{1}{a},\ a \neq 0 \qquad \sqrt{X_n} \xrightarrow{p} \sqrt{a},\ a \geq 0.$$

Using this, we can show that as $n \to \infty$, $\overline{X}^2 \xrightarrow{p} \mu^2$ and:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum X_i^2 - \overline{X}^2\right) \xrightarrow{p} \mathbb{E}[X_1^2] - \mu^2 = \sigma^2.$$

thus the sample variance is a consistent estimator of σ^2.

Consistency according to the definition above may be hard to prove, but it turns out that a sufficient (though not necessary) condition for consistency is that $\mathsf{Bias}(\hat{\theta}) \to 0$ and $\mathbb{V}\mathsf{ar}(\hat{\theta}) \to 0$ as $n \to \infty$.

Definition 5.7 (Consistency in Mean-Squared Error). *If $\hat{\theta}$ is an estimator of θ, then the mean squared error of $\hat{\theta}$ is defined as*

$$\mathsf{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

and $\hat{\theta}$ is said to be consistent in MSE *if* $\mathsf{MSE}(\hat{\theta}) \to 0$ *as the size of the sample on which $\hat{\theta}$ is based increases to infinity.*

It can be shown (for example, by Chebyshev's inequality) that

$$P\left[|\hat{\theta} - \theta| > \epsilon\right] \le \int_{-\infty}^{\infty} \frac{(\hat{\theta} - \theta)^2}{\epsilon^2} f(\hat{\theta})\, d\hat{\theta} = \frac{1}{\epsilon^2}\mathsf{MSE}(\hat{\theta}).$$

where $f(\cdot)$ denotes the density function of the sampling distribution of $\hat{\theta}$.

So $P\left[|\hat{\theta} - \theta| > \epsilon\right] \to 0$ as $n \to \infty$ if $\mathsf{MSE}(\hat{\theta}) \to 0$ *as $n \to \infty$.* i.e. *Consistency in* MSE *implies consistency.*

Exercise 5.3.2. Show that

$$\mathsf{MSE}(\hat{\theta}) = \mathbb{V}\mathsf{ar}(\hat{\theta}) + \left[\mathsf{Bias}(\hat{\theta})\right]^2.$$

Using this equality determine whether the estimators from Exercise 5.3.1 are consistent in MSE.

5.4 Interval Estimation

When a sample is drawn it is unlikely that the value of a point estimate $\hat{\theta}$ is equal to the true θ. In fact if X is continuous then it's generally the case that for any finite sample size $P(\hat{\theta} = \theta) = 0$. An alternative approach is to estimate an interval for θ. This section describes *confidence intervals* which are intervals constructed such that they contain θ with some level of confidence.

Definition 5.8 (Confidence interval). *Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$ where θ is an unknown parameter in the parameter space Ω. If L and U are statistics such that*

$$\mathbb{P}[L \le \theta \le U] = 1 - \alpha$$

then the interval (L, U) is a $100(1-\alpha)\%$ confidence interval for θ.

$1-\alpha$ is known as the confidence coefficient *and α is the* level of significance*.*

Note that the confidence interval defined above is a *random interval* because it is a function of the statistics U and L which are random variables. The parameter θ on the other hand, is fixed. Therefore it is misleading to say that θ lies in the interval with probability $(1-\alpha)$. Rather the probability that the interval contains θ is $(1-\alpha)$. Thus, if we conducted a large number of replications of an experiment which produced such a confidence interval, then we would expect $100(1-\alpha)$ of those confidence intervals to contain θ.

There are several ways in which confidence intervals can be constructed. The basic procedure we shall use to construct a $100(1-\alpha)\%$ confidence interval for a parameter is as follows

1. Select a sample statistic to estimate the parameter.
2. Identify the sampling distribution for the statistic.
3. Determine the bounds within which the sample statistic will reside with probability $1-\alpha$.
4. Invert these bounds to obtain an expression in terms of θ.

For step 2 we shall often invoke the Central Limit Theorem to approximate the sampling distribution of a statistic and hence obtain an *approximate* confidence interval. The following section shows how a confidence interval may be constructed in such a case.

Confidence Intervals based on the CLT

Suppose that we are interested in the population mean θ and we wish to use the sample mean $\overline{X}$ as an estimator for θ. Then if the population density has variance σ^2 the CLT states that

$$Z_n = \frac{\overline{X} - \theta}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} \mathsf{N}\,(0,1)$$

where n is the sample size.

Hence $P(-1.96 < Z < 1.96) = 0.95$ and we can construct a 95% confidence interval as follows

$$\begin{aligned} 0.95 &= P\left(-1.96 < \frac{\sqrt{n}(\overline{X} - \theta)}{\sigma} < 1.96\right) \\ &= P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

In the above derivation it was assumed that the variance σ^2 was known. If the variance is unknown, but we estimate σ with a consistent estimator s, then it can be shown that

$$\frac{\sqrt{n}(\overline{X} - \theta)}{s} \xrightarrow{d} \mathsf{N}\,(0,1)$$

and hence

$$\overline{X} \pm 1.96\frac{s}{\sqrt{n}}$$

would be the endpoints of an approximate 95% CI for θ.

More generally the endpoints of a $100(1-\alpha)\%$ CI for θ are given by

$$\overline{X} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

where $z_{\frac{\alpha}{2}}$ is such that

$$P\left(Z > z_{\frac{\alpha}{2}}\right) = 1 - \Phi\left(z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

The statistic $\frac{s}{\sqrt{n}}$ is known as the *standard error.*

Note that the confidence interval satisfies a number of intuitively logical requirements:

– a higher confidence coefficient leads to a longer CI,
– a larger sample size leads to a shorter CI,
– a larger variance leads to a longer CI.

We now consider four cases where the CI is found using the CLT approach.

CI for the Mean

Let $X_1, \ldots, X_n$ be a random sample from the distribution of an RV X with unknown mean μ and variance σ^2. Let $\overline{X}$ and s^2 denote the sample mean and variance. We have shown that s^2 is a consistent estimator for σ^2 and hence

$$\frac{\sqrt{n}(\overline{X} - \theta)}{s} \xrightarrow{d} \mathsf{N}\,(0,1)\,.$$

Thus the endpoints of an approximate $100(1-\alpha)\%$ CI for μ are

$$\overline{X} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}.$$

CI for a Proportion

Let $X \sim \mathsf{Ber}\,(p)$ and let $X_1, \ldots, X_n$ be a random sample from the distribution of X. Let $\hat{p} = \overline{X}$ be the sample proportion of successes. By the CLT we have

$$\hat{p} \xrightarrow{d} \mathsf{N}\left(p, \frac{p(1-p)}{n}\right).$$

By the weak law of large numbers $\hat{p}$ is a consistent estimator of p and by Theorem 5.6 $\hat{p}(1-\hat{p})$ is a consistent estimator of $p(1-p)$. Hence, an approximate $100(1-\alpha)\%$ CI for p is

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Exercise 5.4.1. In a random sample 136 of 400 people given a flu vaccine experienced some discomfort. Construct a 95% CI for the true proportion of people who will experience discomfort from the vaccine.

CI for Differences in Means

Suppose we have two random variables X and Y and we are interested in comparing their means, μ_1 and μ_2 respectively. We shall derive a CI for $\Delta = \mu_1 - \mu_2$. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be random samples from the distributions of X and Y respectively and let $\sigma_1 = \mathbb{V}\mathsf{ar}[X]$ and $\sigma_2 = \mathbb{V}\mathsf{ar}[Y]$. Suppose we estimate Δ by the unbiased estimator $\overline{X} - \overline{Y}$ and estimate σ_1 and σ_2 by the appropriate sample variances. Then because

$$\overline{X} \xrightarrow{d} \mathsf{N}\left(mu_1, \frac{s_1^2}{n}\right)$$

and

$$\overline{Y} \xrightarrow{d} \mathsf{N}\left(\mu_2, \frac{s_2^2}{n}\right)$$

the endpoints of an approximate $100(1-\alpha)\%$ CI for Δ are

$$\overline{X} - \overline{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

CI for Differences in Proportions

For differences in proportions where $X \sim \mathsf{Ber}\,(p_1)$ and $Y \sim \mathsf{Ber}\,(p_2)$ the approximate CI from the CLT is

$$\hat{p_1} - \hat{p_2} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n} + \frac{\hat{p_2}(1-\hat{p_2})}{n}}.$$

Exercise 5.4.2. If 132 of 200 male voters and 90 out of 159 female voters favour a certain candidate running for governor of Illinois, find a 99% CI for the difference between the actual proportions of male and female voters who favour the candidate.

Note that whenever a CI is obtained via the CLT it holds exactly only asymptotically; such confidence intervals are only approximations for finite samples although the approximations can be extremely good for even moderately large samples.

CIs for Normal Populations

We now consider the special case of sampling from normal populations. In this case we can use the exact sampling distribution to derive confidence intervals as shown in the following examples.

CI for the Mean of a Normal Distribution Suppose $X_1, \ldots, X_n$ are a random sample from a $\mathsf{N}\left(\mu, \sigma^2\right)$ distribution. In Section 5.2 we saw that

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{s} \sim \mathsf{t}_{n-1}.$$

Hence the exact $100(1-\alpha)\%$ CI for μ has endpoints

$$\overline{X} \pm \mathsf{t}_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}.$$

Remember that $\mathsf{t}_{\frac{\alpha}{2}, n-1} \to z_{\frac{\alpha}{2}}$ as $n \to \infty$. In practice, if $n \geq 120$, the quantiles of the standard normal can be used instead (they are essentially indistinguishable from those of the t distribution in this regime).

Exercise 5.4.3. Suppose we observe the following data on lactic acid concentrations in cheese

0.86 1.53 1.57 1.81 0.99 1.09 1.29 1.78 1.29 1.58

Assuming that these data are a random sample from a normal distribution, calculate a 90% CI for the mean of this distribution.

CI for Differences in Means of Normal Populations with Equal Variance

Suppose we have two independent random samples

$$X_1, \ldots, X_{n_1} \sim \mathsf{N}\left(\mu_1, \sigma^2\right) \quad Y_1, \ldots, Y_{n_2} \sim \mathsf{N}\left(\mu_2, \sigma^2\right).$$

In Section 5.2 we saw that

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{s_p\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)} \sim \mathsf{t}_{n_1+n_2-2},$$

where s_p^2 is the pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}$$

which can be shown to be an unbiased estimator of σ^2.

Therefore the exact $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ in this case is

$$\overline{X} - \overline{Y} \pm \mathsf{t}_{\frac{\alpha}{2}, n_1+n_2-2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, s_p.$$

In the case of unequal variances, an exact CI cannot be derived.

Exercise 5.4.4. Consider a study to compare the nicotine contents of two brands of cigarettes. 10 cigarettes of Brand A had an average nicotine content of 3.1 milligrams with a standard error of 0.5 mg, while 8 cigarettes of Brand B had an average nicotine content of 2.7 mg with a standard error of 0.7 mg.

Assuming that the 2 data sets are independent random samples from normal populations with equal variances, construct a 95% confidence interval for the difference between the mean nicotine contents of the two brands.

CI for the Variance of a Normal Population

Suppose $X_1, \ldots, X_n$ are iidnormal with unknown μ and σ^2. From the Corollary of Theorem 5.3 we know that

$$Q = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Therefore

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

where $P(Q < \chi^2_{\alpha/2,n-1}) = 1 - \frac{\alpha}{2}$. Hence an exact $100(1-\alpha)\%$ CI for σ^2 is given by

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

Note that the χ^2 distribution is asymmetric, hence this CI is asymmetric, unlike the previous CIs we have defined. This illustrates an important point: any interval which has the appropriate coverage probability can be used as a confidence interval and there is no unique $(1-\alpha)$ confidence interval for a statistic (imagine shifting the left endpoint of the interval slightly towards the centre; the right endpoint could then be moved a small amount away from the centre to compensate). Symmetry is one of the properties which is often considered desirable when dealing with confidence intervals, but it's certainly not essential.

CI for the Ratio of Variance of Normal Populations

Suppose we have two independent random samples

$$X_1, \ldots, X_m \sim \mathsf{N}\left(\mu_1, \sigma_1^2\right) \quad Y_1, \ldots, Y_n \sim \mathsf{N}\left(\mu_2, \sigma_2^2\right)$$

where μ_1, μ_2, σ_1 and σ_2 are unknown. Let

$$Q = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

In Section 5.2 we found that

$$\frac{s_X^2}{s_Y^2} \sim \mathsf{F}_{n_1-1,n_2-1}$$

Thus we have

$$P(F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} < Q < F_{n_1-1,n_2-1,\alpha/2}) = 1-\alpha$$

$$P\left(\frac{1}{F_{n_2-1,n_1-1,\frac{\alpha}{2}}} < Q < F_{n_1-1,n_2-1,\alpha/2}\right) = 1-\alpha$$

from which we derive the following exact CI for $\frac{\sigma_1^2}{\sigma_2^2}$

$$\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{n_2-1,n_1-1,\frac{\alpha}{2}}}, \frac{s_1^2}{s_2^2}F_{n_1-1,n_2-1,\frac{\alpha}{2}}\right).$$

Note this CI is also asymmetric.

Example 5.3. Consider the study described in Exercise 5.4.4. A 95% CI for the ratio of the variances of the two populations is given by

$$\left(\frac{0.5^2}{0.7^2}\frac{1}{F_{7,9,0.025}}, \frac{0.5^2}{0.7^2}F_{9,7,0.025}\right) = \left(\frac{0.5^2}{0.7^2}\frac{1}{4.197}, \frac{0.5^2}{0.7^2}(4.823)\right) = (0.122, 2.46)$$

which includes 1, so the assumption of equal variances seems reasonable in this case. ◁

5.5 Exercises

Exercise 5.5.1. Suppose that a random sample is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2. How large a random sample must be taken in order that

$$\mathbb{P}[|\overline{X} - \theta| < 0.1] \geq 0.95$$

for every possible value of θ?

Exercise 5.5.2. Suppose that a random sample is to be taken from a Bernoulli distribution for which the value of the parameter p is unknown. Suppose also that it is believed that the value of p is in the neighbourhood of 0.2. Use the central limit theorem to find approximately the size of a random sample that must be taken in order that

$$\mathbb{P}[|\overline{X} - p| < 0.1] \geq 0.75$$

when $p = 0.2$.

Exercise 5.5.3. Let $X_1, \ldots, X_n$ be a random sample from the distribution of X. Let $\mathbb{E}[X] = \mu$ and $\mathbb{V}\mathrm{ar}[X] = \sigma^2$. Consider the following estimators for μ

1. $\hat{\mu_1} = a\overline{X}$, $0 < a < 1$;
2. $\hat{\mu_2} = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$.
3. $\hat{\mu_3} = \frac{1}{n^2} \sum_{i=1}^n X_i$
4. $\hat{\mu_4} = \frac{1}{n-1} \sum_{i=1}^n X_i$

Which of these estimators are unbiased? If they are not unbiased, are they asymptotically unbiased?

Which estimators are consistent in MSE?

Exercise 5.5.4. Let $X_1, \ldots, X_n$ be a random sample from the distribution of X. Let $\mathbb{E}[X] = \mu$ and $\mathbb{V}\mathrm{ar}[X] = \sigma^2$. Prove that $\overline{X}$ is the unbiased linear estimator with the smallest variance.

Note: the class of all linear estimators is

$$\tilde{X} = \sum_{i=1}^n a_i X_i \quad \text{with} \quad \mathbb{V}\mathrm{ar}[\tilde{X}] = \sigma^2 \sum_{i=1}^n a_i^2$$

Exercise 5.5.5. Let X_n, $n = 1, 2, 3, \ldots$ be a sequence of random variables with

$$P(X_n = a) = 1 - \frac{1}{n}, \quad a \in \mathbb{R}$$

$$P(X_n = n^k) = \frac{1}{n}, \quad k > 1, k \text{ fixed.}$$

Show that X_n is a consistent (in probability, Definition 5.5) estimator for a. Is X_n also consistent in MSE?

Exercise 5.5.6. The following table gives the weight in grams of 100 airmail envelopes

weight	frequency	weight	frequency	weight	frequency
1.80	1	1.89	5	1.98	4
1.81	0	1.90	7	1.99	3
1.82	1	1.91	6	2.00	7
1.83	1	1.92	8	2.01	2
1.84	1	1.93	8	2.02	4
1.85	1	1.94	9	2.03	5
1.86	1	1.95	4	2.04	1
1.87	2	1.96	11	2.05	2
1.88	3	1.97	3		

Assuming that these data are a random sample from a normal population, give 95% confidence intervals for

1. μ, with $\sigma = 0.05$,
2. μ, with σ unknown,
3. σ, with μ unknown.

Exercise 5.5.7. A sample was taken of 40 eligible voters. 12 of these said they were going to vote for party A. Give 90% and 99% confidence intervals for the proportion p of voters of party A in the population

1. using the binomial distribution,
2. using the normal distribution as an approximation.

Part II

Core Material

6. Maximum Likelihood Estimation

The construction of estimators for individually problems is made much simpler by the existence of some general techniques for doing so. One of the most widely used approaches in statistics centres around the likelihood function.

6.1 Likelihood Function and ML estimator

Suppose we have a simple random sample $X_1, \ldots, X_n$ from a density $f(x; \theta)$ parameterised by some possibly unknown parameter θ.

The the joint pdf of the entire data set is

$$f(\mathbf{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

with $\mathbf{x} = (x_1, \ldots, x_n)^T$.

The **likelihood function** $L(\theta; \mathbf{x})$ is this joint pdf viewed as a function of θ with $\mathbf{x}$ fixed to the observed data. It is often more convenient to work with the **log-likelihood**

$$l(\theta; \mathbf{x}) = \log \left[L(\theta; \mathbf{x}) \right]$$

which in the simple random sampling case above yields:

$$l(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

The **maximum likelihood estimator (MLE)** is the value $\hat{\theta}$ which maximises $L(\theta; \mathbf{x})$. The MLE also maximises $l(\theta; \mathbf{x})$ because $\log(\cdot)$ is a monotonic function. Usually it is easier to maximise $l(\theta; \mathbf{x})$, so we work with this.

Actually finding the maximum of the likelihood function is an optimization problem. For simple cases we can find closed-form expressions for $\hat{\theta}$ as a function of $\mathbf{x}$. However, this is often not possible and it is necessary to result to numerical approximations: often in the form of iterative numerical optimisation procedures.

It can be useful to plot (log-)likelihood surface to identify potential problems as complicated optimisation algorithms can go rather badly wrong if applied without considering the function being optimised.

Example 6.1. Suppose $X_1, \ldots, X_n$ is a random sample from the exponential distribution with pdf

$$f(x; \theta) = \begin{cases} \theta e^{-\theta x} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$L(\theta;\mathbf{x}) = \theta^n e^{-\theta\sum_{i=1}^n x_i}, \quad \text{so} \quad l(\theta;\mathbf{x}) = n\log(\theta) - \theta\sum_{i=1}^n x_i$$

$$\frac{\partial l}{\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^n x_i = 0 \quad \text{gives the MLE} \quad \hat{\theta} = \frac{1}{\overline{\mathbf{X}}}$$

$$\text{Check:} \quad \frac{\partial^2 l}{\partial\theta^2} = \frac{-n}{\theta^2} < 0, \quad \text{so } \hat{\theta} \text{ does correspond to a maximum.}$$

◁

Exercise 6.1.1. Suppose $X_1, \ldots, X_n$ form a random sample from the Poisson distribution $Poisson(\theta)$. Derive the ML estimator for θ.

A widely accepted idea in statistics, known as the **likelihood principle** is that: all information about the parameters contained within the data is contained within the likelihood function. Although not universally accepted it's a principle which one would at least need a reason to deviate from.

6.2 MLE and Exponential Families of Distributions

Definition 6.1 (The Exponential Family of Distributions). *The RV X belongs to the* ***k-parameter exponential family of distributions*** *iff its pdf can be written in the form*

$$f(x;\theta) = \exp\left\{\sum_{j=1}^k A_j(\theta)B_j(x) + C(x) + D(\theta)\right\} \tag{6.1}$$

where

$A_1(\theta), \ldots, A_k(\theta), D(\theta)$ are functions of θ alone.
$B_1(x), \ldots, B_k(x), C(x)$ are functions of x alone.

Example 6.2.

Exponential $(k=1)$: $\quad \theta e^{-\theta x} = \exp\{(-\theta)(x) + \log\theta\}$

i.e. $A(\theta) = -\theta, B(x) = x, C(x) = 0$ and $D(\theta) = \log\theta$.

Normal $(k=2)$:
$$(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$
$$= \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}$$

i.e. $A_1(\theta) = -\frac{1}{2\sigma^2}, A_2(\theta) = \frac{\mu}{\sigma^2}, B_1(x) = x^2, B_2(x) = x, C(x) = 0$ and $D(\theta) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$. ◁

Table 6.1 shows that many well known distributions belong to the exponential family.

Exercise 6.2.1. Fill in the gaps in table 6.1 for the following:

Poisson: $\dfrac{e^{-\theta}\theta^x}{x!}$.

Gamma (two parameters): $\dfrac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}$.

Distribution	$f(x;\theta)$	$A(\theta)$	$B(x)$	$C(x)$	$D(\theta)$
$k=1$					
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	$\log(p/(1-p))$	x	$ln\binom{n}{x}$	$n\log(1-p)$
Poisson					
Exponential	$\theta\exp^{-\theta x}$	$-\theta$	x	0	$\log\theta$
$N(0,\sigma^2)$	$(2\pi\sigma^2)^{-1/2}\exp^{-\frac{x^2}{2\sigma^2}}$	$-\frac{1}{2\sigma^2}$	x^2	0	$-\frac{1}{2}\log(2\pi sigma^2)$
$N(\mu,1)$	$(2\pi)^{-1/2}\exp^{-\frac{(x-\mu)^2}{2}}$	μ	x	$-\frac{x^2}{2}$	$-\frac{1}{2}\log(2\pi)-\frac{\mu^2}{2}$
Gamma (1 param)	$\theta^r x^{r-1}\exp^{-\theta x}/(r-1)!$	$-\theta$	x	$(r-1)\log x$	$r\log\theta-\log((r-1)!)$
$k=2$					
$N(\mu,\sigma^2)$	$(2\pi\sigma^2)^{-1/2}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$A_1(\theta)=-\frac{1}{2\sigma^2}$ $A_2(\theta)=\frac{\mu}{\sigma^2}$	$B_1(x)=x^2$ $B_2(x)=x$	0	$-\frac{1}{2}\log(2\pi sigma^2)$ $-\frac{1}{2}\mu^2/\sigma^2$
Gamma (2 param)					

Table 6.1. Some members of the exponential family of distributions

Natural Parameterization

Letting $\phi_j = A_j(\theta), j = 1, \ldots, k$, the exponential form becomes

$$f(x;\phi) = \exp\left\{\sum_{j=1}^{k}\phi_j B_j(x) + C(x) + D(\phi)\right\}.$$

The parameters $\phi_1, \ldots, \phi_k$ are called **natural** or **canonical** parameters.

The exponential density in terms of its natural parameter $\phi = -\theta$ is:

$$-\phi e^{\phi x}.$$

Whilst the normal density in terms of its natural parameters $\phi_1 = -\frac{1}{2\sigma^2}, \phi_2 = \frac{\mu}{\sigma^2}$:

$$\exp\left(\phi_1 x^2 + \phi_2 x + \frac{\phi_2^2}{4\phi} - \frac{1}{2}\log\left(-\frac{\pi}{\phi_1}\right)\right).$$

Theorem 6.1 (MLEs of Natural Parameters). *Suppose $X_1, \ldots, X_n$ form a random sample from a distribution which is a member of the k-parameter exponential family with pdf*

$$f(x;\phi) = \exp\left(\sum_{j=1}^{k}\phi_j B_j(x) + C(x) + D(\phi)\right)$$

then the MLEs of $\phi_1, \ldots, \phi_k$ are found by solving the equations

$$t_j = \mathbb{E}[T_j], \quad j = 1, \ldots, k$$

$$\textit{where} \quad T_j = \sum_{i=1}^{n} B_j(X_i), \quad j = 1, \ldots, k \quad \textit{and} \quad t_j = \sum_{i=1}^{n} B_j(x_i).$$

Proof. The likelihood function is

$$
\begin{aligned}
L(\phi;\mathbf{x}) &= \prod_{i=1}^{n} f(x_i;\phi) = \prod_{i=1}^{n} \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x_i) + C(x_i) + D(\phi)\right\} \\
&= \exp\left\{\sum_{j=1}^{k} \phi_j \sum_{i=1}^{n} B_j(x_i) + \sum_{i=1}^{n} C(x_i) + nD(\phi)\right\} \\
&= \exp\left\{\sum_{j=1}^{k} \phi_j t_j + \sum_{i=1}^{n} C(x_i) + nD(\phi)\right\} \\
\Rightarrow l(\phi;\mathbf{x}) &= \text{ constant } + \sum_{j=1}^{k} \phi_j t_j + nD(\phi)
\end{aligned}
$$

The log likelihood function is

$$
\begin{aligned}
l(\phi;\mathbf{x}) &= \text{ constant } + \sum_{j=1}^{k} \phi_j t_j + nD(\phi) \\
\Rightarrow \frac{\partial l}{\partial \phi_j} &= t_j + n\frac{\partial D(\phi)}{\partial \phi_j}
\end{aligned}
$$

Furthermore,

$$
\mathbb{E}\left[\frac{\partial l}{\partial \phi_j}\right] = 0, \quad \textit{so} \quad \mathbb{E}[T_j] = -n\frac{\partial D(\phi)}{\partial \phi_j},
$$

$$
\text{hence} \quad \frac{\partial l}{\partial \phi_j} = t_j - \mathbb{E}[T_j]
$$

$$
\text{and so solving } \frac{\partial l}{\partial \phi_j} = 0 \text{ is equivalent to solving } t_j = \mathbb{E}[T_j].
$$

□

Moreover, it can be shown (not here) that if these equations have a solution then it is the unique MLE (thus there is not need to check second derivatives). See Bickel and Doksum, 1977, “Mathematical Statistics, Basic Ideas and selected Topics”, Holden Day, San Francisco.

Example 6.3. For the $N(\mu, 1)$ distribution,

$$
A(\theta) = \mu \quad \text{and} \quad B(x) = x
$$

Therefore $T = \sum_{i=1}^{n} X_i$ and

$$
\mathbb{E}[T] = n\mathbb{E}[X_i] = n\mu.
$$

Setting $t = \mathbb{E}[T]$ gives

$$
\sum_{i=1}^{n} x_i = n\hat{\mu}
$$

and solving for $\hat{\mu}$ gives the MLE $\hat{\mu} = \overline{\mathbf{X}}$. ◁

Although these *canonical parameterisations* are useful for applying general results concerning exponential family distributions to a particular distribution of interest, the commonly used parameterisations can be rather easier to interpret and are often rather more closely related to a distribution's interesting statistical properties.

6.3 The Cramer-Rao Inequality and Lower Bound

Theorem 6.2 (The Cramer-Rao Inequality and Lower Bound). *Suppose $X_1, \ldots, X_n$ form a simple random sample from the distribution with pdf $f(x;\theta)$. Subject to certain regularity conditions on $f(x;\theta)$, we have that for any unbiased estimator $\hat{\theta}$ for θ,*

$$\mathbb{V}\mathsf{ar}[\hat{\theta}] \geq I_\theta^{-1}$$

where I_θ is the ***Fisher Information about*** *θ*

$$I_\theta = \mathbb{E}\left[\left(\frac{\partial \log[L(\theta;\mathbf{x})]}{\partial\theta}\right)^2\right] = \mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}\right)^2\right].$$

This is known as the ***Cramer-Rao lower bound*** *and I_θ^{-1} is the minimum variance achievable by any unbiased estimator.*

Proof (Outline). For unbiased $\hat{\theta}$

$$\mathbb{E}[\hat{\theta}] = \int_{\mathbf{x}} \hat{\theta} L(\theta;\mathbf{x})\,\mathrm{d}\mathbf{x} = \theta$$

Under regularity conditions (essentially, to allow the interchange of integration and differentiation):

$$\int \hat{\theta}\frac{\partial L}{\partial\theta}\,\mathrm{d}\mathbf{x} = 1$$

Now

$$\frac{\partial l}{\partial\theta} = \frac{\partial \log L}{\partial\theta} = \frac{1}{L}\frac{\partial L}{\partial\theta} \Rightarrow \frac{\partial L}{\partial\theta} = L\frac{\partial l}{\partial\theta}.$$

Therefore

$$1 = \int \hat{\theta}\frac{\partial L}{\partial\theta}\mathrm{d}\mathbf{x} = \mathbb{E}\left[\hat{\theta}\frac{\partial l}{\partial\theta}\right].$$

We can then prove the result using the Cauchy-Schwarz inequality. Let $U = \hat{\theta}$ and $V = \frac{\partial l}{\partial\theta}$. Then

$$\mathbb{E}[V] = \int \frac{\partial l}{\partial\theta}L\,\mathrm{d}\mathbf{x} = \int \frac{\partial L}{\partial\theta} = \frac{\partial}{\partial\theta}\left[\int L\,\mathrm{d}\mathbf{x}\right] = 0$$

Therefore

$$\mathbb{C}\mathsf{ov}[U,V] = \mathbb{E}[UV] - \mathbb{E}[U]\mathbb{E}[V] = \mathbb{E}[UV] = \mathbb{E}\left[\hat{\theta}\frac{\partial l}{\partial\theta}\right] = 1$$

Also

$$\mathbb{V}\mathsf{ar}[V] = \mathbb{E}[V^2] = \mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}\right)^2\right] = I_\theta$$

So by theorem 4.4, we have:

$$(\mathbb{C}\mathsf{ov}[U,V])^2 \leq \mathbb{V}\mathsf{ar}[U]\mathbb{V}\mathsf{ar}[V] \Rightarrow 1 \leq \mathbb{V}\mathsf{ar}(\hat{\theta})I_\theta$$

□

Comments:

The larger I_θ is, the more informative a typical observation is about the parameter θ, and the smaller the attainable variance of $\hat{\theta}$.

Regularity conditions required to justify the exchange integration and differentiation in the proof include that the range of values of X must not depend on θ.

An unbiased estimator $\hat{\theta}$ whose variance attains the Cramer-Rao lower bound is called **efficient**.

Example 6.4. Suppose $X_1, \ldots, X_n$ form a random sample from $N \sim (\mu, \sigma^2)$ with μ unknown.

$$L = \prod_{i=1}^{n} f(x_i, \mu) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

$$\Rightarrow l = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\Rightarrow \frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{n}{\sigma^2}(\overline{\mathbf{x}} - \mu)$$

$$\therefore \hat{\mu} = \overline{\mathbf{X}} \text{ is MLE.}$$

$$I_\theta = \mathbb{E}\left[\left(\frac{\partial l}{\partial \mu}\right)^2\right] = \mathbb{E}\left[\frac{n^2}{\sigma^4}(\overline{\mathbf{X}} - \mu)^2\right] = \frac{n^2}{\sigma^4}\mathbb{E}\left[(\overline{\mathbf{X}} - \mu)^2\right]$$

$$= \frac{n^2}{\sigma^4}\mathbb{V}\text{ar}[\overline{\mathbf{X}}] = \frac{n^2}{\sigma^4}\frac{\sigma^2}{n} = \frac{n}{\sigma^2}.$$

Thus the lower bound is $I_\theta^{-1} = \frac{\sigma^2}{n}$, which is attained by $\hat{\mu} = \overline{\mathbf{X}}$, hence $\hat{\mu}$ is an efficient estimator.

$\hat{\mu}$ may also be referred to as a **minimum variance unbiased estimator (MVUE)** and in some sense it is the best *unbiased* estimator available. ◁

There are a number of situations in statistics in which it is possible to reduce variance at the expense of an increased bias. As was shown in the previous chapter, the mean squared error may be expressed as the sum of the variance and the bias squared. In some settings, it may make more sense to look for an estimator which minimise this quantity rather than an unbiased estimator which minimises the variance. However, the property of unbiasedness is naturally appealing.

Exercise 6.3.1. Under the same regularity conditions as before, and the additional assumption that ℓ is twice continuously differentiable, show that I_θ can be expressed in the more useful form

$$I_\theta = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right].$$

Using this result show that the ML estimator obtained earlier for the parameter of a Poisson distribution attains the Cramer-Rao lower bound.

6.4 Properties of MLEs

The maximum likelihood estimator has a number of useful properties. In particular, it doesn't really matter how we choose to parameterise a distribution as any two parameterisations which really describe the same distribution will lead to the same conclusions.

Theorem 6.3. *Suppose θ and ϕ represent two alternative parameterizations and that ϕ is a one-to-one function of θ (more formally, there exists a bijective mapping, g, from θ to ϕ), so we can write*

$$\phi = g(\theta), \theta = h(\phi)$$

for appropriate g and $h = g^{-1}$.

Then if $\hat{\theta}$ is the MLE of θ, then the MLE of ϕ is $g(\hat{\theta})$.

Proof. Suppose the value of ϕ that maximises L corresponds to $\tilde{\theta} \neq \hat{\theta}$, so that

$$L(g(\tilde{\theta}); \mathbf{x}) > L(g(\hat{\theta}); \mathbf{x})$$

Taking the inverse function $h(.)$ we have, noting that the likelihood of a point does not depend upon the parameterisation used to describe that point, that

$$L(\tilde{\theta}; \mathbf{x}) > L(\hat{\theta}; \mathbf{x})$$

so $\hat{\theta}$ is not the MLE. Which is a contradiction. □

Corollary 6.1 (Invariance of MLE). *Let $\hat{\theta}_1, \ldots, \hat{\theta}_k$ be a MLE for $\theta_1, \ldots, \theta_k$. If*

$$\mathcal{T}(\theta) = (\mathcal{T}_1(\theta), \ldots, \mathcal{T}_r(\theta))$$

is a sufficiently regular transformation of the parameter space Ω, then

$$\mathcal{T}(\hat{\theta}) = (\mathcal{T}_1(\hat{\theta}), \ldots, \mathcal{T}_r(\hat{\theta}))$$

is a MLE of $\mathcal{T}(\theta)$.

Example 6.5. Consider $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, μ, σ^2 both unknown. Then the log-likelihood is

$$l = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Could find $\hat{\sigma}^2$ from

$$\frac{\partial l}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

or

$$\frac{\partial l}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2$$

note: this is okay here: $g : \sigma \to (\sigma^2)$ is a bijection because we require that $\sigma > 0$ but in general it's important to verify that the conditions of theorem 6.3 hold.

Substituting $\hat{\mu} = \overline{\mathbf{X}}$ and setting the second equation to zero:

$$\frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i - \overline{\mathbf{X}})^2 = \frac{n}{\hat{\sigma}}$$

Could solve for $\hat{\sigma}$ and square but easier to solve for $\hat{\sigma}^2$ directly:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{\mathbf{X}})^2}{n}.$$

◁

By invariance, we can find MLEs for parameters of distributions in the exponential family via the natural parameterisation.

Example 6.6 (The Poisson Distribution).

$$A(\theta) = \log \theta \quad \text{and} \quad B(x) = x$$

Therefore $T = \sum_{i=1}^{n} X_i$ and

$$\mathbb{E}[T] = n\mathbb{E}[X_i] = n\theta$$

Setting $t = \mathbb{E}[T]$ gives

$$\sum_{i=1}^{n} x_i = n\hat{\theta}$$

and solving for $\hat{\theta}$ gives the MLE $\hat{\theta} = \overline{\mathbf{X}}$. ◁

Exercise 6.4.1. According to a genetic theory, the 3 blood types A, B and O occur in a population with probabilities $Pr[A] = p^2$, $Pr[B] = 2p(1-p)$ and $Pr[O] = (1-p)^2$. A sample of 200 randomly selected individuals gave the following result:

Blood Type	Frequency
A	60
B	130
O	10

Derive the ML estimates for the three probabilities.

The MLE has been very extensively studied and its properties have been characterised rather well. The following two results provide an example of some useful results; they are presented without proof and we do not dwell upon the regularity conditions under which they hold.

Lemma 6.1. *Suppose there exists an unbiased estimator, $\tilde{\theta}$, which attains the Cramer-Rao bound. Suppose that $\hat{\theta}$, the MLE is a solution to*

$$\frac{\partial l}{\partial \theta} = 0.$$

Then $\tilde{\theta} = \hat{\theta}$.

Lemma 6.2. *Under fairly weak regularity conditions, MLE's are consistent. If $\hat{\theta}$ is the MLE for θ, then asymptotically*

$$\hat{\theta} \sim N(\theta, I_\theta^{-1}).$$

Example 6.7. For the normal distribution $N(\mu, \sigma^2)$

– the MLE $\hat{\mu}$ is an efficient estimator for μ
– the MLE $\hat{\sigma}^2$ is asymptotically efficient for σ^2, but not efficient for finite sample sizes.

◁

6.5 MLE and properties of MLE for the multi-parameter Case

Let $X_1, ..., X_n$ be iid with common pdf $f(x;\theta)$, where $\theta \in \Omega_\theta \subset R^p$.

Likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Under mild regularity conditions the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ solves the vector equations

$$\frac{\partial}{\partial \theta} l(\theta) = \mathbf{0}.$$

The following properties extend to θ from the scalar case:

– The value of θ at which $L(\theta)$ is maximized is an unbiased estimator of the true θ;
– Invariance: let $\eta = \mathbf{g}(\theta)$, then $\hat{\eta} = \mathbf{g}(\hat{\boldsymbol{\theta}})$ is MLE of η.

Theorem 6.4 (Properties of MLE, multi-parameter case). *Under a set of regularity conditions*

1. *Consistency: the likelihood equation*
$$\frac{\partial}{\partial \theta} l(\theta) = \mathbf{0}$$
has a solution $\hat{\boldsymbol{\theta}}_n$ *such that*
$$\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \theta.$$
2. *Asymptotic normality:*
$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \theta) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \mathbf{I}^{-1}(\theta)\right).$$

$\mathbf{I}(\theta)$ *is the Fisher Information matrix with entries*

$$\begin{aligned} I_{ii}(\theta) &:= \mathbb{V}\mathsf{ar}\left[\frac{\partial \log f(X;\theta)}{\partial \theta_i}\right] \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i^2} \log f(X;\theta)\right] \\ I_{jk}(\theta) &= \mathbb{C}\mathsf{ov}\left[\frac{\partial \log f(X;\theta)}{\partial \theta_j}, \frac{\partial \log f(X;\theta)}{\partial \theta_k}\right] \\ &:= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_j \theta_k} \log f(X;\theta)\right] \end{aligned} \tag{6.2}$$

for $i, j, k = 1, ..., p$.

Exercise 6.5.1. Derive the equalities stated in 6.2.

Exercise 6.5.2. Find the MLE and Fisher information for the case in which f is normal with parameter vector $\theta = (\mu, \sigma)'$.

Comments

Cramer-Rao bound in the multi-parameter case:
Let $\hat{\theta}_{n,j}$ be an unbiased estimator of θ_j. Then it can be shown that

$$\mathbb{V}\mathsf{ar}(\hat{\theta}_{n,j}) \geq \frac{1}{n} \mathbf{I}_{jj}^{-1}(\theta).$$

The unbiased estimator is efficient if it attains the lower bound.
$\hat{\theta}_n$ are asymptotically efficient estimators, that is, for $j = 1, ..., p$

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_j) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \mathbf{I}_{jj}^{-1}(\theta)\right).$$

Theorem 6.5 (Transformation). *Let* $\mathbf{g}$ *be a transformation* $\mathbf{g}(\theta) = (g_1(\theta), ..., g_k(\theta))^T$ *such that* $1 \leq k \leq p$ *and that the* $k \times p$ *matrix of partial derivatives*

$$\mathbf{B} = \left[\frac{\partial g_i}{\partial \theta_j}\right], i = 1, ..., k; j = 1, ..., p,$$

has continuous elements and does not vanish in the neighbourhood of θ*. Let* $\hat{\eta} = \mathbf{g}(\eta)$*. Then* $\hat{\eta}$ *is MLE of* $\eta = \mathbf{g}(\theta)$ *and*

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \mathbf{B}\mathbf{I}^{-1}(\theta)\mathbf{B}'\right).$$

Exercise 6.5.3. Use the theorem above to derive the asymptotic variance of the MLE $\widehat{\sigma^2}$ for transformation $\sigma^2 = g(\mu, \sigma)$ in the case of a Normal density.

6.6 Computation of MLEs

As was noted above, determining the MLE for a particular model and data set is simply a matter of finding the parameter value which maximises the likelihood function for that data set. Unfortunately, optimisation is a rather difficult thing to do in general. We have seen that in some simple cases it is possible to obtain, by analysis, general expressions for the MLE as a function of the observed data. Unfortunately, it is often impossible to obtain such an analytic solution and considerable efforts have been made to develop robust numerical techniques for estimating the maximisers. This section summarises some of the simpler techniques which are commonly used. In more difficult problems still it may be possible to resort to Monte Carlo algorithms to obtain estimates (including the simulated annealing method). See "Monte Carlo Statistical Methods", Robert and Casella, Springer, 2004 for an introduction.

All of the techniques described here have limitations and will fail in some situations. It's always important to be aware of the methods that software uses in order to calculate quantities of interest and to be aware of the limitations of those methods: the mere fact that a computer has returned a value can inspire more confidence than is justified.

6.6.1 The Newton-Raphson Method

Let $g(\theta)$ be the gradient of $l(\theta, ; \mathbf{x})$ and let $H(\theta)$ denote the matrix of 2nd derivatives (i.e. the Hessian matrix). Suppose θ_0 is an initial estimate of θ and $\hat{\theta}$ is the MLE. Expanding $g(\theta)$ about θ_0 using the Taylor expansion gives the following multivariate equation:

$$\begin{aligned} g(\theta) &= g(\theta_0) + (\theta - \theta_0)^T H(\theta_0) + \ldots \\ \Rightarrow 0 &= g(\theta_0) + (\hat{\theta} - \theta_0)^T H(\theta_0) + \ldots \end{aligned}$$

Therefore $\hat{\theta}$ is approximated by

$$\theta_1 = \theta_0 - g(\theta_0) H^{-1}(\theta_0) \tag{6.3}$$

This would be exact if g was a linear function of θ (*i.e.* if $l(\theta; (x))$ were quadratic in θ). However, it can be used as an approximation (which becomes good close to the true value of the maximum under regularity conditions which allow the higher order terms in the Taylor expansion to become negligible throughout a neighbour of the optimum). In order to exploit this property, the Newton-Raphson method takes an initial value, θ_0, uses equation 6.3 to obtain a hopefully improved estimate and this improved estimate is used as the starting point for the next iteration. This procedure is repeated until convergence of some sort is observed (actually, this method does not always converge, even to a local maximum).

6.6.2 Fisher's Method of Scoring

A simple modification of N-R in which the $H(\theta)$ are replaced by its expectation

$$\mathbb{E}[H(\theta)] = -I_\theta$$

(where the expectation is respect to the data, $\mathbf{X}$, which is generated from $f(\mathbf{x};\theta)$). Now (under the usual regularity conditions)

$$\begin{aligned} \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] &= -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right] \\ \text{and so} \quad \mathbb{E}\left[\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right] &= -\mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right]. \end{aligned}$$

therefore we need only calculate the **score vector** of 1st derivatives.

Also $\mathbb{E}[H(\theta)]$ is positive definite, thus eliminating possible non-convergence problems of N-R.

6.6.3 Newton-Raphson and the Exponential Family of Distributions

When the likelihood is a member of the exponential family, the Newton-Raphson algorithm and Fisher's method of scoring are equivalent. This can be seen by considering the derivatives of the likelihood with respect to the natural parameterisation:

$$\begin{aligned} l(\phi;\mathbf{x}) &= \text{ constant } + \sum_{j=1}^{k} \phi_j t_j + nD(\phi) \\ \text{thus} \quad \frac{\partial l}{\partial \phi_j} &= t_j + n\frac{\partial D(\phi)}{\partial \phi_j} \\ \text{and} \quad \frac{\partial^2 l}{\partial \phi_i \partial \phi_j} &= n\frac{\partial^2 D(\phi)}{\partial \phi_i \partial \phi_j} \end{aligned}$$

As $D(\phi)$ does not depend on x, $H(\theta)$ and $E(H(\theta))$ are identical.

6.6.4 The Expectation-Maximisation (EM) Algorithm

Suppose data is decomposed into observed (incomplete data) and missing (augmented data) values

$$\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_m)$$

$$\underbrace{L(\theta|\mathbf{x}_0) = g(\mathbf{X}_0|\theta)}_{\text{incomplete data likelihood}} = \int \underbrace{f(\mathbf{x}_0, \mathbf{x}_m|\theta)}_{\text{complete data likelihood}} \, \mathrm{d}\mathbf{x}_m$$

This sort of structure is rather common in statistical problems. We would like to maximise $L(\theta\ \mathbf{x}_0)$ but this may be difficult to obtain (doing so analytically is often impossible).

The EM algorithm maximises $L(\theta\ \mathbf{x}_0)$ by working with $f(\mathbf{x}_0, \mathbf{x}_m|\theta)$.

We have (note that this follows from the definition of the conditional distribution; k is simply the conditional distribution of the missing data given both the observed data and the parameter vector):

$$\begin{aligned} \frac{f(\mathbf{x}_0, \mathbf{x}_m|\theta)}{g(\mathbf{x}_0|\theta)} &= k(\mathbf{x}_m|\theta, \mathbf{x}_0) \\ \text{so} \qquad g(\mathbf{x}_0|\theta) &= \frac{f(\mathbf{x}_0, \mathbf{x}_m|\theta)}{k(\mathbf{x}_m|\theta, \mathbf{x}_0)} \end{aligned}$$

Taking logs

$$\begin{aligned} \log g(\mathbf{x}_0|\theta) &= \log f(\mathbf{x}_0, \mathbf{x}_m|\theta) - \log k(\mathbf{x}_m|\theta, \mathbf{x}_0) \\ \text{or} \qquad l(\theta|\mathbf{x}_0) &= l(\theta|\mathbf{x}_0, \mathbf{x}_m) - \log k(\mathbf{x}_m|\theta, \mathbf{x}_0) \end{aligned}$$

Now taking the expectation with respect to the missing data under $k(\mathbf{x}_m|\theta^i, \mathbf{x}_0)$

$$l(\theta|\mathbf{x}_0) = \mathbb{E}[l(\theta|\mathbf{x}_0, \mathbf{x}_m)|\theta^i, \mathbf{x}_0] - \mathbb{E}[\log k(\mathbf{x}_m|\theta, \mathbf{x}_0)|\theta^i, \mathbf{x}_0]$$

We seek to maximise $\mathbb{E}[l(\theta|\mathbf{x}_0, \mathbf{x}_m)|\theta^i, \mathbf{x}_0]$.

The E-step of the EM algorithm calculates the expected log-likelihood

$$\mathbb{E}[l(\theta|\mathbf{x}_0, \mathbf{x}_m)|\theta^i, \mathbf{x}_0] = \int l(\theta|\mathbf{x}_0, \mathbf{x}_m) k(\mathbf{x}_m|\theta^i, \mathbf{x}_0)\, \mathrm{d}\mathbf{x}_m$$

and the M-step mazimises it w.r.t. θ giving θ^{i+1}.

We then iterate through the E- and M-steps until convergence to the incomplete data MLE. Actually, the EM algorithm will always converge to something but it is possible that it may only reach a local maximum of the likelihood function. It is common practice to consider several runs of an EM algorithm with different initial values in order to investigate its sensitivity. This algorithm is the most common of a class of optimization techniques which allow the maximisation of non-convex functions via a convex surrogate which matches the function of interest locally.

6.7 Exercises

Exercise 6.7.1.

Let $X_1, ..., X_n$ be i.i.d. random variables with probability distribution F. Derive the MLE of the parameters for each of the following distributions:

(a) F is N $(\mu, 1)$.
(b) F is N $\left(\mu, \sigma^2\right)$.
(c) F is Exp (λ).
(d) F is Poi (λ).
(e) F is Ber (p).
(f) F is U$[1, b]$.

Exercise 6.7.2. A garage assumes that the minimal time necessary for an oil change is α. Let X denote the time for an oil change in a garage which varies from client to client. Assume that X is exponentially distributed with pdf

$$f(x) = \exp(\alpha - x) \qquad \text{for } x \geq \alpha$$

In order to estimate α the following data on time to change oil was collected:

4.2 3.1 3.6 4.5 5.1 7.6 4.4 3.5 3.8
3.9 4.1 4.3 4.4 3.7 3.5 3.8 3.9

(a) Derive the MLE for α and give the estimate for the above data.
(b) Derive an MLE for the mean time necessary to change the oil and give an estimate.
(c) Give an ML estimate for the probability that an oil change takes longer than 5 minutes.

7. Hypothesis Testing

7.1 Introduction

A **hypothesis** is a falsifiable claim about the real world; a statement which if true will explain some observable quantity. Statisticians will be interested in hypotheses like:

1. "The probabilities of a male panda or a female panda being born are equal',
2. "The number of flying bombs falling on a given area of London during World War II follows a Poisson distribution",
3. "The mean systolic blood pressure of 35-year-old men is no higher than that of 40-year-old women",
4. "The mean value of $Y = \log(\texttt{systolic blood pressure})$ is independent of $X = \texttt{age}$" ($i.e.$ $\mathbb{E}[Y|X = x] = constant$).

These hypotheses can be translated into statements about parameters within a probability model:

1. '$p_1 = p_2$',
2. '$N \sim \mathsf{Poi}(\lambda)$ for some $\lambda > 0$', $i.e.$: $p_n = \mathbb{P}(N = n) = \lambda^n \exp(-\lambda)/n!$ (within the general probability model $p_n \geq 0\ \forall n = 0, 1, \ldots ; \sum p_n = 1$),
3. '$\theta_1 \leq \theta_2$' and
4. '$\beta_1 = 0$' (assuming the linear model $\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$).

Definition 7.1 (Hypothesis test). *A **hypothesis test** is a procedure for deciding whether to accept a particular hypothesis as a reasonable simplifying assumption, or to reject it as unreasonable in the light of the data.*

Definition 7.2 (Null hypothesis). *The **null hypothesis** H_0 is the default assumption we are considering making.*

Definition 7.3 (Alternative hypothesis). *The **alternative hypothesis** H_1 is the alternative explanation(s) we are considering for the data.*

Ordinarily, the null hypothesis is what we would assume to be true in the absence of data which suggests that it is not. Ordinarily, H_0 will explain the data in at least as simple a way as H_1.

Definition 7.4 (Type I error). *A **type I error** is made if H_0 is rejected when H_0 is true. In some situations this type of error is known as a **false positive**.*

Definition 7.5 (Type II error). *A **type II error** is made if H_0 is accepted when H_0 is false. This may also be termed a **false negative**.*

Example 7.1. – In the first example above (pandas) the null hypothesis is $H_0 : p_1 = p_2$.

– The alternative hypothesis in the first example would usually be $H_1 : p_1 \neq p_2$, though it could also be (for example)

1. $H_1 : p_1 < p_2$,
2. $H_1 : p_1 > p_2$, or
3. $H_1 : p_1 - p_2 = \delta$ for some specified $\delta \neq 0$.

each of these alternative hypotheses makes a slightly different statement about the collection of situations which we believe are possible. A statistician needs to decide which type of hypothesis test is appropriate in any real situation.

◁

7.2 Simple Hypothesis Tests

The simplest type of hypothesis testing occurs when the probability distribution giving rise to the data is specified completely under the null and alternative hypotheses.

Definition 7.6 (Simple hypotheses). *A **simple hypothesis** is of the form $H_k : \theta = \theta_k$, where θ is the parameter vector which parameterises the probabilistic model for the data. A simple hypothesis specifies the precise value of the parameter vector (*i.e. *the probability distribution of the data is specified completely).*

Definition 7.7 (Composite hypotheses). *A **composite hypothesis** is of the form $H_k : \theta \in \Omega_k$,* i.e. *the parameter θ lies in a specified subset Ω_k of the parameter space Ω_Θ. This type of hypothesis specifies an entire collection of values for the parameter vector and so specifies a class of probabilistic models from which the data may have arisen.*

Definition 7.8 (Simple hypothesis test). *A **simple hypothesis test** tests a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$, where θ parameterises the distribution of our experimental random variables $\mathbf{X} = X_1, X_2, \ldots X_n$.*

Although simple hypothesis tests seem appealing, there are many situations in which a statistician cannot reduce the problem at hand to a clear dichotomy between two fully-specified models for the data-generating process.

There may be many seemingly sensible approaches to testing a given hypothesis. A reasonable criterion for choosing between them is to attempt to minimise the chance of making a mistake: incorrectly rejecting a true null hypothesis, or incorrectly accepting a false null hypothesis.

Definition 7.9 (Size). *A **test of size** α is one which rejects the null hypothesis $H_0 : \theta = \theta_0$ in favour of the alternative $H_1 : \theta = \theta_1$ iff*

$$\mathbf{X} \in C_\alpha \qquad \text{where } \mathbb{P}(\mathbf{X} \in C_\alpha \mid \theta = \theta_0) = \alpha$$

for some subset C_α of the sample space S of $\mathbf{X}$.

The size, α, of a test is the probability of rejecting H_0 when H_0 is in fact true; *i.e.* size is the probability of type I error if H_0 is true. We want α to be small ($\alpha = 0.05$, say).

Definition 7.10 (Critical region). *The set C_α in Definition 7.9 is called the **critical region** or **rejection region** of the test.*

Definition 7.11 (Power & power function). *The **power function** of a test with critical region C_α is the function*

$$\beta(\theta) = \mathbb{P}(\mathbf{X} \in C_\alpha \mid \theta),$$

*and the **power** of a simple test is $\beta = \beta(\theta_1)$,* i.e. *the probability that we reject H_0 in favour of H_1 when H_1 is true.*

Thus a simple test of power β has probability $1-\beta$ of a type II error occurring when H_1 is true. Clearly for a fixed size α of test, the larger the power β of a test the better.

However, there is an inevitable trade-off between small size and high power (as in a jury trial: the more careful one is not to convict an innocent defendant, the more likely one is to free a guilty one by mistake).

A hypothesis test typically uses a **test statistic** $T(\mathbf{X})$, whose distribution is known under H_0, and such that extreme values of $T(\mathbf{X})$ are more compatible with H_1 that H_0.

Many useful hypothesis tests have the following form:

Definition 7.12 (Simple likelihood ratio test). *A* ***simple likelihood ratio test (SLRT)*** *of* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta = \theta_1$ *rejects* H_0 *iff*

$$\mathbf{X} \in C_\alpha^* = \left\{ \mathbf{x} \;\middle|\; \frac{L(\theta_0;\mathbf{x})}{L(\theta_1;\mathbf{x})} \leq A_\alpha \right\}$$

where $L(\theta;\mathbf{x})$ *is the likelihood of* θ *given the data* $\mathbf{x}$*, and the number* A_α *is chosen so that the size of the test is* α*.*

Exercise 7.2.1. Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. Show that the likelihood ratio for testing $H_0 : \theta = 0$ against $H_1 : \theta = 1$ can be written

$$\lambda(\mathbf{x}) = \exp\big[n\big(\bar{x} - \tfrac{1}{2}\big)\big].$$

Hence show that the corresponding SLRT of size α rejects H_0 when the test statistic $T(\mathbf{X}) = \bar{x}$ satisfies $T > \Phi^{-1}(1-\alpha)/\sqrt{n}$.

A number of points should be borne in mind:

- For a simple hypothesis test, both H_0 and H_1 are 'point hypotheses', each specifying a particular value for the parameter θ rather than a region of the parameter space.
- In practice, no hypothesis will be precisely true, so the whole foundation of classical hypothesis testing seems suspect! Actually, there is a tendency to overuse hypothesis testing: it is appropriate only when one really does wish to compare competing, well-defined hypotheses. In many problems the use of point estimation with an appropriate confidence interval is much easier to justify.
- Regarding likelihood as a measure of compatibility between data and model, an SLRT compares the compatibility of θ_0 and θ_1 with the observed data $\mathbf{x}$, and accepts H_0 iff the ratio is sufficiently large.
- One reason for the importance of likelihood ratio tests is the following theorem, which shows that out of all tests of a given size, an SLRT (if one exists) is 'best' in a certain sense.

Theorem 7.1 (The Neyman-Pearson lemma). *Given random variables* $X_1, X_2, \ldots, X_n$*, with joint density* $f(\mathbf{x}|\theta)$*, the simple likelihood ratio test of a fixed size* α *for testing* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta = \theta_1$ *is at least as powerful as any other test of the same size.*

Proof. Fix the size of the test to be α.

Let A be a positive constant and C_0 a subset of the sample space satisfying

(a) $\mathbb{P}(\mathbf{X} \in C_0 \mid \theta = \theta_0) = \alpha$,

(b) $\mathbf{X} \in C_0 \iff \dfrac{L(\theta_0;\mathbf{x})}{L(\theta_1;\mathbf{x})} = \dfrac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} \leq A.$

Suppose that there exists another test of size α, defined by the critical region C_1, *i.e.*

$$\text{Reject } H_0 \text{ iff } \mathbf{x} \in C_1, \text{ where } \mathbb{P}(\mathbf{x} \in C_1 | \theta = \theta_0) = \alpha.$$

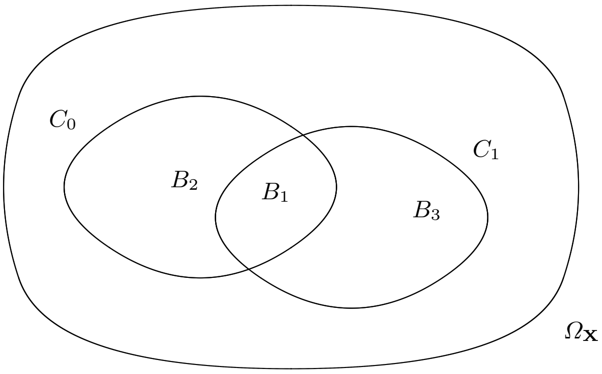


Fig. 7.1. Proof of Neyman-Pearson lemma

Let $B_1 = C_0 \cap C_1$, $B_2 = C_0 \cap C_1^c$, $B_3 = C_0^c \cap C_1$.

Note that $B_1 \cup B_2 = C_0$, $B_1 \cup B_3 = C_1$, and B_1, B_2 & B_3 are disjoint.

Let the power of the likelihood ratio test be $I_0 = \mathbb{P}(\mathbf{X} \in C_0 \mid \theta = \theta_1)$, and the power of the other test be $I_1 = \mathbb{P}(\mathbf{X} \in C_1 \mid \theta = \theta_1)$.
We want to show that $I_0 - I_1 \geq 0$.

But

$$\begin{aligned} I_0 - I_1 &= \int_{C_0} f(\mathbf{x}|\theta_1)d\mathbf{x} - \int_{C_1} f(\mathbf{x}|\theta_1)d\mathbf{x} \\ &= \int_{B_1 \cup B_2} f(\mathbf{x}|\theta_1)d\mathbf{x} - \int_{B_1 \cup B_3} f(\mathbf{x}|\theta_1)d\mathbf{x} \\ &= \int_{B_2} f(\mathbf{x}|\theta_1)d\mathbf{x} - \int_{B_3} f(\mathbf{x}|\theta_1)d\mathbf{x}. \end{aligned}$$

Also $B_2 \subseteq C_0$, so $f(\mathbf{x}|\theta_1) \geq A^{-1} f(\mathbf{x}|\theta_0)$ for $\mathbf{x} \in B_2$, similarly $B_3 \subseteq C_0^c$, so $f(\mathbf{x}|\theta_1) \leq A^{-1} f(\mathbf{x}|\theta_0)$ for $\mathbf{x} \in B_3$,

Therefore

$$\begin{aligned} I_0 - I_1 &\geq A^{-1}\left[\int_{B_2} f(\mathbf{x}|\theta_0)d\mathbf{x} - \int_{B_3} f(\mathbf{x}|\theta_0)d\mathbf{x}\right] \\ &= A^{-1}\left[\int_{C_0} f(\mathbf{x}|\theta_0)d\mathbf{x} - \int_{C_1} f(\mathbf{x}|\theta_0)d\mathbf{x}\right] \\ &= A^{-1}[\alpha - \alpha] \quad = \quad 0 \end{aligned}$$

as required. □

7.3 Simple Null, Composite Alternative

Suppose that we wish to test the simple null hypothesis $H_0 : \theta = \theta_0$ against the composite alternative hypothesis $H_1 : \theta \in \Omega_1$.

The easiest way to investigate this is to imagine the collection of simple hypothesis tests with null hypothesis $H_0 : \theta = \theta_0$ and alternative $H_1 : \theta = \theta_1$, where $\theta_1 \in \Omega_1$. Then, for any given θ_1, an SLRT is the most powerful test for a given size α. The only problem would be if different values of θ_1 result in different SLRTs.

Definition 7.13 (UMP Tests). *A hypothesis test is called a* ***uniformly most powerful test*** *of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, $\theta_1 \in \Omega_1$, if*

1. *There exists a critical region C_α corresponding to a test of size α not depending on θ_1,*
2. *For all values of $\theta_1 \in \Omega_1$, the critical region C_α defines a most powerful test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.*

Exercise 7.3.1. Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}\left(0, \sigma^2\right)$.

(a) Find the UMP test of $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 > 1$.
(b) Find the UMP test of $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 < 1$.
(c) Show that no UMP test of $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 \neq 1$ exists.

There are several points to notice:

- If a UMP test exists, then it is clearly the appropriate test to use. It's the "best" test for comparing H_0 with any alternative consistent with H_1.
- Often UMP tests don't exist! The requirements above are actually rather strong: in particular, requiring that a test is the most powerful test for a complicated class of alternative hypotheses limits the circumstances in which it is possible to specify such a test.
- Sufficient statistics are summaries of data which have the property that the distribution of the data, given the statistic, is conditionally independent of the data itself (*i.e.* $\mathbb{P}[X = x|T(x) = t, \theta] = \mathbb{P}[X = x|T(x) = t]$). A UMP test involves the data only via a likelihood ratio, so is a function of the sufficient statistics.
- As a result of the previous point, when a UMP test does exist, the critical region C_α often has a simple form, and is usually easily found once the distribution of the sufficient statistics have been determined (hence the importance of the χ^2, t and F distributions).
- The above three examples illustrate how important is the form of alternative hypothesis being considered. The first two are **one-sided alternatives** whereas $H_1 : \sigma^2 \neq 1$ is a **two-sided alternative hypothesis**, since σ^2 could lie on either side of 1.

7.4 Composite Hypothesis Tests

The most general situation we'll consider is that in which the parameter space Ω is divided into two subsets: $\Omega = \Omega_0 \cup \Omega_1$, where $\Omega_0 \cap \Omega_1 = \emptyset$, and the hypotheses are $H_0 : \theta \in \Omega_0$, $H_1 : \theta \in \Omega_1$.

For example, one may want to test the null hypothesis that the data come from an exponential distribution against the alternative that the data come from a more general gamma distribution. In this case, the class of exponential distributions corresponds to the class of gamma distributions in which the shape parameter is required to be 1. Note that Ω_0 is a one-dimensional subspace of Ω, whilst Ω_1 is the remainder of the two-dimensional parameter space of the gamma distribution: here, as in many other cases, $\dim(\Omega_0) < \dim(\Omega_1) = \dim(\Omega)$.

One possible approach to this situation is to regard the maximum possible likelihood over $\theta \in \Omega_i$ as a measure of compatibility between the data and the hypothesis H_i $(i = 0, 1)$.

It is convenient to define the following:

- $\widehat{\theta}$ is the MLE of θ over the whole parameter space Ω,
- $\widehat{\theta}_0$ is the MLE of θ over Ω_0, *i.e.* under the null hypothesis H_0, and
- $\widehat{\theta}_1$ is the MLE of θ over Ω_1, *i.e.* under the alternative hypothesis H_1.

Note that $\widehat{\theta}$ must therefore be the same as either $\widehat{\theta}_0$ or $\widehat{\theta}_1$, because $\Omega = \Omega_0 \cup \Omega_1$.

One might consider using the likelihood ratio criterion $L(\widehat{\theta}_1; \mathbf{x})/L(\widehat{\theta}_0; \mathbf{x})$, by direct analogy with the SLRT. However, it is generally easier to use the related ratio $L(\widehat{\theta}; \mathbf{x})/L(\widehat{\theta}_0; \mathbf{x})$. Note that this agrees with the likelihood ratio criterion above when $\widehat{\theta}_1$ has the higher likelihood and is otherwise equal to 1. This avoids the calculation of the MLE under the constraint that $\theta \in \Omega_1$ which will generally be somewhat more difficult than the calculation of the unconstrained MLE.

Definition 7.14 (Likelihood Ratio Test (LRT)). *A* ***likelihood ratio test*** *rejects* $H_0 : \widehat{\theta} \in \Omega_0$ *in favour of the alternative* $H_1 : \widehat{\theta} \in \Omega_1 = \Omega \setminus \Omega_0$ *iff*

$$\lambda(\mathbf{x}) = \frac{L(\widehat{\theta}; \mathbf{x})}{L(\widehat{\theta}_0; \mathbf{x})} \geq \lambda, \tag{7.1}$$

where $\widehat{\theta}$ *is the MLE of* θ *over the whole parameter space* Ω, $\widehat{\theta}_0$ *is the MLE of* θ *over* Ω_0, *and the value* λ *is fixed so that*

$$\sup_{\theta \in \Omega_0} \mathbb{P}(\lambda(\mathbf{X}) \geq \lambda | \theta) = \alpha$$

where α, the size of the test, is some chosen value.

It is possible to define an equivalent criterion in terms of the log likelihood and the **log likelihood ratio test statistic**:

$$r(\mathbf{x}) = \ell(\widehat{\theta}; \mathbf{x}) - \ell(\widehat{\theta}_0; \mathbf{x}) \geq \lambda', \tag{7.2}$$

where $\ell(\theta; \mathbf{x}) = \ell(\theta; \mathbf{x})$, and λ' is chosen to give chosen size $\alpha = \sup_{\theta \in \Omega_0} \mathbb{P}(r(\mathbf{X}) \geq \lambda' | \theta)$. As is often the case, taking logarithms can simplify computations. Equation 7.2 is often easier to work with than Equation 7.1—see the exercises and problems.

As ever, the size α is typically chosen by convention to be 0.05 or 0.01.

High values of the test statistic $\lambda(\mathbf{x})$, or equivalently of $r(\mathbf{x})$, are taken as evidence against the null hypothesis H_0. If $\lambda(\mathbf{x}) = 1$ or equivalently $r(\mathbf{x}) = 0$ then the null hypothesis explains the observed data at least as well as the alternative hypothesis.

A note on terminology: The test given in Definition 7.14 is sometimes referred to as a *generalized* likelihood ratio test, and Equation 7.1 a *generalized* likelihood ratio test statistic. When this terminology is used the *simple* likelihood ratio test defined here in definition 7.12 may be referred to as a likelihood ratio test.

These abstract tests may seem far removed from real statistical problems. However, as the following exercises demonstrate, standard widely-used procedures can be readily obtained as particular cases of the type of procedure introduced above. An understanding of the general case can be invaluable when trying to interpret or apply specific examples.

Exercise 7.4.1 (Paired t-test). Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu, \sigma^2\right)$, and let $\bar{X} = \sum X_i / n$, $S^2 = \sum (X_i - \bar{X})^2 / (n-1)$. What is the distribution of $T = \bar{X} / (S / \sqrt{n})$?

Is the test based on rejecting $H_0 : \mu = 0$ for large T a likelihood ratio test?

Assuming that the observed differences in diastolic blood pressure (after–before) are iid and Normally distributed with mean δ_D, use the captopril data (Table 7.1) to test the null hypothesis $H_0 : \delta_D = 0$ against the alternative hypothesis $H_1 : \delta_D \neq 0$.

This procedure is known as the **paired t test**.

Patient		Systolic			Diastolic	
Number	before	after	change	before	after	change
1	210	201	-9	130	125	-5
2	169	165	-4	122	121	-1
3	187	166	-21	124	121	-3
4	160	157	-3	104	106	2
5	167	147	-20	112	101	-11
6	176	145	-31	101	85	-16
7	185	168	-17	121	98	-23
8	206	180	-26	124	105	-19
9	173	147	-26	115	103	-12
10	146	136	-10	102	98	-4
11	174	151	-23	98	90	-8
12	201	168	-33	119	98	-21
13	198	179	-19	106	110	4
14	148	129	-19	107	103	-4
15	154	131	-23	100	82	-18

Table 7.1. Supine systolic and diastolic blood pressures of 15 patients with moderate hypertension (high blood pressure), immediately before and 2 hours after taking 25 mg of the drug captopril.

Exercise 7.4.2 (Two sample t-test). Suppose $X_1, X_2, \ldots, X_m \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu_X, \sigma^2\right)$ and $Y_1, Y_2, \ldots, Y_n \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu_Y, \sigma^2\right)$.

(a) Derive the LRT for testing $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X \neq \mu_Y$.

(b) Show that the LRT can be based on the test statistic

$$T = \frac{\bar{x} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}. \tag{7.3}$$

where

$$S_p^2 = \frac{\sum_{i=1}^{m}(X_i - \bar{x})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{m + n - 2}. \tag{7.4}$$

(c) Show that, under H_0, $T \sim \mathsf{t}_{m+n-2}$.

(d) Two groups of female rats were placed on diets with high and low protein content, and the gain in weight (grammes) between the 28th and 84th days of age was measured for each rat, with the following results:

High protein diet

134 146 104 119 124 161 107 83 113 129 97 123

Low protein diet

70 118 101 85 107 132 94

Using the test statistic T above, test the null hypothesis that the mean weight gain is the same under both diets.

Comment: this is called the **two sample t-test**, and S_p^2 is the **pooled estimate of variance**.

Exercise 7.4.3 (F-test). Suppose $X_1, X_2, \ldots, X_m \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu_X, \sigma_X^2\right)$ and $Y_1, Y_2, \ldots, Y_n \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu_Y, \sigma_Y^2\right)$, where μ_X, μ_Y, σ_X and σ_Y are all unknown.

Suppose we wish to test the hypothesis $H_0 : \sigma_X^2 = \sigma_Y^2$ against the alternative $H_1 : \sigma_X^2 \neq \sigma_Y^2$.

(a) Let $S_X^2 = \sum_{i=1}^{m}(X_i - \bar{x})^2$ and $S_Y^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.
What are the distributions of S_X^2/σ_X^2 and S_Y^2/σ_Y^2?

(b) Under H_0, what is the distribution of the statistic

$$V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)}?$$

(c) Taking values of V much larger or smaller than 1 as evidence against H_0, and given data with $m = 16$, $n = 16$, $\sum x_i = 84$, $\sum y_i = 18$, $\sum x_i^2 = 563$, $\sum y_i^2 = 72$, test the null hypothesis H_0.

With the alternative hypothesis $H_1 : \sigma_X^2 > \sigma_Y^2$, the above procedure is called an F **test**.

Even in simple cases like this, the null distribution of the log likelihood ratio test statistic $r(\mathbf{x})$ (7.2) can be difficult or impossible to find analytically. Fortunately, there is a very powerful and very general theorem that gives the approximate distribution of $r(\mathbf{x})$:

Theorem 7.2 (Wald's Theorem). *Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(\mathbf{x}|\theta)$ where $\theta \in \Omega$, and let $r(\mathbf{x})$ denote the log likelihood ratio test statistic*

$$r(\mathbf{x}) = \ell(\widehat{\theta}; \mathbf{x}) - \ell(\widehat{\theta}_0; \mathbf{x}),$$

where $\widehat{\theta}$ is the MLE of θ over Ω and $\widehat{\theta}_0$ is the MLE of θ over $\Omega_0 \subset \Omega$.

Then (under reasonable conditions on the pdf (or pmf) $f(\cdot|\cdot)$ which are very often satisfied in practice), the distribution of $2r(\mathbf{x})$ converges to a χ^2 distribution of $\dim(\Omega) - \dim(\Omega_0)$ degrees of freedom as $n \to \infty$.

A proof of this theorem is somewhat beyond the scope of this course, but may be found in e.g. Kendall & Stuart, '*The Advanced Theory of Statistics*', Vol. II.

Wald's theorem implies that, provided the sample size is large, you only need tables of the χ^2 distribution to find the critical regions for a wide range of hypothesis tests.

Exercise 7.4.4. Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, with hypotheses $H_0 : \theta = 0$ and $H_1 : \theta$ arbitrary. Show that $2r(\mathbf{x}) = n\bar{x}^2$, and hence that Wald's theorem holds *exactly* in this case.

Exercise 7.4.5. Suppose now that $X_i \sim \mathsf{N}(\theta_i, 1)$, $i = 1, \ldots, n$ are independent, with null hypothesis $H_0 : \theta_i = \theta \; \forall i$ and alternative hypothesis $H_1 : \theta_i$ arbitrary.

Show that $2r(\mathbf{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2$. and hence (quoting any other theorems you need) that Wald's theorem again holds exactly.

7.5 Exercises

This is a good point to consider the following three problems concerning general hypothesis tests.

Exercise 7.5.1. Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu, \sigma^2\right)$ with null hypothesis $H_0 : \sigma^2 = 1$ and alternative $H_1 : \sigma^2$ is arbitrary. Show that the LRT will reject H_0 for large values of the test statistic $2r(\mathbf{x}) = n(\widehat{v} - 1 - \log \widehat{v})$, where $\widehat{v} = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n$.

Exercise 7.5.2. Let $X_1, \ldots, X_n$ be independent each with density

$$f(x) = \begin{cases} \lambda x^{-2} e^{-\lambda / x} & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where λ is an unknown parameter.

(a) Show that the UMP test of $H_0 : \lambda = \frac{1}{2}$ against $H_1 : \lambda > \frac{1}{2}$ is of the form: 'reject H_0 if $\sum_{i=1}^{n} X_i^{-1} \leq A^*$', where A^* is chosen to fix the size of the test.

(b) Find the distribution of $\sum_{i=1}^{n} X_i^{-1}$ under the null & alternative hypotheses.

(c) You observe values 0.59, 0.36, 0.71, 0.86, 0.13, 0.01, 3.17, 1.18, 3.28, 0.49 for $X_1, \ldots, X_{10}$. Test H_0 against H_1, and comment on the test in the light of any assumptions made.

Exercise 7.5.3. Assume that a particular bus service runs at regular intervals of θ minutes, but that you do not know θ. Assume also that the times you find you have to wait for a bus on n occasions, $X_1, \ldots, X_n$, are independent and identically distributed with density

$$f(x|\theta) = \begin{cases} \theta^{-1} & \text{if } 0 \leq x \leq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Discuss *briefly* when the above assumptions would be reasonable in practice.

(b) Find the likelihood $L(\theta; \mathbf{x})$ for θ given the data $(X_1, \ldots, X_n) = \mathbf{x} = (x_1, \ldots, x_n)$.

(c) Find the most powerful test of size α of the hypothesis $H_0 : \theta = \theta_0 = 20$ against the alternative $H_1 : \theta = \theta_1 > 20$.

7.6 The Multinomial Distribution and χ^2 Tests

Although we have focused on continuous distributions in this section, we can also apply the same ideas to discrete distributions. Recall the Multinomial distribution which was defined in section 4.2.1.

Exercise 7.6.1. By partial differentiation of the likelihood function, show that the MLEs $\hat{\theta}_i$ of the parameters θ_i of the Mult (n, θ) distribution satisfy the equations

$$\frac{x_i}{\hat{\theta}_i} - \frac{x_k}{1 - \sum_{j=1}^{k-1} \hat{\theta}_j} = 0, \qquad (i = 1, \ldots, k-1)$$

and hence that $\hat{\theta}_i = x_i/n$ for $i = 1, \ldots, k$.

7.6.1 Chi-Squared Tests

Suppose one wishes to test the null hypothesis H_0 that, in the multinomial distribution 4.1, θ is some specified function $\theta(\phi)$ of another parameter ϕ. The alternative hypothesis H_1 is that θ is arbitrary.

Exercise 7.6.2. Suppose H_0 is that $X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim}$ Bin $(3, \phi)$. Let Y_i (for $i = 1, 2, 3, 4$) denote the number of observations X_j taking value $i - 1$. What is the null distribution of $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$ (*i.e.* the distribution of this vector under the assumption that the null hypothesis holds)?

The log likelihood ratio test statistic $r(\mathbf{X})$ is given by

$$r(\mathbf{X}) = \sum_{i=1}^{k} Y_i \log \hat{\theta}_i - \sum_{i=1}^{k} Y_i \log \theta_i(\widehat{\phi}) \tag{7.5}$$

where $\hat{\theta}_i = y_i/n$ for $i = 1, \ldots, k$.

By Wald's theorem (Theorem 7.2), under H_0, $2r(\mathbf{X})$ has approximately a χ^2 distribution for sufficiently large samples:

$$2 \sum_{i=1}^{k} Y_i [\log \hat{\theta}_i - \log \theta_i(\widehat{\phi})] \quad \sim \quad \chi^2_{k_1 - k_0} \tag{7.6}$$

where

$\hat{\theta}_i = Y_i/n$,

k_0 is the dimension of the parameter ϕ, and

$k_1 = k - 1$ is the dimension of θ under the constraint $\sum_{i=1}^{k} \theta_i = 1$.

It is straightforward to check, using a Taylor series expansion of the log function, that provided $\mathbb{E}[Y_i]$ is large $\forall\, i$,

$$2 \sum_{i=1}^{k} Y_i [\log \hat{\theta}_i - \log \theta_i(\widehat{\phi})] \approx \sum_{i=1}^{k} \frac{(Y_i - \mu_i)^2}{\mu_i}, \tag{7.7}$$

where $\mu_i = n\theta_i(\widehat{\phi})$ is the expected number of individuals (under H_0) in the ith category.

Definition 7.15 (Chi-squared Goodness of Fit Statistic).

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}, \tag{7.8}$$

where o_i is the observed count in the ith category and e_i is the corresponding expected count under the null hypothesis, is called the χ^2 ***goodness-of-fit statistic****.*

Under H_0, X^2 has approximately a χ^2 distribution with number of degrees of freedom being (number of categories) - 1 - (number of parameters estimated under H_0).

This approximation works well provided all the expected counts are reasonably large (in practice, it's often employed when the expected number of counts are all at least 5 and works tolerably well even here).

This χ^2 test was suggested by Karl Pearson before the theory of hypothesis testing was fully developed.

Test of Independence in a Contingency Table The same test statistic can also be used to test for independence of variables in a contingency table. In this case, the row and column totals are fixed in addition to the grand total, hence we lose $1 + (r-1) + (c-1)$ degrees of freedom, leaving $rc - r - c + 1 = (r-1)(c-1)$ degrees of freedom.

Example 7.2. Use the following 3×3 contingency table to test at the 0.01 level of significance whether a person's ability in mathematics is independent of her/his interest in statistics.

		Ability in Maths			
		Low	Average	High	Totals
Interest in Stats	Low	63	42	15	120
	Average	58	61	31	150
	High	14	47	29	90
	Totals	135	150	75	360

$$\begin{aligned} & H_0: \text{ Ability in maths and interest in statistics are independent.} \\ \text{versus} \quad & H_1: \text{Ability in maths and interest in statistics are not independent.} \end{aligned}$$

Decision rule: reject H_0 if $X^2 > \chi^2_{0.01,4} = 13.277$ where

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}}$$

Under H_0, *i.e.* independence, the expected frequencies are given by the product of corresponding marginal estimated probabilities times the total number of individuals. For example for the first row these are $\left(\frac{120}{360}\right)\left(\frac{135}{360}\right) 360 = 45$, $\left(\frac{120}{360}\right)\left(\frac{150}{360}\right) 360 = 50$, and $120 - 45 - 50 = 25$. Thus

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(63-45)^2}{45} + \frac{(42-50)^2}{50} + \cdots + \frac{(29-18.75)^2}{18.75} = 32.14$$

Since $32.14 > 13.277$ the null hypothesis must be rejected. ◁

7.7 Exercises

Exercise 7.7.1. The random variables $X_1, X_2, \ldots, X_n$ are iid with $\mathbb{P}(X_i = j) = p_j$ for $j = 1, 2, 3, 4$, where $\sum p_j = 1$ and $p_j > 0$ for each $j = 1, 2, 3, 4$.

Interest centres on the hypothesis H_0 that $p_1 = p_2$ and simultaneously $p_3 = p_4$.

(a) Define the following terms
   i. a hypothesis test,
   ii. simple and composite hypotheses, and
   iii. a likelihood ratio test.

(b) Letting $\theta = (p_1, p_2, p_3, p_4)$, $\mathbf{X} = (X_1, \ldots, X_n)^\mathsf{T}$ with observed values $\mathbf{x} = (x_1, \ldots, x_n)^\mathsf{T}$, and letting y_j denote the number of $x_1, x_2, \ldots, x_n$ equal to j, what is the likelihood $L(\theta|\mathbf{x})$?

(c) Assume the usual regularity conditions, *i.e.* that the distribution of $-2\ell(\theta|\mathbf{x})$ tends to χ^2_ν as the sample size $n \to \infty$. What are the dimension of the parameter space Ω_θ and the number of degrees of freedom ν of the asymptotic chi-squared distribution?

(d) By partial differentiation of the log-likelihood, or otherwise, show that the maximum likelihood estimator of p_j is y_j/n.

(e) Hence show that the asymptotic test statistic of $H_0 : p_1 = p_2$ and $p_3 = p_4$ is

$$2r(\mathbf{x}) = 2\sum_{j=1}^{4} y_j \log(y_j/m_j),$$

where $m_1 = m_2 = (y_1 + y_2)/2$ and $m_3 = m_4 = (y_3 + y_4)/2$.

(f) In a hospital casualty unit, the numbers of limb fractures seen over a certain period of time are:

	Side	
	Left	Right
Arm	46	49
Leg	22	32

Using the test developed above, test the hypothesis that limb fractures are equally likely to occur on the right side as on the left side.

Discuss *briefly* whether the assumptions underlying the test appear reasonable here.

Exercise 7.7.2. Suppose 200 students are selected at random at a large university and each student in the sample is classified both according to the faculty in which he/she is enrolled and according to his/her preference for either of two candidates A and B in a forthcoming election.

		Candidate preferred			
		A	B	Undecided	Totals
Curriculum	Engineering & Science	24	23	12	59
	Humanities & Social Science	24	14	10	48
	Fine Arts	17	8	13	38
	Industrial & Public Administration	27	19	9	55
	Totals	92	64	44	200

Test the hypothesis

$$H_0 : p_{ij} = p_{i+}p_{+j} \text{ for } i = 1, \ldots, R; j = 1, \ldots, C$$

versus H_1 : H_0 is not true.

where
p_{ij} : probability that an individual selected at random will be classified with the $i'th$ row and the $j'th$ column of the table,
p_{i+} : marginal probability that individual will be classified in the $i'th$ row,
p_{+j} : marginal probability that individual will be classified in the $j'th$ column.

8. Simulation for Inference

You will already be familiar with the material in this chapter from the first half of the lecture course. This chapter is, intentionally, rather more verbose than much of the rest of the lecture notes. This is to allow students taking ST903 who have not previously encountered Monte Carlo methods to familiarise themselves with the material. You will find that the notes you took in that part of the course cover what you need to know about Monte Carlo methods for the entirety of this course. It is hoped that you will find this brief precis of the fundamentals useful. Rest assured, only a short amount of time will be spent on it.

You may find that section 8.3 is rather different to the simulation-based techniques covered in the first part of the course. It's certainly worth making sure that this material makes sense, even if you're confident that you have a firm grasp of Monte Carlo methods.

The Monte Carlo method is used to estimate features of distributions that we cannot compute analytically. It is particularly well suited to the approximation of expectations with respect to a particular distribution, but can also be used for optimisation, the approximation of quantiles and a host of other tasks.

Values are generated from a specified distribution, simulating a random sample from a specified population. The properties of the sample are then used to approximate the properties of that population. Note that this is in some ways an inversion of the usual statistical use of sampling theory: rather than using knowledge of distributions to make inferences about a realised sample, the same connection is used to write down quantities of interest in probabilistic terms and then to use a related sample to approximate that probabilistic representation.

8.1 Background and History

Monte Carlo methods may be thought of as a collection of computational techniques for the (usually approximate) solution of mathematical problems, which make fundamental use of random samples. Two classes of statistical problems are most commonly addressed within this framework: integration and optimisation. The next few sections will concentrate on the former: it is the (approximate) calculation of integrals using collections of random samples that people usually think of when they refer to the Monte Carlo Method. Monte Carlo methodology is also widely used in the simulation of physical, chemical and biological systems.

In the field of statistics, Monte Carlo methods are most interesting as a computational device for performing statistical inference. Many interesting models have extremely complex structures and cannot easily be dealt with using traditional techniques. Within the Bayesian paradigm, all information upon which inference can be based is encoded within the posterior probability distribution (see chapter 9). Using Monte Carlo methods we are able to characterise these distributions, and calculate expectations under them: the primary inferential technique.

8.1.1 A Simple Example

Monte Carlo methods invert the usual problem of statistics: rather than estimating random quantities in a deterministic manner, random quantities are employed to provide estimates of deterministic quantities.

Example 8.1 (Estimating π). One simple Monte Carlo experiment considers rain which falls uniformly at random (*i.e.* the location of any raindrop may be interpreted as a realisation of a uniformly distributed random variable) over some square region of space, and a circle inscribed within that square. Without reference to any formal probability theory, it is intuitive that the probability of a uniform raindrop falling in any region within the square must be proportional to the area of that region and independent of its location. Consequently, the probability, p, that a raindrop lies within the inscribed circle may be expressed in terms of their areas. If the square has sides of length $2r$ the circle must be of radius r and:

$$p = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4}.$$

In itself this may not seem particularly interesting. However, having expressed π as a function of this probability its estimators can be used to approximate π. Proceeding analytically is not possible: obtaining the probability requires knowledge of π. Intuitively, it is possible to estimate this probability by counting the proportion of raindrops which lie within the circle: if n raindrops are observed and m of those lie within the circle then one may estimate p, using $\hat{p} = m/n$.

Figure 8.1 shows a computer simulation in which 500 raindrops were distributed uniformly over a square — in this case $\hat{p} = 383/500$ and, defining $\hat{\pi}$ via the relationship between p and π, $\hat{\pi} = 383/125 = 3.064$. A poor estimate of π, considering the computational effort used, but an estimate nonetheless. m is, in fact, a realisation of a Bin (n, p) random variable.

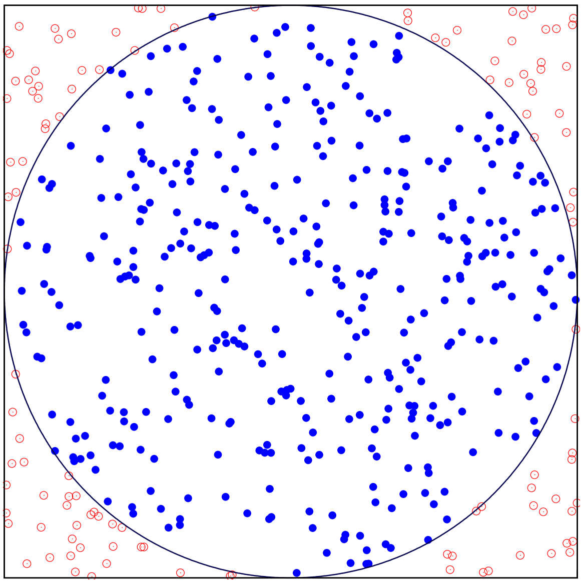


Fig. 8.1. Estimating π: 383 of the 500 samples lie within the circle.

This table shows results of this Monte Carlo simulation for increasing n

n	100	500	1000	10000	100000
$4m$	3.28	3.064	3.132	3.138	3.13828
$1.96 \cdot 4 \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}$	0.308	0.148	0.102	0.032	0.010

Using the fact that m is a realisation of a binomial random variable, M, together with the large values of n used, we can employ a CLT to argue that $4M$ is approximately distributed according to a normal distribution:

$$4M \sim \mathsf{N}\left(\pi, \frac{16\frac{\pi}{4}(1-\pi/4)}{n}\right)$$

leading to the estimate for the standard error used in the table. ◁

8.1.2 History

Whilst there is some debate about the nature of the first Monte Carlo computations ever carried out (with some authors arguing that they date back as far as the times of ancient Babylon), it is generally agreed that the first modern Monte Carlo experiments were carried out in the latter decades of the nineteenth century and, like the above example, were concerned with the estimation of π.

Example 8.2 (Buffon's Needle). Consider dropping a needle of length l uniformly onto an array of parallel lines separated from one another by a distance, $d > l$. Figure 8.2 illustrates this. Originally, the question was *what is the probability that the needle intersects a line?* This question is relatively easy to answer. Assume that the array is large enough that edge effects are negligible, and assume that the x and y coordinates of the needles centre are uniform over the array and the orientation of the needle is uniformly distributed over the interval $[0, \pi)$.

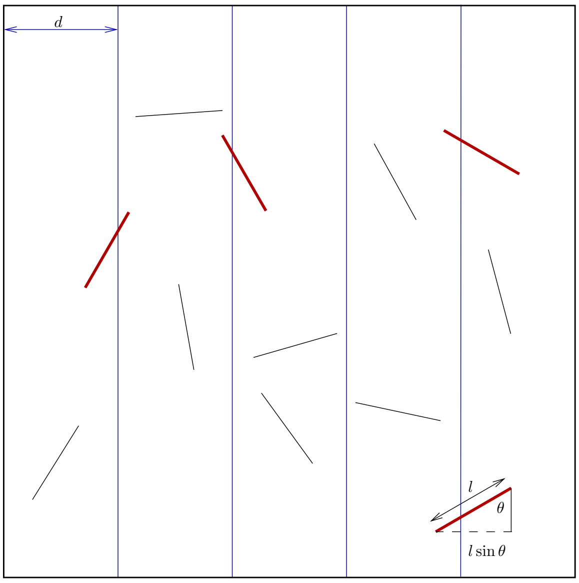


Fig. 8.2. Buffon's needles: thicker needles intersect the grid lines.

As figure 8.2 illustrates, the needle intersecting a line is equivalent to a rectangle of width $l \sin\theta$ intersecting that line. This happens, for given θ, with probability

$$\mathbb{P}[\text{intersection}|\theta] = \frac{l \sin\theta}{d}.$$

The angle θ is uniform on $[0, \pi)$, hence:

$$\begin{aligned}\mathbb{P}[\text{intersection}] =& \int_0^\pi \mathbb{P}[\text{intersection}|\theta] \frac{1}{\pi}\, d\theta \\ =& \frac{l}{\pi d} \int_0^\pi \sin\theta \, d\theta = \frac{2l}{\pi d}.\end{aligned}$$

Consequently, if n needles are dropped then the average number which would intersect a line if the experiment were carried out many times is $2nl/d\pi$.

In a Monte Carlo setting, this means that π may be estimated using:

$$\hat{\pi} = \frac{2nl}{Md},$$

where M is the number of needles crossing a line when n are dropped. In 1901 Mario Lazzarini reported carrying this experiment out using a 2.5cm long needle and lines separated by $d = 3$cm. 1808 of his 3408 needles crossed a line, suggesting

$$\pi \approx \frac{2 \times 3408 \times 2.5}{1808 \times 3} = \frac{355}{133},$$

which, remarkably, corresponds to the best rational approximation to π with denominator below 16,000 leading to suggestions that this was an early example of fabricated data. ◁

Although these early examples of Monte Carlo techniques were in some sense successful, the cost of carrying out the experiments rendered them largely impractical and of only specialist interest. This began to change in Los Alamos in the 1940s, when physicists working on particle transport problems began solving them using something which they termed the "Monte Carlo method". The precise origin of the name varies from one account to another, but there is general agreement that it stems from its relationship with the games of chance played in the casinos of Monte Carlo. The revolutionary step introduced at this time was the use of random number generators (and latterly digital computers producing pseudorandom numbers) in place of physical experiments to perform the calculation. The physics community has continued to contribute to the development of Monte Carlo methodology to the present day.

Perhaps the final revolution in Monte Carlo methodology occurred during the 1980s, during which the increasing use of Bayesian methods (see chapter 9) and the associated need to evaluate complex high-dimensional integrals led to interest in and development of the methods by the statistics community. This was the point at which the use of Monte Carlo methods to approximate general integral expressions became widespread. The increased interest also drove — and continues to drive — the development of methodology and theory amongst researchers outside of the traditional application domains of Monte Carlo integration.

8.2 Basic Monte Carlo Inference

In order to develop some theory, it's useful to have some simple examples to refer to. The following toy example supplements the estimation of π and Buffon's needle experiment described previously.

Example 8.3. Consider an experiment where a fair-sided die is rolled and

$$X = \begin{cases} 1 & \text{if the die shows a 1 or a 2} \\ 0 & \text{otherwise.} \end{cases}$$

Let $U \sim U(0,1)$. Then X can be simulated as

$$X = \begin{cases} 1 & \text{if } 0 < U \leq \frac{1}{3} \\ 0 & \text{if } \frac{1}{3} < U < 1 \end{cases}.$$

For example, we might generate the following

U_i	0.47	0.78	0.55	0.96	0.029	0.84	0.60	0.10	0.05	0.46
X_i	0	0	0	0	1	0	0	1	1	0

From which we can estimate $\bar{x} = 0.3$. Note that $\bar{x}$ is a consistent estimator of $\mu = \mathbb{E}[X]$: $\bar{x} \xrightarrow{p} \mu$. ◁

Exercise 8.2.1. Suppose a computer programme has been written to

1. Draw 60 x values from a distribution uniform between 0 and 100.
2. Count the number g of x values less than 20.
3. Repeat this procedure to get 5000 g values, $g_1, \ldots g_{5000}$.
4. Calculate the average g_{av} and the variance g_{var} of the g values.

(i) What is this Monte Carlo study designed to investigate?
(ii) What number should g_{av} be close to? Explain your logic.
(iii) What number should g_{var} be close to? Explain your logic.

8.2.1 Monte Carlo Integration

Suppose we want to evaluate $\int_a^b g(x)\,\mathrm{d}x$ a continuous function g over the closed and bounded $[a, b]$.

We can use the Monte Carlo method to evaluate this integral numerically when the anti-derivative of g is not available analytically. Note that

$$\int_a^b g(x)\,\mathrm{d}x = (b-a)\int_a^b g(x)\frac{1}{b-a}\,\mathrm{d}x = (b-a)\mathbb{E}[g(X)]$$

where $X \sim U(a, b)$. Thus we proceed as follows

1. Generate a random sample $X_1, \ldots, X_n$ from $U(a, b)$.
2. Compute $Y_i = (b-a)g(X_i)$.
3. Compute $\bar{Y}$ — a consistent estimator of $\int_a^b g(x)\,\mathrm{d}x$.

In fact, this is a particular case of a more general principle. If we can write an integral of interest as the expectation of a function of some random variable then we can use sampled realisations of that random variable in order to approximate the expectation.

If $X \sim f_X$ for some density f_X over a space $\mathcal{X}$ and we have $\varphi : \mathcal{X} \to \mathbb{R}$ then we can approximate

$$\mathbb{E}[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) f_X(x)\mathrm{d}x$$

in the following way:

1. Generate a random sample $X_1, \ldots, X_n$ from f_X.
2. Compute $Y_i = \varphi(X_i)$.
3. Compute $\bar{Y}$ — an unbiased and consistent estimator of $\mathbb{E}[\varphi(X)]$.

Note that, if Y_i has finite second moment, then the CLT (theorem 5.1) also tells us that $\bar{Y}$ is asymptotically normal with known variance.

Example 8.4 (Example 8.1 contd.). This Monte Carlo integration argument can be used to formally justify the approach used to estimate π in example 8.1 by considering a function which takes a value of one within the circle and 0 outside and calculating its expectation under the distribution of the coordinates which is uniform over the square. To see this explicitly, writing this function as

$$\mathbb{I}_{\text{circle}}(x, y) = \begin{cases} 1 & \text{if } x^2 + y^2 \leq r^2 \\ 0 & \text{otherwise} \end{cases}$$

it is clear that the area of the circle of interest is equal to the integral of this function over any region which includes the circle: such as the square in which it is inscribed. Consequently, our representation for π may be written in the form

$$\begin{aligned}\pi/4 = p =& \frac{\int_{-r}^{r}\int_{-r}^{r}\mathbb{I}_{\textsf{circle}}(x,y)\mathrm{d}x dy}{(2r)^2} \\ =& \int_{-r}^{r}\int_{-r}^{r}\frac{1}{(2r)^2}\mathbb{I}_{\textsf{circle}}(x,y)\mathrm{d}x dy = \mathbb{E}[\mathbb{I}_{\textsf{circle}}(X,Y)].\end{aligned}$$

where the expectation is with respect to the uniform distribution over the square $[-r,r]\times[-r,r]$ which has density $(2r)^{-2}$ over that square and 0 elsewhere.

Approximating the final line in this expression via the simple Monte Carlo method would give *exactly* the same estimator as that described above. The integral may be approximated by sampling pairs of points (X,Y) from this distribution and calculating the mean value of the function $\mathbb{I}_{\textsf{circle}}$ at the sampled points. $\triangleleft$

Example 8.5. (A More Realistic Simulation Application) Consider a stock whose current price is S_0 and suppose that the price at u years in the future is

$$S_u = S_0 e^{(r-\sigma^2/2)u+\sigma u^{1/2}Z}$$

where

$$Z \sim N(r-\sigma^2/2, \sigma^2 u)$$

and r is the risk-free interest rate.

Suppose we need to price the option to buy one share for price q at a particular u.

The value of the option at time u will be $h(S_u)$ where

$$h(s) = \begin{cases} s-q & \text{if } s > q, \\ 0 & \text{otherwise.}\end{cases}$$

The **Black-Scholes formula** for pricing options suggests that fair price for the option is the value of $\mathbb{E}[h(S_u)]$, which is $e^{-ru}\mathbb{E}[h(S_u)]$ (the exponential term provides discounting to compensate for the fact that the money used to buy the option could also produce a profit if invested in "risk-free" investments).

For example, suppose that $q = S_0, r = 0.06, u = 1$, and $\sigma = 0.1$.

Then the Black-Scholes formula says the option price should be $0.0726 S_0$.

But what if σ is unknown? Here we consider a simple model for this sort of problem; the mathematical finance literature contains rather more sophisticated approaches to the problem.

Suppose that both Z and σ are independent random variables.

Let $Z \sim N(0,1)$ and let

$$\tau = \frac{1}{\sigma^2} \sim \textsf{Gamma}\,(\alpha,\beta)$$

α and β might be obtained by estimating the variance of stock prices based on historical data combined with expert opinion.

The Black-Scholes formula tells us that the fair price, given σ is just the conditional mean of $e^{-ru}h(S_u)$. If we don't know σ then calculating the expectation of this fair price with respect to σ (which we assumed we knew the distribution of) would tell us what we want to know.

To estimate this marginal mean of $e^{-ru}h(S_u)$, we can simulate a large number of $\sigma^{(i)}$, evaluate the Black-Scholes formula in each case and average the results.

As before, let $q = S_0, r = 0.06$, and $u = 1$, but now

$$\frac{1}{\sigma^2} \sim \textsf{Gamma}\,(2, 0.0127)$$

Thus $\mathbb{E}[\sigma] = 0.1$ but σ has substantial variablity.

We sample 1,000,000 values of σ from this distribution and compute the Black-Scholes formula for each value.

The average, is $0.0756 S_0$ with standard error $1.814 S_0 \times 10^{-5}$, only slightly larger than when we assumed $\sigma = 0.1$.

This method can be rather easily extended. For example, we might have a much more complicated model for σ. As we only need to be able to simulate values of σ to calculate expectations of this sort, this doesn't dramatically increase the complexity of the problem. When the price process S_u is even more complicated, we can simulate S_u directly and estimate the mean of $h(S_u)$. ◁

In order to employ simple Monte Carlo to calculate an expectation with respect to a distribution, it is necessary to be able to sample from the distribution of interest. This requirement is not necessarily easily satisfied: Monte Carlo methods are generally used to deal with complicated distributions which do not admit tractable analytic solutions and it can be very difficult to obtain samples from such distributions. Two generally-applicable methods for obtaining samples from known distributions are introduced below, and this is followed by a method for estimating expectations under a distribution using samples from a *different* distribution.

8.2.2 Inversion Sampling

Assume that the distribution of interest is univariate and has known distribution function $F(x) = \mathbb{P}[X \leq x]$. The generalised inverse of a distribution function, F, may be defined as:

$$F^{-}(p) = \inf \{x : F(x) \geq p\},$$

that is, $F^{-}(p)$ is the smallest value of x such that $F(x)$ is at least p. This allows us to talk about "the inverse" of the distribution function when it isn't bijective (F may map a range of values to the same value if there's a interval with zero probability).

Inversion sampling transforms a random variable, U, with a uniform distribution on the interval $[0, 1]$ (*i.e.* , a random variable with density 1 over $[0, 1]$ and 0 elsewhere), by setting $X = F^{-}(U)$. If F^{-} is the generalised inverse of the distribution function of interest, then X has the distribution of interest:

$$\begin{aligned}\mathbb{P}[X \leq x] =& \mathbb{P}[F^{-}(U) \leq x] \\ =& \mathbb{P}[U \leq F(x)] \text{ by monotonicity} \\ =& F(x).\end{aligned}$$

Note that this method follows essentially automatically from theorem 4.6 and, indeed, provides a solution to exercise 4.9.2. Figure 8.3 illustrates the method. The hollow circles on the vertical axis correspond to three realisations of U and the filled circles on the horizontal axis are the corresponding realisations of X. This method can be used to sample realisations of any one-dimensional real-valued random variable, provided that the inverse of the distribution function is known.

8.2.3 Rejection Sampling

Unfortunately, the conditions given in the final sentence of the previous section very often fail to be met in practice: many interesting distributions are multivariate and it's very often not practical to obtain the inverse of the distribution function even in univariate cases.

A more general approach is motivated by the fact that sampling a collection of random variables according to a given density is equivalent to sampling uniformly in the area under the density graph and discarding the additional dimension. More formally, sampling uniformly from the region $\{(x, u) : u < f(x)\}$ and retaining only x is equivalent to sampling x according to $f(x)$. **Rejection sampling** provides samples uniformly under the graph of the density of interest by sampling from

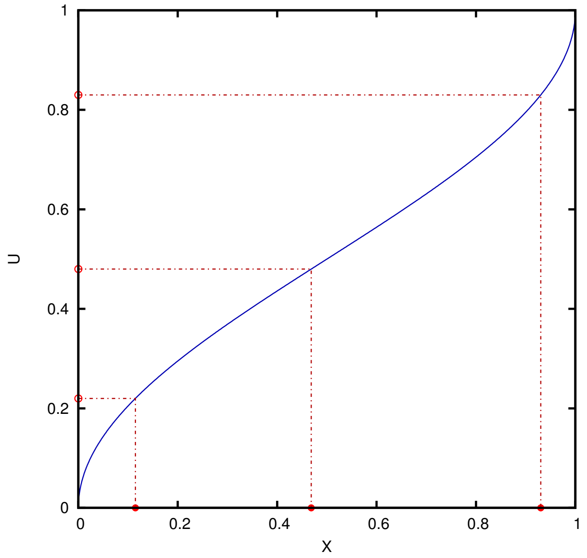


Fig. 8.3. Inversion sampling. Filled circles illustrate samples from the illustrated distribution function.

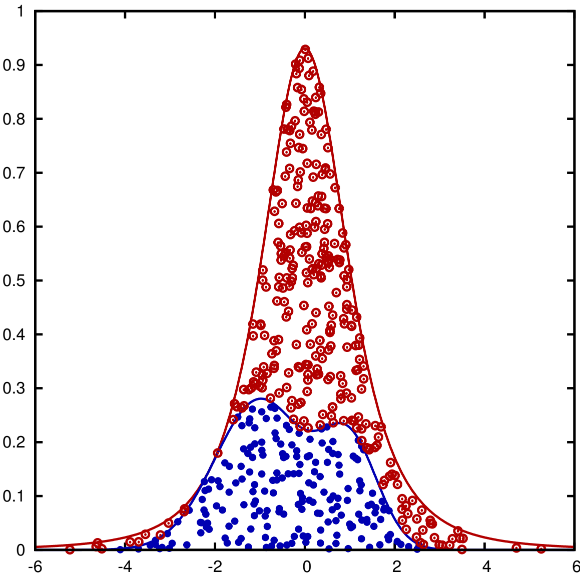


Fig. 8.4. Rejection sampling. Filled circles indicate accepted samples; hollow circles rejected proposals.

a larger area and rejecting those samples which fall outside the region of interest. Formally, given a density, g, from which it is possible to obtain samples and some known constant, $M \geq \sup_x \frac{f(x)}{g(x)}$, one can consider the following algorithm to obtain a sample from f:

1. Sample X from g.
2. Sample U from $\mathcal{U}(0, M)$.
3. If $U > f(X)/g(X)$ reject X and go to 1.
4. Accept X as a sample from f.

Figure 8.4 illustrates the principle: the filled circles indicate accepted samples (which are distributed uniformly beneath the bimodal density) the hollow circles were rejected (samples were generated uniformly in the area under the rescaled unimodal density).

The above algorithm is constructed such that given a sampled value, x, the probability that it will be accepted is:

$$\mathbb{P}[X \text{accepted}|X = x] = \mathbb{P}[U \leq f(x)/g(x)] = \int_0^{f(x)/g(x)} 1/M \mathrm{d}x = f(x)/Mg(x).$$

Notice that, on average 1 sample in M is accepted:

$$\begin{aligned}\mathbb{P}[X \text{Accepted}] =& \mathbb{P}[U \leq f(X)/g(X)] = \int \mathbb{P}(X \text{ Accepted}|X = x)\mathbb{P}(X = x)\mathrm{d}x \\ =& \int g(x) \cdot \frac{f(x)}{Mg(x)}\mathrm{d}x = \frac{1}{M}.\end{aligned}$$

Consequently, for this to be an efficient strategy it must be possible to sample from a distribution for which a small value of this constant can be found.

Proposition 8.1 (Correctness of Rejection Sampling). *The distribution of the accepted samples is, indeed, f.*

Proof. The proof follows by considering the distribution function of the accepted samples.

$$\begin{aligned}\mathbb{P}\left[X \leq x \middle| U \leq \frac{f(X)}{g(X)}\right] =& \mathbb{P}\left[X \leq x, U \leq \frac{f(X)}{g(X)}\right] \bigg/ \mathbb{P}\left[U \leq \frac{f(X)}{g(X)}\right] \\ =& \int_{-\infty}^{x} g(x) \cdot \frac{f(x)}{Mg(x)}\mathrm{d}x \bigg/ \frac{1}{M} \\ =& \frac{1}{M}\int_{-\infty}^{x} f(x)\mathrm{d}x \cdot M = F(x).\end{aligned}$$

The equivalance of distributions with identical distribution functions completes the proof. □

Example 8.6. Suppose we wish to simulate a random variable X having a Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$ distribution, *i.e.*

$$f(x) = \frac{1}{\pi}x^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}} \quad \text{for } 0 < x < 1$$

Note that

$$f(x) \leq \frac{1}{\pi}(x^{-\frac{1}{2}} + (1-x)^{-\frac{1}{2}}) \quad \text{for } 0 < x < 1$$

The RHS can be written as $Mg(x)$ with $M = \frac{4}{\pi}$ and

$$g(x) = \frac{1}{2}\left[\frac{1}{2x^{\frac{1}{2}}} + \frac{1}{2(1-x)^{\frac{1}{2}}}\right].$$

which is a half-and-half mixture of two pdf's g_1, g_2

$$g_1(x) = \frac{1}{2\sqrt{x}}, 0 < x < 1 \quad \text{and} \quad g_2(x) = \frac{1}{2\sqrt{(1-x)}}, 0 < x < 1.$$

It is easy to generate observations from g_1 and g_2 using the probability integral transformation (PIT). Thus we can use the following acceptance/rejection algorithm

1. Simulate 3 independent random variables U_1, U_2, U_3 from $U(0,1)$.
2. If $U_1 \leq \frac{1}{2}$ simulate a value from g_1 using the PIT applied to U_2. Otherwise simulate a value from g_2 using PIT applied to U_2. This gives a proposal Y from pdf g(y).
3. If $U_3 \leq \frac{f(y)}{Mg(y)}$, accept the proposal and set $X = Y$. Otherwise repeat the process.

◁

Exercise 8.2.2. Suppose you want to simulate values of a random variable with pdf

$$f(x) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2}\right)$$

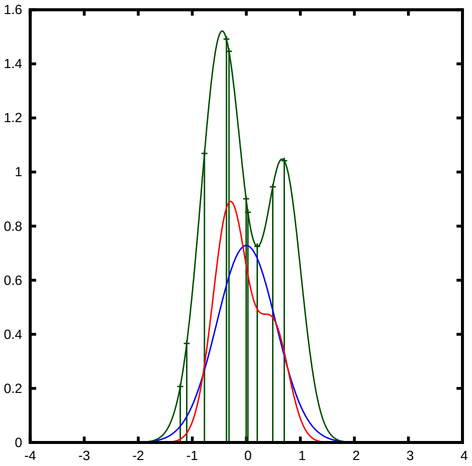


Fig. 8.5. Importance sampling: ten samples and their weights (green vertical lines) from a proposal distribution (blue) weighted according to the weight function (solid line) appropriate for the target distribution (red).

using the Cauchy distribution with pdf

$$g(x) = \pi^{-1}(1 + x^2)^{-1}$$

as the instrumental distribution in an rejection sampler. Describe how to do it. What it the value of M and the probability of acceptance? How would things change if you could sample from $f(x)$ but were interested in $g(x)$?

Various techniques to improve the computational efficiency of the method have been devleoped, but these do not qualitative change the implementation or behaviour of the approach. In some sources rejection sampling is referred to as the **accept-reject method**.

8.2.4 Importance Sampling

Rejection sampling can work well if it is possible to find a proposal g which is everywhere similar to f and such that M can be found and is small. However, it suffers from a number of deficiencies:

- To work efficiently it requires that f is nowhere much greater than g (so that M can be small); this can be difficult to arrange, particularly in multivariate settings.
- An explicit (and ideally tight) bound on f/g must be available; again, this can be all but impossible in high-dimensional problems.
- We must know the normalising constant of both f and g (or, at least, be able to incorporate these into M). As will become apparent in chapter 9 we often know f only up to an unknown normalising constant in problems of statistical interest.
- It seems somewhat wasteful to simply discard all of those reject samples (some other approaches do exist to make use of these samples to some extent).

Another technique which makes use of a distribution other than that of interest, but which is used specifically for calculating expectations, is known as **importance sampling**, as it attaches a weight to each sample based on its significance to the integral of interest. If f is the density of interest, we wish to calculate the expectation of φ under that distribution:

$$\begin{aligned}\mathbb{E}(\varphi(X)) &= \int \varphi(x) f(x) \mathrm{d}x \\ &= \int \varphi(x) \frac{f(x)}{g(x)} g(x) \mathrm{d}x,\end{aligned}$$

provided that $f(x)/g(x) < \infty$ (more formally, we require that f is absolutely continuous with respect to g, meaning that any set which has positive probability under f also has positive probability under g). An importance sampling estimate of the expectation of φ with respect to f is obtained by calculating the simple Monte Carlo estimate of $\varphi f/g$ with samples from g. So, given $X_1, \dots, X_n \sim g$ we have if we set $Y_i = f(X_i)/g(X_i)\varphi(X_i)$ that:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}\varphi(X_i).$$

is an unbiased and consistent estimate of $\mathbb{E}[\varphi(X)]$.

As presented above, it is necessary to know $f(x)/g(x)$ pointwise. A simple modification to the algorithm (which leads to an estimator which is biased for finite samples, but consistent) allows for situations in which $f(x)/g(x)$ is only known up to a normalising constant. If $w(x) = Cf(x)/g(x)$, then $\int w(x)\varphi(x)g(x)\mathrm{d}x = C\mathbb{E}(\varphi)$ and so, if we let $\mathbf{1}$ denote the unit function (the function which maps every point to 1), then we know that:

$$\frac{\int w(x)\varphi(x)g(x)\mathrm{d}x}{\int w(x)\mathbf{1}(x)g(x)\mathrm{d}x} = \frac{C\mathbb{E}(\varphi)}{C\mathbb{E}(\mathbf{1})} = \mathbb{E}(\varphi)$$

and so, by approximating both the numerator and denominator using a collection of samples from g, an estimator for $\mathbb{E}(\varphi)$ is given by:

$$\frac{\sum\limits_{i=1}^{n} w(X_i)\varphi(X_i)}{\sum\limits_{i=1}^{n} w(X_i)}.$$

in this case, the estimate is *biased.* The ratio of two unbiased estimators is not generally itself unbiased. However, the consistency property of the two estimators is transferred to their ratio and this **self-normalised importance sampling estimator** is consistent (and asymptotically unbiased) — this can be proved by considering a multivariate analogue of theorem 5.6. As the two estimators in the ratio are typically positive correlated it's very often the case that the self-normalised estimator has lower variance than the properly normalised form: another example of a bias-variance trade-off. As such, the self-normalised form is often used (and produces estimates with smaller MSE) even when the normalising constants are known.

8.3 Bootstrap Procedures

The previous sections of this chapter have been concerned with techniques known as Monte Carlo methods. Although the techniques discussed in the present section are motivated by similar ideas, they are generally regarded as being qualitatively different. Rather than attempting to calculate expectations with respect to a distribution of interest by employing samples from that distribution, we will consider what we can tell about the distribution of functions of a sample we really have (*e.g.* the sampling distribution of an estimator) based upon subsets of the sample drawn by sampling with replacement from within the original sample.

Suppose we have a sample of i.i.d. RV's $X_1, \dots, X_n$ and suppose inferences about a parameter θ are to be made on the basis of this sample.

Qualitatively, **Bootstrap** procedures simulate additional samples based on the observed sample. From each sample set an estimate of θ is determined and inferences are drawn from these estimates.

The Bootstrap Method

Let F denote the distribution of the population and let $x_1, \dots, x_n$ be the observed data. The (non-parametric) bootstrap method proceeds as follows

1. Construct an **empirical** distribution function, *i.e.* estimate F via $\hat{F}$ with $P(x_i) = \frac{1}{n}$
$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[x_i,\infty)}(x) = \frac{\text{“Number of } x_i \leq x\text{”}}{n}$$
where
$$\mathbb{I}_{(-\infty,x_i]}(x) = \begin{cases} 1 & \text{if } x \in (\infty, x_i] \\ 0 & \text{otherwise} \end{cases}$$
This is a formal way of describing a (random) discrete probability distribution which places a mass of $1/n$ n each of the n values in the sample.
2. For $i = 1, \ldots, B$:
   - Draw samples of size n from $\hat{F}(x)$ $X_1^i, \ldots, X_n^i$, (**bootstrap samples**). This process is sometimes termed **resampling** as it amounts to sampling from a sample.
   - For each bootstrap sample, $X_1^i, \ldots, X_n^i$ evaluate an estimator of θ, to obtain an estimate $\hat{\theta}_i^\star$
3. The set $\{\hat{\theta}_i^\star\}_{i=1}^B$ is a sample from the distribution of $\hat{\theta}^\star$ — the **bootstrap distribution**.

Motivating hypothesis: *T*he relationship between $\hat{\theta}$ and F is similar to the relationship between $\hat{\theta}^\star$ and $\hat{F}$, so that the distribution of $\hat{\theta} - \theta$ is similar to the bootstrap distribution of $\hat{\theta}^\star - \hat{\theta}$.

If this hypothesis is true (which, in some sense it is) we can learn about θ by learning about the bootstrap distribution of $\hat{\theta}^\star$.

Estimating Variance and Bias of $\hat{\theta}$

Suppose we draw B bootstrap samples, giving estimates $\hat{\theta}_i^\star, i = 1, \ldots, n$. We compute
$$\bar{\theta}^\star = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_i^\star$$
$$\text{and} \quad \hat{\sigma}^2(\hat{\theta}^\star) = \frac{1}{B-1}\sum_{i=1}^{B} (\hat{\theta}_i^\star - \bar{\theta}^\star)^2$$
to give estimates of the variance and bias:
$$\mathbb{V}\mathsf{ar}[\hat{\theta}] \approx \hat{\sigma}^2(\hat{\theta}^\star) \quad \text{and} \quad \mathsf{Bias}[\hat{\theta}] \approx \bar{\theta}^\star - \hat{\theta}$$
From this we obtain a bias-adjusted estimate of θ:
$$\hat{\hat{\theta}} := \hat{\theta} - (\bar{\theta}^\star - \hat{\theta}) = 2\hat{\theta} - \bar{\theta}^\star$$

Example 8.7. Consider the following random sample from a population with unknown mean

7.0 19.8 12.8 6.0 15.2 5.1 15.0 7.6

Suppose we are interested in the 25% trimmed mean:
$$\hat{\theta} = \frac{1}{6}(7.0 + 12.8 + 6.0 + 15.2 + 15.0 + 7.6) = 10.6$$
We obtain 10 bootstrap samples by using uniform random numbers to sample with replacement from the observed sample and calculate the 25% trimmed mean in each case.

We obtain the results
$$\hat{\theta}_1^\star = 13.43, \quad \hat{\theta}_2^\star = 9.68, \quad \ldots, \quad \hat{\theta}_{10}^\star = 11.53$$
The mean and variance of these estimates are
$$\bar{\theta}^\star = 10.904 \quad \text{and} \quad \hat{\sigma}^2(\hat{\theta}^\star) = 4.604$$
The latter estimates the variance of $\hat{\theta}$ and the bias-adjusted estimate is
$$2\hat{\theta} - \bar{\theta}^\star = 2 \cdot 10.6 - 10.904 = 10.296$$

◁

Bootstrap Confidence Interval

One common method of forming a bootstrap confidence interval is the **percentile** method which simply equates quantiles of the distribution of $\hat{\theta}$ to the equivalent quantiles of the bootstrap distribution of $\hat{\theta}^{\star}$.

Thus the $(1-\alpha)100\%$ percentile interval is

$$\left(\hat{\theta}^{\star}_{(\frac{\alpha}{2})}, \hat{\theta}^{\star}_{(1-\frac{\alpha}{2})}\right)$$

Under certain mild assumptions it can be shown that the percentile inteval has the correct coverage probability.

Example 8.8. (Example contd) We generate 1000 bootstrap samples and calculate the 25% trimmed mean for each. The 50th smallest and the 50th largest $\hat{\theta}*$ are 7.517 and 14.50 respectively, *i.e.*

$$\hat{\theta}^{\star}(0.05) = 7.517 \quad \text{and} \quad \hat{\theta}^{\star}(0.95) = 14.50$$

Hence the 90% bootstrap confidence interval is (7.517, 14.50). ◁

Hypothesis Testing

The equivalence between confidence intervals and hypothesis tests may be used to implement a (two-sided) bootstrap hypothesis test.

In the example above the 90% CI is

$$(7.517, 14.50)$$

Hence if we were to test

$$\begin{aligned} & H_0 : \theta = 6.5 \\ \text{versus} \quad & H_1 : \theta \neq 6.5 \end{aligned}$$

we would reject H_0 in favour of H_1 at the 10% significance level.

Exercise 8.3.1. Explain how bootstrapping could be used to test whether a population is symmetric based on the sample skewness $\mu_3/\mu_2^{3/2}$.

Example 8.9. (Application: Difference in Means) Frisby and Clatworthy (1975) studied the time that it took people to fuse **random-dot stereograms**: pairs of images that at first appearance seem to be random dots, but when viewed at a certain distance fuse to appear as a recognisable object.

The experimenters were concerned with the extent to which prior information about the recognisable object affected the time it took to fuse the images.

The conducted an experiment in which 43 subjects were not shown a picture of the object before being asked to fuse the images, whilst a second group of 35 were.

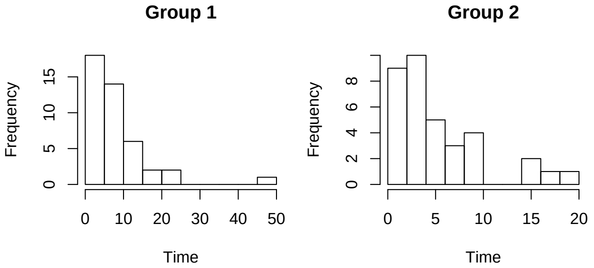
The results of the experiment were as follows:

	Mean	Variance
Group 1	8.560	2745.7
Group 2	5.551	783.9

We would like to test whether or not the mean time taken without prior information is no worse than the mean time taken with prior information.

$$\begin{aligned} & H_0 : \mu_1 \leq \mu_2 \\ \text{versus} \quad & H_1 : \mu_1 > \mu_2 \end{aligned}$$

One approach would be to use a two-sample t-test. This assumes the two samples are drawn from normal populations with equal variances. We plot the data from each group:



Since the two-sample t-test is not appropriate, we use a bootstrap hypothesis test instead. Given the sample data $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ we proceed as follows

1. Draw 1000 bootstrap samples consisting of
   – a sample $\boldsymbol{x}_1^\star$ of size 43 drawn with replacement from $\boldsymbol{x}_1$,
   – a sample $\boldsymbol{x}_2^\star$ of size 35 drawn with replacement from $\boldsymbol{x}_2$.
2. Evaluate $\bar{x}_{1b}^\star$ and $\bar{x}_{2b}^\star$, $b = 1, \ldots, 1000$ and compute

$$\hat{\theta}_b^\star = \bar{x}_{1b}^\star - \bar{x}_{2b}^\star$$

Then to test

$$\begin{array}{rl} & H_0 : \mu_1 \leq \mu_2 \\ \text{versus} & H_1 : \mu_1 > \mu_2 \end{array}$$

at the 5% level we find the 95th percentile of the bootstrap distribution and reject H_0 if $\hat{\mu}_1 - \hat{\mu}_2 = 8.560 - 5.551 = 3.009$ is greater than this statistic. We find the 95th percentile is 5.373, so we accept H_0 at the 5% significance level. ◁

8.4 Notes

Monte Carlo methods have advanced a great deal since their inception. This chapter has provided a brief summary of some of the simplest techniques which are available.

Having motivated the need for more sophisticated methods, we will touch on one class of algorithms which have been very widely used in modern statistics in section 9.6.

9. Elements of Bayesian Inference

Thus far we have considered only the so-called **classical** or **frequentist** approach to statistical inference. In the classical paradigm, unknown parameters are treated as fixed but unknown constants which are to be estimated. Probabilistic statements are made only about "true" random variables and observed data is assumed to correspond to a sampled set of realisations of random variables. **Bayesian** inference provides an alternative approach.

9.1 Introduction

In the Bayesian approach to inference, parameters are treated as **random variables** and hence have a probability distribution.

Prior information about θ is combined with information from **sample data** to estimate the distribution of θ.

This distribution contains all the available information about θ so should be used for making estimates or inferences.

We have prior information about θ given by the **prior distribution**, $p(\theta)$, and information from sample data given by the **likelihood** $L(\theta; x) = f(x; \theta)$. By Bayes Theorem the conditional distribution of θ given $X = x$ is

$$q(\theta|x) = \frac{f(x;\theta)p(\theta)}{h(x)} = \frac{L(\theta;x)p(\theta)}{h(x)} \propto L(\theta;x)p(\theta)$$

where $h(x)$ is the marginal distribution of x. We call $q(\theta|x)$ the **posterior distribution**.

Actually, a Bayesian would most probably have written:

$$q(\theta|x) = \frac{f(x|\theta)p(\theta)}{h(x)}.$$

There is no need to distinguish between parameters and random variables in notation and it's perfectly reasonable to condition upon the parameters within the Bayesian framework.

Note: A Bayesian statistician does not necessarily believe that all parameter values are classical random variables. The Bayesian interpretation of probability itself is different from the frequentist interpretation. Viewing probabilistic statements as quantifications of uncertainty without explicit reference to relative frequencies of occurrence allows their use much more widely. In the subjective Bayesian framework, probabilities are quantifications of personal belief in a statement.

Prior Distributions

The prior distribution $p(\theta)$ quantifies information about θ prior to the gathering of the current data.

Sometimes $p(\theta)$ can be constructed on the basis of past data. More commonly, $p(\theta)$ must be based upon an expert's experience and personal judgement. The following three examples give some simplified examples of the procedures by which such distributions can be obtained.

Example 9.1. Suppose that the proportions θ of defective items in a large manufactured lot is unknown. The prior distribution assigned to θ might be $U(0,1)$, *i.e.*

$$p(\theta) = \begin{cases} 1 & \text{for} < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$

◁

Example 9.2. Suppose that the lifetimes of lamps of a certain type are to be observed. Let X be the lifetime of any lamp and let $X \sim \text{Exp}(\beta)$, where β is unknown.

On the basis of previous experience the prior distribution of β is taken as a gamma distribution with mean 0.0002 and standard deviation 0.0001, i.e.

$$p(\beta) = \begin{cases} \frac{20000^4}{3!}\beta^3 e^{-20000\beta}, & \beta > 0 \\ 0 & \text{otherwise.} \end{cases}$$

◁

Example 9.3. A medical researcher was questioned about θ, the proportion of asthma sufferers who would be helped by a new drug.

She thought that

$$\mathbb{P}[\theta > 0.3] = \mathbb{P}[\theta < 0.3]$$

i.e. that the median $\theta_{0.5} = 0.3$.

Similarly, she thought that

$$\theta_{0.25} = 0.2 \quad \text{and} \quad \theta_{0.75} = 0.45$$

From tables giving quantiles of beta distributions, the researcher's opinion could be represented by $\text{Beta}(\alpha = 2, \beta = 4)$ for which

$$\theta_{0.25} = 0.194, \quad \theta_{0.5} = 0.314 \quad \text{and} \quad \theta_{0.75} = 0.454.$$

◁

Note that fusing information from different sources and extracting the knowledge of domain experts in order to produce prior distributions which genuinely encode the state of knowledge prior to the gathering of a first data set is a difficult problem. A large amount of research has been done on the area which is termed **prior elicitation**. If one conducts a series of experiments concerned with the same parameter then the situation is improved a little as we shall see the posterior distribution $p(\theta|x_1)$ can be used as the prior distribution when the next data set x_2, say, is obtained and this procedure can be applied iteratively as we learn progressively more about the parameter.

9.2 Parameter Estimation

Suppose we wish to estimate a parameter θ. We define a **loss function**

$$L_s(\theta, \hat{\theta})$$

which measures the loss which would be suffered as a result of using $\hat{\theta}$ as an estimator when the true value is θ. This is a formal way of encoding how bad various types of possible error in the estimation are to us. For instance, it can encode the difference between type I and type II errors if we are interested in estimating a binary indicator of disease status.

The **Bayes estimator** minimises the *expected* loss:

$$E\left[L_s(\theta,\hat{\theta})|x\right] = \int_{-\infty}^{\infty} L_s(\theta,\hat{\theta})q(\theta|x)d\theta$$

for the observed value of x.

The form of the Bayes estimator depends on the loss function that is used and the prior that is assigned to θ.

For example, if the loss is the absolute error

$$L_s(\theta,\hat{\theta}) = |\theta - \hat{\theta}|$$

then the Bayes estimator $\hat{\theta}_B(x)$ is the median of the posterior distribution.

For other loss functions the minimum might have to be numerically estimated.

Exercise 9.2.1. A commonly used loss function is the squared error loss function

$$L_s(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$$

Show that the corresponding Bayes estimator $\hat{\theta}_B$ is equal to the mean of the posterior distribution.

$$\hat{\theta}_B(x) = \mathbb{E}[\theta|x]$$

What is the minimum expected loss?

Example 9.4 (Continuing example 9.1.). Suppose that a random sample of n items is taken from the lot of manufactured items. Let

$$X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ item is defective} \\ 0 & \text{otherwise.} \end{cases}$$

then $X_1,\ldots,X_n$ is a sequence of Bernoulli trials with parameter θ. The pdf of each X_i is

$$f(x|\theta) = \begin{cases} \theta^x(1-\theta)^{1-x} & \text{for } x = 0,1 \\ 0 & \text{otherwise.} \end{cases}$$

Then the joint pdf of $X_1,\ldots,X_n$ is

$$f_n(x|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.$$

Since the prior pdf $p(\theta)$ is uniform it follows that

$$f_n(x|\theta)p(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}, \quad 0 \le \theta \le 1.$$

This is proportional to a beta distribution with parameters $\alpha = y+1$ and $\beta = n-y+1$, where $y = \sum x_i$. Therefore the posterior has pdf

$$q(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y}, \quad 0 \le \theta \le 1.$$

◁

Example 9.5. (Example 9.2 cont'd.) Suppose that the lifetimes $X_1, \ldots, X_n$ of a random sample of n lamps are recorded. The pdf of each x_i is

$$f(x_i, \beta) = \begin{cases} \beta e^{-\beta x_i} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The joint pdf of $x_1, \ldots, x_n | \beta$ is

$$f(x|\beta) = \beta^n e^{-\beta y}, \quad \text{where } y = \sum_{i=1}^{n} x_i.$$

with a gamma specified for $p(\beta)$ we have

$$f(x|\beta)p(\beta) \propto \beta^{n+3} e^{(-y+20000)\beta}$$

where a factor that is constant w.r.t. β has been omitted. The RHS is proportional to a Gamma $(n+4, y+20000)$, hence

$$q(\beta|x) = \frac{(y+20000)^{n+4}}{(n+3)!} \beta^{n+3} e^{-(y+20000)\beta}.$$

◁

9.3 Conjugate Prior Distributions

A **conjugate prior distribution** when combined with the likelihood function, produces a posterior distribution in the same family as the prior.

If we find a conjugate prior distribution which adequately fits our prior beliefs regarding θ, we should use it because it will simplify computations considerably. However, one should not employ a conjugate prior distribution for computational convenience if it does not represent those prior beliefs reasonably closely.

Example 9.6 (Sampling from a Bernoulli Distribution). Suppose $X_1, \ldots, X_n$ are a random sample from Ber (θ) $, 0 < \theta < 1$. Let $p(\theta)$ be Beta (α, β).

Then $q(\theta|x)$ is Beta $\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right)$.

The proof of this claim is analogous to Example 9.4 (note U$[0,1] \equiv$ Beta $(1,1)$). ◁

Exercise 9.3.1 (Sampling from a Poisson Distribution). Suppose $X_1, \ldots, X_n$ are a random sample from Poi (θ) Let $p(\theta)$ be Gamma (α, β).

Show that the posterior density, $q(\theta|x)$, is

$$\text{Gamma}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right).$$

Example 9.7 (Sampling from a Normal Distribution with known σ^2). Suppose $X_1, \ldots, X_n$ are a random sample from N (θ, σ^2). with σ^2 known. Let $p(\theta)$ be N (ϕ, τ^2).

Then $q(\theta|x)$ is

$$\text{N}\left(\frac{\phi\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2}, \left(\frac{\sigma^2 + n\tau^2}{\sigma^2\tau^2}\right)^{-1}\right)$$

Proof.

$$\begin{aligned}
q(\theta|x) &\propto p(\theta)L(\theta;x)\\
&= (2\pi\tau^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\theta-\phi}{\tau}\right)^2\right\}\\
&\qquad\times\prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{x_i-\theta}{\sigma}\right)^2\right\}\\
\therefore q(\theta|x) &\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)-2\theta\left(\frac{\phi}{\tau^2}+\frac{\sum x_i}{\sigma^2}\right)\right]\right\}\\
&= \exp\left\{-\frac{\sigma^2+n\tau^2}{2\sigma^2\tau^2}\left[\theta-\frac{\phi\sigma^2+n\bar{x}\tau^2}{\sigma^2+n\tau^2}\right]^2+\text{constant}\right\}
\end{aligned}$$

i.e. $q(\theta|x)$ is the pdf of $\mathsf{N}\left(\frac{\phi\sigma^2+n\bar{x}\tau^2}{\sigma^2+n\tau^2},\left(\frac{\sigma^2+n\tau^2}{\sigma^2\tau^2}\right)^{-1}\right)$ as required. □

Precision Note that the posterior variance

$$\left(\frac{\sigma^2+n\tau^2}{\sigma^2\tau^2}\right)^{-1}=\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-1}$$

i.e. the reciprocal of the sum of the reciprocals of the prior variance and the variance of the sample mean, respectively.

Because of this reciprocal relationship

$$\text{precision}=\frac{1}{\text{variance}}$$

is sometimes quoted instead of the variance.

Posterior Mean The posterior mean

$$\frac{\phi\sigma^2+n\bar{x}\tau^2}{\sigma^2+n\tau^2}=\frac{\frac{\phi}{\tau^2}+\frac{\bar{x}}{\sigma^2/n}}{\frac{1}{\tau^2}+\frac{1}{\sigma^2/n}}$$

i.e. a weighted average of the prior mean ϕ and the sample mean $\bar{x}$, with weights proportional to the prior precision and the precision of the sample mean.

This type of relationship holds for several sampling distributions when a conjugate prior is used. ◁

Example 9.8 (Sampling from an Exponential Distribution). Suppose $X_1,\ldots,X_n$ are a random sample from Exp (θ). Let $p(\theta)$ be Gamma (α,β).

Then $q(\theta|x)$ is

$$\text{Gamma}\left(\alpha+n,\beta+\sum_{i=1}^{n}x_i\right).$$

See example 9.5 for a proof of this claim. ◁

9.4 Uninformative Prior Distributions

One possible criticism of Bayesian inference is that it's not clear what to do when we don't have any prior information. In fact, Bayesian inference is really just a rule for updating our beliefs in light of new data. One way to deal with this is to attempt to employ priors which encode our ignorance.

An **uninformative** prior is one which attempts to encode as little information as possible about the value of a parameter

The simplest approach to the construction of uninformative priors is the **flat** prior which has $p(\theta) = \text{constant} \ \forall\, \theta$.

Flat priors can sometimes be obtained as special or limiting cases of conjugate priors, e.g.

– using a $\mathsf{Beta}\,(1,1) \equiv \mathsf{U}[0,1]$ prior for the Bernoulli parameter
– letting $\tau^2 \to \infty$ in the $\mathsf{N}\left(\phi,\tau^2\right)$ prior for the mean of $\mathsf{N}\left(\mu,\sigma^2\right)$ with known σ^2.

Other types of uninformative prior can also sometimes be obtained by procedures such as these.

If the prior is approximately constant over the range of θ for which the likelihood is appreciable, then approximately

$$q(\theta|x) \propto L(\theta;x)$$

and inference becomes similar to ML estimation. Note that even here there is a difference as a Bayesian would make parametric inference on the basis of a loss function and it is not necessarily (or in fact often) the case that minimising expected loss occurs when one chooses the maximum of the posterior density as the estimator.

In fact, the estimator which takes the maximum of the posterior density is the Bayesian estimator obtained as the limit of a sequence of loss functions and corresponds essentially to dealing with a loss which is zero if one estimates the parameter exactly correctly and one if there is any error at all. There are perhaps some situations in which this is justifiable, but one must think carefully about the choice of loss function when making decisions or inferences in a Bayesian framework.

Problems with Flat Priors

If the prior range of θ is infinite, a flat prior cannot integrate to 1. Such an **improper** prior may lead to problems in finding a "proper" posterior (*i.e.* one which can be normalised to integrate to unity as any probability density must).

We usually have *some* prior knowledge of θ and if we do we should use it, rather than claiming ignorance. In fact, one could argue that there are essentially no cases in which we really believe that any value in an unbounded set is not only possible but equally plausible.

Jeffreys Prior

Another issue is whether an informative prior should be flat for θ or some function of θ, say θ^2 or $\log\theta$. We will draw different inferences for a prior which is flat over any one of these functions. In some sense, this demonstrates that flat priors do not really encode complete ignorance at all.

One solution is to construct a prior which is flat for a function $\phi(\theta)$ whose Fisher information I_ϕ is constant. This leads to the **Jeffreys prior** which is proportional to

$$I_\theta^{\frac{1}{2}} = E\left[\left(\frac{\partial \ln[L(\theta;x)]}{\partial\theta}\right)^2\right]^{\frac{1}{2}} = \left[-E\left(\frac{\partial^2 \ln[L(\theta;x)]}{\partial\theta^2}\right)\right]^{\frac{1}{2}}.$$

Example 9.9. Suppose we are sampling from a Bernoulli distribution so that the likelihood is binomial. Show that in this case, the Jeffreys prior is proportional to

$$[\theta(1-\theta)]^{-\frac{1}{2}}$$

which is a $\mathsf{Beta}\left(\frac{1}{2},\frac{1}{2}\right)$ distribution. ◁

Note: Harold Jeffreys was a physicist who made substantial contributions to the theory and philosophy of Bayesian inference. His original motivation when developing this class of prior distributions was to develop a way of encoding a belief that the distribution should be invariant under a particular type of transformation: location parameters should have a prior invariant under shifts; scale parameters should have a prior invariant under scaling etc..

Perhaps the biggest problem with uninformative priors is that there's really no way to represent total ignorance. Saying that a prior is flat over the real line is arguably a very strong statement. It says that *a priori* you believe there's a very significant possibility that the value is arbitrarily large, for example.

9.5 Hierarchical Models

Bayesian inference treats unknown parameters as variables with a probability distribution. **Hierarchical models** exploit the flexibility this gives.

Consider the following **three-stage hierarchical model**. The data x have density

$$f(x;\theta) \quad \text{[stage 1]}$$

where θ is unknown. Denote the prior distribution for θ by

$$p(\theta,\psi) \quad \text{[stage 2]}$$

where ψ is also unknown, with prior distribution (sometimes known as a **hyper-prior**)

$$g(\psi) \quad \text{[stage 3]}.$$

Example 9.10. Suppose $\theta_1,\ldots,\theta_k$ are the average reading abilities of 7-year-old children in each of k different schools.

Samples of 7-year-olds are to be given reading tests to estimate θ_i. Let $X_{i1},\ldots,X_{in_i}$ be the reading abilities of a sample of n_i children from school i.

Suppose

$$X_{ij}|\theta_i \sim \mathsf{N}\left(\theta_i,\sigma^2\right) \quad \text{[stage 1]}$$

where σ^2 is the same for all schools and is assumed known.

Then let each θ_i be normally distributed, so that

$$p(\theta_1,\ldots,\theta_k;\psi) = \prod_{i=1}^{k}(2\pi\tau^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\tau^2}(\theta_i-\phi)^2\right\} \quad \text{[stage 2]}$$

where $\psi=(\phi,\tau)$. Finally assign an uninformative prior to ϕ and τ^2

$$g(\phi,\tau^2)=g_1(\phi)g_2(\tau^2)=\text{constant} \quad \text{[stage 3]}.$$

◁

Note that information about some of the θ_i provides information about the remainder (**borrowing strength**).

For example, if we had data from $k-1$ schools, then we could estimate the average reading abilities $\theta_1,\ldots,\theta_{k-1}$. Hence we could estimate ϕ and τ^2, the mean and variance of the distribution of θ_k.

As usual

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Here the unknown parameters are θ and ψ and their prior is

$$p(\theta, \psi)g(\psi).$$

The sampling distribution of the data depends only on θ, not ψ

$$L(\theta, \psi; x) = f(x; \theta).$$

Hence the joint posterior distribution of the parameters is

$$q(\theta, \psi|x) \propto f(x; \theta)p(\theta; \psi)g(\psi).$$

From the joint posterior

$$q(\theta, \psi|x) \propto f(x; \theta)p(\theta; \psi)g(\psi)$$

we can obtain the posterior density of θ by integrating w.r.t. ψ.

Further integration yields marginal distributions of individual components of θ.

For many hierarchical models the relevant integrations cannot be done analytically and the standard approach is to use a class of stochastic algorithms known as Markov chain Monte Carlo methods to approximate the integration numerically.

9.6 Inference using Markov Chain Monte Carlo (MCMC) Methods

The material in this section is a very brief introduction to the area. Of course, the notes you have from the first half of the lecture course cover this and much more. This section may prove useful to refresh you memory but is intended predominantly for those students attending ST903 who are not studying a formal Monte Carlo methods course.

The basic problem of Bayesian parameter inference is that we need to obtain the posterior distribution for a particular problem (and, indeed, optimise the expected loss function with respect to that posterior).

Suppose θ is a vector of unknown parameters. As usual

$$q(\theta|x) \propto f(x|\theta)p(\theta)$$

and to perform inference we need to obtain or approximate $q(\theta|x)$. We can say that

$$q(\theta|x) = cf(x|\theta)p(\theta)$$

with

$$c = \left\{ \int f(x|\theta)p(\theta)d\theta \right\}^{-1}$$

We cannot generally calculate c analytically. For many problems simple numerical integration techniques fail because the dimension of the space can be large and we may not know which regions of the space contain the most probability mass (in dimensions much above 3 it is surprisingly difficult). We may be able to use simple Monte Carlo integration methods, e.g. importance sampling to find c, but this is often difficult. Obtaining good importance sampling proposals, for example, requires a detailed understanding of the structure of the posterior — something which is not often available.

Markov chain Monte Carlo methods allows us to obtain a collection of dependent samples which marginally have distribution close to $q(\theta|x)$ and with which we can obtain consistent estimates of expectations with respect to $q(\theta|x)$ without knowing c.

9.6.1 Foundations

Before talking about methods, a few simple definitions are needed.

Definition 9.1. *(Discrete Time) Markov Chain A Markov chain is a discrete time stochastic process (*i.e. *a sequence of random variables),* $X_1, X_2, \ldots$ *with the following conditional independence property:*

$$\forall n, A, x_{1:n-1}: \quad \mathbb{P}(X_n \in A | X_{1:n-1} = x_{1:n-1}) = \mathbb{P}(X_n \in A | X_{n-1} = x_{n-1})$$

i.e. *Given the present the future is independent of the past. This property is known as the (weak) Markov property.*

The law of a time-homogeneous Markov chain may be specified by the initial density, $q(x_1)$, and the transition probability (or kernel) $K(x_n|x_{n-1})$. The joint density of the first n elements of the chain may then be written straightforwardly:

$$p(x_{1:n}) = q(x_1) \prod_{i=2}^{n} K(x_i|x_{i-1}).$$

Definition 9.2 (Invariance). *A Markov chain, X, is f-invariant if:*

$$\int f(x)K(y|x)dx = f(y)$$

If a Markov chain is f-invariant and $X_m \sim f$ for some m then it's clear that the marginal distribution of any X_n for $n \geq m$ must also be f.

If a Markov chain satisfies a condition known as *detailed balance* with respect to a distribution, then it is reversible (in the sense that the statistics of the time-reversed process match those of the original process) and hence invariant with respect to that distribution. The detailed balance condition states, simply, that the probability of starting at x and moving to y is equal to the probability of starting at y and moving to x. Formally, given a distribution f and a kernel K, one requires that $f(x)K(y|x) = f(y)K(x|y)$ and simple integration of both sides with respect to x proves invariance with respect to f under this condition.

The principle of most MCMC algorithms is that, if a Markov Chain has an invariant distribution, f, and (in some suitable sense) forgets where it has been, then using its sample path to approximate integrals with respect to f is a reasonable thing to do. This can be formalised under technical conditions to provide an analogue of the law of large numbers (often termed the ergodic theorem) and the central limit theorem. The first of these results tells us that we can expect the sample average to converge to the expectation with probability one as the number of samples becomes large enough; the second tells us that the estimator we obtain is asymptotically normal with a particular variance (which depends upon the covariance of the samples obtained, demonstrating that it is important that the Markov chain forgets where it has been reasonably fast). These conditions are not always easy to verify in practice, but they are important: it is easy to construct examples which violate these conditions and have entirely incorrect behaviour.

In order to use this strategy to estimate expectations of interest, it is necessary to construct Markov chains with the correct invariant distribution. There are two common approaches to this problem.

9.6.2 Gibbs Sampling

Let $q_i(\theta_i|\theta_{\setminus i}, x)$ denote the posterior probability density of θ_i given values of $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k$ where the parameter vector of interest $\theta = \theta_1, \ldots, \theta_k$.

The **Gibbs sampler** requires that for each $i = 1, \ldots, k$ these **full conditional densities** are ones that we can (ideally easily) sample from. The Hamersley-Clifford theorem (which is outside the remit of this course) tells us that if we know all of the full conditional distributions associated with a joint distribution then we can completely characterise that joint distribution so it won't be too surprising if we can construct an algorithm based around these full conditionals.

The Gibbs sampling algorithm aims to obtain a random sample from $q(\theta|x)$ by iteratively and successively sampling from the individual $q_i(\theta_i|\theta_{\setminus i}, x)$.

The algorithm is simple to apply:

```
– Initialise $\theta$, i.e. find starting values $\theta_i^{(1)}$, $i = 1, \dots, k$.
– For $j = 1, \dots, M$
  1. Draw $\theta_1^{(j+1)}$ from $q_1(\theta_1|\theta_{\setminus 1}^{(j)}, x)$; $\theta_{\setminus 1}^{(j)} = (\theta_2^{(j)}, \theta_3^{(j)}, \dots, \theta_k^{(j)})$.
  2. Draw $\theta_2^{(j+1)}$ from $q_2(\theta_2|\theta_{\setminus 2}^{(j)}, x)$; $\theta_{\setminus 2}^{(j)} = (\theta_1^{(j+1)}, \theta_3^{(j)}, \dots, \theta_k^{(j)})$.
  3. ...
  4. Draw $\theta_k^{(j+1)}$ from $q_k(\theta_k|\theta_{\setminus k}^{(j)}, x)$; $\theta_{\setminus k}^{(j)} = (\theta_1^{(j+1)}, \theta_2^{(j+1)}, \dots, \theta_{k-1}^{(j+1)})$.
  5. Put $\theta^{(j+1)} = (\theta_1^{(j+1)}, \theta_2^{(j+1)}, \dots, \theta_k^{(j+1)})$. Set $j = j + 1$.
```

As $j \to \infty$, under suitable regularity conditions (examples of which you saw in the first part of this module), the limiting distribution of the vector $\theta^{(j)}$ is the required posterior $q(\theta|x)$, i.e. for large j, $\theta^{(j)}$ is a random observation from $q(\theta|x)$.

The sequence $\theta^{(1)}, \theta^{(2)}, \dots$ is one realisation of a Markov Chain, since the probability of $\theta^{(j+1)}$ is only dependent on $\theta^{(j)}$.

We motivated the idea of MCMC by noting that once one sample form an f-invariant Markov chain has distribution f, the marginal distribution of all future samples must also be f. In the above algorithm we initialised our chain arbitrarily. We rely upon an idea known as **ergodicity** which tells us that, under regularity conditions, a Markov chain forgets its initial conditions. In essence, it won't matter where we start the chain as long as we run it for long enough.

We need to run the Markov chain until it has converged to its invariant distribution — all observations from a **burn-in** phase (in which the initialisation is still being forgotten) are discarded.

Suppose we have generated a large random sample $\theta^{[1]}, \theta^{[2]}, \dots, \theta^{[n]}$ using the Gibbs sampler. Inferences about a single θ_i would be based on this sample. For example the posterior mean and variance of θ_i given x would be

$$\bar{\theta}_i = \frac{1}{n}\sum_j \theta_i^{[j]} \quad \text{and} \quad \frac{1}{n}\sum_j (\theta_i^{[j]} - \bar{\theta}_i)^2.$$

The output from MCMC can also be used to approximate other integrals, *e.g.* for marginalisation or calculation of expected loss.

Exercise 9.6.1. Let $X_1, \dots, X_n$ be a random sample. Consider the following hierarchical Bayesian model:

$$\begin{aligned} X_1, \dots, X_n &\sim \mathsf{N}\left(\theta, \sigma^2\right) \quad \sigma^2 \text{ is known,} \\ \theta &\sim \mathsf{N}\left(0, \tau^2\right), \\ \frac{1}{\tau^2} &\sim \mathsf{Gamma}\left(a, b\right) \quad \text{where a, b are known.} \end{aligned}$$

Derive a Gibbs sampler for this model. I.e. how can we sample from the posterior distribution of the unknown parameters given a realisation of $x_1, \dots, x_n$?

Exercise 9.6.2. Let $X_1, \dots, X_n$ be a random sample. Consider the following hierarchical Bayesian model of failure rates:

$$\begin{aligned} X_1, \dots, X_n &\sim \mathsf{Poi}\left(\theta_i t_i\right) \quad \text{for fixed time } t_i, \\ \theta_i &\sim \mathsf{Gamma}\left(\alpha, \beta\right), \\ \alpha &\sim \mathsf{Exp}\left(a_0\right) \quad \text{for known } a_0, \\ \beta &\sim \mathsf{Gamma}\left(c, b_0\right) \quad \text{for known } c, b_0, \end{aligned}$$

where α and β are independent. Show that the conditionals for θ_i and β are Gamma distributions, whilst

$$q(\alpha|\beta,\theta) \propto \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^n \theta_i\right)^{\alpha-1} \exp(-a_0\alpha).$$

9.6.3 Metropolis Hastings Algorithm

Although Gibbs sampling is intuitive and simple it requires the availability of full conditional distributions. Various extensions exist which consider, for example, simulating more than one component of θ at a time or randomising the order of operations. However, they all require the ability to sample from a collection of full conditional distributions.

The Metropolis-Hastings Algorithm is an extension of Gibbs sampling which uses a rejection step to allow the use of a **proposal** distribution other than the full conditionals but which preserves invariance with respect to the target distribution of interest.

Intuitively, the algorithm proceeds at each iteration by sampling a proposed value, and this value is accepted with a probability which corrects for the relative probability of the state under the proposal and target distributions. If rejection occurs the existing state is replicated — notice this difference from the simple rejection sampling algorithm. More formally, the algorithm for sampling from p using a proposal q proceeds at iteration t as:

1. Sample $X' \sim q(\cdot|X_{t-1})$.
2. Calculate
$$\alpha(X_{t-1}, X') = \frac{p(X')q(X_{t-1}|X')}{p(X_{t-1})q(X'|X_{t-1})}.$$
3. Sample $U \sim \mathcal{U}[0,1]$.
4. If $U \leq \alpha$, set $X_t = X'$.
5. Otherwise, set $X_t = X_{t-1}$.

The acceptance probability α depends upon only the ratio of the target probability density at the proposed point to that at the current point (in addition to the ratio of proposal densities). This has the significant consequence that it is only necessary to evaluate the density up to a normalising constant. One common choice of q is a symmetric random-walk kernel (such as a Gaussian distribution centred on the previous value). In this case, the acceptance probability α simplifies to $p(X')/p(X_{t-1})$ due to the symmetry of the proposal kernel. This approach is often referred to as the random-walk Metropolis algorithm and corresponds to the first MCMC algorithm to be proposed. Another choice, leading to something termed the independence sampler, is to employ a proposal distribution which is entirely independent of the previous position. Of course, situations between these two extremes exist and often produce better results (for example, using a proposal comprising a mixture of a diffuse proposal distribution and a local random walk). It is also noteworthy that Gibbs sampling may be interpreted as a Metropolis-Hastings algorithm in which a particular choice of proposal kernel is made, guaranteeing that $\alpha = 1$ at all times (i.e. every proposed value is accepted). Although the Metropolis-Hastings algorithm allows a broad range of models to be explored, care is still required to design algorithms with proposals which allow a thorough exploration of the space and to assess their convergence. In order for a sampler to perform well, it is necessary for it to explore the entire support of the distribution and to move around it quickly. This can only be achieved through the use of good proposal distributions.

By way of an example, consider the random-walk Metropolis-Hastings algorithm. In contrast to the rejection sampling case, it is not desirable to maximise the acceptance probability of proposed moves in such an algorithm. Using a very small proposal variance leads to moves which are almost always accepted (α is always close to one if the density is continuous and proposed moves are small) but movements are all very small, leading to poor exploration of the distribution on a global scale. In contrast, if a very large proposal variance is used then most moves are proposed in regions of very little probability and many moves are rejected, leading to the sampler sticking with a single value for many iterations. Somewhere in between, good performance can often be obtained.

Without explaining the Metropolis-Hastings algorithm in detail, it might be informative to condsider the form of α. It consists essentially of two parts. The first is $p(X')/p(X_{t-1})$: the ratio of the target distribution at the proposed value to that at the old value. This term favours moves to regions of higher posterior density. The second is $q(X', X_{t-1})/q(X_{t-1}, X')$: the ratio of the density of the reverse of the current proposed move to that of the move itself. The second component penalises moves which are disproportionately probable. This component is uniformly 1 if the proposal distribution is symmetric.

9.6.4 Extensions NE

There is an enormous literature on MCMC and related methodology. It wouldn't be feasible to attempt to summarise even its key features here. However, there are a number of ideas which you are likely to encounter in the future. This section simply mentions a few of the most important and gives a very brief summary of their key features.

Simulated Annealing. One common MCMC algorithm, albeit one which is often considered in isolation, is simulated annealing (SA). This is an *optimisation*, rather than integration, algorithm which simulates a time inhomogeneous Markov Chain which is intended to provide samples from the regions of parameter space which minimise some objective function. Given a function, $H(x)$, which one wishes to find the minima of (usually the function of interest, but it can be any other function which shares its minimisers), SA provides a numerical method for locating its minima.

Modelled upon the physical annealing processes, SA employs densities proportional to $\exp(-\beta H(x))$ in order to find the value(s), $x^\star$, of its argument which minimise $H(x)$. Corresponding to an inverse temperature in physical annealing, β controls how concentrated these distributions are about the optima: if $\beta \ll 1/H(x^\star)$ then the distribution is very flat; as $\beta \to \infty$ the distribution becomes singular with all of its mass located at the optima.

The simulated annealing algorithm simulates a Markov chain with one such distribution as its invariant distribution at each iteration using the same mechanism as the Metropolis-Hastings algorithm. However, β is gradually increased as the algorithm runs leading to a time-inhomogeneous chain. This allows a single simulation run to be initiated at small β, allowing the chain to explore the space well, and to reach the minimisers which have increasing mass as β increases. Selecting the sequence of values taken by β is nontrivial; if it increases too quickly the simulation will become trapped in a local optimum but the more slowly it increases the more expensive each simulation becomes.

Reversible Jump. In many settings, often those involving model selection it is useful to have an algorithm which can deal with a collection of densities defined over spaces of different dimension (for example, consider choosing not just the parameters of a mixture model, but also the number of mixture components). Doing this is technically nontrivial, as the densities are not comparable (just as the mass per unit area of a sheet of steel cannot be compared directly with the density of a steel block). One solution to this problem is termed Reversible Jump Markov Chain Monte Carlo and, loosely speaking, it employs a dimension-matching construction in order to produce a Markov Chain which explores these spaces of differing dimension whilst retaining reversibility.

Adaptive MCMC. Several approaches have been developed to reduce the amount of work required to design an efficient sampler. One is to adaptively tune the parameters of the proposal distribution as a sampler is run (ensuring that these lead to the correct distribution introduces nontrivial technical difficulties and this should not be done without thorough consideration of these issues). The other involves using the entire history of the Markov Chain as a representation of the distribution of interest and incorporating this into the proposal distribution (again, this is nontrivial to implement in a theoretically-justifiable manner).

Perfect Simulation. Perfect simulation is a technique which has been developed to combat the difficulties associated with convergence assessment and the use of a collection of variables which are not iid. Under particular (and, unfortunately, somewhat restrictive) circumstances it is possible to obtain a single sample with the correct distribution by simulating a Markov Chain in such a way that its final value is independent of its beginnings. In order to prevent bias associated with the time at which the sample is produced, considerable care is required in the design of such algorithms.

Population-based MCMC / Sequential Monte Carlo. Finally, much research has been done recently in the area of population based methods (going under various names: annealed importance sampling, population Monte Carlo, and sequential Monte Carlo – the last of these proposed frameworks encompasses almost all proposed techniques to date). These approaches use a collection of many samples at each iteration and propagate this population around using importance sampling techniques, combined with elements of MCMC. These techniques show great promise and real applications have been developed in the past few years.

10. Linear Statistical Models

In many areas of mathematical modelling, a surprising amount can be gained by considering simple linear models in which the relationship between quantities of interest are assumed the be linear. Statistics is no exception to this (although, as in many other areas, one should consider whether a linear relationship is reasonable *a priori* and whether it is supported by the data).

10.1 Regression

In a regression we are interested in modelling the conditional distribution of a random variable Y, called the **response variable**, for given values of k other variables $x_1, \ldots, x_k$, called **explanatory variables**.

The explanatory variables may be random variables whose values are observed in an experiment along with the values of Y or they may be **control** variables whose values are set by the experimenter. In either case, when modelling the conditional distribution of Y, we consider the values of $x_1, \ldots, x_k$ to be **fixed**.

10.2 The Linear Model

In a **linear model** we assume that the conditional expectation of Y given $x_1, \ldots, x_k$ is a linear function of some unknown parameters, $\beta_i, i = 1, \ldots, k$, called **regression coefficients**

$$\mathbb{E}[Y|x_1, \ldots, x_k] = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

We suppose that n observations of each variable are available. Then the model for the i^{th} observation is given by

$$\mathbb{E}[Y_i|x_{1i}, x_{2i}, \ldots, x_{ki}] = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}.$$

We also assume that the random variation of Y about the regression line is additive. Thus the model for a set of observed $y_i, i = 1, \ldots, n$ is

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + e_i,$$

which we can write in matrix form

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{21} & \ldots & x_{k1} \\ x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{kn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

or $\qquad \mathbf{y} = \mathbf{X}\beta + \mathbf{e}$

The matrix $\mathbf{X}$ is called the **design matrix**.

The definition of the linear model is completed by the assumption that the vector of errors, $\mathbf{e}$ is an unobservable $n \times 1$ vector of i.i.d. real random variables with mean vector 0 and covariance matrix $\sigma^2 \mathbf{I}_n$.

If we also assume that $\mathbf{e}$ is normal, i.e.

$$\mathbf{e} \sim \mathsf{N}\left(0, \sigma^2 \mathbf{I}_n\right)$$

then the model is called the **normal linear model (NLM)**.

Example 10.1. Consider the relationship between the height H and the weight W of individuals in a certain city. Certainly there's no functional relationship between H and W, but there does seem to be some kind of relation.

We consider them as random variables and postulate that (H, W) has a bivariate normal distribution. Then

$$\mathbb{E}[W|H = h] = \beta_0 + \beta_1 h$$

where β_0 and β_1 are functions of the parameters in a bivariate normal density. Note that β_0, β_1 and h are all constants. We may write

$$W = \beta_0 + \beta_1 h + E$$

where the error E is a normally distributed random variable with mean zero. Thus if we observe the heights and weights of a sample of n people, the model for the weights $w_i, \ldots, w_n$ is given by

$$w_i = \beta_0 + \beta_1 h_i + e_i, \quad i = 1, \ldots, n$$

where the error $e_i \sim \mathsf{N}\left(0, \sigma^2\right)$. This is a **simple linear regression model**: a linear model with one explanatory variable. In matrix form,

$$\mathbf{w} = \mathbf{X}\beta + \mathbf{e}$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} 1 & h_1 \\ 1 & h_2 \\ \vdots & \vdots \\ 1 & h_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

◁

Note that the linear model is **linear in the parameters**, not necessarily linear in the explanatory variables. *Thus the following are examples of linear models:*

The polynomial model:

$$\mathbb{E}[Y_i|x_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k, \quad i = 1, \ldots, n.$$

The response surface model:

$$\mathbb{E}[Y_i|x_{1i}, x_{2i}] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{1i} x_{2i} + \beta_5 x_{2i}^2, \quad i = 1, \ldots, n.$$

Whilst the following is an example of a nonlinear model

The simple exponential model

$$\mathbb{E}[Y_i|x_i] = \alpha e^{\beta x_i}, \quad i = 1, \ldots, n.$$

10.2.1 Selected Extensions NE

You may find yourself in a situation in which you require a slightly more general model than the NLM but would like to retain some of its simplicity and tractability. There are a number of directions in which you could consider extending the model.

Generalised Linear Models (GLMs) consider the case in which a more flexible than linear relationship between response and parameters is considered by the introduction of a link function.

Linear Mixed Models consider a generalisation in another direction. They permit additional subject-specific random effects.

Generalised Linear Mixed Models (GLMMs) unsurprisingly combine both of these extensions.

There are, of course, numerous other possible modelling strategies, but each of these classes has found widespread use in the literature in recent years.

10.3 Maximum Likelihood Estimation for NLMs

In the NLM the assumption that $\mathbf{e} \sim \mathsf{N}\left(0, \sigma^2 \mathbf{I}_n\right)$ implies that $\mathbf{y}$ is multivariate normal with mean vector $\mathbf{X}\beta$ and covariance $\sigma^2 \mathbf{I}_n$:

$$\mathbf{Y}|\mathbf{X} \sim \mathsf{N}\left(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n\right).$$

Thus, the density function for a particular y_i is

$$f(y_i|\mathbf{x}_i = x_{1i}, x_{2i}, \ldots, x_{ki}) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{ -\frac{(y_i - \mathbf{x}_i'\beta)^2}{2\sigma^2} \right\}.$$

The joint density is given by

$$f(y_1, \ldots, y_n|\mathbf{X}) = f(y_1|\mathbf{x}_1) f(y_2|\mathbf{x}_2) \ldots f(y_n|\mathbf{x}_n),$$

so

$$f(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \right\}.$$

Once a sample is drawn for $\mathbf{y}$, the joint density can be expressed as a function of the unknown parameters given the data

$$L(\beta, \sigma^2|\mathbf{X}, \mathbf{y}) \equiv f(\mathbf{y}|\mathbf{X}, \beta, \sigma^2).$$

Thus the log-likelihood is given by

$$l(\beta, \sigma^2|\mathbf{X}, \mathbf{y}) = -\left(\frac{n}{2}\right) \ln 2\pi - \left(\frac{n}{2}\right) \ln \sigma^2 - \frac{(\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta)}{2\sigma^2}$$

where we have expanded

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = (\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta).$$

The ML estimates are obtained by taking partial derivatives w.r.t β and σ^2, setting them equal to zero and solving the resulting set of equations:

We have

$$\begin{aligned} \frac{\partial l(\beta, \sigma^2|\mathbf{X}, \mathbf{y})}{\partial \beta_1} &= \frac{1}{\sigma^2}(\mathbf{x}_1'\mathbf{y} - \mathbf{x}_1'\mathbf{X}\beta) = 0 \\ &\vdots \\ \frac{\partial l(\beta, \sigma^2|\mathbf{X}, \mathbf{y})}{\partial \beta_k} &= \frac{1}{\sigma^2}(\mathbf{x}_k'\mathbf{y} - \mathbf{x}_k'\mathbf{X}\beta) = 0 \end{aligned}$$

which is a set of k equations that can be expressed in matrix form as

$$\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta) = 0.$$

Solving for β yields the ML estimator for the regression coefficients:

$$\hat{\beta}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

(It can be shown that the 2nd order conditions are fulfilled.)

Note that maximising the log-likelihood w.r.t. β is equivalent to minimising the sum of squared errors

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \sum_{i=1}^{n}(y_i - \mathbf{X}_i'\beta)^2$$

and thus $\hat{\beta}_{ML}$ is the same as the estimator yielded by the **Least Squares criterion**, i.e.

$$\hat{\beta} = \hat{\beta}_{ML} = \hat{\beta}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Sampling Distribution of $\hat{\boldsymbol{\beta}}$ The expectation of β is

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}]\\ &= \beta\end{aligned}$$

therefore $\hat{\beta}$ is an unbiased estimator of β. Also

$$\begin{aligned}\mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] &= \mathbb{E}[\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\}\{\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\}]\\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

and thus

$$\hat{\beta} \sim \mathsf{N}\left(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right).$$

The **Gauss Markov Theorem** states that $\hat{\beta}$ is a **best linear unbiased estimator(BLUE)**. That is, out of the class of linear unbiased estimators for β in the linear model, the ML estimator $\hat{\beta}$ is best in the sense of having a minimum sampling variance.

10.3.1 ML estimation for $\boldsymbol{\sigma^2}$

$$\begin{aligned}\frac{\partial l(\beta, \sigma^2|\mathbf{X}, \mathbf{y})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = 0\\ \Rightarrow \hat{\sigma}^2_{ML} &= \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{n} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n}\end{aligned}$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}$ are estimated errors (called **residuals**). It can be shown that

$$\mathbb{E}[\hat{\sigma}^2_{ML}] = \frac{\sigma^2(n-k)}{n}$$

and thus the ML estimator for the variance is biased, although $\hat{\sigma}^2_{ML}$ is asymptotically unbiased. However, we can see that the estimator

$$\hat{\sigma}^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$$

is an unbiased estimator of the variance. It can be shown that the estimators $\hat{\sigma}^2$ and $\hat{\beta}$ are independent.

Exercise 10.3.1. Give the Cramér-Rao lower bound for an unbiased estimator of β and σ^2 and verify whether $\hat{\beta}$ and the unbiased estimator of σ^2 attain it.

Exercise 10.3.2 (Dependence and heteroscedasticity). Suppose that $\mathbf{Y}|\mathbf{X}$ is multivariate normal: $\mathsf{N}\left(\mathbf{X}\beta, \sigma^2 \Phi\right)$ where Φ is a known positive definite matrix. Show that the MLE of β is given by the *generalised least squares estimator*

$$\tilde{\beta} = (\mathbf{X}^{\mathsf{T}}\Phi^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\Phi^{-1}\mathbf{y}.$$

Example 10.2. In the simple linear regression model,

$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}.$$

Hence

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \\ \Rightarrow (\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}. \end{aligned}$$

Also

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

From which we can derive the following

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}$$

where

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\,\bar{y}$$

and

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

the sample variance of the observed $x_i, i = 1, \dots, n$. ◁

Exercise 10.3.3. In an investigation of the properties of a synthetic rubber, varying amounts and types of compounding material were used to prepare a number of specimens. The following table lists the abrasion loss (y) and hardness (x) of thirty different specimens. Given that

$$\begin{aligned} &\sum x_i = 2108 \qquad &&\sum x_i^2 = 152422 \\ &\sum y_i = 5263 \qquad &&\sum y_i^2 = 1148317 \\ &\sum x_i y_i = 346867 && \end{aligned}$$

find the coefficients of the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \dots, 30$$

for these data.

Specimen	x	y	Specimen	x	y
1	45	372	16	68	196
2	55	206	17	75	128
3	61	175	18	83	97
4	66	154	19	88	64
5	71	136	20	59	249
6	71	112	21	71	219
7	81	55	22	80	186
8	86	45	23	82	155
9	53	221	24	89	114
10	60	166	25	51	341
11	64	164	26	59	340
12	68	113	27	65	283
13	79	82	28	74	267
14	81	32	29	81	215
15	56	228	30	86	148

10.4 Confidence Intervals

For a single element of $\hat{\beta}$, say $\hat{\beta}_j$, we have

$$\begin{aligned} \mathbb{E}[\hat{\beta}_j] &= \beta_j \\ \text{and} \quad \mathbb{V}\mathsf{ar}[\hat{\beta}_j] &= \sigma^2 c_{jj} \end{aligned}$$

where $c_{jj} = (\mathbf{X}'\mathbf{X})^{-1}_{jj}$. Thus

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}} \sim \mathsf{N}\left(0,1\right)).$$

However because σ^2 is unknown and estimated by $\hat{\sigma}^2$ we have

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim \mathsf{t}_{(n-k)}.$$

Therefore a $(1-\alpha)$ level confidence interval for β_j is given by

$$\hat{\beta}_j \pm \mathsf{t}_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 c_{jj}}$$

where $\mathsf{t}_{\frac{\alpha}{2},n-k}$ is such that

$$\mathbb{P}\left(\mathsf{t}_{n-k} > \mathsf{t}_{\frac{\alpha}{2},n-k}\right) = \frac{\alpha}{2}.$$

In the case of σ^2 we have that

$$\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k)}.$$

Therefore a $(1-\alpha)$ level confidence interval for σ^2 is given by

$$\left(\frac{(n-k)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-k}}, \frac{(n-k)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-k}}\right).$$

10.5 Hypothesis Tests for the Regression Coefficients

We can use the sampling distribution of β_j to formulate a hypothesis test to determine whether x_j is "important" in explaining the variation in Y. We test

$$\begin{aligned} & H_0 : \beta_j = 0, \\ \text{versus} \quad & H_1 : \beta_j \neq 0, \end{aligned}$$

using the test statistic

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim \mathsf{t}_{n-k}$$

which follows a t_{n-k} distribution under the null hypothesis.

Our decision rule at the α significance level is:

Do not reject H_0 if

$$-\mathsf{t}_{\frac{\alpha}{2},n-k} < T_j < \mathsf{t}_{\frac{\alpha}{2},n-k}$$

otherwise reject H_0 in favour of H_1.

The result of the test needs careful interpretation. If we reject H_0 then β_j is significantly different from zero **in the presence of all the other terms**. Or equivalently x_j contributes significantly to the variation in Y after the contribution of all the other explanatory variables has been taken into account. Given the nature of the interpretation, such a test is known as a **partial t-test**.

Example 10.3. For simple linear regression, we have already seen that

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

Hence

$$\hat{\mathbb{V}\mathsf{ar}}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\hat{\sigma}^2}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

But

$$\mathbb{V}\mathsf{ar}(\hat{\beta}) = \begin{pmatrix} \mathbb{V}\mathsf{ar}(\hat{\beta}_0) & \mathbb{C}\mathsf{ov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathbb{C}\mathsf{ov}(\hat{\beta}_0, \hat{\beta}_1) & \mathbb{V}\mathsf{ar}(\hat{\beta}_1) \end{pmatrix}.$$

So

$$\hat{\mathbb{V}\mathsf{ar}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}}.$$

Therefore our test statistic for the slope in a simple linear regression is

$$T_1 = \frac{\hat{\beta}_1}{\hat{\sigma}/s_x}$$

where s_x is the standard deviation of the observed $x_i, i = 1, \ldots, n$. ◁

Exercise 10.5.1. For the rubber samples data in exercise 10.3.3, test the significance of the slope parameter at the 5% significance level.

10.6 Prediction

The predicted mean response $\hat{y}_0$ at a given set of values for the explanatory variables, $\mathbf{x}_0$ is

$$\hat{y}_0 = \mathbf{x}_0'\hat{\beta}.$$

Now

$$\mathbb{E}[\hat{y}_0] = \mathbf{x}_0'\beta$$

and

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\hat{y}_0) &= \mathbb{V}\mathrm{ar}(\mathbf{x}_0'\hat{\beta}) \\ &= \mathbf{x}_0'\mathbb{V}\mathrm{ar}(\hat{\beta})\mathbf{x}_0 \\ &= \mathbf{x}_0'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\end{aligned}$$

hence

$$\hat{y}_0 \sim \mathsf{N}\left(\mathbf{x}_0'\beta, \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right).$$

Again, we estimate σ^2, by $\hat{\sigma}$, so a $1-\alpha$ confidence interval for $y_0 = \mathbf{x}_0'\beta$ is given by

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}.$$

Since the regression model is based on the observed data, care should be taken when making predictions outside the observed range of the explanatory variables.

Example 10.4. In the case of simple linear regression

$$\begin{aligned}\widehat{\mathbb{V}\mathrm{ar}}(\hat{y}) &= \begin{pmatrix}1\\x_0\end{pmatrix}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\begin{pmatrix}1 & x_0\end{pmatrix} \\ &= \frac{\hat{\sigma}^2}{nS_{xx}}\begin{pmatrix}1\\x_0\end{pmatrix}\begin{pmatrix}\sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n\end{pmatrix}\begin{pmatrix}1 & x_0\end{pmatrix} \\ &= \frac{\hat{\sigma}^2}{nS_{xx}}\left(\sum_{i=1}^n x_i^2 - x_0\sum_{i=1}^n x_i - x_0\sum_{i=1}^n x_i + nx_0^2\right) \\ &= \frac{\hat{\sigma}^2}{S_{xx}}\left(\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 + x_0^2 - 2x_0\frac{\sum_{i=1}^n x_i}{n} + \bar{x}^2\right) \\ &= \frac{\hat{\sigma}^2}{S_{xx}}\left(\frac{S_{xx}}{n} + (x_0 - \bar{x}^2)\right).\end{aligned}$$

Thus we can estimate the variance of $\hat{y}$ by

$$\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_x^2}\right).$$

◁

Exercise 10.6.1. For the rubber samples data in exercise 10.3.3, use the simple linear regression model to predict the abrasion loss for a specimen of rubber with a hardness measurement of 70. Give a 95% confidence interval for this prediction.

10.7 Test for Significance of a Regression

We can test whether the regression is significantly better than an intercept-only model, by testing

$$\begin{aligned} & H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0, \\ \text{versus} \quad & H_1 : \beta_j \neq 0 \quad \text{for at least one } j. \end{aligned}$$

Our test statistic is based on

$$SS_R = (\mathbf{X}\hat{\beta} - \bar{\mathbf{y}})'(\mathbf{X}\hat{\beta} - \bar{\mathbf{y}})$$

which is the regression sum of squares. This statistic compares the fitted values under the regression model, $\mathbf{X}\hat{\beta}$, with the fitted values under the intercept-only model, $\bar{\mathbf{y}}$. It can be shown that if H_0 is true

$$SS_R \sim \sigma^2 \chi^2_{k-1}$$

Since we don't know σ^2 we need to adjust this statistic in some way to obtain a test statistic with a known distribution. We have seen that

$$\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k)}$$

and hence

$$(n-k)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{(n-k)}.$$

Now

$$(n-k)\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = SS_E,$$

the sum of squared errors. If H_0 is true

$$F = \frac{SS_R/k-1}{SS_E/n-k} \sim \mathsf{F}_{k-1,n-k}.$$

Note that when H_0 is true,

$$\mathbb{E}\left[\frac{SS_R}{k-1}\right] = \frac{1}{k-1}\mathbb{E}[SS_R] = \sigma^2.$$

However when H_1 is true,

$$\mathbb{E}\left[\frac{SS_R}{k-1}\right] > \sigma^2.$$

Hence it is appropriate to use a one sided test here. Our decision rule is then:
Reject H_0 if

$$F > F_{\frac{\alpha}{2},k-1,n-k}$$

otherwise we cannot reject H_0 at significance level α.

10.7.1 Analysis of Variance for a Regression

Analysis of variance is a general method that partitions the total sum of squares

$$SS_T = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}),$$

(which is proportional to the variance of the observed $y_i, i = 1, \ldots, n$) into components explained by a statistical model. These components and related statistics are presented in a **analysis of variance (ANOVA) table**.

The ANOVA table for a regression is as follows

Source	SS	df	MS	F
Regression	SS_R	$k-1$	MS_R	MS_R/MS_E
Residual	SS_E	$n-k$	MS_E	
Total	SS_T	$n-1$		

Where MS stands for **mean square**, i.e. the sum of squares SS divided by the degrees of freedom df. Thus MS_R/MS_E is the F statistic defined in the previous section. Also $MS_E = \hat{\sigma}^2$.

Example 10.5. In the case of simple linear regression, there is only one explanatory variable, so the F test is equivalent to the partial t-test on the slope. There are $k = 2$ parameters, so the numerator of the F statistic, becomes

$$SS_R = (\mathbf{X}\hat{\beta} - \bar{\mathbf{y}})'(\mathbf{X}\hat{\beta} - \bar{\mathbf{y}}).$$

By substituting

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}$$

we can show that

$$SS_R = \frac{S_{xy}^2}{S_{xx}}.$$

As for all models, the denominator of the F statistic is $\hat{\sigma}^2$, so we have

$$F = \frac{S_{xy}^2/S_{xx}}{\hat{\sigma}^2} = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} = T_1^2.$$

Thus the F statistic for a simple linear regression model is the square of the t statistic for the slope parameter. A one-sided test based on the F statistic will give the same p-value (exact significance level) as a two-sided test based on the t statistic. ◁

10.8 The Coefficient of Multiple Determination ($\mathbf{R^2}$)

The **coefficient of (multiple) determination ($\mathbf{R}^2$)** is

$$R^2 = 1 - \frac{SS_E}{SS_T}, \quad 0 \leq R^2 \leq 1.$$

It can be interpreted as the proportion of variation in the response variable explained by the regression. Some software packages provide the **adjusted $\mathbf{R}^2$**:

$$R_{adj}^2 = 1 - \frac{SS_E/n-k}{SS_T/n-1}.$$

This adjusts for the number of fitted parameters, to allow comparison over models with different numbers of parameters.

10.9 Residual Analysis

Residual analysis allows us to

1. check the assumptions of the model,
2. check the adequacy of the model,
3. detect outliers.

The residuals are given by

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}.$$

Before examining the residuals, it is common to standardise them as follows

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}_{e_i}}$$

where $\hat{\sigma}_{e_i}$ is the estimated standard error for $\hat{e}_i$. Then under the assumptions of the NLM, the r_i should be **i.i.d. N(0, 1)** random variables. We can check this using graphical methods.

Assumption of Normality

The normality assumption can be checked using a q-q plot. Plot $r_{(i)}$, the ith rank order residuals against

$$\Phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right)$$

the "expected normal deviate". If the r_i are normally distributed, the points should fall on a straight line at a 45 degree angle.

Assumption of Constant Variance

Plot r_i versus $\hat{y}_i$ and look for systematic changes in spread about $r = 0$.

Model Adequacy

We need check if there are any systematic trends in the residuals.
Plot r_i versus explanatory variables in the model.
Plot r_i versus explanatory variables **not** in the model.
Plot r_i versus $\hat{y}_i$.
Plot r_i versus time (if applicable).

Outliers

If the residuals have been standardised we expect the majority to fall in the range (-2, 2). We can check the plot of r_i vs. y_i for any observations with "large" r_i. Such observations should be checked.

Residual analysis is a general technique which can be applied to most statistical models. We should always consider whether the residuals are consistent with the assumed model for random variation: if they are not then this must cast some doubt upon the model being used and any conclusions which are drawn from it.

A. Solutions to Exercises

1.3.1 One option is to identify the number of subsets of each possible size. A set of size M can have subsets of size $0, 1, \dots, M$. Let N_i denote the number of subsets of size i and N denote the total number of subsets.

Using the definition of the binomial coefficients, $N_i = \binom{M}{i}$. And, of course, $N = \sum_{i=1}^{M} N_i$:

$$N = \sum_{i=1}^{N} \binom{M}{i} = \sum_{i=1}^{N} \binom{M}{i} 1^i 1^{M-i} = (1+1)^M = 2^M$$

Alternatively, note that we can represent any subset of M as a binary string of length M. Element m of this string takes a value of 1 if element m of M is a member of the subset and zero otherwise. It's clear that there is a one-to-one correspondence between such strings and the subsets of M and there are 2^M different strings.

1.5.1 Probability space is $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathcal{A} = 2^{\Omega}$ and $\mathbb{P}(A) = |A|/4$ (whenever a probability is uniform it is proportional to the number of elements in any set for which it is evaluated).

(i) $A = \{(H, H)\}$ and $B = \{(H, H), (H, T)\}$. As $A \subset B$, $\mathbb{P}(A \cap B) = \mathbb{P}(A)$. $\mathbb{P}(A) = 1/4$ and $\mathbb{P}(B) = 2/4$.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

(ii) $C = \{(H, H), (H, T), (T, H)\}$ and $\mathbb{P}(C) = 3/4$. Hence:

$$\mathbb{P}(A|C) = \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)} = \frac{\mathbb{P}(A)}{\mathbb{P}(C)} = \frac{1/4}{3/4} = 1/3.$$

1.5.2 Let U_i denote the event that urn i is selected. Let D denote the event a defective ball is selected. We assume that $\mathbb{P}(U_i) = 1/5$ for $i = 1, \dots, 5$. We are told that $\mathbb{P}(D|U_i) = i/10$.

(i) The probability of selecting a defective ball can be found by summing over the possibility of selecting a defective ball and a particular urn (using the partition formula):

$$\begin{aligned}\mathbb{P}(D) = \sum_{i=1}^{5} \mathbb{P}(D \cap U_i) =& \sum_{i=1}^{5} \mathbb{P}(D|U_i)\mathbb{P}(U_i) \\ =& \sum_{i=1}^{5} \frac{1}{5}\mathbb{P}(D|U_i) = \frac{1}{5}\sum_{i=1}^{5} \frac{i}{10} \\ =& \frac{1}{5}\frac{15}{10} = 15/50 = 3/10.\end{aligned}$$

Notice that this coincides with the proportion of balls which are defective as you would expect (there are equal numbers of balls in each urn and urns are selected uniformly at random and so every ball has the same probability of selection).

(ii) This is a simple application of Bayes formula:

$$\begin{aligned}\mathbb{P}(U_i|D) =&\frac{\mathbb{P}(U_i \cap D)}{\mathbb{P}(D)}\\ =&\frac{\mathbb{P}(D|U_i)\mathbb{P}(U_i)}{\mathbb{P}(D)}\\ =&\frac{i/10 \times 1/5}{15/50}\\ =&i/15.\end{aligned}$$

In this case, the probability of having selected an urn is simply the proportion of the defective balls which it contains.

1.5.3 $A =$ “odd total”, $B =$ “6 on 1st die” and $C =$ “total is 7”.

(i) A and B are independent. $\mathbb{P}(A) = \mathbb{P}(\text{Total in } \{3, 5, 7, 9, 11\}) = (2 + 4 + 6 + 4 + 2)/36 = 1/2$, and $\mathbb{P}(B) = 1/6$. Whilst $\mathbb{P}(A \cap B) = \mathbb{P}(\{(6, 1), (6, 3), (6, 5)\}) = 3/36 = \mathbb{P}(A)\mathbb{P}(B)$.

(ii) $C = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ and $\mathbb{P}(C) = 1/6$. $\mathbb{P}(A \cap C) = \mathbb{P}(C)$. Hence $\mathbb{P}(A \cap C) \neq \mathbb{P}(A)\mathbb{P}(C)$: A and C are not independent. This isn't surprising: $A \supset C$.

(iii) $B \cap C = \{(6, 1)\}$. Thus $\mathbb{P}(B \cap C) = 1/36 = 1/6 \times 1/6 = \mathbb{P}(B) \times \mathbb{P}(C)$. So, B and C are independent.

1.5.4 Establish that $\mathbb{P}(A_1) = \frac{1}{2}$, $\mathbb{P}(A_2) = \frac{1}{2}$ and $\mathbb{P}(A_3) = \frac{1}{2}$.

Additionally, $\mathbb{P}(A_1|A_2) = \frac{1}{2}$ and $\mathbb{P}(A_2|A_1) = \frac{1}{2}$; $\mathbb{P}(A_3|A_1) = \frac{1}{2}$ as this is the probability of an even face on the second die, and similarly $\mathbb{P}(A_3|A_2) = \frac{1}{2}$.

The joint probabilities are $\mathbb{P}(A_1 \cap A_2) = \frac{1}{4}$, $\mathbb{P}(A_1 \cap A_3) = \frac{1}{4}$ and $\mathbb{P}(A_2 \cap A_3) = \frac{1}{4}$ by enumeration.

Hence $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) \times \mathbb{P}(A_2)$, $\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1) \times \mathbb{P}(A_3)$ and $\mathbb{P}(A_2 \cap A_3) = \mathbb{P}(A_2) \times \mathbb{P}(A_3)$ and these events are pairwise independent.

However, $\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0 \neq \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$ as no pair of odd numbers has an odd sum. The events are not independent and pairwise independent is a weaker concept than independence.

2.1.1 Consider $Y(i, j) = |i - j|$:

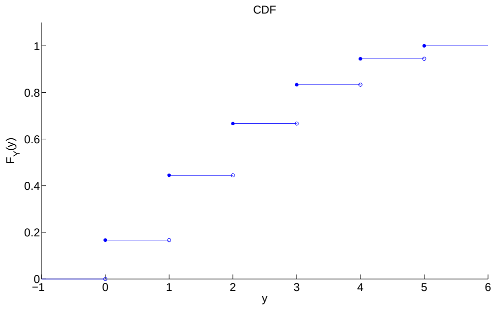
i/j	1	2	3	4	5	6
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

Hence, as $\mathbb{P}(\{(i, j)\}) = \frac{1}{36}$ for any valid pair i, j, the probability mass function and associated distribution function is:

y	0	1	2	3	4	5
$f_Y(y)$	6/36	10/36	8/36	6/36	4/36	2/36
$F_Y(y)$	6/36	16/36	24/36	30/36	35/36	36/36

where $f_Y(y) = \mathbb{P}(Y = y) = |\{(i, j) : |i - j| = y\}|/36$ and $F_Y(y) = \sum_{y_j \leq y} f_Y(y_j)$.

The CDF is:



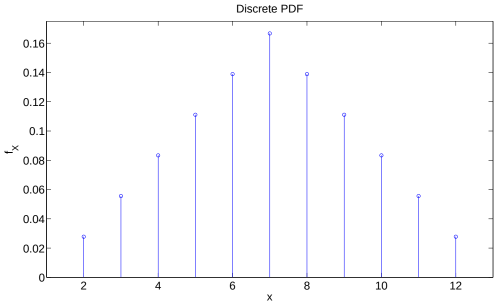
2.2.1 Again, enumerating the possible (equally probable) outcomes, $X(i, j) = i + j$:

i/j	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The probability density for X is proportional to the number of ways in which X can take each possible value (and normalised such that it sums to one):

x	2	3	4	5	6	7	8	9	10	11	12
$f_X(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

And this may be represented graphically as:



$f_Y(y)$ for $Y(i, j) = |i - j|$ was given in the answer to 2.1.1.

2.3.1 With the same X and Y as in the previous questions:

$$\begin{aligned}
\mathbb{E}[X] =& \sum_{x_j} x_j \cdot f_X(x_j) \\
=& \sum_{x=2}^{12} f_x(x) \cdot x \\
=& \frac{1}{36} [1 \cdot (2 + 12) + 2 \cdot (3 + 11) + 3 \cdot (4 + 10) + 4 \cdot (5 + 9) + 5 \cdot (6 + 8) + 6 \cdot 7] \\
=& \frac{1}{36} (15 \cdot 14 + 3 \cdot 14) = 7
\end{aligned}$$

And:

$$\begin{aligned}\mathbb{E}[Y] =& \sum_{y_j} y_j \cdot f_Y(y_j)\\ =&\sum_{y=0}^{5} f_Y(y)\cdot y = \sum_{y=1}^{5} f_Y(y)\cdot y\\ =&\frac{1}{36}\left[10\cdot 1 + 8\cdot 2 + 6\cdot 3 + 4\cdot 4 + 2\cdot 5\right]\\ =&70/36 = 35/18\end{aligned}$$

2.3.2 Calculating the expectation explicitly:

$$\begin{aligned}E(X) &= \int_{-\infty}^{\infty} u f_X(u)\mathrm{d}u\\ &= \int_{0}^{\infty} u\lambda e^{-\lambda u}\mathrm{d}u\end{aligned}$$

Using *integration by parts*, we obtain:

$$\begin{aligned}&= \left[u\cdot\lambda\frac{e^{-\lambda u}}{-\lambda}\right]_0^\infty - \int_0^\infty \frac{\lambda e^{-\lambda u}}{-\lambda}\mathrm{d}u\\ &= \int_0^\infty e^{-\lambda u}\mathrm{d}u\\ &= \left[\frac{e^{-\lambda u}}{-\lambda}\right]_0^\infty = 1/\lambda\end{aligned}$$

The distribution function can be calculated directly from the definition of the density function:

$$\begin{aligned}F_x(x) &= \int_{-\infty}^{x} f_x(u)\mathrm{d}u\\ &= \int_0^x \lambda e^{-\lambda u}\mathrm{d}u\\ &= \left[\frac{\lambda e^{-\lambda u}}{-\lambda}\right]_0^x = 1 - e^{-\lambda x}\end{aligned}$$

2.6.1 The moment generating function is:

$$\begin{aligned}m_X(t) =& \mathbb{E}\left[\exp(Xt)\right]\\ =&\sum_{x=0}^{n} \mathbb{P}(X=x)\cdot\exp(xt)\\ =&\sum_{x=0}^{n}\binom{n}{x}p^x(1-p)^{n-x}(e^t)^x\\ =&\sum_{x=0}^{n}\binom{n}{x}(p\cdot e^t)^x(1-p)^{n-x}\\ =&(pe^t+(1-p))^n\end{aligned}$$

where the final line follows from the binomial theorem. Hence $m_X(t) = (p(e^t-1)+1)^n$.

To calculate the mean (i.e. the expectation) of X, we need to calculate the first derivative of the MGF. By the chain rule:

$$\begin{aligned}\frac{\partial m_X}{\partial t} =& n(pe^t+(1-p))^{n-1}\frac{\partial}{\partial t}(pe^t+(1-p))\\ =& n(pe^t+(1-p))^{n-1}pe^t\end{aligned}$$

And evaluating this derivative at $t = 0$, we obtain $n(pe^0 + 1 - p)^{n-1}pe^0 = n \cdot 1 \cdot p \cdot 1 = np$. Using the product rule, we can calculate the second derivative:

$$\begin{aligned}
\frac{\partial^2 m_X}{\partial t^2} =& \frac{\partial}{\partial t} n(pe^t + (1-p))^{n-1}pe^t \\
=& n(pe^t + (1-p))^{n-1}\frac{\partial}{\partial t}pe^t + pe^t\frac{\partial}{\partial t}n(pe^t + (1-p))^{n-1} \\
=& n(pe^t + (1-p))^{n-1}pe^t + pe^t n(n-1)(pe^t + (1-p))^{n-2}\frac{\partial}{\partial t}(pe^t + (1-p)) \\
=& n(pe^t + (1-p))^{n-1}pe^t + pe^t n(n-1)(pe^t + (1-p))^{n-2}pe^t \\
=& n(pe^t + (1-p))^{n-1}pe^t + p^2e^{2t}n(n-1)(pe^t + (1-p))^{n-2}
\end{aligned}$$

And, evaluating this at $t = 0$:

$$\begin{aligned}
\mathbb{E}[X^2] =& n(pe^0 + (1-p))^{n-1}pe^0 + p^2e^0n(n-1)(pe^0 + (1-p))^{n-2} \\
=& n(p + (1-p))^{n-1}p + p^2n(n-1)(p + (1-p))^{n-2} \\
=& np + n(n-1)p^2
\end{aligned}$$

From the usual expression for variance:

$$\begin{aligned}
\mathbb{V}\mathsf{ar}[X] =& \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
=& np + n(n-1)p^2 - n^2p^2 \\
=& np + n(n-1-n)p^2 \\
=& np - np^2 \\
=& np(1-p).
\end{aligned}$$

3.1.1 For the Poisson distribution, the moment generating function is:

$$\begin{aligned}
m(t) =& \mathbb{E}[\exp(Xt)] \\
=& \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \times e^{xt} \\
=& \sum_{x=0}^{\infty} \frac{e^{-\lambda}(e^t\lambda)^x}{x!} \\
=& e^{-\lambda}\sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!}
\end{aligned}$$

Recall that the exponential function has series expansion $e^u = \sum_{i=0}^{\infty} \frac{u^i}{i!}$, and so:

$$m(t) = e^{-\lambda}\exp(e^t\lambda) = \exp([e^t - 1]\lambda)$$

Note that this is the exponential of an exponential.

To obtain the expectation we require the first derivative of m:

$$\begin{aligned}
\frac{\partial m}{\partial t} =& \exp([e^t - 1]\lambda)\frac{\partial m}{\partial t}([e^t - 1]\lambda) \\
=& \exp([e^t - 1]\lambda)e^t\lambda \\
=& \lambda\exp([e^t - 1]\lambda + t)
\end{aligned}$$

Evaluating this at $t = 0$ yields the first moment:

$$\mathbb{E}[X] = \lambda \exp([e^0 - 1]\lambda + 0) = \lambda$$

We require the second derivative in order to evaluate the second moment and hence the variance. Via the chain rule:

$$\begin{aligned}\frac{\partial^2 m}{\partial t^2} =& \frac{\partial}{\partial t}\lambda \exp([e^t - 1]\lambda + t)\\ =&\lambda \exp([e^t - 1]\lambda + t)\frac{\partial}{\partial t}\left([e^t - 1]\lambda + t\right)\\ =&\lambda \exp([e^t - 1]\lambda + t)\left(\lambda e^t + 1\right)\end{aligned}$$

The second moment is obtained by evaluating this expression at t=0:

$$\begin{aligned}\mathbb{E}[X^2] =&\lambda \exp([e^0 - 1]\lambda + 0)\left(\lambda e^0 + 1\right)\\ =&\lambda(1)(\lambda + 1) = \lambda(\lambda + 1)\end{aligned}$$

And the variance is then:

$$\mathbb{V}\mathsf{ar}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda.$$

3.1.2 If $X \sim \mathsf{Poi}\,(3)$:

–

$$\begin{aligned}\mathbb{P}(X \leq 3) =&\sum_{x=0}^{3}\mathbb{P}(X = x)\\ =&e^{-3}\sum_{x=0}^{3}3^x/x!\\ =&e^{-3}[1 + 3 + 9/2 + 27/6] = 0.64723\end{aligned}$$

–

$$\begin{aligned}\mathbb{P}(X > 1) =&1 - \mathbb{P}(X \not> 1) = 1 - \mathbb{P}(X \leq 1)\\ =&1 - [\mathbb{P}(X = 0) + \mathbb{P}(X = 1)]\\ =&1 - e^{-3}[1 + 3] = 0.80085\end{aligned}$$

–

$$\begin{aligned}\mathbb{P}(2 \leq X \leq 4) =&\sum_{x=2}^{4}\mathbb{P}(X = x)\\ =&e^{-3}[3^2/2! + 3^3/3! + 3^4/4!]\\ =&e^{-3}[9/2 + 27/6 + 81/24] = 0.61611\end{aligned}$$

–

$$\begin{aligned}F(4.2) =&\mathbb{P}(X \leq 4.2) = \mathbb{P}(X \leq 4) \text{ as X takes only integer values}\\ =&\mathbb{P}(X \leq 1) + \mathbb{P}(2 \leq X \leq 4)\\ =&0.19915 + 0.61611 = 0.81526\end{aligned}$$

– It would seem likely that the density of larvae would vary over the plate – for example, it is likely that regions close to the edge of the colony would have lower density.

3.1.3 The moment generating function of a negative binomial random variable can be obtained directly with a degree of lateral thinking:

$$\begin{aligned}
m(t) =&\mathbb{E}[\exp(Xt)] \\
=&\sum_{x=0}^{\infty} \binom{x+r-1}{r-1} p^r (1-p)^x e^{xt} \\
=&p^r \sum_{x=0}^{\infty} \binom{x+r-1}{r-1} [e^t(1-p)]^x \\
=&p^r \sum_{x=0}^{\infty} \binom{x+r-1}{x} [e^t(1-p)]^x \\
=&\frac{p^r}{(1-e^t(1-p))^r}
\end{aligned}$$

where the final line follows by considering a power series expansion of $g(z) = (1-z)^{-s}$. $g^{(k)}(z) = \frac{\prod_{i=0}^{k-1}(s+i)}{(1-z)^{-(s+k)}}$. And, by Taylor's theorem, expanding about $z^\star = 0$:

$$\begin{aligned}
(1-z)^{-s} =&\sum_{i=1}^{\infty} \frac{1}{i!} g^{(i)}(z^\star)(z-z^\star)^i \\
=&\sum_{i=1}^{\infty} \binom{s+i-1}{i} (1-z^\star)^{-(s+i)} (z-z^\star)^i \\
=&\sum_{i=1}^{\infty} \binom{i+s-1}{i} z^i
\end{aligned}$$

where $z = e^t(1-p)$ in the particular case considered here.

However, it's simpler to represent the negative binomial distribution as that of the sum of r independent geometric random variables (it is possible to prove that the density of such a sum matches the density of the negative binomial distribution by summing over all combinations of values that lead to a particular sum and iteratively applying Vandermonde's identity). In this case, if $X = \sum_{i=1}^{r} X_i$, with the X_i iid geometric random variables with parameter p:

$$\begin{aligned}
m_{X_i}(t) =&\mathbb{E}[\exp(tX_i)] \\
=&\sum_{x_i=0}^{\infty} p(1-p)^{x_i} e^{x_i t} \\
=&p \sum_{x_i=0}^{\infty} [e^t(1-p)]^{x_i} \\
=&\frac{p}{1-e^t(1-p)}
\end{aligned}$$

where the last line follows by recognising a geometric series.

Then, as the X_i are iid:

$$\begin{aligned}
m_X(t) =&\prod_{i=1}^{r} m_{X_i}(t) = m_{X_1}(t)^r \\
=&\left[\frac{p}{(1-e^t(1-p))}\right]^r
\end{aligned}$$

In order to calculate the derivatives of m_X with respect to t, it's simplest to rearrange it as $m_X(t) = [p/(1-p)]^r[1-e^t(1-p)]^{-r}$, such that (making use of the chain and product rules):

$$\begin{aligned}
m'_X(t) = \frac{\partial m_x(t)}{\partial t} =& p^r(-r)[1-e^t(1-p)]^{-(r+1)}(-(1-p)e^t) \\
=& r(1-p)p^r e^t[1-e^t(1-p)]^{-(r+1)} \\
m''_X(t) = \frac{\partial m_x^2(t)}{\partial t^2} =& r(1-p)p^r \Big[e^t \times -(r+1)(1-e^t(1-p))^{-(r+2)} \times (-(1-p)e^t) + \\
& e^t[1-e^t(1-p)]^{-(r+1)}\Big] \\
=& r(1-p)p^r e^t \left[(r+1)(1-p)e^t(1-e^t(1-p))^{-(r+2)} + (1-e^t(1-p))^{(-r+1)}\right]
\end{aligned}$$

Thus, we obtain:

$$\begin{aligned}
\mathbb{E}[X] =& m'_X(0) = r(1-p)p^r p^{-(r+1)} \\
=& r(1-p)/p \\
\mathbb{E}[X^2] =& m''_X(0) = r(1-p)p^r \left[(r+1)(1-p)p^{-(r+2)} + p^{-(r+1)}\right] \\
=& r(1-p)p^r \left[(r+1-pr-p)p^{-(r+2)} + pp^{-(r+2)}\right] \\
=& r(1-p)/p^2[r+1-pr] \\
\mathbb{V}\mathsf{ar}[X] =& \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
=& \frac{r(1-p)}{p^2}(r+1-pr) - \frac{r(1-p)}{p^2}(r(1-p)) \\
=& \frac{r(1-p)}{p^2}
\end{aligned}$$

3.2.1 To calculate the fourth moment of a standard Normal random variable, it's convenient to use the moment generating function.

If $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$, then

$$\begin{aligned}
m_X(t) =& \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) e^{tx} \mathrm{d}x \\
=& \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2-2\mu x+\mu^2}{2\sigma^2} + xt\right) dx \\
=& \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2-2(\mu+\sigma^2 t)x+\mu^2}{2\sigma^2}\right) dx \\
=& \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-(\mu+\sigma^2 t))^2-(\mu+\sigma^2 t)^2+\mu^2}{2\sigma^2}\right) dx \\
=& \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}\right) \exp\left(\frac{+(\mu+\sigma^2 t)^2-\mu^2}{2\sigma^2}\right) dx \\
=& \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}\right) \mathrm{d}x \exp\left(\frac{2\mu\sigma^2 t+\sigma^4 t^2}{2\sigma^2}\right) \\
=& \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)
\end{aligned}$$

Where the integral is equal to 1 as it is the integral of the density of a $\mathsf{N}\left(\mu+\sigma^2 t, \sigma^2\right)$ random variable over it's entire codomain.

We're interested in the particular case of $Z \sim \mathsf{N}(0,1)$, so $m_Z(t) = e^{\frac{1}{2}t^2}$. Consider it's first 4 derivatives with respect to t:

$$\begin{aligned}
m'_Z(t) =& \frac{1}{2} 2t \exp\left(\frac{1}{2}t^2\right) \\
=& t \exp\left(\frac{1}{2}t^2\right) \\
m''_Z(t) =& \exp\left(\frac{1}{2}t^2\right) + t^2 \exp\left(\frac{1}{2}t^2\right) \\
=& \exp\left(\frac{1}{2}t^2\right) + t^2 \exp\left(\frac{1}{2}t^2\right) \\
=& \left(t^2+1\right) \exp\left(\frac{1}{2}t^2\right) \\
m^{(3)}_Z(t) =& 2t \exp\left(\frac{1}{2}t^2\right) + \left(t^2+1\right) t \exp\left(\frac{1}{2}t^2\right) \\
=& [t^3+3t] \exp\left(\frac{1}{2}t^2\right) \\
m^{(4)}_Z(t) =& [3t^2+3] \exp\left(\frac{1}{2}t^2\right) + [t^4+3t^2] \exp\left(\frac{1}{2}t^2\right) \\
=& [t^4+6t^2+3] \exp\left(\frac{1}{2}t^2\right)
\end{aligned}$$

Evaluating these at zero, the first 4 non-central moments are clearly $0, 1, 0, 3$ and hence the Kurtosis of the standard Gaussian is 3 (as the first moment is zero and so central and non-central moments coincide).

3.3.1 Let X_i be a random variable which takes value 0 if the i^{th} chocolate bar is a Yorkie and 1 if it is a Mars bar.

(a) $\mathbb{P}(\text{Mars, Mars, Yorkie, Yorkie}) = \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0)$. We can expand this as:

$$\begin{aligned}
&\mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1|X_1 = 1)\mathbb{P}(X_3 = 0|X_1 = 1, X_2 = 1)\mathbb{P}(X_4 = 0|X_3 = 0, X_2 = X_1 = 1) \\
=&\mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1|X_1 = 1)\mathbb{P}(X_3 = 0|X_1 = 1)\mathbb{P}(X_4 = 0|X_3 = 0)
\end{aligned}$$

as the choice of chocolate bar depends only upon the previously chosen bar. In this case $\mathbb{P}(X_1 = 1) = \frac{1}{2}$ and $\mathbb{P}(X_i = X_{i-1}) = \frac{1}{3}$ and so we arrive at $\frac{1}{2} \times \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} = 2/54 = 1/27$.

(b) If we're only interested in the probability that the *second* purchase is a Yorkie, then we're interested in the marginal probability that $X_2 = 0$.

$$\begin{aligned}
\mathbb{P}(X_2 = 0) =& \sum_{x_1} \mathbb{P}(X_1 = x_1, X_2 = 0) \\
=& \sum_{x_1} \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = 0|X_1 = 1) \\
=& \mathbb{P}(X_1 = 0)\mathbb{P}(X_2 = 0|X_1 = 0) + \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 0|X_1 = x_1) \\
=& \frac{1}{4}\frac{1}{3} + \frac{3}{4}\frac{2}{3} \\
=& \frac{1}{12} + \frac{1}{2} = 7/12.
\end{aligned}$$

3.3.2 Again, we can simply decompose the probability of a faulty yoghurt as the combination of a faulty yoghurt and a particular production line and then sum over the production lines. Let A denote a faulty yoghurt and B denote production line 1, then as there are only two yoghurt-states (faulty and not faulty) and two production lines:

$$\begin{aligned}\mathbb{P}(A) =&\mathbb{P}(A \cap B \cup A \cap \bar{B})\\ =&\mathbb{P}(A \cap B) + \mathbb{P}(A \cap \bar{B}) \text{ by disjointness}\\ =&\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|\bar{B})\mathbb{P}(\bar{B})\\ =&0.01 \times 0.55 + 0.025 \times 0.45 = 0.01675 = 1.675\%.\end{aligned}$$

Given a faulty yoghurt, the probability that it came from a particular production line can be found by Bayes' rule:

$$\begin{aligned}\mathbb{P}(\bar{B}|A) =&\mathbb{P}(A|\bar{B})\mathbb{P}(\bar{B})/\mathbb{P}(A)\\ =&0.025 \times 0.45/0.01675\\ =&0.67 = 67\%.\end{aligned}$$

(a) **3.3.3** (a) There are 500 identical, independent Bernoulli-type trials and we are interested in the distribution of the number of successes. Thus, this is a binomial distribution (with $n = 500, p = 0.02$ given the specified parameters). Consequently, $\mathbb{E}[X] = np = 10$ and $\mathbb{V}\mathsf{ar}[X] = np(1-p) = 9.8$.

(b) To approximate a binomial distribution with a Poisson, it is necessary that n is large and $p \ll 1$ such that $np \approx np(1-p)$; the moment-matched Poisson distribution has $\lambda = np = 10$.
In order to find $\mathbb{P}(5 \leq X \leq 15) = F(15) - F(4)$ we note that the distribution function of a Poisson random variable can only be expressed in closed form in terms of the incomplete gamma distribution and it's rather easier to use the density function directly:

$$\begin{aligned}\mathbb{P}(5 \leq X \leq 15) =&e^{-10} \sum_{x=5}^{15} \frac{10^x}{x!}\\ =&0.92201.\end{aligned}$$

(c) Chebyshev's inequality can be employed with less calculation. It tells us that:

$$\mathbb{P}(|X - \mathbb{E}[X]| > r \times \sqrt{\mathbb{V}\mathsf{ar}(X)}) \leq 1/r^2$$

In our case, were interested in a deviation of at least 5 and so, $r = 5/\sqrt{9.8} = 1.5972$ and the bound on the probability of interest is $1 - 1.5972^{-2} = 0.608$. In this case the bound is rather loose.

3.3.4 In this case $X \sim N(10, 4^2)$ if X is the rate of return as a percentage.

(a) To calculate $\mathbb{P}(X \leq 0)$, it's convenient to note that $Z = (X - 10)/4$ is a standard normal (i.e. $N(0, 1^2)$) random variable and $\{\omega : X(\omega) \leq 0\} = \{\omega : Z(\omega) \leq -2.5\}$. Hence, $\mathbb{P}(X \leq 0) = \mathbb{P}(Z \leq -2.5) = \Phi(-2.5)$ which can be found tabulated, or via software, to be approximately 0.006207.

(b) Similarly, $\mathbb{P}(X \geq 15) = \mathbb{P}(Z \geq 1.25) = 1 - \Phi(1.25)$ which is approximately 0.10565.

(c) Let Y denote the performance of this second portfolio. $Y \sim N(12, 5^2)$. In this case:

$$\begin{aligned}\mathbb{P}(Y \leq 0) =&\mathbb{P}(Z \leq -12/5 = 2.4) = \Phi(2.4)\\ \approx&0.0082\\ \mathbb{P}(Y \geq 15) =&\mathbb{P}(Z \geq 3/5 = 0.6) = 1 - \Phi(0.6)\\ \approx&0.27425\end{aligned}$$

Whether this change is desirable is a rather personal decision. The average performance is significantly better; the variability is a little higher, but not dramatically so. Most people would probably prefer the second fund to the first. However, as we can't assume that the two funds are independent it is difficult to assess the likely difference in performance.

4.1.1 3 coins, which we assume are fair, are tossed. The underlying probability space consists of all ordered triples of the elements H and T: $\Omega = \{(c_1, c_2, c_3) : c_1, c_2, c_3 \in \{H, T\}\}$.

Let $I_H(c) = 1$ if $c = H$ and 0 otherwise.

$X_1((c_1, c_2, c_3)) = I_H(c_1)$ and $X_2((c_1, c_2, c_3)) = I_H(c_2) + I_H(c_3)$.

Consider the joint density function, $f_{X_1,X_2}(x_1, x_2)$. Clearly, $x_1 \in \{0, 1\}$ and $x_2 \in \{0, 1, 2\}$. The probability density can be established by enumerating the number of equally probable elementary events which correspond to each outcome.

$X_1 \vert X_2$	0	1	2
0	1/8	2/8	1/8
1	1/8	2/8	1/8

Evaluating various quantities follows by considering how these quantities can be expressed in terms of the values which the random variables can take.

$$\begin{aligned}
F(0.4, 1.3) =& \mathbb{P}(X_1 \leq 0.4, X_2 \leq 1.3) \\
=& \mathbb{P}(X_1 = 0, X_2 \leq 1) \\
=& \mathbb{P}(X_1 = 0, X_2 = 0) + \mathbb{P}(X_1 = 0, X_2 = 1) = 3/8 \\
F(0, 0) =& \mathbb{P}(X_1 \leq 0, X_2 \leq 0) \\
=& \mathbb{P}(X_1 = 0, X_2 = 0) = 1/8 \\
F(1.4, 2.1) =& \mathbb{P}(X_1 \leq 1.4, X_2 \leq 2.1) \\
=& \mathbb{P}(X_1 \leq 1, X_2 \leq 2) = 1 \\
F(-1, 2) =& \mathbb{P}(X_1 \leq -1, X_2 \leq 2) = 0 \\
\mathbb{P}(X_1 = 1, X_2 \geq 1) =& \mathbb{P}(X_1 = 1, X_2 = 1) + \mathbb{P}(X_1 = 1, X_2 = 2) \\
=& 3/8
\end{aligned}$$

4.1.2 To show that $f(x_1, x_2) = (6 - [x_1 + x_2])/8$ for $0 \leq x_1 \leq 2, 2 \leq x_2 \leq 4$ is a density function requires the verification of two things.

It must be non-negative everywhere. As the function is zero outside the specified region, this requires only that we verify that it is never negative in this region:

$$\begin{aligned}
(6 - [x_1 + x_2])/8 \geq & \min_{(x_1,x_2)\in[0,2]\times[2,4]} \frac{6 - [x_1 + x_2]}{8} \\
=& 0
\end{aligned}$$

It must integrate to unity:

$$\begin{aligned}
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2 =& \frac{1}{8} \int_2^4 \int_0^2 [6 - x_1 - x_2] \mathrm{d}x_1 \mathrm{d}x_2 \\
=& \frac{1}{8} \int_2^4 \left[(6 - x_2)x_1 - \frac{1}{2} x_1^2 \right]_{x_1=0}^2 \mathrm{d}x_2 \\
=& \frac{1}{8} \int_2^4 [10 - 2x_2] \mathrm{d}x_2 \\
=& \frac{1}{8} [10x_2 - x_2^2]_2^4 \\
=& \frac{1}{8} [10 \times (4 - 2) - 4^2 + 2^2] \\
=& \frac{1}{8} [20 - 16 + 4] = 1.
\end{aligned}$$

Thus $f(x_1, x_2)$ describes a probability density.

If we're interested in the distribution function, we can calculate it explicitly explicitly:

$$\int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 = \frac{1}{8}\int_{\max(2,x_2)}^{\min(4,x_2)}\int_{\max(0,x_1)}^{\min(2,x_1)}[6-x_1-x_2]\mathrm{d}x_1\mathrm{d}x_2$$

NB strictly the integrals for x_1 should be over the region $\max(0,\min(2,x_1))$ and $\min(2,\max(x_1,0))$ but the notation becomes rather cumbersome if this is made explicit; similar for x_2.

In which case we'd need to consider the various regimes of x_1 and x_2 in, above and below the feasible region. But we can simply calculate the probabilities of interest explicitly from the density:

–

$$\begin{aligned}
F(1,3) &= \int_{-\infty}^{3}\int_{-\infty}^{1} f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&= \int_{2}^{3}\int_{0}^{1}\frac{6-[x_1+x_2]}{8}\mathrm{d}x_1\mathrm{d}x_2 \\
&= \int_{2}^{3}\left[\frac{(6-x_2)x_1-\frac{1}{2}x_1^2}{8}\right]_0^1 \mathrm{d}x_2 \\
&= \int_{2}^{3}\frac{-x_2+11/2}{8}\mathrm{d}x_2 \\
&= \frac{1}{8}\left[\frac{-x_2^2}{2}+\frac{11x_2}{2}\right]_2^3 \\
&= \frac{1}{8}\left[\frac{4-9}{2}+\frac{11(3-2)}{2}\right] = 3/8
\end{aligned}$$

– As $\mathbb{P}(X_2 \leq 1) = 0$: $F(0,1) = 0$.
– As $\mathbb{P}(X_1 \leq 3) = \mathbb{P}(X_2 \leq 5) = 1$: $F(3,5) = 1$.

4.3.1 Y_1 and Y_2 are two RVs with joint density f_{Y_1,Y_2} as shown:

	0	1	2
0	q^3	pq^2	0
1	pq^2	pq	p^2q
2	0	p^2q	p^3

– The marginal densities are:

$$\begin{aligned}
f_{Y_1}(y_1) &= \sum_{y_2} f_{Y_1,Y_2}(y_1,y_2) \\
f_{Y_2}(y_2) &= \sum_{y_1} f_{Y_1,Y_2}(y_1,y_2)
\end{aligned}$$

and so we can simply sum the rows and columns of the table showing the joint density:

$Y_1 \vert Y_2$	0	1	2	f_{Y_1}
0	q^3	pq^2	0	$q^2(p+1)$
1	pq^2	pq	p^2q	$pq(p+q+1)$
2	0	p^2q	p^3	$p^2(p+q)$
f_{Y_2}	$(p+q)q^2$	$pq(p+q+1)$	$p^2(p+q)$	

– The conditional density function of y_2 given y_1 is

$$f_{Y_2|Y_1}(y_2|y_1) = f_{Y_1,Y_2}(y_1,y_2)/f_{Y_1}(y_1)$$

and so the full conditional density can be shown in a table in which the joint densities are divided by the marginal density of the conditioning variable:

$Y_1\|Y_2$	0	1	2
0	$\frac{q}{p+1}$	$\frac{p}{p+1}$	0
1	$\frac{q}{p+q+1}$	$\frac{1}{p+q+1}$	$\frac{p}{p+q+1}$
2	0	$\frac{q}{p+q}$	$\frac{p}{p+q}$

Each row corresponds to the conditional density of Y_2 given a particular value of Y_1.

– Evaluating these quantities:

1. $\mathbb{E}[Y_1 - Y_2] = \mathbb{E}[Y_1] - \mathbb{E}[Y_2] = 0$ by symmetry.
2. $\mathbb{E}[Y_1 + Y_2] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] = 2\mathbb{E}[Y_1]$ by symmetry.
3. $\mathbb{E}[Y_1] = pq(p+q+1) + 2p^2(p+q)$.

4.4.1 If random variable (X_1, X_2) has joint density

$$f(x_1,x_2) = \begin{cases} \frac{1}{8}(6 - X_1 - X_2) & \text{for } 0 \leq X_1 \leq 2, 2 \leq X_2 \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

– The conditional densities are:

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) =& \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} \\ =& \frac{f_{X_1,X_2}(x_1,x_2)}{\int f_{X_1,X_2}(x_1',x_2)\mathrm{d}x_1'} \\ =& \frac{6 - x_1 - x_2}{\left[(6-x_2)x_1 - \frac{1}{2}x_1^2\right]_{x_1=0}^{2}} \\ =& \frac{6 - x_1 - x_2}{10 - 2x_2} \end{aligned} \qquad \begin{aligned} f_{X_2|X_1}(x_2|x_1) =& \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} \\ =& \frac{f_{X_1,X_2}(x_1,x_2)}{\int f_{X_1,X_2}(x_1,x_2')\mathrm{d}x_2'} \\ =& \frac{6 - x_1 - x_2}{\left[(6-x_1)x_2 - \frac{1}{2}x_2^2\right]_{x_2=2}^{4}} \\ =& \frac{6 - x_1 - x_2}{6 - 2x_1} \end{aligned}$$

for $0 \leq X_1 \leq 2, 2 \leq X_2 \leq 4$.

– The distribution functions are obtained from their definition as:

$$\begin{aligned} F_{X_1|X_2}(x_1|x_2) &= \int_{-\infty}^{x_1} f_{X_1|X_2}(x_1'|x_2)\mathrm{d}x_1' \\ &= \int_{\min(2,\max(0,x_1)}^{\max(0,\min(x_1,2))} \frac{6 - x_1' - x_2}{10 - 2x_2}\mathrm{d}x_1' \\ &= \begin{cases} 0 & x_1 < 0 \\ \frac{6x_1 - \frac{1}{2}x_1^2 - x_1x_2}{10 - 2x_2} & 0 \leq x_1 \leq 2 \\ 1 & 2 < x_1 \end{cases} \end{aligned}$$

(assuming that $x_2 \in [2,4]$ and undefined otherwise). Similarly:

$$\begin{aligned} F_{X_2|X_1}(x_2|x_1) &= \int_{-\infty}^{x_2} f_{X_2|X_1}(x_2'|x_1)\mathrm{d}x_2' \\ &= \int_{\min(4,\max(2,x_2)}^{\max(2,\min(x_2,4))} \frac{6 - x_1 - x_2'}{6 - 2x_1}\mathrm{d}x_2' \\ &= \begin{cases} 0 & x_2 < 2 \\ \frac{6(x_2-2) - \frac{1}{2}(x_2^2 - 2^2) - x_1(x_2-2)}{6 - 2x_1} & 2 \leq x_2 \leq 4 \\ 1 & 4 < x_2 \end{cases} \end{aligned}$$

– The conditional expectation is:

$$
\begin{aligned}
\mathbb{E}\left[X_1|X_2=x_2\right] &= \int x_1 f_{X_1|X_2}(x_1|x_2)\mathrm{d}x_1\\
&= \int x_1 \frac{6-x_1-x_2}{10-2x_2}\mathrm{d}x_1\\
&= \int_0^2 \frac{(6-x_2)x_1-x_1^2}{10-2x_2}\mathrm{d}x_1\\
&= \frac{\left[\frac{1}{2}(6-x_2)x_1^2-\frac{1}{3}x_1^3\right]_{x_1=0}^2}{10-2x_2}\\
&= \frac{\left[2(6-x_2)-\frac{8}{3}\right]}{10-2x_2}
\end{aligned}
$$

Again, the conditional expectation is undefined if $x_2 \notin [2,4]$.

4.4.2 Proof of the tower property of conditional expectation for continuous random variables:

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[X_1|X_2]] =& \mathbb{E}\left[\int f_{X_1|X_2}(x_1|X_2)x_1\mathrm{d}x_1\right]\\
=& \int\int f_{X_1|X_2}(x_1|x_2)x_1\mathrm{d}x_1 f_{X_2}(x_2)\mathrm{d}x_2\\
=& \int\int f_{X_1|X_2}(x_1|x_2)x_1 f_{X_2}(x_2)\mathrm{d}x_1\mathrm{d}x_2 \text{ by integrability}\\
=& \int\int f_{X_1,X_2}(x_1,x_2)\mathrm{d}x_2 x_1\mathrm{d}x_1 \text{ by integrability}\\
=& \int f_{X_1}(x_1)x_1\mathrm{d}x_1 = \mathbb{E}[X_1].
\end{aligned}
$$

4.4.3 $X \sim \mathsf{U}[0,1]; Y|X=x \sim \mathsf{U}[x,1]$. The conditional expectation of Y is:

$$
\begin{aligned}
\mathbb{E}[Y|X=x] =& \int y f_{Y|X}(y|x)\mathrm{d}y\\
=& \int_x^1 \frac{1}{1-x}\cdot y\mathrm{d}y \text{ density of conditional}\\
=& \frac{1}{1-x}\left[\frac{1}{2}y^2\right]_x^1\\
=& \frac{(1-x^2)}{2(1-x)} = \frac{(1+x)(1-x)}{2(1-x)} = \frac{1+x}{2}.
\end{aligned}
$$

Hence, the expectation of Y is:

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[Y|X]] &= \int f_X(x)\mathbb{E}[Y|X=x]dx\\
&= \int_0^1 \frac{1+x}{2}dx\\
&= \left[\frac{1}{2}x+\frac{1}{4}x^2\right]_0^1 = \frac{3}{4}.
\end{aligned}
$$

4.4.4 As stated, $f_\Theta(\theta)=1$ for $\theta \in [0,1]$; $f_{X|\Theta}(x|\theta) = \mathsf{Bin}\,(x;2,\theta)$:

$$
\begin{aligned}
\mathbb{E}[X|\Theta] =& \sum_{x=0}^{2} x \cdot f_{X|\Theta}(x|\Theta) \\
=& \mathbb{P}(X = 1|\Theta) + 2\mathbb{P}(X = 2|\Theta) \\
=& 2\Theta(1 - \Theta) + 2\Theta^2 \\
=& 2\Theta \\
\mathbb{E}[X] =& \int f_{\Theta}(\theta)\mathbb{E}[X|\Theta = \theta]d\theta \\
=& \int_0^1 1 \cdot 2\theta d\theta \\
=& [\theta^2]_0^1 = 1.
\end{aligned}
$$

4.5.1 For any random variables X_1 and X_2, and for any function $h(\cdot)$ for which the expectations are well defined,

$$E\big[\mathbb{E}[h(X_1)|X_2]\big] = \mathbb{E}[h(X_1)].$$

Proof for the discrete case:

$$
\begin{aligned}
E\big[\mathbb{E}[h(X_1)|X_2]\big] =& \sum_{x_2} f_{X_2}(x_2)\mathbb{E}[h(X_1)|X_2 = x_2] \\
=& \sum_{x_2} f_{X_2}(x_2) \sum_{x_1} h(x_1) f_{X_1|X_2}(x_1|x_2) \\
=& \sum_{x_2} \sum_{x_1} f_{X_2}(x_2) h(x_1) f_{X_1|X_2}(x_1|x_2) \\
=& \sum_{x_1} \sum_{x_2} f_{X_1,X_2}(x_1, x_2) h(x_1) \\
=& \sum_{x_1} f_{X_1}(x_1) h(x_1).
\end{aligned}
$$

4.5.2 Proof of conditional variance decomposition for the continuous case.

$$
\begin{aligned}
\mathbb{E}[\mathbb{V}\mathsf{ar}[X_1|X_2]] =& \mathbb{E}[\mathbb{E}[X_1^2|X_2] - \mathbb{E}[X_1|X_2]^2] \\
=& \int f_{X_2}(x_2) \left[\int x_1^2 f_{X_1|X_2}(x_1|x_2) \mathrm{d}x_1 - \left(\int x_1 f_{X_1|X_2}(x_1|x_2) \right)^2 \right] \mathrm{d}x_2 \\
=& \int f_{X_2}(x_2) \int x_1^2 f_{X_1|X_2}(x_1|x_2) \mathrm{d}x_1 \mathrm{d}x_2 - \int \left(\int x_1 f_{X_1|X_2}(x_1|x_2) \right)^2 f_{X_2}(x_2) \mathrm{d}x_2 \\
\mathbb{V}\mathsf{ar}[\mathbb{E}[X_1|X_2]] =& \mathbb{E}[\mathbb{E}[X_1|X_2]^2] - \mathbb{E}[\mathbb{E}[X_1|X_2]]^2 \\
=& \int \left[\int x_1 f_{X_1|X_2}(x_1|x_2) \right]^2 f_{X_2}(x_2) - \left(\int f_{X_2}(x_2) \int x_1 f_{X_1|X_2}(x_1|x_2) \right)^2
\end{aligned}
$$

Summing these two terms, noting that the second term in the expectation of the conditional variance is the negation of the first term in the variance of the conditional expectation:

$$
\begin{aligned}
\mathbb{E}[\mathbb{V}\mathsf{ar}[X_1|X_2]] + \mathbb{V}\mathsf{ar}[\mathbb{E}[X_1|X_2]] =& \int f_{X_2}(x_2) \int x_1^2 f_{X_1|X_2}(x_1|x_2) \mathrm{d}x_1 \mathrm{d}x_2 - \left(\int f_{X_2}(x_2) \int x_1 f_{X_1|X_2}(x_1|x_2) \right. \\
=& \int f_{X_1}(x_1) x_1^2 \mathrm{d}x_1 - \left(\int f_{X_1}(x_1) x_1 \mathrm{d}x_1 \right)^2 \\
=& \mathbb{V}\mathsf{ar}[X_1].
\end{aligned}
$$

Note that this can actually be done without expanding the expectations at all to provide a general proof a little more elegantly.

4.5.3 In 4.4.4. we established that $\mathbb{E}[X|\Theta] = 2\Theta$. Hence:

$$\begin{aligned}\mathbb{V}\mathsf{ar}[\mathbb{E}[X|\Theta]] =&\mathbb{E}[\mathbb{E}[X|\Theta]^2] - \mathbb{E}[\mathbb{E}[X|\Theta]]^2\\ =&\int \mathbb{E}[X|\Theta=\theta]^2 f_\Theta(\theta)d\theta - E(X)^2\\ =&\int_0^1 4\theta^2 d\theta - [1]^2\\ =&4/3 - 1 = 1/3\end{aligned}$$

Similarly:

$$\begin{aligned}\mathbb{E}[\mathbb{V}\mathsf{ar}[X|\Theta]] =&\mathbb{E}[\mathbb{E}[X^2|\Theta] - \mathbb{E}[X|\Theta]^2]\\ =&\mathbb{E}[\mathbb{E}[X^2|\Theta]] - \mathbb{E}[\mathbb{E}[X|\Theta]^2]\\ =&\int_0^1 \mathbb{E}[X^2|\Theta=\theta]d\theta - \int_0^1 4\theta^2 d\theta\\ =&\int_0^1 [1^2\cdot 2\theta(1-\theta) + 2^2\cdot\theta^2]d\theta - [4/3]\\ =&[\theta^2 - 2\theta^3/3]_0^1 + [4\theta^3/3]_0^1 - 4/3\\ =&[1-2/3] + 4/3 - 4/3 = 1/3\end{aligned}$$

Summing these, $\mathbb{V}\mathsf{ar}[X] = 2/3$. On average, observing Θ halves the variability in X ($\mathbb{E}[\mathbb{V}\mathsf{ar}[X|\Theta]] = 1/3$). Once the parameter of the binomial is known, the conditional distribution of X is more concentrated — the parameter provides information about the distribution of X.

4.6.1 The bivariate normal has density:

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

In the case $\rho = 0$:

$$\begin{aligned}f(x_1,x_2) =&\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-0^2}}\exp\left\{-\frac{1}{2(1-0)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\cdot 0\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}\\ =&\frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma^2}\right)^2\right]\right)\\ =&\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right)\cdot\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)\end{aligned}$$

which is precisely the product of two univariate normal distributions.

Considering $F(x_1,x_2) = 1-e^{-x_1}-e^{-x_2}+e^{-x_1-x_2-\rho x_1 x_2}$, we know that X_1 and X_2 are independent provided this distribution function can be factorised as a distribution function over X_1 and a second over X_2.

$$\begin{aligned}F(x_1,x_2) =&1 - e^{-x_1} - e^{-x_2} + e^{-x_1-x_2-\rho x_1 x_2}\\ =&(1-e^{-x_1})(1-e^{-x_2}) + [e^{-x_1-x_2-\rho x_1 x_2} - e^{-x_1-x_2}]\end{aligned}$$

if (and only if) $\rho = 0$ then the final term vanishes are we're left with:

$$F(x_1, x_2) = (1 - e^{-x_1})(1 - e^{-x_2}) = F_{X_1}(x_1)F_{X_2}(x_2)$$

with $F_{X_1}(x_1) = 1 - e^{-x_1}$ and similarly for X_2. These are the distribution functions of exponential random variables of rate 1.

4.7.1 Let

$$f(x_1, x_2) = \begin{cases} x_1 + x_2 & \text{for } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

Now,

$$\begin{aligned} f(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_2 \\ &= \int_0^1 [x_1 + x_2]\mathrm{d}x_2 \\ &= [x_1 x_2 + \frac{1}{2}x_2^2]_0^1 \\ &= x_1 + \frac{1}{2} \text{ provided } x_1 \in [0, 1] \end{aligned}$$

and by symmetry $f(x_2) = x_2 + \frac{1}{2}$ for $x_2 \in [0, 1]$.

However,

$$\begin{aligned} f(x_1|x_2) &= \frac{f(x_1, x_2)}{f(x_2)} \\ &= \frac{x_1 + x_2}{x_2 + \frac{1}{2}} \neq f(x_1) \end{aligned}$$

and so x_1 and x_2 are dependent.

The covariance can be calculated from $\mathbb{E}[X_1]$, $\mathbb{E}[X_2]$ and $\mathbb{E}[X_1 \cdot X_2]$. These are:

$$\begin{aligned} \mathbb{E}[X_1] &= \int f(x_1)x_1\mathrm{d}x_1 \\ &= \int_0^1 [x_1 + \frac{1}{2}]x_1\mathrm{d}x_1 \\ &= \int_0^1 [x_1^2 + \frac{1}{2}x_1]\mathrm{d}x_1 \\ &= [\frac{1}{3}x_1^3 + \frac{1}{4}x_1^2]_0^1 \\ &= \frac{7}{12} \end{aligned}$$

By symmetry: $\mathbb{E}[X_2] = 7/12$. Whilst:

$$\begin{aligned}
\mathbb{E}[X_1 \cdot X_2] =& \int \int f(x_1, x_2) x_1 x_2 \mathrm{d}x_2 \mathrm{d}x_1 \\
=& \int_0^1 \int_0^1 [x_1 + x_2] x_1 x_2 \mathrm{d}x_2 \mathrm{d}x_1 \\
=& \int_0^1 \int_0^1 [x_1^2 x_2 + x_1 x_2^2] \mathrm{d}x_2 \mathrm{d}x_1 \\
=& \int_0^1 \left[\frac{1}{2} x_1^2 x_2^2 + \frac{1}{3} x_1 x_2^3\right]_{x_2=0}^1 \mathrm{d}x_1 \\
=& \int_0^1 \left[\frac{1}{2} x_1^2 + \frac{1}{3} x_1\right] \mathrm{d}x_1 \\
=& \left[\frac{1}{6} x_1^3 + \frac{1}{6} x_1^2\right]_{x_1=0}^1 \\
=& \frac{1}{3}.
\end{aligned}$$

Thus, $\mathbb{C}\mathsf{ov}(X_1, Y_1) = \mathbb{E}[X_1 \cdot X_1] - \mathbb{E}[X_1]\mathbb{E}[X_2] = 1/3 - 49/144 = -1/144$; there is a negative correlation but it would seem to be a weak one.

Formalising this, we need to calculate $\mathbb{V}\mathsf{ar}(X_1)$ and $\mathbb{V}\mathsf{ar}(X_2)$. By symmetry, these must be equal and so it suffices to calculate one of them:

$$\begin{aligned}
\mathbb{V}\mathsf{ar}[X_1] =& \int f(x_1) x_1^2 \mathrm{d}x_1 - \mathbb{E}[X_1]^2 \\
=& \int_0^1 [x_1 + \frac{1}{2}] x_1^2 \mathrm{d}x_1 - 49/144 \\
=& [\frac{1}{4} x_1^4 + \frac{1}{6} x_1^3]_0^1 - 49/144 \\
=& 11/144
\end{aligned}$$

and, finally, $\rho = \mathbb{C}\mathsf{ov}(X_1, Y_1)/\sqrt{\mathbb{V}\mathsf{ar}(X_1)\mathbb{V}\mathsf{ar}(X_2)} = [-1/144]/[11/144] = -1/11$ a small negative correlation.

4.7.2 If we consider the suggested expectation, we find (non-negativity follows as we have the expectation of a non-negative function):

$$\begin{aligned}
0 \leq \mathbb{E}\left[(tX - Y)^2\right] =& \mathbb{E}[t^2 X^2 - 2tXY + Y^2] \\
0 \leq& t^2 \mathbb{E}[X^2] - 2t\mathbb{E}[XY] + \mathbb{E}[Y^2] \\
2t\mathbb{E}[XY] \leq& t^2 \mathbb{E}[X^2] + \mathbb{E}[Y^2]
\end{aligned}$$

If we consider $t = \mathbb{E}[XY]/\mathbb{E}[X^2]$ we find:

$$\begin{aligned}
2\mathbb{E}[XY]\mathbb{E}[XY]/\mathbb{E}[X^2] \leq& \mathbb{E}[XY]^2 \mathbb{E}[X^2]/\mathbb{E}[X^2]^2 + \mathbb{E}[Y^2] \\
2\mathbb{E}[XY]^2/\mathbb{E}[X^2] \leq& \mathbb{E}[XY]^2/\mathbb{E}[X^2] + \mathbb{E}[Y^2] \\
\mathbb{E}[XY]^2/\mathbb{E}[X^2] \leq& \mathbb{E}[Y^2] \\
\mathbb{E}[XY]^2 \leq& \mathbb{E}[X^2]\mathbb{E}[Y^2]
\end{aligned}$$

thus proving the Cauchy-Schwarz inequality.

Note that $Corr(X, Y) = \mathbb{C}\mathsf{ov}(X, Y)/\sqrt{\mathbb{V}\mathsf{ar}(X)\mathbb{V}\mathsf{ar}(Y)}$. Let $\bar{X} = X - \mathbb{E}[X]$ and $\bar{Y} = Y - \mathbb{E}[Y]$. Now: $\mathbb{V}\mathsf{ar}(X) = \mathbb{E}[\bar{X}^2]$, $\mathbb{V}\mathsf{ar}(Y) = \mathbb{E}[\bar{Y}^2]$ and $\mathbb{C}\mathsf{ov}(X, Y) = \mathbb{E}[\bar{X}\bar{Y}]$. Hence:

$$\begin{aligned}
Corr(X,Y) =& \frac{\mathbb{E}[\bar{X}\bar{Y}]}{\sqrt{\mathbb{E}[\bar{X}^2]\mathbb{E}[\bar{Y}^2]}} \\
|Corr(X,Y)| =& \sqrt{\frac{\mathbb{E}[\bar{X}\bar{Y}]^2}{\mathbb{E}[\bar{X}^2]\mathbb{E}[\bar{Y}^2]}} \\
\leq& 1
\end{aligned}$$

By the Cauchy-Schwarz inequality (and monotonicity of the square root).

4.8.1 As stated, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ with the X_i iid replicates of a random variable X. $\mathbb{E}[X] = \mu$, $\mathbb{V}\mathsf{ar}(X) = \sigma^2$.

$$\begin{aligned}
\mathbb{E}[\bar{X}] =& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
=& \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] \text{ linearity} \\
=& \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \\
\mathbb{V}\mathsf{ar}[\bar{X}] =& \mathbb{E}[(\bar{X} - \mathbb{E}[X])^2] \\
=& \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mathbb{E}[X])(X_j - \mathbb{E}[X])\right] \\
=& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(X_i - \mathbb{E}[X])(X_j - \mathbb{E}[X])\right]
\end{aligned}$$

It is convenient to separate the terms in the double sum in which $i = j$ and those for which $i \neq j$ as these behave in a qualitatively different way:

$$\begin{aligned}
\mathbb{V}\mathsf{ar}[\bar{X}] =& \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[(X_i - \mathbb{E}[X])^2] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathbb{E}\left[(X_i - \mathbb{E}[X])(X_j - \mathbb{E}[X])\right] \\
=& \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathsf{ar}[X_i] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathbb{E}\left[(X_i - \mathbb{E}[X])\right]\mathbb{E}\left[(X_j - \mathbb{E}[X])\right]
\end{aligned}$$

With the decomposition of the expectation following via independence. Clearly, whatever i or j, $\mathbb{E}[X_i - \mathbb{E}[X]] = \mathbb{E}[X_j - \mathbb{E}[X]] = 0$. For all i, $\mathbb{V}\mathsf{ar}[X_i] = \sigma^2$. Consequently:

$$\mathbb{V}\mathsf{ar}[\bar{X}] = \frac{1}{n^2}n\sigma^2 + 0 = \sigma^2/n.$$

4.9.1 The original random variable has density $f_X(x) = \alpha e^{-\alpha x}$ for $x \geq 0$.

1 We only need to consider $g(x)$ for $x \geq 0$ as the original random variable is non-negative with probability 1. In this regime, $g(x) = 1 - e^{-\alpha x}$, which is injective.
If $y = g(x)$, then

$$\begin{aligned}
y =& 1 - \exp(-\alpha x) \\
\exp(-\alpha x) =& 1 - y \\
\alpha x =& -\log(1-y) \\
x =& -\frac{\log(1-y)}{\alpha}
\end{aligned}$$

and so $g^{-1}(y) = -\frac{\log(1-y)}{\alpha}$ for $y \in (0,1)$ which is the range of values which y can take if it is obtained by transforming $x \in (0,\infty)$.
We also need the derivative of g^{-1} to calculate the transformed density.

$$\begin{aligned}\frac{\partial g^{-1}}{\partial y} =& -\frac{1}{\alpha}\frac{1}{1-y}(-1) \\ =& \frac{1}{\alpha(1-y)}\end{aligned}$$

Using the density transformation formula:

$$\begin{aligned}f_Y(y) =& f_X(g^{-1}(y)) \cdot \frac{\partial g^{-1}}{\partial y} \\ =& f_X(-\log(1-y)/\alpha)\frac{1}{\alpha(1-y)} \\ =& \alpha\exp(-\alpha[-\log(1-y)/\alpha])\frac{1}{\alpha(1-y)} \\ =& \alpha\exp(\log(1-y))/\alpha(1-y) \\ =& 1\end{aligned}$$

Recall that $f_Y(y)$ is non-zero only over $[0,1]$ and so this corresponds to a uniform random variable over the unit interval. This should be no surprise; in this case, $g(x) = F_X(x)$.

2 In this case $g(x) = x^{1/\beta}$ for $\beta > 0$. Clearly, $g^{-1}(y) = y^\beta$ and $g^{-1\prime}(y) = \beta y^{\beta-1}$. This is an injective mapping over the range of X. Hence:

$$\begin{aligned}f_Y(y) =& f_X(g^{-1}(y))|g^{-1\prime}(u)| \\ =& [\alpha exp(-\alpha y^\beta)]\beta y^{\beta-1} \\ =& \alpha\beta y^{\beta-1}\exp(-\alpha y^\beta)\end{aligned}$$

3 This case requires more thought. As we know that $X \geq 0$ with probability 1, we need consider only two cases: $g(X) = \begin{cases} x & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$

This function is not injective: it maps many values of x to 1. As X is an exponential random variable, we know that $F_X(x) = 1 - e^{-\alpha x}$, hence

$$\mathbb{P}(X \geq 1) = 1 - F_X(1) = 1 - 1 - e^{-\alpha} = e^{-\alpha}.$$

Hence, $\mathbb{P}(Y = 1) = \exp(-\alpha)$. However, $g(x)$ is injective over $(0,1)$. Indeed, it clearly doesn't change x over this interval, so, the distribution of Y, given that $X < 1$ is the same as the distribution of X conditional upon the same event.
Hence:

$$\begin{aligned}\mathbb{P}(Y = 1) =& \mathbb{P}(X \geq 1) = e^{-\alpha} \\ \mathbb{P}(Y < 1) =& \mathbb{P}(X < 1) = 1 - e^{-\alpha} \\ \mathbb{P}(Y \in [a,b], Y < 1) =& \int_a^b f_X(x)dx \text{ for } a \leq b \in [0,1] \\ \mathbb{P}(Y \in [a,b]|Y < 1) =& \int_a^b f_X(x)dx/[1 - e^{-\alpha}]\end{aligned}$$

as such, Y does not have a simple density function in the sense it has been defined in this course (nor does it have a density function with respect to Lebesgue measure). Y is a mixed discrete-continuous random variable. It can be viewed as what happens if a Bernoulli random variable

with success probability $e^{-\alpha}$ is drawn and, if a success occurs, Y is set deterministically to 1 (i.e. it is a realisation of a degenerate random variable with mass 1 on point 1) otherwise, it is drawn from X subject to the condition that $X < 1$ which has density $f_X(x)/[1 - e^{-\alpha}]$ over $[0, 1)$.
Such discrete mixtures may seem esoteric and unimportant, but they occur rather frequently. In the present case, it may be that a quantity which is viewed as taking distribution X is measured extremely accurately by a device which can only measure values between 0 and 1, leading to measurements of the form Y, for example.

4.9.2 To use the inversion method described, we need to know the inverse of the distribution function of the target distribution.

$$\begin{aligned}
F_X(x) =& \int_{-\infty}^{x} f_x(u)du \\
=& \int_{0}^{x} [u + \frac{1}{2}]du \text{ for } x \in [0, 1] \\
=& \left[\frac{1}{2}u^2 + \frac{1}{2}u\right]_0^x \\
=& \frac{1}{2}[x^2 + x]
\end{aligned}$$

and, of course, $F_X(x) = 0$ for $x \leq 0$ and $F_X(x) = 1$ for $x \geq 1$.

If $y = F_X(x)$, then:

$$\begin{aligned}
y =& \frac{1}{2}[x^2 + x] \\
x^2 + x - 2y =& 0 \\
x =& \frac{-1 \pm \sqrt{1 - 4(-2y \cdot 1)}}{2} \\
=& \frac{-1 + \sqrt{1 + 8y}}{2} \text{ as } \mathbb{P}(X \geq 0) = 1 \\
F^{-1}(y) =& \frac{\sqrt{1 + 8y} - 1}{2} \text{ for } y \in [0, 1].
\end{aligned}$$

We're given two realisations of a uniform $[0, 1]$ (pseudo)random variable, $u_1 = 0.25, u_2 = 0.46$ and we can use these to generate 2 pseudorandom numbers with the desired target distribution, by setting $x_i = F^{-1}(u_i)$, hence we obtain:

$$\begin{aligned}
x_1 = F^{-1}(0.25) =& \frac{\sqrt{1 + 8 \cdot 0.25} - 1}{2} \\
=& [\sqrt{3} - 1]/2 = 0.3660 \\
x_2 = F^{-1}(0.46) =& \frac{\sqrt{1 + 8 \cdot 0.46} - 1}{2} \\
=& [\sqrt{4.68} - 1]/2 = 0.5817
\end{aligned}$$

4.10.1 In 3.1.1 we established that the MGF associated with a Poisson distribution of parameter λ was $m(t) = \exp([e^t - 1]\lambda)$.

If $X_1, \ldots, X_n$ are independent Poisson random variables drawn from distributions with parameters $\lambda_1, \ldots, \lambda_n$, and $X = \sum_{i=1}^n X_i$, then:

$$\begin{aligned}
m_X(t) =&\mathbb{E}[\exp(t\sum_{i=1}^{n} X_i)] = \mathbb{E}[\prod_{i=1}^{n} \exp(tX_i)] \\
=&\prod_{i=1}^{n} \mathbb{E}[\exp(tX_i)] \text{ by independence} \\
=&\prod_{i=1}^{n} m_{X_i}(t) = \prod_{i=1}^{n} \exp([e^t-1]\lambda_i) \\
=&\exp(\sum_{i=1}^{n} [e^t-1]\lambda_i) = \exp\left([e^t-1]\sum_{i=1}^{n}\lambda_i\right)
\end{aligned}$$

This is exactly the MGF of a Poisson distribution of parameter $\sum_{i=1}^{n} \lambda_i$ and so, by uniqueness of the MGF, X must have such a distribution.

4.11.1 In this setting, X and Y are independent realisations of a random variable with discrete pdf $(p_0, \ldots, p_5)$ where $p_0 = 0.05$, $p_1 = 0.10$, $p_2 = 0.20$, $p_3 = 0.30$, $p_4 = 0.25$, and $p_5 = 0.10$.

Hence, for $x, y \in \{0, \ldots, 5\}$, $f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = p_x \cdot p_y$, and for any other values of x, y the density takes the value 0.

If the vector $\underline{p} = (p_0, \ldots, p_5)$ then the matrix $\underline{pp}^{\mathrm{T}}$ contains the joint probability distribution, which in tabular form is:

x/y	0	1	2	3	4	5
0	0.00250	0.00500	0.01000	0.01500	0.01250	0.00500
1	0.00500	0.01000	0.02000	0.03000	0.02500	0.01000
2	0.01000	0.02000	0.04000	0.06000	0.05000	0.02000
3	0.01500	0.03000	0.06000	0.09000	0.07500	0.03000
4	0.01250	0.02500	0.05000	0.07500	0.06250	0.02500
5	0.00500	0.01000	0.02000	0.03000	0.02500	0.01000

Probabilities can be calculated directly from this matrix:

$$\begin{aligned}
\mathbb{P}(X = Y) =&\sum_{x=0}^{5} \mathbb{P}(X = Y = x) \\
=&\mathrm{Trace}\left[\underline{pp}^{\mathrm{T}}\right] = 0.215
\end{aligned}$$

Whilst $\mathbb{P}(X > Y) = \sum_{x=0}^{4} \sum_{y=x+1}^{5} f_{X,Y}(x, y)$ which is the sum of all elements below the diagonal in the matrix. Summing these terms, we get 0.392.

4.11.2 In order for

$$f(x, y) = \begin{cases} \frac{(2y)^x}{x!} e^{-3y} & \text{if } y > 0 \text{ and } x = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

to be a joint density, it must be non-negative everywhere, which is clearly the case as it's the product of positive terms over the region in which it is nonzero.

It must also have total mass one; summing over the discrete components and integrating over the continuous ones must give unity.

$$\begin{aligned}
&\int_0^\infty \sum_{x=0}^{\infty} \frac{(2y)^x}{x!} e^{-3y} \mathrm{d}y \\
=&\int_0^\infty \exp(2y) e^{-3y} \mathrm{d}y
\end{aligned}$$

by recognising the series representation of the exponential function, and this is simply:

$$\int_0^\infty \exp(2y)e^{-3y}\mathrm{d}y = \int_0^\infty e^{-y}\mathrm{d}y = [-e^{-y}]_0^\infty - = 1$$

To find $\mathbb{P}(X = 0)$, we need to evaluate the marginal (discrete) density of X at zero, which is:

$$\begin{aligned} f_X(x)|_{x=0} &= \int_0^\infty \frac{(2y)^x}{x!}e^{-3y}\mathrm{d}y|_{x=0} \\ &= \int_0^\infty e^{-3y}\mathrm{d}y \\ &= \left[-\frac{1}{3}e^{-3y}\right]_0^\infty = \frac{1}{3}. \end{aligned}$$

4.11.3 The marginal densities are found by summing over the irrelevant variable in the joint distribution, which corresponds to summing rows and columns in this table.

$f_{X,Y}(x,y)$	1 (Bbb)	2 (Bb)	3 (B)	$f_Y(y)$
8.5	0.26	0.10	0.00	0.36
11.5	0.04	0.28	0.04	0.36
17.5	0.00	0.02	0.26	0.28
$f_X(x)$	0.30	0.40	0.30	1

Where the bottom right entry is just a check that both $f_X(x)$ and $f_Y(y)$ sum to unity.

The specified expectations can be found most easily from the marginal densities:

$$\begin{aligned} \mathbb{E}[X] =& \sum_{x=1}^{3} f_x(x) \cdot x = 2.0 \\ \mathbb{E}[Y] =& \sum_y f_Y(y) \cdot y \\ =& 8.5 \cdot 0.36 + 11.5 \cdot 0.36 + 17.5 \cdot 0.28 = 12.1 \end{aligned}$$

The bond rating and yield are not independent; consider $x = 1, y = 17.5$ for example.

To calculate the covariance, we also require $\mathbb{E}[XY]$:

$$\begin{aligned} \mathbb{E}[XY] =& \sum_x \sum_y f_{X,Y}(x,y)xy \\ =& \sum_x x \sum_y y f_{X,Y}(x,y) \\ =& 1\left[0.26 \cdot 8.5 + 0.04 \cdot 11.5\right] + 2\left[0.10 \cdot 8.5 + 0.28 \cdot 11.5 + 0.02 \cdot 17.5\right] + 3\left[0.04 \cdot 11.5 + 0.26 \cdot 17.5\right] \\ =& 1 \cdot 2.67 + 2 \cdot 4.42 + 3 \cdot 5.01 \\ =& 26.54 \end{aligned}$$

And then:

$$\begin{aligned} \mathsf{Cov}(X,Y) =& \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ =& 26.54 - 2.0 \cdot 12.1 = 2.34 \end{aligned}$$

To further calculate the correlation, we require the variances of X and Y, which we can obtain via the second moments:

$$
\begin{aligned}
\mathbb{E}[X^2] =& \sum_{x=1}^{3} x^2 f_X(x) = 4.6 \\
\mathbb{E}[Y^2] =& \sum_{y} y^2 f_Y(y) = 159.37 \\
\mathbb{V}\mathsf{ar}[X] =& \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 4.6 - 2.0^2 = 0.6 \\
\mathbb{V}\mathsf{ar}[Y] =& \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = 159.37 - 12.1^2 = 12.96
\end{aligned}
$$

And so:

$$
\begin{aligned}
\rho(X, Y) =& \frac{\mathbb{C}\mathsf{ov}(X, Y)}{\sqrt{\mathbb{V}\mathsf{ar}(X)\mathbb{V}\mathsf{ar}(Y)}} \\
=& \frac{2.34}{\sqrt{0.6 \cdot 12.96}} \\
=& 0.84
\end{aligned}
$$

which tells us that there is a strong positive correlation between the bond rating and its yield. This matches the pattern seen in the table.

Finally, $\mathbb{E}[Y|X = 1]$, is the expectation with respect to the conditional density of Y.

$$
\begin{aligned}
\mathbb{E}[Y|X = 1] =& \sum_{y} y f_{Y|X}(y|1) \\
=& \sum_{y} y \frac{f_{X,Y}(1, y)}{f_X(1)} \\
=& \frac{1}{f_X(1)} \sum_{y} y f_{X,Y}(1, y) \\
=& \frac{1}{0.30} [0.26 \cdot 8.5 + 0.04 \cdot 11.5] \\
=& 2.67/0.30 = 8.9
\end{aligned}
$$

4.11.4 Two models:

a If $X_1 \sim \mathsf{Poi}\,(\lambda_1)$ and $X_1 \sim \mathsf{Poi}\,(\lambda_2)$ are independent then the joint density is:

$$
\begin{aligned}
f^{a}_{X_1,X_2}(x_1, x_2) =& f_{X_1}(x_1) f_{X_2}(x_2) \\
=& \frac{\lambda_1^{x_1}}{x_1!} e^{-\lambda_1} \frac{\lambda_2^{x_2}}{x_2!} e^{-\lambda_2} \\
=& \frac{\lambda_1^{x_1} \lambda_2^{x_2}}{x_1! x_2!} e^{-(\lambda_1 + \lambda_2)}
\end{aligned}
$$

b If $N \sim \mathsf{Poi}\,(\lambda)$ and $X_1|N \sim \mathsf{Bin}\,(N, \theta)$ then the joint density of N and X_1 is:

$$
\begin{aligned}
f_{N,X_1}(n, x_1) =& f_N(n) f_{X_1|N}(x_1|n) \\
=& \frac{\lambda^n}{n!} e^{-\lambda} \binom{n}{x_1} \theta^{x_1} (1 - \theta)^{n - x_1}
\end{aligned}
$$

In order to produce comparable results from the two models, one approach is to match moments. The expected number of male and female piglets under model a is λ_1 and λ_2 and the total expected number is $\lambda_1 + \lambda_2$. Under model b, the expected number is λ and the expected number of male piglets is $\theta\lambda$. So, making the identifications:

$$
\lambda = \lambda_1 + \lambda_2 \qquad\qquad \theta = \lambda_1 / [\lambda_1 + \lambda_2]
$$

model b becomes:

$$\begin{aligned}f_{N,X_1}(n,x_1) =& \frac{[\lambda_1+\lambda_2]^n}{n!}e^{-(\lambda_1+\lambda_2)}\binom{n}{x_1}\left[\frac{\lambda_1}{\lambda_1+\lambda_2}\right]^{x_1}\left[\frac{\lambda_2}{\lambda_1+\lambda_2}\right]^{n-x_1}\\ =& \frac{1}{n!}e^{-[\lambda_1+\lambda_2]}\frac{n!}{x_1!(n-x_1)!}\lambda_1^{x_1}\lambda_2^{n-x_1}\\ =& \frac{\lambda_1^{x_1}\lambda_2^{n-x_1}}{x_1!(n-x_1)!}e^{-[\lambda_1+\lambda_2]}\end{aligned}$$

And, as there is a one-to-one correspondence between pairs (x_1, x_2) and (x_1, n) given by setting $n = x_1 + x_2$ this also defines a density over (x_1, x_2) which can be obtained by substitution:

$$\begin{aligned}f^b_{X_1,X_2}(x_1,x_2) =& f_{N,X_1}(x_1+x_2,x_1)\\ =& \frac{\lambda_1^{x_1}\lambda_2^{x_2}}{x_1!x_2!}e^{-[\lambda_1+\lambda_2]}\end{aligned}$$

which is exactly the same as model a. The two models are equivalent and indistinguishable.

4.11.5 This is a transformation of a random variable. $T = 1/Z$. As $f_Z(z) = 5e^{-5z}$, and $T = g(Z)$ with $g(z) = 1/z$, for which $g^{-1}(t) = 1/t$ and $\partial g^{-1}/\partial t = -1/t^2$ we have (for positive t as all positive z will yield positive t and any positive t can be obtained from some positive z):

$$\begin{aligned}f_T(t) =& f_Z(g^{-1}(t))\left|\frac{\partial g^{-1}}{\partial_t}\right|_t\\ =& 5\exp(-5/t)\cdot\frac{1}{t^2} \text{ for } t \geq 0.\end{aligned}$$

whilst, $f_T(t) = 0 \forall t < 0$.

The easy solution to the second part is to say that:

$$\mathbb{P}(T \geq t) = \mathbb{P}(Z \leq 1/t) = 1 - e^{-5/t}$$

and so, $\mathbb{P}(T > 5) = 1 - 1/e$ and $\mathbb{P}(T > 10) = 1 - 1/e^{1/2}$.

More directly, the integral $\int_t^\infty t^{-2}\exp(-5/t)dt$ can be done by making the substitution $s = 1/t$ but this amounts to an algebraic version of the same argument.

4.11.6 Supposing that the random variables X and Y have a continuous joint distribution, with pdf $f(x, y)$, means μ_X & μ_Y respectively, variances σ_X^2 & σ_Y^2 respectively, and correlation ρ and also that $\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$:

(a)

$$\begin{aligned}\int_{-\infty}^{\infty} yf(x,y)\mathrm{d}y =& \int_{-\infty}^{\infty} yf(y|x)\mathrm{d}yf_x(x)\\ =& \mathbb{E}[Y|x]f_X(x)\end{aligned}$$

(b)

$$\begin{aligned}\mu_Y =& \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]\\ =& \int f_X(x)\mathbb{E}[Y|X=x]dx\\ =& \int f_X(x)[\beta_0+\beta_1 x]dx\\ =& \beta_0 + \beta_1\int f_X(x)xdx\\ =& \beta_0+\beta_1\mathbb{E}[X] = \beta_0+\beta_1\mu_X\end{aligned}$$

(c)

$$\begin{aligned}\rho\,\sigma_X\sigma_Y + \mu_X\mu_Y =& \frac{\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}{\sigma_X\sigma_Y}\sigma_X\sigma_Y + \mu_X\mu_Y\\ =&\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|X]]\\ =&\mathbb{E}[X\mathbb{E}[Y|X]]\\ =&\mathbb{E}[X(\beta_0+\beta_1 X)]\\ =&\beta_0\mu_X + \beta_1(\sigma_X^2+\mu_X^2)\end{aligned}$$

(d) Parts (b) and (c) give two equations in the unknown quantities β_0 and β_1:

$$\begin{aligned}\mu_Y =&\beta_0 + \beta_1\mu_X \Rightarrow \beta_0 = \mu_Y - \beta_1\mu_X\\ \rho\,\sigma_X\sigma_Y + \mu_X\mu_Y =&\beta_0\mu_X + \beta_1(\sigma_X^2+\mu_X^2)\end{aligned}$$

Substituting the expression for β_0 from the first of these equations into the second, we obtain:

$$\begin{aligned}\rho\,\sigma_X\sigma_Y + \mu_X\mu_Y =&(\mu_Y - \beta_1\mu_X)\mu_X + \beta_1(\sigma_X^2+\mu_X^2)\\ =&\mu_X\mu_Y + \beta_1\sigma_X^2\\ \rho\sigma_Y/\sigma_X =&\beta_1\end{aligned}$$

Thus $\beta_1 = \rho\sigma_Y/\sigma_X$ and substituting this back into the expression for β_0 in terms of β_1, we obtain:

$$\begin{aligned}\beta_0 =&\mu_Y - \beta_1\mu_X\\ =&\mu_Y - \rho\sigma_Y/\sigma_X\mu_X\end{aligned}$$

5.2.1 If $X_i \overset{iid}{\sim} N(20, 2^2)$ and the sample size, $n = 25$ then $\bar{X} \sim N(20, 2^2/25)$ and $2^2/25 = 0.4^2$. Hence,

$$\begin{aligned}\mathbb{P}(\bar{X} > 21) =&1 - \mathbb{P}(\bar{X} > 21)\\ =&1 - P\left(\frac{\bar{X}-20}{0.4} > \frac{21-20}{0.4}\right)\\ =&1 - P\left([\bar{X}-20]/0.4 > 2.5\right)\\ =&1 - \varPhi(2.5) = 0.006.\end{aligned}$$

Which is much smaller than the probability that a single car achieves a value greater than 21.

5.2.2 The model described for the total number of heads, $X = \sum_{i=1}^{900} X_i$ is $\mathsf{Bin}\,(900, 0.5)$ — it's the sum of 900 independent Bernoulli variables of parameter 0.5.

It's impractical to evaluate the probability of obtaining more than 495 heads exactly. However, we can use the CLT to argue that for such a large sample the number of heads is approximately normal with mean $900 \cdot 0.5$ and variance $0.25 \cdot 900$ (where 0.25 is the variance of a single toss).

If $X \overset{approx}{\sim} \mathsf{N}\,(450, 225)$ then the standard deviation is 15. Hence $\mathbb{P}(X \geq 495) \approx 1 - \varPhi([495 - 450]/15)$ which is 0.0013.

5.2.3 The MGF of a χ_k^2 random variable is $m_{\chi_k^2}(t) = (1-2t)^{-k/2}$. If $X_i \sim \chi_{r_i}^2$, independently, and $X = \sum_{i=1}^n X_i$, then it has MGF:

$$
\begin{aligned}
m_X(t) =& \mathbb{E}[\exp(t\sum_{i=1}^{n} X_i)] \\
=& \prod_{i=1}^{n} m_{X_i}(t) \\
=& \prod_{i=1}^{n} (1-2t)^{-r_i/2} \\
=& (1-2t)^{-\sum_{i=1}^{n} r_i/2}
\end{aligned}
$$

which is the MGF of a $\chi^2_{\sum_{i=1}^n r_i}$ distribution. By uniqueness of the MGF, X must have such a distribution.

5.2.4 The MGF of a χ^2_k random variable is $m_{\chi^2_k}(t) = (1-2t)^{-k/2}$. If $X_i \sim \chi^2_{r_i}$, independently, and $X = \sum_{i=1}^n X_i$, then it has MGF:

$$
\begin{aligned}
m_X(t) =& \mathbb{E}[\exp(t\sum_{i=1}^{n} X_i)] \\
=& \prod_{i=1}^{n} m_{X_i}(t) \\
=& \prod_{i=1}^{n} (1-2t)^{-r_i/2} \\
=& (1-2t)^{-\sum_{i=1}^{n} r_i/2}
\end{aligned}
$$

which is the MGF of a $\chi^2_{\sum_{i=1}^n r_i}$ distribution. By uniqueness of the MGF, X must have such a distribution.

5.1

$$\mathbb{P}[\mathsf{t}_9 < (8.2 - 8.6)/(\sqrt{0.4}/\sqrt{10})] = \mathbb{P}[\mathsf{t}_9 < -2] = 0.038$$

5.2.5 Assuming that the population means *and* variances are equal, we have observations $\bar{X} = 17.89, \bar{Y} = 16.97$ the observed sample variances are $S_X^2 = 3.43$ and $S_Y^2 = 4.76$. $n_x = n_Y = 15$.

From the lecture notes, we know that:

$$
\begin{aligned}
T =& \frac{\bar{X} - \bar{Y} - [\mu_X - \mu_Y]}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \\
\text{with } s_p^2 =& \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}
\end{aligned}
$$

has a $\mathsf{t}_{n_X+n_Y-2}$ distribution.

If $\mu_X = \mu_Y$, we have:

$$T = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim \mathsf{t}_{n_X+n_Y-2}$$

and in our case $s_p^2 = \frac{1}{2}[s_X^2 + s_Y^2] = 4.095$ and so $s_p = 2.0236$.

So:

$$\mathbb{P}(\bar{X} - \bar{Y} > 0.92) = P\left(\frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{2}{14}}} > 0.92/s_p\sqrt{\frac{2}{14}}\right)$$

We have $s_p\sqrt{2/14} = 0.76485$ and $0.92/s_p\sqrt{2/14} = 1.2028$. Hence:

$$\mathbb{P}(\bar{X} - \bar{Y} > 0.92) = \mathbb{P}(\mathsf{t}_{28} > 1.2028) = 0.12.$$

By symmetry, the probability that $\mathbb{P}(|\bar{X} - \bar{Y}| > 0.92) = 0.24$: there's almost a 1 in 4 chance of observing a difference in the means as large as this one if the two population means and variances are equal.

5.3.1 If $X_1, \ldots, X_n \overset{iid}{\sim} f$ with $\mathbb{E}[X_i = \mu]$ and $\mathbb{V}\mathsf{ar}[X_i] = \sigma^2$, then the two estimators are both unbiased:

$$\begin{aligned}
\mathbb{E}[\hat{\mu}_1] =& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
=& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\
=& \frac{1}{n} n\mu = \mu \\
\mathbb{E}[\hat{\mu}_2] =& \mathbb{E}[X_1 + X_2]/2 = 2\mu/2 = \mu
\end{aligned}$$

However, the variance of the first estimator will be much smaller than that of the second for $n \gg 2$. Unbiasedness is one desirable property, but it's not the only consideration when choosing an estimator.

5.3.2 The decomposition of MSE as variance plus squared bias follows from the following:

$$\begin{aligned}
MSE(\hat{\theta}) =& \mathbb{E}[(\hat{\theta} - \theta)^2] \\
=& \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\
=& \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta) + (\mathbb{E}[\hat{\theta}] - \theta)^2] \\
=& \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])](\mathbb{E}[\hat{\theta}] - \theta) + \mathbb{E}[(\mathbb{E}[\hat{\theta}] - \theta)^2] \\
=& \mathbb{V}\mathsf{ar}(\hat{\theta}) + 0 + [\mathbb{E}[\hat{\theta}] - \theta]^2 \\
=& \mathbb{V}\mathsf{ar}(\hat{\theta}) + Bias(\hat{\theta})^2
\end{aligned}$$

5.4.1 This is a proportion. Assuming that we may treat the people as identical and independent (which may not be a good assumption under the circumstances) we can estimate $\hat{p} = 136/400 = 0.34$ and $\hat{p}(1 - \hat{p}) = 0.224$. Then using the standard CLT-justified approximate confidence interval, we seek:

$$\begin{aligned}
&\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2} \\
=& 0.34 \pm \sqrt{\frac{0.224}{400}} \cdot 1.96 \\
=& 0.34 \pm 0.023685 \cdot 1.96 \\
=& 0.34 \pm 0.0464 = [0.294, 0.386]
\end{aligned}$$

5.4.2 This time we have a difference in proportions (and can use the expression which immediately precedes the exercise).

We may use the sample means to estimate the proportions for male and female voters, respectively:

$$\hat{p}_m = 132/200 = 0.61 \qquad \hat{p}_f = 90/159 = 0.566$$

And the variances are estimated in the usual way as:

$$p_m(1-p_m) \approx \hat{p}_m(1-\hat{p}_m) = 0.61 \cdot 0.39 = 0.2379 \quad p_f(1-p_f) \approx \hat{p}_f(1-\hat{p}_f) = 0.566 \cdot 0.434 = 0.2456$$

Asymptotically, both sample means have a normal distribution and so do their differences. Consistency of the variance estimators allow us to replace the population variances with their estimators to obtain an approximate confidence interval of:

$$\begin{aligned}
&\hat{p}_m - \hat{p}_f \pm z_{\frac{1-0.99}{2}} \sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{n_m} + \frac{\hat{p}_f(1-\hat{p}_f)}{n_f}} \\
=&0.61 - 0.566 \pm z_{0.005} \sqrt{\frac{0.2379}{200} + \frac{0.2456}{159}} \\
=&0.044 \pm 2.578 \cdot \sqrt{0.00273} \\
=&0.044 \pm 0.135
\end{aligned}$$

As 0 lies within this interval, we might think that a result like this one could arise even if male and female voters have the same probability of voting for this candidate.

5.4.3 This time we have a difference in proportions (and can use the expression which immediately precedes the exercise).

We may use the sample means to estimate the proportions for male and female voters, respectively:

$$\hat{p}_m = 132/200 = 0.61 \qquad \hat{p}_f = 90/159 = 0.566$$

And the variances are estimated in the usual way as:

$$p_m(1-p_m) \approx \hat{p}_m(1-\hat{p}_m) = 0.61 \cdot 0.39 = 0.2379 \quad p_f(1-p_f) \approx \hat{p}_f(1-\hat{p}_f) = 0.566 \cdot 0.434 = 0.2456$$

Asymptotically, both sample means have a normal distribution and so do their differences. Consistency of the variance estimators allow us to replace the population variances with their estimators to obtain an approximate confidence interval of:

$$\begin{aligned}
&\hat{p}_m - \hat{p}_f \pm z_{\frac{1-0.99}{2}} \sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{n_m} + \frac{\hat{p}_f(1-\hat{p}_f)}{n_f}} \\
=&0.61 - 0.566 \pm z_{0.005} \sqrt{\frac{0.2379}{200} + \frac{0.2456}{159}} \\
=&0.044 \pm 2.578 \cdot \sqrt{0.00273} \\
=&0.044 \pm 0.135
\end{aligned}$$

As 0 lies within this interval, we might think that a result like this one could arise even if male and female voters have the same probability of voting for this candidate.

5.4.4 An exact confidence interval for the difference can be obtain provided that we assume that the two sets of data are simple random samples from normal distributions with common variances.

Summarising the available data:

$$\begin{aligned}
n_A &= 10 & n_B &= 8 \\
\bar{X}_A &= 3.1 & \bar{X}_B &= 2.7 \\
\frac{s_A}{\sqrt{n_A}} &= 0.5 & \frac{s_B}{\sqrt{n_B}} &= 0.7
\end{aligned}$$

The difference in means divided by the pooled standard deviation is t-distributed about the sample mean with $n_A + n_B - 2 = 16$ degrees of freedom. A 95% confidence interval is:

$$\bar{X}_A - \bar{X}_B \pm \mathsf{t}_{0.025,16} \cdot \sqrt{\left(\frac{s_A^2}{n_A}\right) + \left(\frac{s_B^2}{n_B}\right)} = 0.4 \pm 2.12 \cdot \sqrt{0.5^2 + 0.7^2}$$
$$= 0.4 \pm 1.8236$$

which again suggests that data like this could arise quite often even if the two sample means are equal.

5.5.1 If $X \sim N(\theta, 2^2)$ and $X_1, X_2, \ldots$ is a simple random sample of replicates of X then in order to ensure that:

$$\mathbb{P}(|\bar{X} - \theta| > 0.1) \leq 0.05$$

we need n to be large enough that a normal random variable of mean 0 and variance $2^2/n$ has probability at least 0.95 of being at most 0.1.

As $z_{0.025} = 1.96$, we need $\sqrt{2^2/n} \cdot 1.96 \leq 0.1$ i.e.:

$$\begin{aligned} 2 \cdot 1.96/\sqrt{n} \leq& 0.1 \\ 20 \cdot 1.96 \leq& n \\ 39.2 \leq& n \end{aligned}$$

and so a sample of size at least 40 will be required (as the number of samples must be an integer).

5.5.2 If $X_i \overset{iid}{\sim} Bernoulli(p)$ and $p \approx 0.2$ so that we may assume $\mathbb{V}\mathsf{ar}(X_i) \approx p(1-p) = 0.16$ (if $p \in [0.15, 0.25]$ then $\mathbb{V}\mathsf{ar}(p) \in [0.1275, 0.1875]$).

The CLT tells us that $\sqrt{n}\bar{X} \xrightarrow{d} N(p, p(1-p))$ if we may assume that $p = 0.2$ for the purposes of calculating the variance, we can easily solve:

$$\begin{aligned} \mathbb{P}(|\bar{X} - p| < 0.1) \geq& 0.75 \\ \mathbb{P}(\bar{X} - p < 0.1) \geq& 0.875 \text{ by symmetry} \\ \mathbb{P}(\frac{\bar{X} - p}{\sigma/\sqrt{n}} < 0.1\sqrt{n}/\sigma) \geq& 0.875 \\ \mathbb{P}(Z < 0.1\sqrt{n}/\sigma) \geq& 0.875 \\ 0.1\sqrt{n}/\sigma >& 1.15(\text{ normal df}) n > \qquad (11.5\sigma)^2 \\ >& 21.16 \Rightarrow n \geq 22. \end{aligned}$$

Of course, we could be more conservative by incorporating some uncertainty in the variance here.

5.5.3 Consider these four estimators in turn:

(a) $\hat{\mu}_1 = a\bar{X}, 0 < a < 1$ is biased, it's expectation is $a\mathbb{E}[\bar{X}] = a\mu$ and $a \neq 1$. It's asymptotically biased as the bias doesn't decrease with sample size. It's not consistent in MSE as the asymptotic bias is nonzero.

(b) This estimator includes $\bar{X}$ as a special case (consider $a_i = 1/n$), in fact:

$$\begin{aligned} \mathbb{E}[\hat{\mu}_2] =& \mathbb{E}[\sum_{i=1}^{n} a_i X_i] \\ =& \sum_{i=1}^{n} a_i \mathbb{E}[X_i] \\ =& \sum_{i=1}^{n} a_i \mu = \mu \end{aligned}$$

It is unbiased for any sequence a_i which sums to one. It may or may not be consistent in MSE depending upon the sequence of sequences a_i. Trivially, if $a_1 =$ and $a_i = 0 \forall i > 1$ then it isn't, for example, whilst if $a_i = 1/n$ then we recover the usual estimator.

(c) This seems a rather poor estimator:

$$\begin{aligned}\mathbb{E}[\hat{\mu}_3] =&\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} X_i\right]\\ =&\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[X_i]\\ =&\frac{1}{n^2}n\mu = \mu/n\end{aligned}$$

which is biased for any $n > 1$. It's not asymptotically unbiased as:

$$\lim_{n\to\infty}\mathbb{E}[\hat{\mu}_3] = \lim_{n\to\infty}\mu/n = 0$$

and it cannot be consistent in mean squared error as it is asymptotically biased.

(d) $\hat{\mu}_4 = \frac{n}{n-1}\bar{X}$. As such, it is biased for finite n with expected value $n\mu/(n-1)$. However,

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}[\hat{\mu}_4] = &\lim_{n\to\infty}\frac{n}{n-1}\mu\\ =&\mu\cdot\lim_{n\to\infty} 1 + \frac{1}{n-1} = \mu\end{aligned}$$

and so it is asymptotically unbiased.

As the estimator is asymptotically equal to the sample mean, the standard argument also shows that it is consistent in MSE.

5.5.4 Consider the class of linear estimators of the sample mean:

$$\tilde{X} = \sum_{i=1}^{n} a_i X_i.$$

In order for such an estimator to be unbiased:

$$\begin{aligned}\mu = \mathbb{E}[\tilde{X}] =&\sum_{i=1}^{n} a_i\mathbb{E}[X_i]\\ =&\mu\sum_{i=1}^{n} a_i \Rightarrow \sum_{i=1}^{n} a_i = 1.\end{aligned}$$

To determine the minimum variance *unbiased* linear estimator we need to incorporate this constraint (otherwise, we can minimise the variance by setting $a_i = 0\forall i$ but this estimator wouldn't be much use).

The constraint may be written in the form $\sum_{i=1}^{n} a_i - 1 = 0$, leading to the Lagrange-multiplier problem (if you're not familiar with the method of Lagrange multipliers then this would be a very good time to find out how it works):

$$\begin{aligned}\frac{\partial}{\partial a_j}\left[\sigma^2\sum_{i=1}^{n} a_i - \lambda\left(\sum_{i=1}^{n} a_i - 1\right)\right] =&0\\ 2\sigma^2 a_j - \lambda =&0\\ a_j =&\lambda/2\sigma^2\end{aligned}$$

To identify the value of λ we use the constraint:

$$\begin{aligned}\sum_{i=1}^{n} a_i =&1\\ \sum_{i=1}^{n} \lambda/2\sigma^2 =&1\\ \lambda = &2\sigma^2/n\end{aligned}$$

and so, for each j, $a_j = 1/n$ to extremise the constrained variance. It's easy to verify that any deviation from this value leads to a higher variance and so it is a minimum.

5.5.5 The sequence of RVs, X_n, $n \geq 1$ is defined via:

$$\begin{aligned}\mathbb{P}(X_n = a) =&1 - 1/n, a \in \mathbb{R}\\ \mathbb{P}(X_n = n^k) =&1/n, k > 1\end{aligned}$$

It is reasonably clear that X_n is consistent, in probability, as an estimator of a:

$$\begin{aligned}\mathbb{P}(|X_n - a| > \epsilon) \leq&\mathbb{P}(X_n \neq a)\\ =&1/n \text{ for any } \epsilon > 0\end{aligned}$$

and $lim_{n\to\infty}1/n = 0$ and so we have the convergence we seek.

The MSE of the estimator, on the other hand is:

$$\begin{aligned}\mathbb{E}[(X_n - a)^2] =&\mathbb{P}(X_n = n^k)(n^k - a)\\ =&\frac{n^k - a}{n} \to \infty\end{aligned}$$

and so the estimator is *not* consistent in MSE. Convergence in mean squared error is strictly stronger than convergence in probability.

5.5.6 Summarising the given data:

$$n =100 \qquad \bar{X} =1.944 \qquad s^2 =0.0028 \qquad s =0.053$$

(a) If $\sigma = 0.05$ is known, then a 95% confidence interval for μ is:

$$\bar{X} \pm z_{0.025} \cdot \sqrt{\frac{\sigma^2}{n}} =1.944 \pm 1.96 \cdot 0.005 = 1.944 \pm 0.0098.$$

(b) On the other hand, if the variance is estimated, then we must use the estimated variance and also take into account that the distribution is now t_{99} rather than normal, although the difference is rather small:

$$\bar{X} \pm \mathsf{t}_{0.025,99} \cdot \sqrt{\frac{s^2}{n}} =1.944 \pm 1.9842 \cdot 0.0053 = 1.944 \pm 0.0105.$$

(c) If neither mean nor variance are known but we're interested in estimating σ, then we can use the known distribution of the sample variance to construct a confidence interval (see page 55) of the following form for σ^2:

$$\begin{aligned}\left[\frac{(n-1)s^2}{\chi^2_{0.025,n-1}}, \frac{(n-1)s^2}{\chi^2_{0.975,n-1}}\right] =&\left[\frac{99 \cdot 0.0028}{128.42}, \frac{99 \cdot 0.0028}{73.61},\right]\\ =&[0.0022, 0.0038]\end{aligned}$$

and so, the interval $[0.0465, 0.0615]$ with probability 0.95.

5.5.7 To calculate the binomial interval exactly, it's necessary to find $\{p : \mathbb{P}(\mathsf{Bin}\,(n,p) \leq X) \geq \alpha/2\}$ and $\{p : \mathbb{P}(\mathsf{Bin}\,(n,p) \geq X) \geq \alpha/2\}$ and take their intersections.

It's rather simpler to employ the CLT-based normal approximation.

6.1.1 If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Poi}\,(\lambda)$ then, with $\mathbf{x} = (x_1, \ldots, x_n)$, then the likelihood is:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

and the log-likelihood is:

$$\begin{aligned}\ell(\theta; \mathbf{x}) =& \sum_{i=1}^{n} [x_i \log\theta - \theta - \log(x_i!)] \\ =& \log\theta \sum_{i=1}^{n} x_i - n\theta - \sum_{i=1}^{n} \log(x_i!).\end{aligned}$$

Differentiating with respect to the parameter, and setting the derivative equal to zero, we obtain:

$$\begin{aligned}0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta; \mathbf{x}) =& \frac{1}{\theta}\sum_{i=1}^{n} x_i - n \\ n\theta =& \sum_{i=1}^{n} x_i \Rightarrow \theta = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.\end{aligned}$$

Checking the second derivative:

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\theta}\ell(\theta; \mathbf{x}) = -\frac{1}{\theta^2} < 0$$

and so this is, indeed, a maximiser of the log-likelihood and, by monotonicity, of the likelihood.

6.2.1 First consider the Poisson distribution:

$$\begin{aligned}f(x;\theta) =& \frac{e^{-\theta}\theta^x}{x!} \\ =& \exp\left(-\theta + x\log\theta - \log(x!)\right) \\ =& \exp\left(\underbrace{\log(\theta)}_{A(\theta)} \underbrace{x}_{B(x)} + \underbrace{(-\log(x!))}_{C(x)} - \underbrace{\theta}_{D(\theta)}\right).\end{aligned}$$

Whilst, for the two-parameter Gamma distribution:

$$\begin{aligned}f(x;\alpha,\beta) =& \frac{\beta^\alpha x^{\alpha-1}\exp(-\beta x)}{\varGamma(\alpha)} \\ =& \exp\left(\alpha\log(\beta) + (\alpha-1)\log(x) - \beta x - \log\left(\varGamma(\alpha)\right)\right) \\ =& \exp\left(\underbrace{\alpha}_{A_1(\theta)}\underbrace{\log(x)}_{B_1(x)} - \underbrace{\beta}_{A_2(\theta)}\underbrace{x}_{B_2(x)} + \underbrace{(-\log(x))}_{C(x)} + \underbrace{\alpha\log(\beta) - \log\left(\varGamma(\alpha)\right)}_{D(\theta)}\right).\end{aligned}$$

6.3.1 The simplest approach is to begin by considering $\mathbb{E}\left[\frac{\partial \log L}{\partial\theta}\right]$:

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \ell}{\partial \theta}\right] &= \int L(\theta;x)\frac{\partial \log L(\theta;x)}{\partial \theta}\mathrm{d}x \\ &= \int \frac{\partial L(\theta;x)}{\partial \theta}\mathrm{d}x \\ &= \frac{\partial}{\partial \theta}\int L(\theta;x)\mathrm{d}x = 0.\end{aligned}$$

Differentiating both sides (using the product rule within the integral) yields:

$$\begin{aligned}\frac{\partial}{\partial \theta}\mathbb{E}\left[\frac{\partial \ell(\theta;x)}{\partial \theta}\right] &=0 \\ \frac{\partial}{\partial \theta}\int \frac{\partial \log L(\theta;x)}{\partial \theta}L(\theta;x)\mathrm{d}x &=0 \\ \int \frac{\partial}{\partial \theta}\left(\frac{\partial \log L(\theta;x)}{\partial \theta}L(\theta;x)\right)\mathrm{d}x &=0 \\ \int \left(\frac{\partial^2 \log L(\theta;x)}{\partial \theta^2}L(\theta;x) + \frac{\partial \log L(\theta;x)}{\partial \theta}\frac{\partial L(\theta;x)}{\partial \theta}\right)\mathrm{d}x &=0 \\ \int \left(\frac{\partial^2 \log L(\theta;x)}{\partial \theta^2}L(\theta;x) + \frac{\partial \log L(\theta;x)}{\partial \theta}\frac{\partial \log L(\theta;x)}{\partial \theta}L(\theta;x)\right)\mathrm{d}x &=0 \\ \mathbb{E}\left[\frac{\partial^2 \log L(\theta;x)}{\partial \theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial \log L(\theta;x)}{\partial \theta}\right)^2\right] &=0\end{aligned}$$

and the result follows.

In the case of the Poisson distribution, the Fisher information can be thus calculated as:

$$\begin{aligned}I_n = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta;\mathbf{X})}{\partial \theta^2}\right] &= \mathbb{E}\left[\frac{\partial^2}{\partial \theta^2}\sum_{i=1}^{n}\log\left(e^{-\theta}\theta^{X_i}/X_i!\right)\right] \\ &= -\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2}\left(-\theta + X_i\log\theta - \log(X_i!)\right)\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[X_i/\theta^2\right] = n/\theta.\end{aligned}$$

The Fisher information is $I_n = n/\theta$ implying a Cramer-Rao lower bound for the estimator variance of θ/n.

The variance of a Poisson random variable is θ and hence that of the sample mean of a sample of size n is θ/n. Thus the estimator obtained previously, *i.e.* the sample mean, achieves the Cramer-Rao lower bound.

6.4.1 We could view this problem as sampling from a multinomial distribution with probability vector (p_A, p_B, p_C) with $p_A = p^2$, $p_B = 2p(1-p)$ and $p_O = 1-(p_A+p_B)$.

Consulting equation 4.1 we find the density of $\mathbf{X} = (X_A, X_B, X_O))$, given parameter vector $\theta = (p_A, p_B, p_O)$ and $n = 200$ is

$$f_{\mathbf{X};\mathbf{p}}(x_A, x_B) = \frac{n!}{x_A!x_B!x_O!}p_A^{x_A}p_B^{x_B}p_O^{x_O}$$

with $x_O = n-(x_A+x_B)$.

Consequently, we may write:

$$f_{\mathbf{X};\mathbf{p}}(x_A, x_B) = \exp\left(\log(n!) - \log(x_A!x_B!x_O!) + x_A\log(p_A) + x_B\log(p_B) + x_O\log(p_O)\right)$$

Which is in the canonical form (cf. equation 6.1), with:

$$(A_1(\theta), A_2(\theta), A_3(\theta)) = (\log(p_A), \log(p_B), \log(p_O)) \qquad C(x) = -\log(x_A!x_B!x_O!)$$
$$(B_1(x), B_2(x), B_3(x)) = (x_A, x_B, x_O) \qquad D(\theta) = \log(n!)$$

The generic MLE for exponential family models can consequently be used, and we have:

$$B_i(\mathbf{x}) = \mathbb{E}\left[B_i(\mathbf{X})\right]$$

and so:

$$\begin{aligned} x_A = \mathbb{E}[X_A] &= np_A \Rightarrow \hat{p}_A = x_A/n \\ x_B = \mathbb{E}[X_B] &= np_B \Rightarrow \hat{p}_B = x_B/n \\ x_O = \mathbb{E}[X_O] &= np_O \Rightarrow \hat{p}_O = x_O/n \end{aligned}$$

a not very surprising result.

However, we wished to estimate p, not p_A, p_B and p_O. Formally, we can't use the transformation result here because we don't have a bijection.

Writing down the likelihood directly in terms of p, however, we would have obtained:

$$f_{\mathbf{X};p}(x_A, x_B) = \frac{n!}{x_A!x_B!x_O!} p^{2x_A}[2p(1-p)]^{x_B}(1-p)^{2x_O}$$

and hence:

$$L(p;\mathbf{x}) \propto p^{2x_A+x_B}(1-p)^{2x_O+x_B}$$

And differentiating with respect to the parameter yield:

$$\frac{\partial}{\partial p}L(p;\mathbf{x}) \propto (2x_A+x_B)p^{2x_A+x_B-1}(1-p)^{2x_O+x_B} - (2x_O+x_B)p^{2x_A+x_B}(1-p)^{2x_O+x_B-1}$$

setting this equal to zero:

$$\begin{aligned} (2x_A+x_B)p^{2x_A+x_B-1}(1-p)^{2x_O+x_B} &= (2x_O+x_B)p^{2x_A+x_B}(1-p)^{2x_O+x_B-1} \\ (2x_A+x_B)(1-p) &= (2x_O+x_B)p \\ 2x_A+x_B &= (2x_O+x_B+2x_A+x_B)p & \Rightarrow p &= \frac{2x_A+x_B}{2(x_A+x_B+x_O)} \\ && &= \frac{2x_A+x_B}{2n}. \end{aligned}$$

In the case of the data here, we obtain:

$$\hat{p} = \frac{2\times 60+130}{2\times 200} = 5/8 \approx 0.625.$$

Checking the sign of the second derivative:

$$\begin{aligned} \frac{\partial^2}{\partial p^2}L(p;\mathbf{x}) \propto & (2x_A+x_B)(2x_A+x_B-1)p^{2x_A+x_B-2}(1-p)^{2x_O+x_B} \\ & - 2(2x_A+x_B)(2x_O+x_B)p^{2x_A+x_B-1}(1-p)^{2x_O+x_B-1} \\ & + (2x_O+x_B)(2x_O+x_B-1)p^{2x_A+x_B}(1-p)^{2x_O+x_B-2} \\ \propto & (2x_A+x_B)(2x_A+x_B-1)(1-p)^2 + (2x_O+x_B)(2x_O+x_B-1)p^2 \\ & - 2(2x_O+xB)(2x_A+x_B)p(1-p) \qquad < 0. \end{aligned}$$

6.5.1 This is just a multivariate version of exercise 6.3.1.

Begin by noting that if we set $\nabla = (\partial_1, \ldots, \partial_p)^T$ then we may write the Fisher Information Matrix as

$$\begin{aligned} I(\theta) =& \mathsf{Cov}(\nabla \log f(X;\theta), \nabla \log f(X;\theta)) \\ =& \mathbb{E}\left[(\nabla \log f(X;\theta))(\nabla \log(f(X;\theta))^T\right] - \mathbb{E}\left[\nabla \log f(X;\theta)\right] \mathbb{E}\left[\nabla \log f(X;\theta)\right]^T \\ =& \mathbb{E}\left[(\nabla \log f(X;\theta))(\nabla \log(f(X;\theta))^T\right] \end{aligned}$$

where the rightmost term in the second line is identically zero, just as in the univariate case.

Proving this collection of equalities now amounts to showing that the result of exercise 6.3.1 extends to this multivariate setting.

Considering $\mathbb{E}\left[\nabla^T \ell\right]$:

$$\begin{aligned} \mathbb{E}\left[\nabla^T \ell\right] =& \int L(\theta;x) \nabla^T \log L(\theta;x) \mathrm{d}x \\ =& \int \nabla^T L(\theta;x) \mathrm{d}x \\ =& \nabla^T \int L(\theta;x) \mathrm{d}x = 0. \end{aligned}$$

Differentiating both sides (using the product rule within the integral) just as in the univariate case but using the integral operator ∇ we obtain the Hessian matrix, the matrix of second derivatives:

$$\begin{aligned} \nabla \mathbb{E}\left[\nabla^T \ell(\theta;X)\right] =& 0 \\ \nabla \int L(\theta;x) \nabla^T \log L(\theta;x) \mathrm{d}x =& 0 \\ \int \nabla \left(L(\theta;x) \nabla^T \log L(\theta;x)\right) \mathrm{d}x =& 0 \\ \int \left(\nabla\nabla^T \log L(\theta;x) + (\nabla L(\theta;x))(\nabla^T \log L(\theta;x))\right) \mathrm{d}x =& 0 \\ \int \left(\nabla\nabla^T \log L(\theta;x) + (\nabla \log L(\theta;x))(\nabla^T \log L(\theta;x))\right) L(\theta;x) \mathrm{d}x =& 0 \\ \mathbb{E}\left[\nabla\nabla^T \log L(\theta;x)\right] + \mathbb{E}\left[(\nabla \log L(\theta;x))(\nabla^T \log L(\theta;x)\right] =& 0 \end{aligned}$$

and the result follows.

It's also possible to prove the inequalities one term at a time, but the matrix form is perhaps a little more elegant.

6.5.2 In this case:

$$\begin{aligned} L(\theta;x) =& \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2}(x-\mu)^2/\sigma^2) \\ \ell(\theta;x) =& -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2}(x-\mu)^2/\sigma^2 \end{aligned}$$

The *score vector* (i.e. the vector of partial derivatives of the log-likelihood) has components:

$$\frac{\partial \ell}{\partial \mu} = (x-\mu)/\sigma^2 \qquad \frac{\partial \ell}{\partial \sigma} = -\frac{1}{\sigma} + (x-\mu)^2/\sigma^3$$

The collection of second derivatives is:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -1/\sigma^2 \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = \frac{1}{\sigma^2} - 3(x-\mu)^2/\sigma^4 \qquad \frac{\partial^2 \ell}{\partial \sigma \partial \mu} = -2(x-\mu)/\sigma^3$$

Yielding:

$$I(\theta) = -\mathbb{E}\left[\begin{array}{cc} -1/\sigma^2 & -2(X-\mu)/\sigma^3 \\ -2(X-\mu)/\sigma^3 & 1/\sigma^2 - 3(X-\mu)^2/\sigma^4 \end{array}\right] = \left[\begin{array}{cc} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{array}\right]$$

This has the interesting property that it is diagonal. This tells us that the parameters are *orthogonal.*

You can also consider the parameterisation as being (μ, σ^2) in which case a slightly different expression is obtained.

6.5.3 If $g(\mu, \sigma) = \sigma^2$, then:

$$\mathbf{B} = \left[\frac{\partial g}{\partial \mu}, \frac{\partial g}{\partial \sigma}\right] = [0, 2\sigma]$$

The asymptotic variance in question is, by theorem 6.5:

$$[0, 2\sigma]\mathbf{I}^{-1}\left[\begin{array}{c} 0 \\ 2\sigma \end{array}\right]$$

From exercise 6.5.2 we have that

$$\mathbf{I} = \left[\begin{array}{cc} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{array}\right]$$

and consequently that

$$\mathbf{I}^{-1} = \left[\begin{array}{cc} \sigma^{2} & 0 \\ 0 & \frac{1}{2}\sigma^{2} \end{array}\right]$$

Hence, the asymptotic variance of interest is:

$$(2\sigma)\left(\frac{1}{2}\sigma^2\right)(2\sigma) = 2\sigma^4.$$

6.7.1 Broadly the same strategy applies to most of these cases.

(a) In the $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, 1)$ case $\theta = \mu$:

$$\begin{aligned} L(\mu; \mathbf{x}) =& (2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ \ell(\mu; \mathbf{x}) =& \text{constant} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ \frac{\partial \ell}{\partial \mu}(\mu; \mathbf{x}) =& \sum_{i=1}^{n}(x_i-\mu) \\ \frac{\partial^2 \ell}{\partial \mu^2}(\mu; \mathbf{x}) =& -n \end{aligned}$$

Setting the derivative equal to zero, we obtain $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the second derivative is negative confirming that this is a maximum.

(b) In the $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma)$ case $\theta = (\mu, \sigma)^T$:

$$\begin{aligned}
L(\mu,\sigma;\mathbf{x}) =&(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right)\\
\ell(\mu,\sigma;\mathbf{x}) =&\text{constant} - n\log\sigma - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\\
\frac{\partial\ell}{\partial\mu}(\mu,\sigma;\mathbf{x}) =&\sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2}\\
\frac{\partial\ell}{\partial\sigma}(\mu,\sigma;\mathbf{x}) =&-n/\sigma+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3}\\
\frac{\partial^2\ell}{\partial\mu^2}(\mu,\sigma;\mathbf{x}) =&-n/\sigma^2\\
\frac{\partial^2\ell}{\partial\sigma^2}(\mu,\sigma;\mathbf{x}) =&n/\sigma^2 - 3\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^4}\\
\frac{\partial^2\ell}{\partial\mu\partial\sigma}(\mu,\sigma;\mathbf{x}) = \frac{\partial^2\ell}{\partial\sigma\partial\mu}(\mu,\sigma;\mathbf{x}) =&-2\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^3}
\end{aligned}$$

Setting the derivative with respect to μ equal to zero, we obtain $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. And considering the derivative with respect to σ, evaluated using the $\hat{\mu}$ estimate of μ, we obtain $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2}$. Both of the pure second derivatives are negative at these values and the determinant of the Hessian (*i.e.* the matrix of second derivatives) at this point is:

$$\begin{aligned}
&-\frac{n}{\hat{\sigma}^2}\left[\frac{n}{\hat{\sigma}^2} - 3\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\hat{\sigma}^4}\right] - \underbrace{\left[-2\sum_{i=1}^{n}\frac{x_i-\hat{\mu}}{\hat{\sigma}^3}\right]^2}_{=0}\\
=&-\frac{n}{\hat{\sigma}^2}\left[\frac{n}{\hat{\sigma}^2}(1-3)\right]\\
=&+2n^2/\hat{\sigma}^4 > 0
\end{aligned}$$

The Hessian result tells us that this is either a maximum or a minimum; as both pure second derivatives are negative we can conclude that it's a maximum point.

(c) In the $X_i \overset{\text{iid}}{\sim} \mathsf{Exp}(\lambda)$ case:

$$\begin{aligned}
L(\lambda;\mathbf{x}) =&\prod_{i=1}^{n}\lambda\exp(-\lambda x_i) & \ell(\lambda;\mathbf{x}) =&n\log\lambda - \sum_{i=1}^{n}\lambda x_i\\
\frac{\partial\ell}{\partial\lambda}(\lambda;\mathbf{x}) =&n/\lambda - \sum_{i=1}^{n}\mathbf{x}_i & \frac{\partial^2\ell}{\partial\lambda^2}(\lambda;\mathbf{x}) =&-n/\lambda^2
\end{aligned}$$

Setting the first derivative equal to zero we obtain $\hat{\lambda} = \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]$ (*i.e.* the reciprocal of the sample mean) and checking the second derivative we find that this is indeed a maximum.

(d) If $X_i \overset{\text{iid}}{\sim} \mathsf{Poi}(\lambda)$, then:

$$\begin{aligned}
L(\lambda;\mathbf{x}) =&\prod_{i=1}^{n}\frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} & \ell(\lambda;\mathbf{x}) =&-n\lambda + \sum_{i=1}^{n}\left[x_i\log\lambda - \log(x_i!)\right]\\
\frac{\partial\ell}{\partial\lambda}(\lambda;\mathbf{x}) =&-n + \frac{1}{\lambda}\sum_{i=1}^{n}x_i & \frac{\partial^2\ell}{\partial\lambda^2}(\lambda;\mathbf{x}) =&-\frac{1}{\lambda^2}\sum_{i=1}^{n}x_i
\end{aligned}$$

Again, we simply set the first derivative equal to zero to obtain $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$, the sample mean again, and verify that the second derivative is negative.

(e) If $X_i \overset{\text{iid}}{\sim} \mathsf{Ber}\,(p)$, then:

$$L(p;\mathbf{x}) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \qquad \ell(p;\mathbf{x}) = \log p \sum_{i=1}^n x_i + \log(1-p)\left[n - \sum_{i=1}^n x_i\right]$$

$$\frac{\partial \ell}{\partial p}(p;\mathbf{x}) = \frac{1}{p}\sum_{i=1}^n x_i - \frac{1}{1-p}\left[n - \sum_{i=1}^n x_i\right] \quad \frac{\partial^2 \ell}{\partial p^2}(p;\mathbf{x}) = -\frac{1}{p^2}\sum_{i=1}^n x_i - \frac{1}{(1-p)^2}\left[n - \sum_{i=1}^n x_i\right]$$

Letting $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and setting the first derivative equal to zero:

$$n\bar{x}/p = n(1-\bar{x})/(1-p)$$
$$\bar{x}(1-p) = (1-\bar{x})p \Rightarrow \hat{p} = \bar{x}$$

i.e. the maximum likelihood estimator of a Bernoulli parameter is simply the observed proportion of successes. The second derivative is again negative, confirming that this is a maximum.

Actually, we must be careful here. If $\bar{x} = 0$ or $\bar{x} = 1$ then calculus does not give us an answer to this problem. In that case the likelihood is monotone in p and is maximised at the boundary of the parameter space with $p = 0$ or $p = 1$, respectively. Although this coincides with the $p = \bar{x}$ solution above, the argument given above does not hold in that case.

(f) If $X_i \overset{\text{iid}}{\sim} \mathsf{U}[1, b]$ then the situation is slightly more interesting. Now:

$$L(b;\mathbf{x}) = \prod_{i=1}^n \frac{1}{b-1}\mathbb{I}_{[1,b]}(x_i)$$

$$\text{where } \mathbb{I}_{[1,b]}(x) = \begin{cases} 1 & \text{if } x \in [1,b] \\ 0 & \text{otherwise} \end{cases}$$

If $\min(x_1, \ldots, x_n) < 1$ then the likelihood is identically zero and there is no meaningful MLE so we shall assume that the smallest observation is at least 1 (otherwise the data contradicts the model). As the likelihood achieves a value of zero for some parameters, taking the logarithm of the likelihood doesn't much help: it simply introduces values of $-\infty$ whenever the likelihood is zero.

Similarly, differentiating with respect to the parameters isn't very sensible: the likelihood is not everywhere differentiable: it may be written in the form:

$$L(b;\mathbf{x}) = b^{-n}\prod_{i=1}^n \mathbb{I}_{[1,b]}(x_i)$$
$$= b^{-n}\mathbb{I}_{[1,\infty)}(\min(\mathbf{x}))\mathbb{I}_{(-\infty,b]}(\max(\mathbf{x}))$$

If the largest value observed is at most b then, under our assumption that no observation is inferior to 1, the likelihood is b^{-n} which is a strictly decreasing function of b. Otherwise the likelihood is zero. Consequently, the likelihood is maximised by taking for b the smallest value such that no observation exceeds b: $\hat{b} = \max(\mathbf{x})$.

This case illustrates the importance of remembering what we're trying to achieve rather than following formulaic strategies without thinking!

6.7.2 If $X_i \overset{\text{iid}}{\sim} f(\cdot)$ for $i = 1 \ldots 17$ with $f(x;\alpha) = \exp(\alpha - x)$ for $x \geq \alpha$:

(a) The likelihood is $L(\alpha;\mathbf{x}) = \prod_{i=1}^n \exp(\alpha - x_i)\mathbb{I}_{[\alpha,\infty)}(x_i)$ hence provided that $\alpha \leq \min(\mathbf{x})$:

$$\ell(\alpha;\mathbf{x}) = \sum_{i=1}^n (\alpha - x_i)$$

which is an increasing function of α. The maximiser is consequently the largest value of alpha compatible with the constraint that $\alpha \leq \min(\mathbf{x})$ *i.e.* $\hat{\alpha} = \min(\mathbf{x})$.

In the case of this data, $\hat{\alpha} = 3.1$.

(b) The distribution $f(x;\alpha)$ has mean $\alpha+1$:

$$\begin{aligned}\mathbb{E}[X] =& \int_{\alpha}^{\infty} x\exp(\alpha-x)dx\\ =& \int_{0}^{\infty}(x'+\alpha)\exp(-x')dx'\\ =&\alpha+\int_{0}^{\infty} x'\exp(-x')dx'\\ =&\alpha+1.\end{aligned}$$

The MLE is therefore $\hat{\bar{x}} = 1+\min(\mathbf{x})$ which in the case of the current data set is 4.1.

(c) The probability that an oil change takes more than 5 minutes in the case of this particular model is, provided that $\alpha < 5$:

$$\begin{aligned}\mathbb{P}\left(X\geq 5\right) = \int_{5}^{\infty} f_X(x)dx =& \int_{5}^{\infty}\exp(\alpha-x)dx\\ =& \int_{5-\alpha}^{\infty}\exp(-x')dx' \text{ substituting } x'=x-\alpha\\ =& \left[-\exp(-x')\right]_{5-\alpha}^{\infty} = \exp(\alpha-5)\end{aligned}$$

and the MLE of this transformation of the parameter α is the transformation of the MLE, $\hat{\alpha}$:

$$\exp(\hat{\alpha}-5) = \exp(\min(\mathbf{x})-5)$$

If $\hat{\alpha} < 5$ then this is correct; otherwise the assumption above is violated and the MLE of this probability is trivially 1 as the model in that case assigns a probability of 0 to any oil change taking less than some $\alpha > 5$ minutes.

7.2.1 In this case:

$$\begin{aligned}L(\theta_0;\mathbf{x}) =& \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_i^2}{2}\right)\\ L(\theta_1;\mathbf{x}) =& \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i-1)^2}{2}\right)\end{aligned}$$

The likelihood ratio is consequently:

$$\begin{aligned}\lambda(\mathbf{x}) = \frac{L(\theta_1;\mathbf{x})}{L(\theta_0;\mathbf{x})} =& \prod_{i=1}^{n}\frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_i^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i-1)^2}{2}\right)}\\ =& \exp\left(\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-1)^2-x_i^2\right]\right)\\ =& \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[x_i^2-2x+1-x_i^2\right]\right)\\ =& \exp\left[\frac{1}{2}n(2\bar{x}-1)\right] \qquad =\exp\left(n\left(\bar{x}-\frac{1}{2}\right)\right).\end{aligned}$$

$\lambda(\boldsymbol{X})$ will typically take larger values if H_1 is true than if H_0 is true; the critical region for $T(\mathbf{X}) = \bar{X}$ is $C_\alpha = [x_\alpha^\star, \infty)$ where C_α is chosen such that $\mathbb{P}(T(\mathbf{X}) \in C_\alpha) = \alpha$.

Under H_0: $T(\mathbf{X}) = \bar{X} \sim \mathsf{N}\left(0, 1/n\right)$ and so:

$$\begin{aligned}\mathbb{P}(T(\mathbf{X}) \geq x) =&\mathbb{P}(\sqrt{n}\bar{X} \geq \sqrt{n}x)\\ =&1 - \Phi(\sqrt{n}x)\end{aligned}$$

At the critical value, this probability is equal to α:

$$\begin{aligned}\alpha =&1 - \Phi(\sqrt{n}x^\star_\alpha)\\ \sqrt{n}x^\star_\alpha =&\Phi^{-1}(1-\alpha)\\ \mathbf{x}^\star_\alpha =&\Phi^{-1}(1-\alpha)/\sqrt{n}.\end{aligned}$$

7.3.1 Consider the likelihood ratio for any of these three scenarios:

$$\begin{aligned}\lambda_{\sigma^2}(\mathbf{x}) =&\frac{L(\sigma^2;\mathbf{x})}{L(1;\mathbf{x})}\\ =&\prod_{i=1}^{n} \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{x_i^2}{\sigma^2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x_i^2\right)}\\ =&\sigma^{-n}\exp\left(\frac{1}{2}\sum_{i=1}^{n} x_i^2(1-\sigma^{-2})\right)\end{aligned}$$

This ratio exceeds unity if:

$$\begin{aligned}0 < \log\lambda_{\sigma^2}(\mathbf{x}) =&-n\log(\sigma) + \frac{1}{2}(1-\sigma^{-2})n\overline{x^2}\\ =&n\left[\overline{x^2}(1-\sigma^{-2})/2 - \log(\sigma)\right]\\ \Rightarrow 0 <&\overline{x^2}(1-\sigma^{-2})/2 - \log(\sigma)\end{aligned}$$

where $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$. Under the null hypothesis, $\overline{x^2}$ is a realisation of a χ^2_n random variable.

(a) For any $\sigma^2 > 1$ the first term is positive and the second negative; the inequality holds if:

$$\overline{x^2} > \log(\sigma^2)/(1-\sigma^{-2})$$

and the likelihood ratio is an increasing function of $\overline{x^2}$. Thus we need a value $\widehat{x^2}$ such that $\mathbb{P}(\overline{X^2} > \widehat{x^2}|H_0) = \alpha$. As $\overline{X^2} \sim \chi^2_n$ under the null hypothesis:

$$\alpha =\mathbb{P}(\overline{X^2} > \widehat{x^2}|H_0) \Rightarrow \widehat{x^2} = \chi^2_{n,1-\alpha}$$

and the critical region becomes $[\chi^2_{n,1-\alpha}, \infty)$.
For any $\sigma > 1$ the value of the likelihood ratio is larger at every point in this region than it is at any point outside it.

(b) In the $\sigma^2 < 1$ case the first term is negative and the second positive; the inequality holds if:

$$\begin{aligned}\overline{x^2}(1-\sigma^{-2}) >&\log(\sigma^2)\\ \overline{x^2} <&-\log(\sigma^2)/(\sigma^{-2}-1)\end{aligned}$$

and the likelihood ratio is a decreasing function of $\overline{x^2}$.
Thus we need a value $\widehat{x^2}$ such that $\mathbb{P}(\overline{X^2} < \widehat{x^2}|H_0) = \alpha$. As $\overline{X^2} \sim \chi^2_n$ under the null hypothesis:

$$\alpha =\mathbb{P}(\overline{X^2} < \widehat{x^2}|H_0) \Rightarrow \widehat{x^2} = \chi^2_{n,\alpha}$$

and the critical region becomes $[0, \chi^2_{n,\alpha}]$.
For any $\sigma < 1$ the value of the likelihood ratio is larger at every point in this region than it is at any point outside it.

(c) In this case, the most powerful tests for alternatives $\sigma < 1$ and $\sigma > 1$ have different critical regions and no test exists which achieves equal power for both hypotheses which is most powerful for either.

7.4.1 Under the assumption that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu, \sigma^2\right)$:

$$\begin{aligned}
\bar{X} =& \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathsf{N}\left(\mu, \sigma^2/n\right) \\
(n-1)S^2/\sigma^2 =& \sum (X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1} \\
\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) \Big/ \left(\frac{\sqrt{n-1}S/\sigma}{\sqrt{n-1}}\right) &\sim \mathsf{t}_{n-1} \\
\left(\frac{\bar{X}-\mu}{S/\sqrt{n}}\right) &\sim \mathsf{t}_{n-1}
\end{aligned}$$

Where we have used the fact that the sample average of n iid normal random variates is a normal random variate of the same mean as that of the individual samples and a variance of $1/n-1$ times that of the total (see definition 5.2).

Consequently

$$T =\bar{X}/(S/\sqrt{n})$$

is the sum of μ and a t_{n-1} random variable. This distribution is sometimes referred to as the non-central t distribution of $n-1$ degrees of freedom and non-centrality parameter μ.

The likelihood under $H_0 : \mu = 0$ is:

$$\begin{aligned}
L_0(\theta_0; x) =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x_i^2}{2\sigma^2}\right) \\
\ell_0(\theta_0; x) =& \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma}\sum_{i=1}^{n} x^2
\end{aligned}$$

Whilst that under an alternative hypothesis, $\mu \neq 0$, is:

$$\begin{aligned}
L_1(\theta_1; x) =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
\ell_1(\theta_1; x) =& \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

Under H_0, the maximum likelihood estimate of σ^2, $\hat{\sigma_0^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

Under H_1, the maximum likelihood estimate of μ, $\hat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and that of σ^2,

$$\hat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu_1})^2.$$

Hence the likelihood ratio test statistics is:

$$
\begin{aligned}
r(x) =&\ell_1(\hat{\theta}_1;x) - \ell_0(\hat{\theta}_0;x)\\
=&-\frac{n}{2}(\log(\hat{\sigma}_1^2) - \log\widehat{\sigma_0^2}) + \frac{1}{2\widehat{\sigma_0^2}}\sum_{i=1}^{n} x_i^2 - \frac{1}{2\widehat{\sigma_1^2}}\sum_{i=1}^{n}(x_i - \hat{\mu_1})^2\\
=&-\frac{n}{2}\log\left(\frac{(n-1)S^2/n - \bar{x}^2}{(n-1)S^2/n}\right) + \frac{(n-1)S^2}{2(n-1)S^2/n} - \frac{1}{2}\frac{(n-1)S^2 - n\bar{x}^2}{(n-1)S^2/n - \bar{x}^2}\\
=&-\frac{n}{2}\log\left(\frac{(n-1)S^2 - n\bar{x}^2}{(n-1)S^2}\right) + \frac{n}{2} - \frac{n}{2}\\
=&-\frac{n}{2}\log\left(\frac{(n-1)S^2 - n\bar{x}^2}{(n-1)S^2}\right)\\
=&-\frac{n}{2}\log\left(1 - \frac{n\bar{x}^2}{(n-1)S^2}\right) = -\frac{n}{2}\log(1-T^2)
\end{aligned}
$$

This is a monotonically increasing function of T; hence rejecting H_0 on the basis of a large value of the log-likelihood ratio statistic is equivalent to rejecting on the basis of a large value of the T statistic. This test *is* a likelihood ratio test.

Under the stated null hypothesis, we have that $T \sim \mathsf{t}_{14}$. With the data provided for the change in diastolic blood pressure in table 7.1):

$$
n = 15 \qquad \sum_{i=1}^{15} x_i = -139 \qquad \sum_{i=1}^{15} x_i^2 = 2327
$$

Hence we calculate:

$$
\bar{x} = -139/15 = -9.27 \qquad s = \sqrt{2327/14} = 12.89 \qquad t = \frac{-9.27}{12.89/\sqrt{15}} = -2.78
$$

A t_{14} distribution has probability 0.0073 of producing a value less than this and a commensurate probability of producing a positive value of at least this magnitude. Hence we reject the null hypothesis in favour of H_1 at the 5% significance level.

7.4.2 (a) The log-likelihood under the null hypothesis is:

$$
\begin{aligned}
\ell_0(\mu_X,\sigma^2;\mathbf{x},\mathbf{y}) =&\text{const.} + \sum_{i=1}^{m}\left[-\frac{1}{2}\log\sigma^2 - \frac{(x_i-\mu_X)^2}{2\sigma^2}\right] + \sum_{i=1}^{n}\left[-\frac{1}{2}\log\sigma^2 - \frac{(y_i-\mu_X)^2}{2\sigma^2}\right]\\
=&c - \frac{m+n}{2}\log\sigma - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(X_i-\mu_X)^2 + \sum_{i=1}^{n}(Y_i-\mu_X)^2\right]
\end{aligned}
$$

Whilst under the alternative hypothesis, $\mu_Y \neq \mu_X$ and:

$$
\begin{aligned}
\ell_1(\mu_X,\mu_Y,\sigma^2;\mathbf{x},\mathbf{y}) =&\text{const.} + \sum_{i=1}^{m}\left[-\frac{1}{2}\log\sigma^2 - \frac{(x_i-\mu_X)^2}{2\sigma^2}\right] + \sum_{i=1}^{n}\left[-\frac{1}{2}\log\sigma^2 - \frac{(y_i-\mu_Y)^2}{2\sigma^2}\right]\\
=&c - \frac{m+n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(X_i-\mu_X)^2 + \sum_{i=1}^{n}(Y_i-\mu_Y)^2\right]
\end{aligned}
$$

The log-likelihood-ratio test statistic is:

$$
2r(\mathbf{x},\mathbf{y}) = 2\left[\ell_1(\hat{\mu}_{X,1},\hat{\mu}_{Y,1},\hat{\sigma}_1;\mathbf{x},\mathbf{y}) - \ell_0(\hat{\mu}_{X,0},\hat{\sigma}_0;\mathbf{x},\mathbf{y})\right]
$$

By standard arguments:

$$\hat{\mu}_{X,0} = \frac{1}{m+n}\left[\sum_{i=1}^{m} X_i + \sum_{i=1}^{n} Y_i\right] \qquad \hat{\mu}_{X,1} = \frac{1}{m}\left[\sum_{i=1}^{m} X_i\right] \qquad \hat{\mu}_{Y,1} = \frac{1}{n}\left[\sum_{i=1}^{n} Y_i\right]$$

and setting the derivatives of the log-likelihoods to zero we can easily obtain:

$$\begin{aligned}
\hat{\sigma}_0^2 =& \frac{1}{m+n}\left[\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,0})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{X,0})^2\right] \\
\hat{\sigma}_1^2 =& \frac{1}{m+n}\left[\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,1})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{Y,1})^2\right]
\end{aligned}$$

Hence (noting that the respective sums of squares divided by estimated variances are equal under the two hypotheses):

$$\begin{aligned}
r(\mathbf{x}, \mathbf{y}) =& -2\left[\frac{m+n}{2}(\log(\hat{\sigma}_1^2) - \log(\hat{\sigma}_0^2)) + \right] \\
=& (m+n)\log(\hat{\sigma}_0^2/\hat{\sigma}_1^2) \\
=& (m+n)\log\left[\frac{\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,0})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{X,0})^2}{\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,1})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{Y,1})^2}\right] \\
=& (m+n)\log\left[1 + \frac{\left(\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,0})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{X,0})^2\right) - \left(\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,1})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{Y,1})^2\right)}{\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,1})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{Y,1})^2}\right]
\end{aligned}$$

Now, noting that $\hat{\mu}_{X,0} = (m\bar{X} + n\bar{Y})/(m+n)$, that $\hat{\mu}_{X,1} = \bar{X}$ and that $\hat{\mu}_{Y,1} = \bar{Y}$ we can expand the numerator of the fraction as:

$$\begin{aligned}
&\left(\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,0})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{X,0})^2\right) - \left(\sum_{i=1}^{m}(X_i - \hat{\mu}_{X,1})^2 + \sum_{i=1}^{n}(Y_i - \hat{\mu}_{Y,1})^2\right) \\
=& \sum_{i=1}^{m}\left[(X_i - \hat{\mu}_{X,0})^2 - (X_i - \hat{\mu}_{X,1})^2\right] + \sum_{i=1}^{n}\left[(Y_i - \hat{\mu}_{X,0})^2 - (Y_i - \hat{\mu}_{Y,1})^2\right] \\
=& \sum_{i=1}^{m}\left[\left(X_i - \left[\frac{m\bar{X} + n\bar{Y}}{m+n}\right]\right)^2 - (X_i - \bar{X})^2\right] + \sum_{i=1}^{n}\left[(Y_i - \left[\frac{m\bar{X} + n\bar{Y}}{m+n}\right])^2 - (Y_i - \bar{Y})^2\right] \\
=& m\left[\left[\frac{m\bar{X} + n\bar{Y}}{m+n}\right]^2 - 2\bar{X}\left[\frac{m\bar{X} + n\bar{Y}}{m+n}\right] - \bar{X}^2 + 2\bar{X}\bar{X}\right] + \\
& n\left[\left[\frac{m\bar{X} + n\bar{Y}}{m+n}\right]^2 - 2\bar{Y}\left[\frac{m\bar{X} + n\bar{Y}}{m+n}\right] - \bar{Y}^2 + 2\bar{Y}\bar{Y}\right] \\
=& m\bar{X}^2 + n\bar{Y}^2 - \frac{m^2\bar{X}^2 + n^2\bar{Y}^2 + 2mn\bar{X}\bar{Y}}{m+n} \\
=& \frac{mn(\bar{X}^2 + \bar{Y}^2) - 2mn\bar{X}\bar{Y}}{m+n} = \frac{mn}{m+n}(\bar{X} - \bar{Y})^2
\end{aligned}$$

whilst identifying the denominator as $(m+n-2)S_p$ where

$$S_p^2 = \frac{\sum_{i=1}^{m}(X_i - \bar{x})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{m+n-2}.$$

we find that:

$$2r(\mathbf{x},\mathbf{y}) = \log\left[1+T^2\right]$$

with T as defined in equation 7.3. As the LRT statistic depends upon the data only through T, this statistic contains all of the information provided by the data about our test.

(b) Under H_0, $\sum_{i=1}^m (X_i - \bar{X})^2$ and $\sum_{i=1}^n (Y_i - \bar{Y})^2$ are independent of one another and by theorem 5.3 we know that the first has a χ^2_{m-1} distribution and the second has a χ^2_{n-1} distribution. By exercise 5.2.3 the sum of two chi-squared random variables is itself a chi-squared random variable with a number of degrees of freedom equal to the sum of the number of degrees of freedom of the individual random variables.

(c) This is simple calculation:

$$\begin{aligned}
\mathbf{x} =& (134, 146, 104, 119, 124, 161, 107, 83, 113, 129, 97, 123)^T & m =& 12\\
\mathbf{y} =& (70, 118, 101, 85, 107, 132, 94)^T & n =& 7\\
\bar{x} =& 120 & \sum_i (x_i - \bar{x})^2 =& 5032\\
\bar{y} =& 101 & \sum_i (y_i - \bar{y})^2 =& 2552\\
S_p =& \sqrt{7584/17} = 21.1215
\end{aligned}$$

Finally,

$$\begin{aligned}
T =& \frac{\bar{x} - \bar{y}}{\sqrt{\frac{m+n}{mn}}S_p}\\
=& \frac{120 - 101}{\sqrt{19/84} \cdot 21.1215}\\
=& \frac{19}{10.0453} \approx 1.8914
\end{aligned}$$

The probability that a t_{17} random variable exceeds 1.8914 is 0.0379.
At the 5% level this wouldn't justify rejecting the null hypothesis as there is an equal probability that the difference would be at least as large in the opposite sense.

7.4.3 (a) Under H_0, S_X^2/σ_X^2 is a χ^2_{m-1} random variable and S_Y^2/σ_Y^2 is a χ^2_{n-1} random variable (this follows directly from theorem 5.2).

(b) As, under H_0, S_X^2 and S_Y^2 are independent chi-squared random variables, the ratio of their renormalised forms, $V = (S_X^2/(m-1))/(S_Y^2/(n-1))$ follows a $\mathsf{F}_{m-1,n-1}$ from definition 5.3.

(c) From the supplied data we can calculate:

$$\begin{aligned}
S_X^2 =& \sum_{i=1}^m x_i^2 - \left(\sum_{i=1}^m x_i\right)^2/m\\
=& 563 - 84^2/16 = 122\\
S_Y^2 =& 72 - 18^2/16 = 51.75\\
V =& \frac{122/16}{51.75/16} = 2.3575
\end{aligned}$$

At the 5% significance level, we reject the null hypothesis if V doesn't lie between $\mathsf{F}_{0.025;15,15} = 0.3494$ and $\mathsf{F}_{0.975;15,15} = 2.861$. As V lies between these two values, we do not have sufficient evidence to reject the null hypothesis.

7.4.4 Under H_0 the parameter space is degenerate with dimension 0; under H_1 there is a univariate parameter space and $\hat{\theta} = \bar{x}$:

$$\begin{aligned}
2r(\mathbf{x}) =& 2\ell_1(\mathbf{x};\hat{\theta}) - 2\ell_0(\mathbf{x}) \\
=& \left[-\frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2}\right] - \left[-\frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^n x_i^2}{2}\right] \\
=& \sum_{i=1}^n x_i^2 - (x_i - \bar{x})^2 = 2\bar{x}\sum_{i=1}^n x_i - n\bar{x}^2 = n\bar{x}^2
\end{aligned}$$

Under H_0, $\sqrt{n}\bar{x}$ is a standard normal random variable and so $n\bar{x}^2$ is a χ^2_1 random variable: Wald's theorem holds exactly.

7.4.5 Under the null hypothesis, we have a univariate parameter space and $\hat{\theta} = \bar{x} = \sum_{i=1}^n x_i/n$.

Under the alternative hypothesis we have a parameter for every observation and the full parameter space is n dimensional. The MLE for the parameters is trivially: $\hat{\theta}_i = x_i$.

We can calculate the log likelihood ratio test statistic directly:

$$\begin{aligned}
2r(x) =& 2\left[\ell_1(\theta_1, \dots, \theta_n; \mathbf{x}) - \ell_0(\theta; \mathbf{x})\right] \\
=& \left[-\frac{n}{2}\log(2\pi) - \sum_{i=1}^n \frac{(x_i - x_i)^2}{2}\right] - \left[-\frac{n}{2}\log(2\pi) - \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{2}\right] \\
=& \sum_{i=1}^n (x_i - \bar{x})^2.
\end{aligned}$$

Under the null hypothesis the x_i are iidnormal random variables with unit variance. By theorem 5.3 this is a χ^2_{n-1} random variable: Wald's theorem holds exactly.

7.5.1 Under H_0 the only unknown parameter is μ and $\hat{\mu}_0 = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ by the usual argument.

Under H_1 there are two unknown parameters, μ and σ with estimators $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$.

Substituting these standard estimators into the log-likelihood ratio test statistic:

$$\begin{aligned}
r(\mathbf{x}) =& \ell_1(\hat{\mu}, \hat{\sigma}^2; \mathbf{x}) - \ell_0(\hat{\mu}_0; \mathbf{x}) \\
=& \left[-\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (x_i - \bar{x})^2\right] - \left[-\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\right] \\
=& -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2}(\hat{\sigma}^{-2} - 1)\sum_{i=1}^n (x_i - \bar{x})^2
\end{aligned}$$

Thus the test statistic of interest is:

$$\begin{aligned}
2r(\mathbf{x}) =& -n\log(\hat{\ }\widehat{\sigma^2}) - (1/\widehat{\sigma^2} - 1)n\widehat{\sigma^2} \\
=& n(\hat{\sigma}^2 - 1 - \log(\hat{\sigma}^2))
\end{aligned}$$

and the LRT will reject the null hypothesis for sufficiently large values of this statistic.

7.5.2 (a) Consider testing $H_0 : \lambda = \frac{1}{2}$ against any hypothesis of the form $H' : \lambda = \lambda', \lambda' > \frac{1}{2}$.

The simple likelihood ratio test statistic is:

$$\begin{aligned}
\Lambda(\mathbf{x}) =& \frac{L(\lambda'; \mathbf{x})}{L(\lambda; \mathbf{x})} \\
=& \prod_{i=1}^n \frac{\lambda' x_i^{-2}\exp(-\lambda'/x_i)}{\lambda x_i^{-2}\exp(-\lambda/x_i)}
\end{aligned}$$

and its logarithm is:

$$\begin{aligned}\lambda(\mathbf{x}) =& n\log(\lambda'/\lambda) - \sum_{i=1}^{n}(\lambda'-\lambda)/x_i \\ =& n\log(\lambda'/\lambda) - (\lambda'-\lambda)\sum_{i=1}^{n} 1/x_i\end{aligned}$$

The first term is independent of the data and the second is a decreasing function of $\sum_i 1/x_i$ for *any* λ'.

Consequently, the null hypothesis will be rejected in favour of any $\lambda' > 1/2$ if the data is such that $\sum_i 1/x_i \leq A$ for an appropriately chosen constant, A. As the value of A is chosen on the basis of the distribution of $\mathbf{x}$ under the null hypothesis, we will obtain the same critical region for any $\lambda' > 1/2$ and this critical region characterises a UMP test between H_0 and $H_1 : \lambda > \frac{1}{2}$.

(b) Under the null hypothesis $1/X_i$ has a distribution which can be obtained by applying theorem 4.5 with $g(x) = 1/x$, $g^{-1}(y) = 1/y$ and $\partial g^{-1}/\partial y = -1/y^2$, hence, setting $Y_i = g(X_i)$:

$$\begin{aligned}f_Y(y_i) =& f_X(g^{-1}(y_i))\left|\frac{\partial g^{-1}}{\partial y}\right|(y_i) \\ =& \lambda(1/y_i)^{-2}e^{-\lambda/(1/y_i)} \cdot 1/y^2 \\ =& \lambda y_i^2 \exp(-\lambda y_i)/y_i^2 = \exp(-\lambda y_i)\end{aligned}$$

and so, under the null hypothesis incorporating the independence of the x_i, we know that the y_i are independent, identically distributed $\mathsf{Exp}\,(1/2)$ random variables. The sum of n such random variables is a $\mathsf{Gamma}\left(n, \frac{1}{2}\right)$ random variable.

It isn't really meaningful to talk about the distribution of this quantity under the *compound* alternative hypothesis; but under any simple hypothesis compatible with H_1 it will have a $\mathsf{Gamma}\,(n, \lambda')$ distribution for some $\lambda' > \frac{1}{2}$.

(c) In the case of the data provided:

$$\begin{aligned}x =& (0.59, 0.36, 0.71, 0.86, 0.13, 0.01, 3.17, 1.18, 3.28, 0.49)^T \\ \sum_{i=1}^{n} 1/x =& 118.2449\end{aligned}$$

The probability that an $\mathsf{Gamma}\,(10, 0.5)$ random variable takes a value below 118.2449 is *extremely* close to 1. The null hypothesis cannot be rejected.

If we'd considered a double-sided alternative hypothesis we would have arrived at a very different conclusion. The 95% quantile of the $\mathsf{Gamma}\,(10, 0.5)$ distribution is 31.410 and the 99.99% quantile is 52.39. The data is incredibly improbable under the null hypothesis but isn't rejected because it's even more improbable under the alternative hypothesis.

One would perhaps be concerned that this looks like a strong suggestion that if this distribution is at all correct then $\lambda < 1/2$ would fit the data somewhat better.

7.5.3 (a) The model makes some quite strong statements, in particular:

- The model doesn't take into account major incidents: with probability 1 a bus must arrive after a time of at most θ.

If your arrival time at the bus stop is uniform over a long period of time and the buses run exactly to the timetable then this model might be reasonable; somehow the strongest assumption is that the probability of waiting for more than a time θ is exactly zero. This would require that the bus company immediately sends a new bus to the route if one is running even a second late.

(b) The likelihood is:

$$\begin{aligned}L(\theta;\mathbf{x}) =& \prod_{i=1}^{n} f(x_i;\theta)\\ =& \begin{cases} \theta^{-n} & \text{if for all } i,\ 0 < x_i < \theta \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

(c) Consider the hypothesis $H' : \theta = \theta' > \theta_0$, the SLRT between H_0 and H' involves checking:

$$\begin{aligned}\Lambda(\mathbf{x}) =& \frac{L(\theta';\mathbf{x})}{L(\theta_0;\mathbf{x})}\\ =& \frac{\theta_0^n}{\theta'^n} \prod_{i=1}^{n} \frac{\mathbb{I}_{[0,\theta']}(x_i)}{\mathbb{I}_{[0,\theta_0]}(x_i)}\end{aligned}$$

provided that all of the observations are smaller than θ_0 we can't reject the null hypothesis (if any observations lie between θ_0 and θ' we must reject the null hypothesis as the data is impossible under the null hypothesis but possible under H' and if any observation exceeds θ' the test is meaningless as the observed data is impossible under both models; in fact the test we're really conducting is against a compound alternative hypothesis so this isn't an issue).

This is interesting: the only case in which Λ is greater than 1 is the case in which there are observed values greater than θ_0. This is true for all $\theta > \theta_0$ and the probability of generating any observation exceeding θ_0 is exactly zero under the null hypothesis. Hence, a test of size $\leq \alpha$ is obtained for any α and any θ' with rejection iff the largest observation exceeds θ_0. For any other set of data, the data is *more* probable under the null hypothesis.

The rejection region is the same for all $\theta' > \theta_0$ and so the UMP test is to reject the null hypothesis if the data set includes any measured times in excess of θ.

Notice that we have been able to produce tests of size 0 and *only* of size 0 in the case of this model. These tests have a rejection region which depend on only one of the observations. This is extremely brittle; depending only upon the largest observation. This is a consequence of the extremely strong statement in the model that there is a sharp transition in the probability density to exactly zero at this point.

7.6.1 The only trick here is to notice that $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$. Hence the likelihood and its logarithm are:

$$\begin{aligned}L(\theta;\mathbf{x}) =& \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i}\\ \ell(\theta;\mathbf{x}) =& \log(n!) - \sum_{i=1}^{k} \log(x_i!) + \sum_{i=1}^{k} x_i \log \theta_i\\ \ell(\theta;\mathbf{x}) =& \text{Const.} + \sum_{i=1}^{k-1} x_i \log \theta_i + x_k \log(1 - \sum_{i=1}^{k-1} \theta_i)\end{aligned}$$

Looking at the partial derivatives with respect to the individual θ_i:

$$\frac{\partial \ell}{\partial \theta_i}(\theta;\mathbf{x}) = x_i/\theta_i - x_k/(1 - \sum_{i=1}^{k-1} \theta_I) = x_i/\theta_i - x_k/\theta_k.$$

provided that x_i and x_k are greater than zero this yields the answer (otherwise the likelihood is monotone in at least some of its arguments and a little care is required to establish that $\hat{\theta}_i = 0$ whenever $x_i = 0$, $\hat{\theta}_i = 1$ in the event that only x_i is non-zero and all other values may be obtained via calculus).

7.6.2 Under the null hypothesis, $\mathbf{Y}$ is a vector whose components each correspond to the number of a collection of n independent, identically distributed samples which take a particular value. Therefore,

$$\mathbf{Y} \sim \mathsf{Mult}\,(n, (p_1, p_2, p_3))$$

where $p_i = \mathbb{P}(X_1 = i-1)$ (as the X_i are iid we can use the first one as a prototypical example).

In this case $p_1 = (1-\phi)^3$, $p_2 = 3\phi(1-\phi)^2$ and $p_3 = 3\phi^2(1-\phi)$ and the implicit $p_4 = \phi^3$ such that $\sum\limits_{i=1}^{4} p_i = 1$.

7.7.1 (a) These terms are defined in the notes.

(b) The likelihood is:

$$\begin{aligned} L(\theta; x) =& \prod_{i=1}^{n} P(X_i = x_i|\theta) \\ =& \prod_{i=1}^{n} p_{x_i} = \prod_{j=1}^{4} \prod_{\{i:x_i=j\}} p_{x_i} = \prod_{j=1}^{4} p_j^{y_j}. \end{aligned}$$

(c) The parameter space Ω_θ is a 3 dimensional simplex in $[0,1]^4$ as $p_4 = 1 - \sum\limits_{j=1}^{3} p_j$. Hence $\nu = 3$.
Under H_0 a further two degrees of freedom are lost as $p_1 = p_2$ and $p_3 = p_4$ which with the additional constraint tells us that $p_3 = p_4 = \frac{1}{2}(1-2p_1)$.
Hence the asymptotic χ^2 distribution has 2 degrees of freedom.

(d) Let $\ell(\theta; x) = \sum\limits_{j=1}^{3} y_j \log p_j + y_4 \log(1 - (p_1 + p_2 + p_3))$.
Now:

$$\frac{\partial \ell}{\partial p_j} = \frac{y_j}{p_j} - y_4 \frac{1}{1 - (p_1 + p_2 + p_3)}$$

Setting the partial derivatives with respect to p_1, p_2 and p_3 to zero simultaneously:

$$\frac{p_j}{p_4} = \frac{y_j}{y_4}$$

assuming that all of the $y_j \neq 0$. If any of the y_j are zero then the likelihood is monotonically decreasing in p_j and setting that p_j (or those p_j) to zero will maximise the likelihood. In the most extreme case all but one of the y_j are zero and monotonicity arguments tell us that we must set the accompanying $p_j = 1$ in this case.

In all cases, $\hat{p}_j = y_j / \sum\limits_{i=1}^{4} y_i = y_j/n$

(e) Under H_0 our likelihood is further restricted and we obtain:

$$\begin{aligned} \ell_0(\theta; x) =& (y_1 + y_2) \log p_1 + (y_3 + y_4) \log(\frac{1}{2}[1 - 2p_1]) \\ \frac{\partial \ell_0}{\partial p_1} =& (y_1 + y_2)/p_1 - (y_3 + y_4)/[\frac{1}{2}(1 - 2p_1)] \end{aligned}$$

as $2p_3 = 1 - 2p_1$. We conclude that the MLE for p_1 under this model is $\hat{p}_1 = (y_1 + y_3)/2(y_1 + y_2 + y_3 + y_4) = m_1/n$ and the result for $\hat{p}_3$ is automatic.
Substituting the MLE under H_0 and H_1 into the log-likelihoods, we find the log-likelihood test statistic is:

$$\begin{aligned}2r(x) =&2[\ell(\hat{\theta};x) - \ell_0(\hat{\theta}_0;x)]\\ =&2\left[\sum_{i=1}^{4} y_i \log\left(\frac{y_j}{n}\right) - (y_1+y_3)\log\left(\frac{y_1+y_3}{2n}\right) - (y_2+y_4)\log\left(\frac{y_2+y_4}{2n}\right)\right]\\ =&2\sum_{i=1}^{4} y_i \frac{y_i/n}{m_j/n} = 2\sum_{i=1}^{4} y_i \log\left(\frac{y_i}{m_i}\right)\end{aligned}$$

with $m_1 = m_2 = (y_1+y_2)/2$ and $m_3 = m_4 = (y_3+y_4)/2$.

(f) In this case we have $y_1 = 46$ and $y_2 = 49$; whilst $y_3 = 22$ and $y_4 = 32$. $n = 149$.
The MLE under H_0 is $\hat{p}_1 = 95/149 = 0.638$ and $\hat{p}_3 = 54/149 = 0.362$.
The MLE under H_1 is $\hat{p}_1 = 46/149 = 0.309$, $\hat{p}_2 = 0.329$, $\hat{p}_3 = 0.148$ and $\hat{p}_4 = 0.215$.
The test statistic is:

$$\begin{aligned}2r(x) =&2\left[46\log(46/47.5) + 49\log(49/47.5) + 22\log(22/27) + 32\log(32/27)\right]\\ =&2[-1.4761 + 1.5234 - 4.5055 + 5.4368] = 1.9573\end{aligned}$$

The probability of obtaining a value at least this large under the asymptotic distribution of the test statistic is 0.38 and so we cannot reject the null hypothesis at, say, a 5% level.

(g) The question of independence arises: have patients with multiple fractures been admitted? The sample size isn't *very* small and so the asymptotic approximation may be justifiable.

7.7.2 We can begin by determining the probabilities we need:

$$p_{1+} = \frac{59}{200} \qquad p_{2+} = \frac{48}{200} \qquad p_{3+} = \frac{38}{200} \qquad p_{4+} = \frac{55}{200}$$

$$p_{+1} = \frac{99}{200} \qquad p_{+2} = \frac{64}{200} \qquad p_{+3} = \frac{44}{200}$$

Under H_0 the expected number of candidates in row i and column j is simply the product $200p_{i+}p_{j+}$ because of the independence assumption. In our case this yields:

		Candidate preferred		
		A	B	Undecided
Curriculum	Engineering & Science	$(59\times 92)/200$	$(59\times 64)/200$	$(59\times 44)/200$
	Humanities & Social Science	$(48\times 92)/200$	$(48\times 64)/200$	$(48\times 44)/200$
	Fine Arts	$(38\times 92)/200$	$(38\times 64)/200$	$(38\times 44)/200$
	Industrial & Public Administration	$(55\times 92)/200$	$(55\times 64)/200$	$(55\times 44)/200$

Which a little computation tells us is simply:

		Candidate preferred		
		A	B	Undecided
Curriculum	Engineering & Science	27.1400	18.8800	12.9800
	Humanities & Social Science	22.0800	15.3600	10.5600
	Fine Arts	17.4800	12.1600	8.3600
	Industrial & Public Administration	25.3000	17.6000	12.1000

Our test statistic is:

$$\begin{aligned}X^2 =&\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{((200p_{i+}p_{+j})-\text{obs}_{ij})^2}{200p_{i+}p_{+j}}\\ =&6.6849\end{aligned}$$

This should be a realisation of a chi-squared random variable with $(R-1)(C-1)=6$ degrees of freedom. A χ^2_6 random variable will exceed 6.6849 with probability 0.351. We cannot reject the null hypothesis at, say, a 5% significance level.

8.2.1 (i) Each g gives a simulated value of the number of successes observed under a $\mathsf{Bin}\,(60, 0.2)$ sampling regime. Repeating the experiment a large number of times suggests that this number is the quantity of interest and the experimenter is investigating the properties of this number of successes.

(ii) The average value should be close to the quantity being estimated:

$$g_{av} = \frac{1}{5000}\sum_{j=i}^{5000} g_i$$

Each g_i has expectation $60 \cdot \mathbb{P}(X < 20) = 60 \cdot 20/100 = 12$ and so $g_{av} \approx 12$ by the law of large numbers or similar. In fact each g_i has distribution $\mathsf{Bin}\,(60, 0.2)$ and hence expectation $60 \cdot 0.2 = 12$. Averaging 5000 random variables with this expectation and reasonably small variance we would expect to arrive at a value close to 12.

(iii) As the g_i are really $\mathsf{Bin}\,(60, 0.2)$ random variables, they each have variance $60{\cdot}0.2{\cdot}(1{-}0.2) = 9.6$. One would expect the variance of the g_i to be close to 9.6 (note that the variance of g_{av} would be a factor of 5,000 smaller than this due to the effect of averaging and the independence of the samples).

8.2.2 To deduce M, consider the ratio of the target distribution to the instrumental distribution:

$$\begin{aligned}M \geq \sup_x f(x)/g(x) =&\sup_x \frac{(2\pi)^{-1/2}\exp(-\frac{x^2}{2})}{(\pi(1+x^2))^{-1}}\\ =&\sup_x \frac{\sqrt{\pi}(1+x^2)\exp(-\frac{x^2}{2})}{\sqrt{2}}\\ =&\sqrt{\frac{\pi}{2}}\sup_x(1+x^2)\exp\left(-\frac{1}{2}x^2\right)\end{aligned}$$

Let $a(x) = (1+x^2)\exp(-x^2/2)$ and:

$$\frac{\partial a}{\partial x} = 2x\exp(-x^2/2) - x(1+x^2)\exp(-x^2/2)$$

Setting this to zero, we obtain:

$$\begin{aligned}2x\exp(-x^2/2) =&x(1+x^2)\exp(-x/2)\\ 2 =&1+x^2 \Rightarrow x = \pm 1\end{aligned}$$

At the stationary points, $x = \pm 1$. Considering the second derivative at these points:

$$\begin{aligned}\frac{\partial^2 a}{\partial x^2} =&2\exp(-x^2/2) - 2x^2\exp(-x^2/2) - (1+x^2)\exp(-x/2) - 2x^2\exp(-x^2/2)\\ &+x^2(1+x^2)\exp(-x^2/2)\\ =&\exp(-1/2)\,[2-2-2-2+2] < 0\end{aligned}$$

Indicating that these are maxima. At $x = \pm 1$, $a(x) = 2\mathrm{e}^{-1/2}$ and so M must be chosen to be at least $\sqrt{\frac{\pi}{2}}2\mathrm{e}^{-1/2} = \sqrt{2\pi/\mathrm{e}}$.

As making M larger than necessary simply increases the computational cost by increasing the probability of rejecting any sample, we choose M equal to this bound. As demonstrated in lectures, the marginal probability of a sample being accepted is simply $1/M = \sqrt{\mathrm{e}/2\pi} \approx 0.66$.

No upper bound would exist if the instrumental and target distributions were exchanged. We couldn't implement such an algorithm. This is because the Cauchy distribution has much "thicker tails" than the Normal distribution, decaying only polynomial as $|x|$ increases away from the origin, rather than exponentially.

8.3.1 In order to test the symmetry of the population using a bootstrap technique based around the sample skewness, one would need to estimate the sampling distribution of the skewness and then calculate an appropriate confidence interval for the skewness.

If we draw B bootstrap samples and calculate for each bootstrap sample it's mean, m, variance, v, and 3rd central moments, u:

$$m_i^\star = \frac{1}{n}\sum_{j=1}^{n} X_j^i \qquad v_i^\star = \frac{1}{n}\sum_{j=1}^{n}(X_j^i - m_i^\star)^2 \qquad u_i^\star = \frac{1}{n}\sum_{j=1}^{n}(X_j^i - m_i^\star)^3$$

(We have used biased estimators here; large enough samples are required to justify the bootstrap technique at all that it's not worth the additional complication arising from unbiased estimation of skewness).

We can calculate the skewness for each bootstrap sample:

$$s_i^\star = u_i^\star/(v_i^\star)^{3/2}$$

To obtain a confidence interval of $1-\alpha$ we simply take the values of the $\alpha/2B^{\text{th}}$ and $(1-\alpha/2)B^{\text{th}}$ largest values of $s_i^\star$ as the endpoints of our confidence interval.

If that confidence interval includes 0 then we cannot reject the null hypothesis (i.e. that the skewness is 0 and the distribution symmetric) at the $\alpha \cdot 100\%$ level. If the interval does not include 0 then we do reject our null hypothesis.

9.2.1 Remember that the expectation is with respect to the parameter, θ, conditional upon the value of the observed data. Given the data, the estimators are deterministic functions and so it is *only* θ that we must consider random when taking this expectation.

The expected squared error loss is, making explicit the dependence upon the data of the estimators and the expectations:

$$\begin{aligned}
\mathbb{E}\left[L_s(\theta,\hat{\theta})|x\right] =& \mathbb{E}\left[(\theta - \hat{\theta}(x))^2|x\right] \\
=& \mathbb{E}\left[(\theta - \mathbb{E}\left[\theta|x\right] + \mathbb{E}\left[\theta|x\right] - \hat{\theta}(x))^2|x\right] \\
=& \mathbb{E}\left[(\theta - \mathbb{E}\left[\theta|x\right])^2 + (\mathbb{E}\left[\theta|x\right] - \hat{\theta}(x))^2 + 2(\theta - \mathbb{E}\left[\theta|x\right])(\mathbb{E}\left[\theta|x\right] - \hat{\theta}(x))\right] \\
=& \mathbb{E}\left[(\theta - \mathbb{E}\left[\theta|x\right])^2\right] + (\mathbb{E}\left[\theta|x\right] - \hat{\theta}(x))^2 + 2\mathbb{E}[(\theta - \mathbb{E}\left[\theta|x\right])(\mathbb{E}\left[\theta|x\right] - \hat{\theta}(x))]
\end{aligned}$$

The first term is the same for any estimator $\hat{\theta}(x)$ as it's independent of the estimator. The second is non-negative and is minimised at 0 by $\hat{\theta}(x) = \hat{\theta}_B$.

Consider, then the final term:

$$\begin{aligned}
2\mathbb{E}[(\theta - \mathbb{E}\left[\theta|x\right])(\mathbb{E}\left[\theta|x\right] - \hat{\theta}(x))|x] =& 2\mathbb{E}[(\theta - \mathbb{E}\left[\theta|x\right]|x](\mathbb{E}\left[\theta|x\right] - \hat{\theta}(x))] \\
=& 2\mathbb{E}[\theta - \mathbb{E}[\theta|x]|x](\mathbb{E}\left[\theta|x\right] - \hat{\theta}(x)) = 0.
\end{aligned}$$

Where the final equality follows because $\mathbb{E}[\theta - \mathbb{E}[\theta|x]|x] = \mathbb{E}[\theta|x] - \mathbb{E}[\theta|x]$.

Consequently, the estimator which minimises the expected loss is the estimator which minimises the second term in this expansion.

Thus we have:

$$\begin{aligned}\mathbb{E}\left[L_s(\theta,\hat{\theta})|x\right] =&\mathbb{E}\left[(\theta-\mathbb{E}\left[\theta|x\right])^2\right]+(\mathbb{E}\left[\theta|x\right]-\hat{\theta}(x))^2\\ \geq&\mathbb{E}\left[(\theta-\mathbb{E}\left[\theta|x\right])^2\right]=\mathbb{E}\left[L_s(\theta,\hat{\theta}_B)\middle|x\right]\end{aligned}$$

The expected loss of any estimator is at least that of the posterior mean — and any different estimator will have a larger expected loss as $(\hat{\theta}-\hat{\theta}_B)^2>0$ for any $\hat{\theta}\neq\hat{\theta}_B$ (for at least some, and quite possibly all, data sets).

In the case of this minimum MSE estimator, the expected MSE is:

$$\mathbb{E}[(\theta-\mathbb{E}[\theta|x])^2|x]=\mathbb{V}\mathsf{ar}\left[\theta|x\right]$$

i.e. the minimum achievable mean squared error is equal to the the posterior variance of the parameter.

9.3.1 As we know that the posterior is a properly normalised probability distribution we need not worry about the constant of proportionality that is the marginal distribution of θ under the joint distribution of x and θ. We may also drop any other multiplicative constants which do not depend upon θ:

$$\begin{aligned}q(\theta|x)\propto p(\theta)f(x|\theta)\propto&\frac{\beta}{\Gamma(\alpha)}(\beta\theta)^{\alpha-1}e^{-\beta\theta}\cdot\prod_{i=1}^{n}\frac{\exp(-\theta)\theta^{x_i}}{x_i!}\\ \propto&(\beta\theta)^{\alpha-1}e^{-\beta\theta}\cdot\frac{\exp(-n\theta)\theta^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}\\ \propto&(\beta\theta)^{\alpha-1}\theta^{\sum_{i=1}^{n}x_i}\exp(-(n+\beta)\theta)\\ \propto&\theta^{\sum_{i=1}^{n}x_i+\alpha-1}\exp(-(n+\beta)\theta)\end{aligned}$$

Viewed as a function of θ, this is clearly proportional to a $\mathsf{Gamma}\left(\theta;\alpha+\sum_{i=1}^{n}x_i,\beta+n\right)$ density and as both are properly normalised probability density functions this suffices to prove that it *is* exactly that density.

9.6.1 First consider the posterior distribution (there are only two unknown parameters in our model):

$$\begin{aligned}&q(\theta,\tau^{-2}|x,\sigma^2)\propto p(\tau^2)p(\theta|1/\tau^2)f(\mathbf{x}|\theta,\sigma^2)\\ =&\mathsf{Gamma}\left(\tau^{-2};a,b\right)\mathsf{N}\left(\theta;0,\tau^2\right)\prod_{i=1}^{n}\mathsf{N}\left(x_i;\theta,\sigma^2\right)\\ =&\frac{b}{\Gamma(a)}\left(\frac{b}{\tau^2}\right)^{a-1}\exp(-\frac{b}{\tau^2})\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\theta^2}{2\tau^2}\right)\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}\right)\end{aligned}$$

In order to implement a simple Gibbs sampler, we require the full conditional distribution of each unknown parameter (*i.e.* the distribution of that parameter conditional upon the observed data and all the other parameters).

To obtain these, we can exploit the fact that they're proportional to the full posterior distribution and must be properly normalised probability distributions:

$$\begin{aligned}q(\theta|x,\sigma^2,\tau^{-2})\propto&\frac{b}{\Gamma(a)}\left(\frac{b}{\tau^2}\right)^{a-1}\exp(-\frac{b}{\tau^2})\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\theta^2}{2\tau^2}\right)\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}\right)\\ \propto&\exp\left(-\frac{\theta^2}{2\tau^2}\right)\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}\right)\\ \propto&\exp\left(-\frac{1}{2}\left[\frac{\theta^2}{\tau^2}+\frac{n\theta^2-2\theta\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}x_i^2}{\sigma^2}\right]\right)\end{aligned}$$

As this is the exponential of a quadratic function of θ, if it is proportional to a probability distribution for θ, it must be a normal distribution. Consequently, all that we need to do to work out the distribution is to complete the square in the exponent:

$$\begin{aligned} q(\theta|x, \sigma^2, \tau^{-2}) &\propto \exp\left(-\frac{1}{2}\left[\frac{\sigma^2\theta^2 + n\tau^2\theta^2 - 2\tau^2\theta\sum_{i=1}^n x_i}{\sigma^2\tau^2}\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\frac{\theta^2 - (2\tau^2\theta\sum_{i=1}^n x_i)/(\sigma^2\theta^2 + n\tau^2)}{\sigma^2\tau^2/(\sigma^2\theta^2 + n\tau^2)}\right]\right) \\ &\propto \mathsf{N}\left(\theta; \frac{2\tau^2\sum_{i=1}^n x_i}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2\theta^2 + n\tau^2}\right) = \qquad \mathsf{N}\left(\theta; \frac{2n\tau^2\bar{x}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2\theta^2 + n\tau^2}\right) \end{aligned}$$

where $\bar{x}$ is the empirical mean of the sample.

We can use a similar argument to obtain the full conditional distribution of the other parameter; it is more convenient to work with τ^{-2} than τ itself:

$$\begin{aligned} q(\tau^{-2}|x, \sigma^2, \theta) &\propto \frac{b}{\Gamma(a)}\left(\frac{b}{\tau^2}\right)^{a-1}\exp(-\frac{b}{\tau^2})\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\theta^2}{2\tau^2}\right)\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^n\frac{(x_i-\theta)^2}{\sigma^2}\right) \\ q(\tau^{-2}|x, \sigma^2, \theta) &\propto \left(\frac{b}{\tau^2}\right)^{a-1}\exp(-\frac{b}{\tau^2})\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\theta^2}{2\tau^2}\right) \\ &\propto (\tau^{-2})^{a-\frac{1}{2}}\exp\left(-\frac{b+\theta^2/2}{\tau^2}\right) \\ &\propto \mathsf{Gamma}\left(\tau^{-2}; a+\frac{1}{2}, b+\frac{\theta^2}{2}\right) \end{aligned}$$

A simple Gibbs sampler will sample iteratively from these full conditional distributions, given the current value of the other parameter. The simulated Markov chain will have as an invariant distribution the posterior distribution given above.

9.6.2 The posterior distribution is of the form:

$$q(\theta, \alpha, \beta|\mathbf{x}) \propto \prod_{i=1}^n\left[\mathsf{Poi}\left(x_i; \theta_i t_i\right)\mathsf{Gamma}\left(\theta_i; \alpha, \beta\right)\right]\mathsf{Exp}\left(\alpha; \alpha_0\right)\mathsf{Gamma}\left(\beta; c, b_0\right)$$

Noting that the full conditional distributions are proportional to this posterior and that to deduce the full conditional density of a particular parameter we need consider only multiplicative terms which include that parameter we obtain:

$$\begin{aligned} q(\theta_i|\alpha, \beta, \mathbf{x}) &\propto \mathsf{Poi}\left(x_i; \theta_i t_i\right)\mathsf{Gamma}\left(\theta_i; \alpha, \beta\right) \\ &\propto \frac{e^{-\theta_i t_i}(\theta_i t_i)^{x_i}}{x_i!}(\beta\theta_i)^{\alpha-1}e^{-\beta\theta_i} \\ &\propto \exp(-\theta_i(t_i+\beta))\theta_i^{x_i+\alpha-1} \propto \mathsf{Gamma}\left(\theta_i; \alpha + x_i - 1, \beta + t_i\right) \end{aligned}$$

as the θ_i are independent of one another the joint distribution of θ conditional upon all of the other parameters is simply the product of these distributions.

Via the same strategy:

$$\begin{aligned} q(\beta|\theta, \alpha, \mathbf{x}) &\propto \mathsf{Gamma}\left(\beta; c, b_0\right)\prod_{i=1}^n\mathsf{Gamma}\left(\theta_i; \alpha, \beta\right) \\ &\propto (b_0\beta)^{c-1}\exp(-b_0\beta)\prod_{i=1}^n\beta(\beta\theta_i)^{\alpha-1}\exp(-\beta\theta_i) \\ &\propto \beta^{c+n\alpha-1}\exp\left(-\beta\left[b_0+\sum_{i=1}^n\theta_i\right]\right) \propto \mathsf{Gamma}\left(\beta; c+n\alpha, b_0+\sum_{i=1}^n\theta_i\right) \end{aligned}$$

Finally:

$$\begin{aligned}
q(\alpha|\theta,\beta,\mathbf{x}) \propto& \mathsf{Exp}\,(\alpha;\alpha_0)\prod_{i=1}^{n}\mathsf{Gamma}\,(\theta_i;\alpha,\beta)\\
\propto& \exp(-\alpha\alpha_0)\prod_{i=1}^{n}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta_i^{\alpha-1}\\
\propto& \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n\exp(-\alpha\alpha_0)\left(\prod_{i=1}^{n}\theta_i\right)^{\alpha-1}
\end{aligned}$$

10.3.1 To determine the Cramér-Rao bound, we need to compute the diagonal elements of the Fisher information matrix. Using the partial derivatives calculated in section 10.3:

$$\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2\ell}{\partial\beta_i^2}\right] =& -\mathbb{E}\left[\frac{\partial}{\partial\beta_i}\left[\frac{1}{\sigma^2}(\mathbf{x}_i'\mathbf{y}-\mathbf{x}_i'\mathbf{X}\beta)\right]\right]\\
=& \mathbf{x}_i'\mathbf{x}_i/\sigma^2
\end{aligned}$$

The minimum variance unbiased estimator of each β_i must has, at best, variance $\sigma^2/\mathbf{x}_i'\mathbf{x}_i$ which is exactly the variance of the MLE obtained above.

Whilst:

$$\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2\ell}{\partial[\sigma^2]^2}\right] =& -\mathbb{E}\left[\frac{\partial}{\partial\sigma^2}\left[-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)\right]\right]\\
=& \mathbb{E}\left[-\frac{n}{2\sigma^4}+\frac{1}{\sigma^6}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)\right]\\
=& -\frac{n}{2\sigma^4}+\frac{n\sigma^2}{\sigma^6}=\frac{n}{2\sigma^4}
\end{aligned}$$

Hence the minimum variance unbiased estimator of the regression variance has a sampling variance of $2\sigma^4/n$.

We can apply the same argument that proves theorem 5.3 to demonstrate that the renormalised unbiased estimator, $(n-k)\hat{\sigma}^2/\sigma^2$ is a χ^2_{n-k} random variable. Consequently, it has a variance of $2(n-k)$ and $\mathbb{V}\mathsf{ar}[\hat{\sigma}^2]=2(n-k)(\sigma^2/(n-k))^2=2\sigma^4/(n-k)$ Hence this estimator also achieves the Cramér-Rao bound.

10.3.2 In the case described here:

$$\mathbf{Y}\sim\mathsf{N}\left(\mathbf{X}\beta,\sigma^2\Phi\right)$$

the likelihood is:

$$\begin{aligned}
L(\beta,\sigma^2\Phi;\mathbf{X},\mathbf{Y}) =& \frac{1}{(2\pi)^{n/2}|\sigma^2\Phi|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)^T\Phi^{-1}(\mathbf{Y}-\mathbf{X}\beta)\right)\\
\ell(\beta,\sigma^2\Phi;\mathbf{X},\mathbf{Y}) =& \text{const.}-\frac{1}{2}\log|\sigma^2\Phi|-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)^T\Phi^{-1}(\mathbf{Y}-\mathbf{X}\beta)
\end{aligned}$$

Expanding the quadratic form and differentiating with respect to each of the β_i, we have:

$$\begin{aligned}
\frac{\partial\ell}{\partial\beta_i} =& -\frac{1}{2\sigma^2}\frac{\partial}{\partial\beta_i}\left[-2\beta^T\mathbf{X}^T\Phi^{-1}\mathbf{Y}+\beta^T\mathbf{X}^T\Phi^{-1}\mathbf{X}\beta\right]\\
=& -\frac{1}{2\sigma^2}\left[-2\mathbf{x}_i^T\Phi^{-1}\mathbf{Y}+2\mathbf{x}_1^T\Phi^{-1}\mathbf{X}\beta\right]\\
=& \frac{1}{\sigma^2}\left[\mathbf{x}_i^T\Phi^{-1}\mathbf{Y}-\mathbf{x}_1^T\Phi^{-1}\mathbf{X}\beta\right]\\
=& \frac{1}{\sigma^2}\left[\mathbf{x}_i^T\Phi^{-1}(\mathbf{Y}-\mathbf{X}\beta)\right]
\end{aligned}$$

Setting these equal to zero simultaneously, we find that:

$$\mathbf{X}^T\Phi^{-1}(\mathbf{X}\beta - \mathbf{Y}) = 0 \Rightarrow \beta = (\mathbf{X}^T\Phi^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Phi^{-1}\mathbf{Y}$$

note that we need to be careful when premultiplying by the matrix inverse.

The second order conditions can be verified, but it's a little tedious.

10.3.3 From example 10.2

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}$$

In the current question:

$$\begin{aligned} S_{xx} =& \sum x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 4299.9 \\ S_{xy} =& \sum x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = -22946 \end{aligned}$$

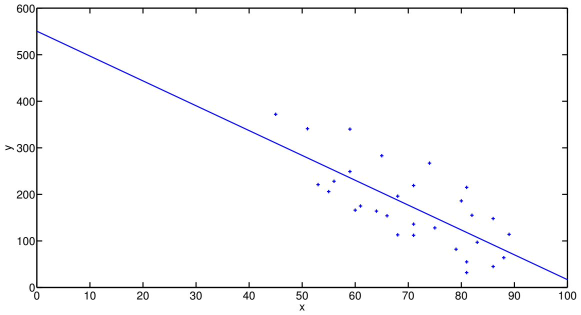
Hence in the notation of this question:

$$\begin{aligned} \beta_0 =& \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \\ =& \frac{5263}{30} - \frac{-22946}{4299.9}\frac{2108}{30} \\ =& \frac{1}{30}5263 + 22946 \times 2108/4299.9 = 550.406 \end{aligned}$$

and:

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = \frac{-22946}{4299.9} = -5.34$$

By way of illustration:



10.5.1 We can estimate the variance of the residuals as:

$$\hat{\sigma}^2 = \frac{1}{n-k}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \frac{1}{30-2}102560 = 3662.7$$

and

$$s_x^2 = 4299.9$$

Thus the test statistic is:

$$T = \frac{-5.3364}{\sqrt{3662.7/4300}} = -5.78$$

The probability of a t_{28} random variable taking a value ≤ -5.78 is approximately $1.65 \times 10^{-6}\%$ and so we can reject the null hypothesis in this case (at essentially any reasonable significance level): the slope is significant. This linear regression does fit the data significantly better than a simple constant model.

10.6.1 The simple prediction is:

$$\hat{y} = \beta_0 + \beta_1 x = 550.4046 - 5.3364 \times 70 = 176.8$$

Following example 10.4 we obtain a confidence interval by calculating:

$$\begin{aligned} v =& \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(70 - \bar{x})}{s_x^2} \right) \\ =& 3662.7 \left(\frac{1}{30} + \frac{(70 - 70.267)^2}{4299.9} \right) = 122.15 \end{aligned}$$

And the calculating the confidence interval:

$$\begin{aligned} I =& 176.8 + [\mathsf{t}_{0.025,28}\sqrt{v}, \mathsf{t}_{0.975,28}\sqrt{v}] \\ =& 176.8 + [-2.048 \times 11.05, 2.048 \times 11.05] = [154.2, 199.4]. \end{aligned}$$